e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
( Peer-Reviewed, Open Access, Fully Refereed International Journal )
Volume:05/Issue:01/January-2023 Impact Factor- 6.752 www.irjmets.com

# A REVIEW ON REAL TIME EMOTION DETECTION SYSTEM USING MACHINE LEARNING ALGORITHMS

**Anshul*1, Manish*2**

*1,2Chandigarh Engineering College Landran, Mohali, Punjab, India.

DOI : https://www.doi.org/10.56726/IRJMETS33027

## ABSTRACT

Emotion detection classification may contribute in making Emotion recognition process more accurate. It appears that automated emotional speech classification will soon replace man-machine interaction. An emotion classification framework is intended to distinguish the emotional state being experienced by the speaker. The emphasis is generally on how something is expressed, not what is expressed. In addition to approaches focusing only on analysing the speaker's voice, a variety of methods can be used to identify emotional states. Some approaches include analysis of voice and spoken words while others focus only on facial expressions. Some people study the brain's reactions to various emotional states. The emotional states will inevitably be a part of human-computer interaction given recent technological advances and expanding study fields including machine learning (ML), audio processing, and speech processing.

Keywords: Neural Network, Feature Extraction, Emotion Recognition.

## I. INTRODUCTION

The three key components of the notion of intelligent behaviour are frequently human capacities for perception, adaption, and environment learning. Numerous research conducted over the past few decades contend that this definition of intelligent behaviour omits one crucial ingredient. Emotional intelligence is that combination. The capacity to sense, express, control, and handle one's own emotions as well as those of others is known as emotional intelligence. Psychology defines an emotional state as a complicated state that causes psychological and physiological changes that have an impact on how we act and think. Humans are fundamentally affected by their emotions, which influence perception and daily tasks including learning, communicating, and making decisions. They communicate using speech, body language, gestures, and another nonverbal cue. The term "emotion detection through speech" refers to the study of vocal behaviour as a sign of affect, with an emphasis on speech's nonverbal components. Its fundamental premise is that voice has a set of objectively measured qualities that accurately reflect the affective state being expressed at any given time.

**Emotion detection through speech based on Machine learning**

The recent improvements of the era and the ongoing studies regions like gadget mastering machine learning, audio and speech processing, the emotional states may be inevitable a part of the human-laptop interaction. There are greater and greater research which can be running on presenting the system with competencies like recognizing interpretation and simulation of emotional states. A popular SER detection device primarily based totally on is fashioned with the aid of using extraction of capabilities vector this sis fashioned with the aid of using extraction of capabilities.

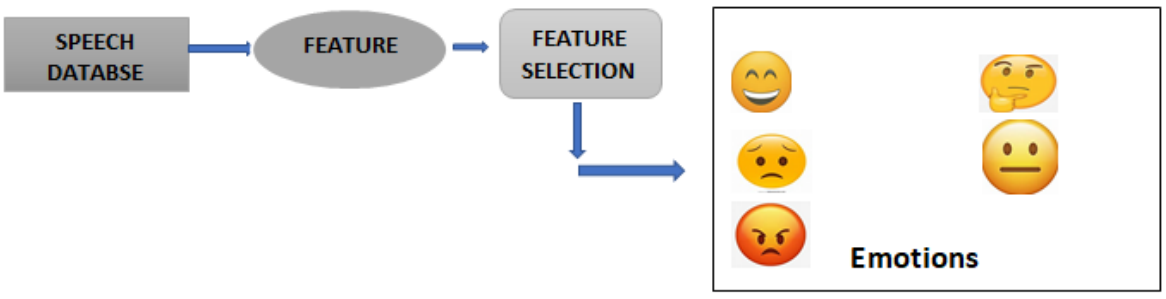


**Fig 1.** ML method for emotion detection

**The full process of ML based Speech emotion detection is defined below in detail:**

a. Several emotive speech databases, including those in English, German, Japanese, Spanish, Chinese, Russian, Dutch, and others, are frequently utilised in literature. The kind of emotions that are communicated in the

