iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 12 **Issue:** IV **Month of publication:** April 2024

DOI: https://doi.org/10.22214/ijraset.2024.60806

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 12 Issue IV Apr 2024- Available at www.ijraset.com

Advances in Speech Emotion Recognition and Analysis: A Review of Applied Machine Learning Methodologies

Ankit Kumar[1], Kshitiz Singh[2], Anmol Sharma[3], Sachi Gupta[4]
[1, 2, 3, 4]*Galgotias college of engineering and technology, Greater Noida, Uttar Pradesh, 201306*

Abstract: ***The most important marketing tactics in the current environment is emotion detection based on individual's interests, allowing extensive customization to cater to the varied needs of customers. For instance, contact centers may start playing music while a caller is on the line if they are agitated. A smart car that slows down when the driver is angry or scared is another illustration.The recognition of speech emotion has become a vital task in the field of computer science department . A very important aspect of human emotional state information is speech emotion that is the source of communication between human and computer interactions . The recognition of speech emotion is currently a very hot topic in the field of machine learning and deep learning and many systems have been build and are currently working on the recognition of these speech emotions using deep learning and neural networks algorithms. In this paper, various types of models and features like Mel-frequency Cepstral Coefficients (MFCC), Modulation spectral features (MSFs), Low-Level Descriptors(LLDs) for speech emotion recognition are categorized. Ultimately, the algorithms like recurrent neural network(RNN), support vector machine(SVM) used in different models are tabulated presenting their advantages and disadvantages along with their features and efficiency.***
Keywords: MFCC, RNN, MSF, Feature Extraction, Speech Emotion Detection, SVM

I. INTRODUCTION

Emotions performs a very important part in our all-day communications with other individuals [1,12 and 13]. By the help of emotions, we could understand other human's feelings by expressing ourselves and replying to others in response to their emotions. Emotions conveys the information about a human's mental condition [1,12 and 13]. It has developed a fresh research area known as Automatic Emotion Recognition (ASR), which includes some aim of understanding and retrieving emotions. In recent researches, many functions had been covered to detect and understand the emotions of humans like, physiological signals [7], facial expressions, speech [2, 12 and 13], etc. Many advantages make signals from speech a tool for the affective computing. The goal of Speech Emotion Recognition (SER) is to detect the internally hidden emotions of human from speech, that is the desired component for better interaction between human and computer [13].

In this paper, an extensive study and research on this technology has been done and a system has been proposed that may have all the quality features of previous models with some more features in it so that the working efficiency of SER can be improved thoroughly. By collecting the vast information by studying different developments on speech emotion recognition, it can be observed that the accuracy rate of classifying emotion directly from the speech can be increased if we use an auto encoder system with prosodic and acoustic features [13] such as MFCC [2, 12] and TEO and HNR. These features offer key information about the emotional content of human speech, simplifying the system's ability to separate various emotions. Applying classifier models such as Support Vector Machine (SVM) and Convolutional Neural Network (CNN) proves to be a thoughtful approach for efficiently categorizing these features into emotions like anger, disgust, happiness, neutrality, etc.

II. BASIC FEATURES AND MODELS USED IN SPEECH EMOTION RECOGNITION

A. Mel-frequency Cepstral Coefficients (MFCC)

The MFCC is extracted by the following process- Each frame of the recorded voice is applied on the Fourier transformation and the energy spectrum were estimated, the energy spectrum is then mapped into the Mel-frequency scale. Then the discrete cosine transform (DCT) of the Mel log energies was estimated, and the first 12 DCT coefficients provided the MFCC value which is used for classification. Resultant from the logarithmically spread-out Mel-frequency scale, MFCCs signify the spectral features of a speech signal, catching vital information about the distribution of energy in diverse frequency groups.

B. Modulation spectral features (MSFs)

Auditory-inspired long-term Spectro-temporal representation is used to extract this feature. This is the main feature for emotion recognition. The Spectro-temporal (ST) processing which is performed in the human system and consider regular acoustic frequency jointly with modulation frequency helps in extract these features. Modulation spectral features comprise the investigation of amplitude differences in different frequency groups over time. Hilbert envelopes, derivative of the Hilbert transform, are mainly beneficial in taking the prompt amplitude of a signal. Together, these methods offer a inclusive understanding of temporal dynamics, helping applications such as speech processing and audio signal classification.

C. Low-Level Descriptors (LLDs)

This is the main feature used for Graph Convolutional Network (GCN) algorithm for emotion recognition. These features first extracted from the raw audio signal then are given to the deep learning model to get the discrete motions label.

D. Wav2vec-2.0

It is the framework. This framework is useful in self-supervised learning of speech representation. This framework helps in masking the speech input in the latent space and solves a contrastive task which are defined over a quantization of the latent representations which are jointly learned. This model helps in providing the methods to perform the feature extraction and classification in one step. The output is in the form of logits.