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
( Peer-Reviewed, Open Access, Fully Refereed International Journal )
Volume:05/Issue:01/January-2023 Impact Factor- 6.752 www.irjmets.com

speech—whether they are simulated or taken from actual life situations—is one of the key characteristics of an emotional speech database. The benefit of using a simulated speech is that the researcher has total control over the emotion exhibited in it as well as the audio quality. However, the drawback is that the level of spontaneity and naturalness is lost. In contrast, the speech in the non-simulated emotional databases is taken from real-world contexts including call centres, interviews, meetings, movies, short videos, and other circumstances where the spontaneity and naturalness are preserved. The drawback is that there is limited control over the emotions displayed in these databases. The audio's poor quality can also be a problem.

b. Feature extraction: Numerous factors in the voice signal that convey emotional traits are present. What features should be employed is one of the challenging issues in emotion recognition. LPC, MFCC and modulation spectral features are just a few of the frequent features that have recently been retrieved in study.

c. Mel-frequency cestrum coefficient (MFCC): MFCC is the easiest way to detection of human voice. It can catch the human voice frequency and can categories it according to the frequency of the tone. It can generate the MFCC values used in the classification procedure.

d. Modulation spectral features (MSFs): Some characteristics are acquired by simulating the human auditory system's Spectro temporal (ST) processing, which considers both the regular acoustic frequency and the modulation frequency. An auditory filter bank first decomposes the spoken input before producing the ST representation. The modulation signals are created by computing the Hilbert envelopes of the critical-band outputs. On the bases of outcomes, it measures the voice frequencies and implements it.

e. Feature selection: To lower the number of features utilised to characterise a dataset so as to improve a learning algorithm's performance on a particular task," is the goal of feature selection in machine learning (ML). Maximizing classification accuracy in a particular task for a specific learning method will be the goal; as a side effect, Recursive feature elimination (RFE) selects either the best- or worst-performing feature and then eliminates it using a model (such as linear regression or SVM). The prediction strength of each feature is then determined after the estimator has been trained on the original set of features. Then, the existing set of features is reduced in size by eliminating the least crucial features.

f. Classification Methods: The feature vector database is created using the extracted, standardised, and chosen features. Each sample of data in the database represents an instance, or feature vector, that is used to classify the data. supervised classification methods are utilised since each event is assigned the correct emotion. Various machine learning algorithms have been applied to categorise discrete emotions. These algorithms want to classify new observations using the knowledge they get from learning from training examples. There is no one right solution when it comes to selecting a learning algorithm because each method has pros and cons of its own.

## II. LITERATURE REVIEW

| AUTHOR | YEAR | TECHNIQUE USED | FINDINGS | LIMITATIONS |
|---|---|---|---|---|
| S.R KADIRI | 2020 | An approach based on Automatic detection of emotions through voice. | Two speech databases of speech based on emotions are used to gauge the effectiveness of system. | Detection of the similarities between feature distribution and normal speech based on emotions. |
| OLIVEIRA | 2021 | A technique based on framework to detect the emotions through voice. | Comparison between the high pitch and low pitch voice frequencies of arousal and valence. | Approached a new method to use framework on voice recognition. |
| M.S HOSSAIN | 2019 | Approached IOT MODULE devices for voice and image and detection. | This system detects the Signals as input based on the voices. Each signal input processed separately. | Such technology can be used in smart home scenarios. |
| J. HEREDIA | 2022 | Emotion recognition architecture having | Management of missing data and improvising in | Approach is more adapt and presence of quality |

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
( Peer-Reviewed, Open Access, Fully Refereed International Journal )
Volume:05/Issue:01/January-2023 Impact Factor- 6.752 www.irjmets.com

| | | | | |
|---|---|---|---|---|
| | | flexibility which can work with multiple sources and modules. | quality. | in the modules. |
| R. THIRUMUR | 2021 | Feature based on modulation spectral for efficient emotion recognition. | High resolution property exploited to extract the amplitude of human voice. | Three-dimensional voice accuracy has been around 85%, 59%, 65.78% of module. |
| V.R. REDDY | 2019 | Two different level of multi way classifier applied into a single classifier. | Created numbers in the form of matrix to show results. | Shows the results in confusion matrix form based on voice recognition for emotion detection. |
| M.S NAIR | 2022 | Method based on time delay neural network | Uses the knowledge learned from a domain and applies it in another domain with fewer data. | Identifies emotions through languages with non-labeled samples. |
| M. ZEESHAN | 2021 | Proposed a method based on novel integration of spectral integration. | Used for complex audio system through novel convolutional neural network. | Results of accuracy 94.99%, precision 94.96%, recall 94.98%existed in this model. |
| A.A.A ZAMIL | 2019 | Mel frequency Cepstrum Coefficient (MFCC) technique was used. | 13 Dimensional features of vector were extracted. | Through this model accuracy count was 70% in 7 different languages. |
| J. PRIBIL | 2019 | Gaussian mixture model technique was used. | This system had ability to detect one or more than one artefact in synthetic speech product. | This system could turn words to into speech. |
| Y. BHANGDIA | 2021 | A Bot which has ability of making appropriate respond according to the human emotion. | This Model could detect three emotions happy, sad, angry and anxious. | The accuracy of the model is between 83% to 92%. |
| T. WANG | 2021 | Contextual attention neural network based on the multi model framework technique is used. | Feature proposition is being used to unite the information extracted from multiple modalities. | Experimental reports on corpus ability is 64.6%. |

## III. PROPOSED METHODOLOGY

Following are the steps of the proposed model: -

A. Data Acquisition: The data is gathered from different groups to perform experiments.

B. Data preprocessing: Data pre-processing is done to apply machine learning algorithms so that completeness can be introduced, and a meaningful analysis can be accomplished on the data. In order to improve the effectiveness of the training model, this stage removes superfluous attributes from the dataset, delivering clean and denoised data for the feature selection process.

C. Feature selection: A subgroup with incredibly unique traits is used in this step to create predictions. The existing class of features is related to these chosen features. The MFCC model is used in the suggested way to pick features. The voice signal is a quasi-stationary or slowly time-varying signal. Speech analysis over a brief enough time span is required for stable acoustic features. Speech analysis must therefore always be performed on brief portions where the speech signal is presumed to be steady. Short-term spectral measurements are

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
( Peer-Reviewed, Open Access, Fully Refereed International Journal )
Volume:05/Issue:01/January-2023 Impact Factor- 6.752 www.irjmets.com

frequently performed using 20 ms windows and 10ms increments. Individual speech sounds' temporal properties can be tracked by moving the time window forward by 10ms, and a 20ms analysis window is typically long enough to discern major temporal characteristics while still giving these sounds acceptable spectral resolution. The goal of the overlapping analysis is to ensure that each speech sound in the input sequence is roughly centred at a given frame. A window is applied to each frame to taper the signal approaching the frame boundaries. Henning or Hamming windows are typically used. While applying the DFT to the signal, this is done to improve the harmonics, soften the edges, and lessen the edge effect. The Fourier processed signal is run through the Mel-filter bank, a collection of band-pass filters, to compute the Mel spectrum. A Mel is a unit of measurement based on the perceived frequency by human ears. As it appears that the human auditory system does not detect pitch linearly, it does not correspond directly to the tonal frequency's physical frequency. The frequency spacing for the Mel scale is roughly linear below 1 kHz and logarithmic above 1 kHz. Both the frequency domain and the time domain can support filter banks. Filter banks are typically built in the frequency domain for MFCC calculations. On the frequency axis, the filters' centre frequencies are typically uniformly spaced. However, the warped axis, in accordance with the nonlinear function, to mimic the human ears' perception. Triangular shapers are the most common type of filter shaper, and Henning filters are occasionally employed as well. Because the vocal tract is smooth, there is a tendency for adjacent bands' energy levels to correlate. When the converted Mel frequency coefficients are subjected to the DCT, a set of cepstral coefficients are generated.