E. Support Vector Machine (SVM)

Support Vector Machine (SVM) is a very useful supervised machine learning algorithm used for both classification and regression. The main objective of the SVM algorithm is to find the optimal hyper-plane in an N-dimensional space that can separate the data points in different classes in the feature space. The hyper-plane tries that the margin between the closest points of different classes should be as maximum as possible. The dimension of the hyper-plane depends upon the number of features. If the number of input features is two, then the hyper-plane is just a line. SVMs outclass in managing high-dimensional feature spaces mined from speech signals, providing a vigorous framework for modeling multipart relationships between acoustic features and different emotions.

III. APPROACHED USED FOR SPEECH EMOTION RECOGNITION

A. Automatic Speech Emotion Recognition

B Leila Kerkeni et al [1], proposed a SER system employing a Recurrent Neural Network (RNN) machine learning model. In this system, the input speech signal is recorded and the features like MFCC and MS are extracted and apply RNN neural network model on it to classify the seven emotions. After wards the other models like SVM and MLR are also applied on the feature to compare the efficiency. It has been observed that the model shows the efficiency of 83% in Berlin and 93% in Spanish data set. The 93% efficiency is achieved by the RNN model so it is considered as the good model for SER. In this paper the SER system is defined in 4 major steps these steps are related to each other-(1)collection of the voice sample (2)The feature vector is made from the extracted features of voice sample (3)Now determine which feature is most relevant for differentiate emotions (4)The last step is to apply some models on these features for classify the emotions.

In this paper only few features such as MFCC and MS are used or we can say that only two features are extracted from the voice. For extracting the MFCC feature, each frame of the recorded voice is applied on the Fourier transformation and generated the energy spectrum, the energy spectrum is then mapped into the Mel-frequency scale. Then Mel log energies generated the discrete cosine transform, and provided the MFCC value which is used for classification.

B. Compact Graph Architecture for Speech Emotion Recognition

In 2021, Amir Shirian and Tanaya Guha [2] proposed a deep graph approach for Speech Emotion Recognition. In this model, the data is represented in the form of a graph that consists of nodes and edges where the nodes have the required data. This paper uses the GCN (graph Convolution Network) [11] based model or architecture for recognition of emotions from the voice samples. In graph classification approach, firstly a graph from each sample speech is created then GCN architecture is developed which is responsible for assigning the discrete emotions to each graph. A simple frame to node transformation strategy is followed for constructing the graph where M nodes in graph G are generated by the M frames. Now GCN architecture is applied on the set of graphs which are represented as one hot vector. There are two convolution layers in the GCN model.

The first layer is a pooling layer which produces a graph-level embedding vector, and second layer is a fully connected layer which yields distinct labels of emotion. The performance of the approach is evaluated on some well-known dataset which are IEMOCAP and MSP-IMPROV. With this approach the Result demonstrate the Maxpool 61%, Meanpool 62.5% and Sumpool 65%.

C. Speech Emotion Recognition Using Deep Learning

In 2019, Hadhami Aouani and Yassine Ben Ayed [3] introduced an emotion recognition system which is based on two level frameworks, one is feature extraction and another is classification engine. They made a system of emotion recognition using these parameters and classify them into emotions using a support vector machine classifier (SVM). After that, they used an Auto encoder to enhance the accuracy of their system. During first stage which is feature extraction they extracted prosodic feature and MFCC and save them into their classification algorithm as feature vectors.To calculate MFCC, they have used algorithm Inverse Fast Fourier Transform that helps in extracting MFCC from input audio and store it as features vector in the algorithm for further processes. Other features are also extracted from the input speech then feature dimension reduction process is done and, in this process. Then auto encoder is used in feature selection. In this approach, two models have been used- basic AE and stacked AE. The stacked AE has two more hidden layers after that SVM classification model is used to classify between these reduced features and classify them into several emotions. In this paper, RML dataset is employed which contain 720 audiovisual emotional expression sample. Their proposed system was compared with other advanced techniques as well and the results were coming out to be the best for their proposed system as the proposed system gives the accuracy rate of 65% without feature selection and 72.83% by using Basic AE and 74.07% accuracy rate while using Stacked Auto encoder.