D. Classification: However, by building the model with the use of training data, the desired value is projected. Only the test data's features are present in this data. SVM is one of the most used algorithms for classifying texts. It chooses the ideal hyperplane for correctly classifying problem scenarios. Decision Tree, a learning method, employs a decision-making model in the form of a tree. The rule that reflects the DT obtained from a disorganized class in an asymmetrical instance is what the classifier depends on. It depends on the feature value, like when using a sorting algorithm to classify data. The routes, leaves, decision nodes, and branches make up the tree. Beginning with the root node, instance classification uses the feature value of that node to categorize the instances. Still on the bases of frequencies being generated through the voice have space uniformly. Through which the numbers of frequencies can be classified differently with respect to the pitch of voice. Generally, the classification depends on the classes. Where it contains different objects in different class. The classification occurs between the classes and the objects or data set present in those classes. The results depend on the bases of observation of data sets. The confusion matrix also known as error matrix results depend on the bases of classifies results.

Classification method formula:

$$\text{Accuracy} = \frac{\text{True positive + true negative}}{\text{Total population}}$$

True positive can be represent as TP where True negative can be shown as TN.

## IV. OBJECTIVES

**a.** To study various machine learnings depends on the speech recognition techniques.
**b.** To develop a classification hybrid design for speech recognition techniques.
**c.** To develop more accuracy in the result of graphs depends on frequencies.
**d.** To develop more accuracy in matric results depends on numerical outputs.

## V. RESULTS AND DISCUSSION

Computational linguistics, artificial intelligence, and computer science all have a foundation in natural language processing, which is important for the relationship between processors and natural languages. There are many NLP tests, including ones for natural language familiarity, making it easier for computers to make understanding of specific facts about natural language usage, and dregs-engrossing NLP. Twitter data is typically used by several accessible techniques to gauge the popularity of movies and products. But the foundation of each of these plans is machine learning or statistical methods. These methods employ language to convey viewpoints. Some classifiers to predict the accuracy of machine learning are as follow: F.Measure Precision Recall

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
( Peer-Reviewed, Open Access, Fully Refereed International Journal )
Volume:05/Issue:01/January-2023 Impact Factor- 6.752 www.irjmets.com

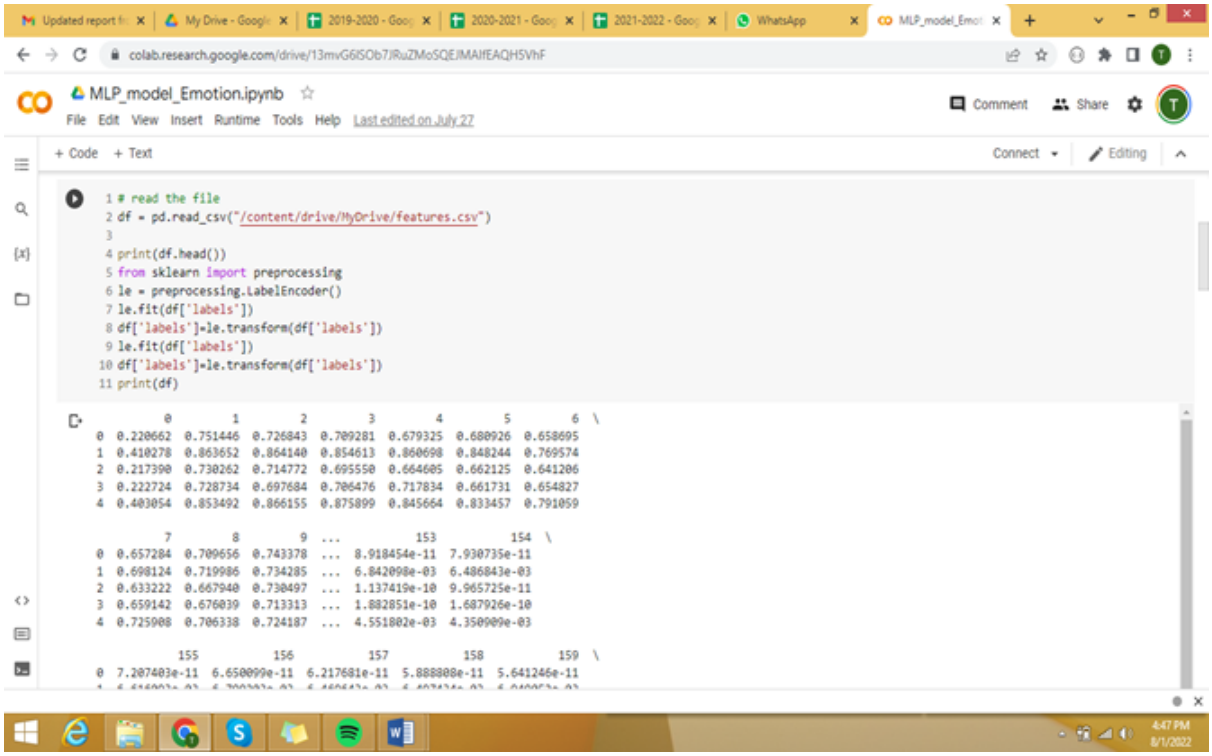


**Figure 2.** Feature Extraction Phase

As shown in figure 3, the features of the input data is extracted for the classification. The MFCC algorithm is applied which can detect various type of features of the speech signal. The dataset will be spitted into training and testing for the classification.

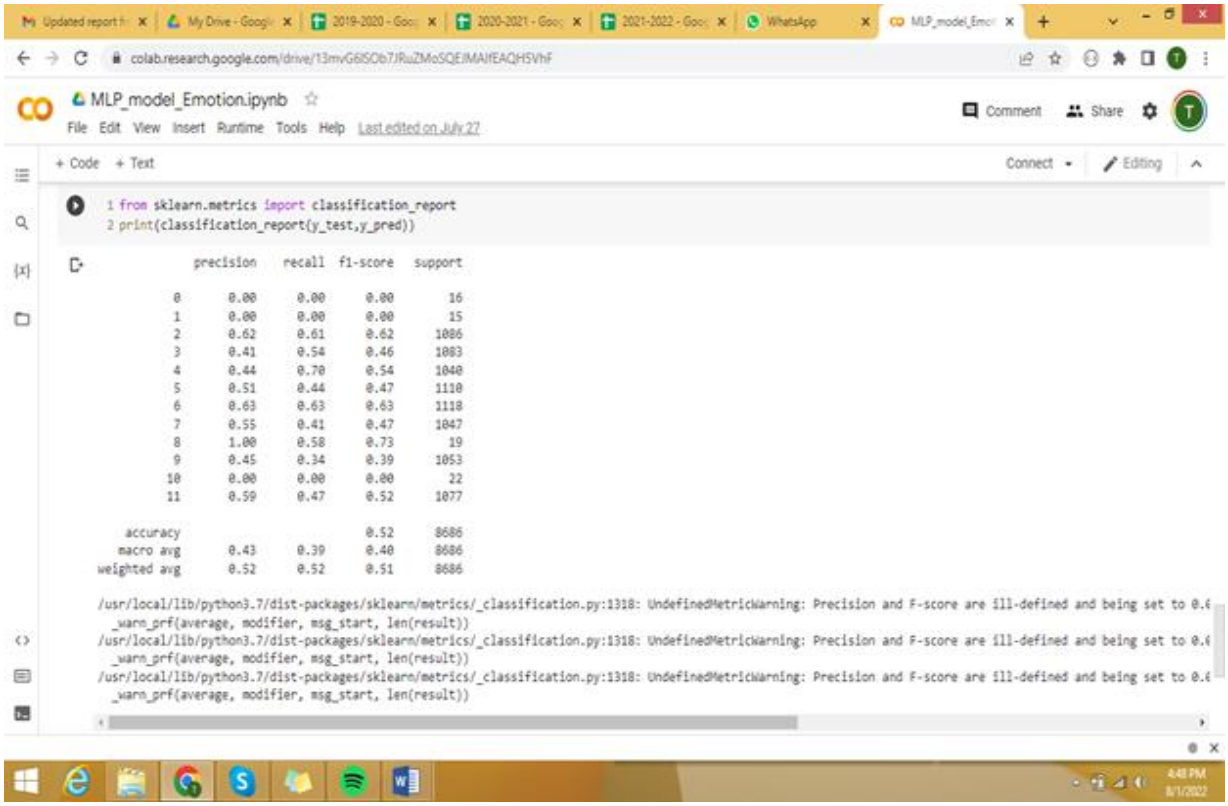


**Fig 3.** Performance Analysis

In fig. the voting classifier is used for the prediction analysis. The voting classifier is the pattern of multiple classifiers. The implementation of the model is analysed in three terms.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
( Peer-Reviewed, Open Access, Fully Refereed International Journal )
Volume:05/Issue:01/January-2023 Impact Factor- 6.752 www.irjmets.com

**Table 1:** Precision Analysis

| Classifier | Precision |
|---|---|
| SVM Classifier | 62 percent |
| KNN Classifier | 59 percent |
| Naïve Bayes | 46 percent |
| Decision Tree | 59 percent |
| Proposed Model | 99 percent |

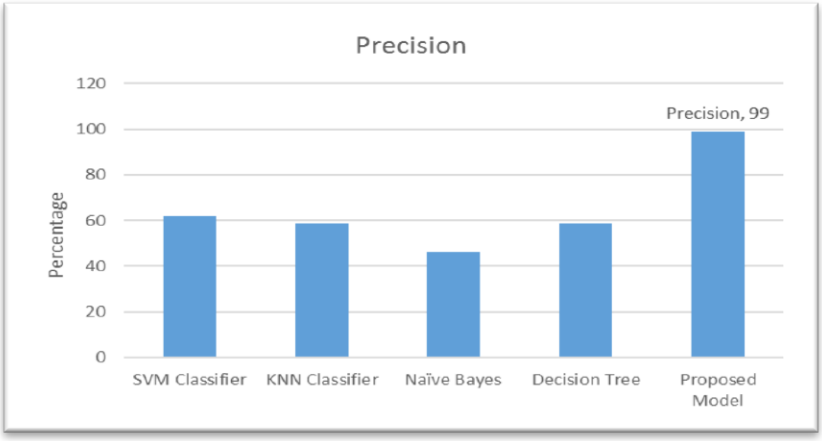


**Fig 4.** Precision Study

As shown in fig. the precision value of the current algorithms such as SVM (Support vector machine, KNN (K's nearest neighbour), Naïve bayes, the decision tree values are same as proposed model. The values present in the model are high as to other classifiers.

## VI. CONCLUSION

The current technological developments and evolving research hotspots such as machine learning, audio handling and speech processes have made emotional environments an integral element of man-machine communication. There are ever more studies relating to provisioning computers with multiple capabilities such as recognition, understanding and simulation of emotional states.

## VII. REFERENCES

[1] J. Wang, Y. Chin, B. Chen, C. Lin and C. Wu, "Speech Emotion Verification Using Emotion Variance Modeling and Discriminant Scale-Frequency Maps," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 23, no. 10, pp. 1552-1562, Oct. 2015

[2] R. V. Darekar and A. P. Dhande, "Improving emotion detection with speech by enhanced approach," 2016 3rd International Conference on Signal Processing and Integrated Networks (SPIN), 2016, pp. 364-369

[3] Z. Qing, W. Zhong and W. Peng, "Research on Speech Emotion Recognition Technology Based on Machine Learning," 2020 7th International Conference on Information Science and Control Engineering (ICISCE), 2020, pp. 1220-1223,

[4] T. Ramakrishna and G. Krishna, "Significance of Accurate Vowel Region Detection for Speech based Emotion Recognition," 2021 IEEE 6th International Conference on Computing, Communication and Automation (ICCCA), 2021, pp. 345-349

[5] O. A. Mohammad and M. Elhadef, "Arabic Speech Emotion Recognition Method Based on LPC And PPSD," 2021 2nd International Conference on Computation, Automation and Knowledge Management (ICCAKM), 2021, pp. 31-36

[6] H. Nishizaki and K. Watase, "Emotion classification of spontaneous speech using spoken term detection," 2017 IEEE 6th Global Conference on Consumer Electronics (GCCE), 2017, pp. 1-5

[7] K. Huang, C. Wu, M. Su and H. Fu, "Mood detection from daily conversational speech using denoising autoencoder and LSTM," 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2017, pp. 5125-5129

[8] A. Harimi, A. Shahzadi and A. Ahmadyfard, "Recognition of emotion using non-linear dynamics of speech," 7'th International Symposium on Telecommunications (IST'2014), 2014, pp. 446-451

[9] M. Y. Alva, M. Nachamai and J. Paulose, "A comprehensive survey on features and methods for speech emotion detection," 2015 IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT), 2015, pp. 1-6

[10] H. K. Vydana, P. Vikash, T. Vamsi, K. P. Kumar and A. K. Vuppala, "Detection of emotionally significant regions of speech for emotion recognition," 2015 Annual IEEE India Conference (INDICON), 2015, pp. 1-6,

[11] Z. Huang and J. Epps, "Detecting the instant of emotion change from speech using a martingale framework," 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2016, pp. 5195-5199

[12] J. C. Vásquez-Correa, N. García, J. F. Vargas-Bonilla, J. R. Orozco-Arroyave, J. D. Arias-Londoño and M. O. L. Quintero, "Evaluation of wavelet measures on automatic detection of emotion in noisy and telephony speech signals," 2014 International Carnahan Conference on Security Technology (ICCST), 2014, pp. 1-6

[13] R. Lotfian and C. Busso, "Emotion recognition using synthetic speech as neutral reference," 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2015, pp. 4759-4763

[14] Leh Luoh, Yu-Zhe Su and Chih-Fan Hsu, "Speech signal processing based emotion recognition," 2010 International Conference on System Science and Engineering, 2010, pp. 487-490

[15] K. Mohan Kudiri, A. Md Said and M. Y. Nayan, "Emotion detection using relative amplitude-based features through speech," 2012 International Conference on Computer & Information Science (ICCIS), 2012, pp. 522-525

[16] Y. Fan, M. Xu, Z. Wu and L. Cai, "Automatic emotion variation detection using multi-scaled sliding window," 2014 International Conference on Orange Technologies, 2014, pp. 232-236

[17] C. Busso, S. Mariooryad, A. Metallinou and S. Narayanan, "Iterative Feature Normalization Scheme for Automatic Emotion Detection from Speech," in IEEE Transactions on Affective Computing, vol. 4, no. 4, pp. 386-397, Oct.-Dec. 2013

[18] R. S. Sudhakar and M. C. Anil, "Analysis of Speech Features for Emotion Detection: A Review," 2015 International Conference on Computing Communication Control and Automation, 2015, pp. 661-664

[19] Z. Huang, "An investigation of emotion changes from speech," 2015 International Conference on Affective Computing and Intelligent Interaction (ACII), 2015, pp. 733-736

[20] R. V. Darekar and A. P. Dhande, "Enhancing effectiveness of emotion detection by multimodal fusion of speech parameters," 2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT), 2016, pp. 3242-3246

[21] S. R. Kadiri and P. Alku, "Excitation Features of Speech for Speaker-Specific Emotion Detection," in IEEE Access, vol. 8, pp. 60382-60391, 2020

[22] R. Lotfian and C. Busso, "Lexical Dependent Emotion Detection Using Synthetic Speech Reference," in IEEE Access, vol. 7, pp. 22071-22085, 2019

[23] S. Zhang, S. Zhang, T. Huang and W. Gao, "Speech Emotion Recognition Using Deep Convolutional Neural Network and Discriminant Temporal Pyramid Matching," in IEEE Transactions on Multimedia, vol. 20, no. 6, pp. 1576-1590, June 2018

[24] J. Oliveira and I. Praça, "On the Usage of Pre-Trained Speech Recognition Deep Layers to Detect Emotions," in IEEE Access, vol. 9, pp. 9699-9705, 2021

[25] Z. Zhao et al., "Exploring Deep Spectrum Representations via Attention-Based Recurrent and Convolutional Neural Networks for Speech Emotion Recognition," in IEEE Access, vol. 7, pp. 97515-97525, 2019

[26] M. T. Teye, Y. M. Missah, E. Ahene and T. Frimpong, "Evaluation of Conversational Agents: Understanding Culture, Context and Environment in Emotion Detection," in IEEE Access, vol. 10, pp. 24976-24984, 2022.