D. Domain Adversarial Neural Network for Multimodal Emotion Recognition

In 2020, Zheng Lian and et al [4], used a Domain Adversarial Neural Network (DANN) to recognize the emotion in a speech. One of the problems occur in recognition is to learn representations which are different for distinct speakers. Hence people start working on data split strategies in training set which reduces speaker overlapping. Another problem is how to collectively use the distinct features with context information because human themselves recognize emotions based on the contextual information and single feature is not enough to meet the accuracy hence multimodal features are used for recognition. For recognition of speech, the framework used in DANN contains three stages: first one is featuring encoder which extracts context representations for which it uses Self Attention based Gated Recurrent Unit (SA-GRU) and Audio Text fusion (AT-fusion), second is domain classifier which gives the features which are not dependent on speaker and last is emotion classifier which classify the emotion data from the extracted representations. In this research, OpenSMILE toolkit is used to extract acoustic features like energy, MFCC, spectral etc. This experiment results with 82.68% accuracy with IEMOCAP dataset.

E. Multimodal Multi-task Learning for Speech Emotion Recognition

Multimodal Emotion Recognition (MMER) is proposed by Sreyan Ghosh and et al in 2023 [5] for the recognition of emotions in a particular speech. It uses text and acoustic modalities collectively and performs three auxiliary tasks which helps in understanding the emotions in a simple speech. This proposed model uses self-supervised models to extract the contextual representation for each utterance. The three tasks solved by MMER to classify emotions are: first, it does an Automatic speech recognition job with the minimized CTC loss which gives overview of the semantic data present in the utterance. Then it performs two similar works: first is supervised contrastive learning in which they understand instance discrimination with the representations based on real time labels so that it gains better knowledge about emotions from multimodal information. Second is Augmented contrastive learning to make the model more powerful and ensure that the trained data is speaker independent and to do so they used wav2vec-2.0 which is a pre-trained encoder that generates contextualized representation for a speech. In this research, The MMER model is applied with the IEMOCAP dataset and it results with the 81.2% of accuracy for a data.

IV. COMPARISON AND EVALUATION

In the first paper, the features like MFCC and MS are extracted from the Berlin database and then the model like SVM and RNN is applied on this which gives the accuracy of 83% while in the second paper the graph-based approach is used to extract features and to predict the emotion in this paper the IEMO-CAP and MSP-IMPROV dataset is used. The reported accuracy is 65.29% in the graph-based approach. The GCN model is used in the third paper where the author has proposed the use of traditional features like MFCCs and other prosodic features like ZCR, HNR and then used the SVM model to classify them into emotions.

After the use of auto encoder to filter features, SVM model on RML database is used which gives the accuracy of 69.3%. The recognition accuracy is 72.03 % with basic auto encoder while 74.03% with the use of stacked auto encoder. In paper four, Domain adversarial neural network (DANN) model is used to extract the MFCC and OpenSMILE toolkit features on the IEMO-CAP database (same as paper 2 database) which gives 82.68% accuracy. In the last paper uses MMER, supervised contrastive learning (SCL), Augmented contrastive learning (ACL) models to extract Wav2vec-2.0 features on the IEMO-CAP database that gives 81.2% accuracy. The objective of sixth paper [6] is to use a 3-stage SVM classifier to categorize seven diverse emotions existing in the Berlin Emotional Records. For classification, MFCC features from all the 535 files existing in the database are mined. Nine statistical quantities are accomplished over these features from each frame of a sentence. For training and testing of data, 10-fold cross-validation is performed. Performance study is done by using the confusion matrix and the accuracy found is 68%. Divya Sree GS and el [7] at proposed the NN and GMM-UBM based classifier for speech emotion recognition and the FFT, MFCC and DWT features are used for vector extraction such as minimum, maximum, mean, entropy, energy, TEO (Teager Energy Operator). The SVM classifier used to differentiate feelings which include happiness and anger state. In year 2012, Pan Y [8] proposed his work to recognize three emotional states: happy, neutral, and sad. The discovered features include: linear predictive spectrum coding (LPCC), MFCC, energy, pitch, and mel-energy spectrum dynamic coefficients (MEDC). Lastly results for diverse grouping of the features and on diverse databases are equated and clarified. The complete investigational results expose that the feature combination of MFCC+ Energy+ MEDC has the maximum accuracy rate on both Chinese emotional databases i.e., 91.3% and Berlin emotional database i.e., 95.1%. In 2017, Lim W [9] proposed a SER technique built on concatenated CNNs and RNNs without using any old-style hand-crafted features. By applying the planned approaches to an emotional speech database, the classification result was proved to have better accurateness than that attained using conventional classification approaches. Sathit P in 2015 [10], proposed a SER system, based on diverse classifiers and diverse approaches for features extraction. MFCC and MS features are extracted from the speech indicators and used to train diverse classifiers. A RNN classifier is used primarily to classify seven emotions. This study illustrates that for Berlin database all classifiers attain an accurateness of 83% when a speaker normalization (SN) and a feature selection are used to the features. For Spanish database, the best accuracy 94 % is achieved by RNN classifier without SN and with FS. A comparative analysis of the listed papers is shown in Table 4.1.

Table 4.1: A comparative analysis of various Speech Emotion Recognition Systems

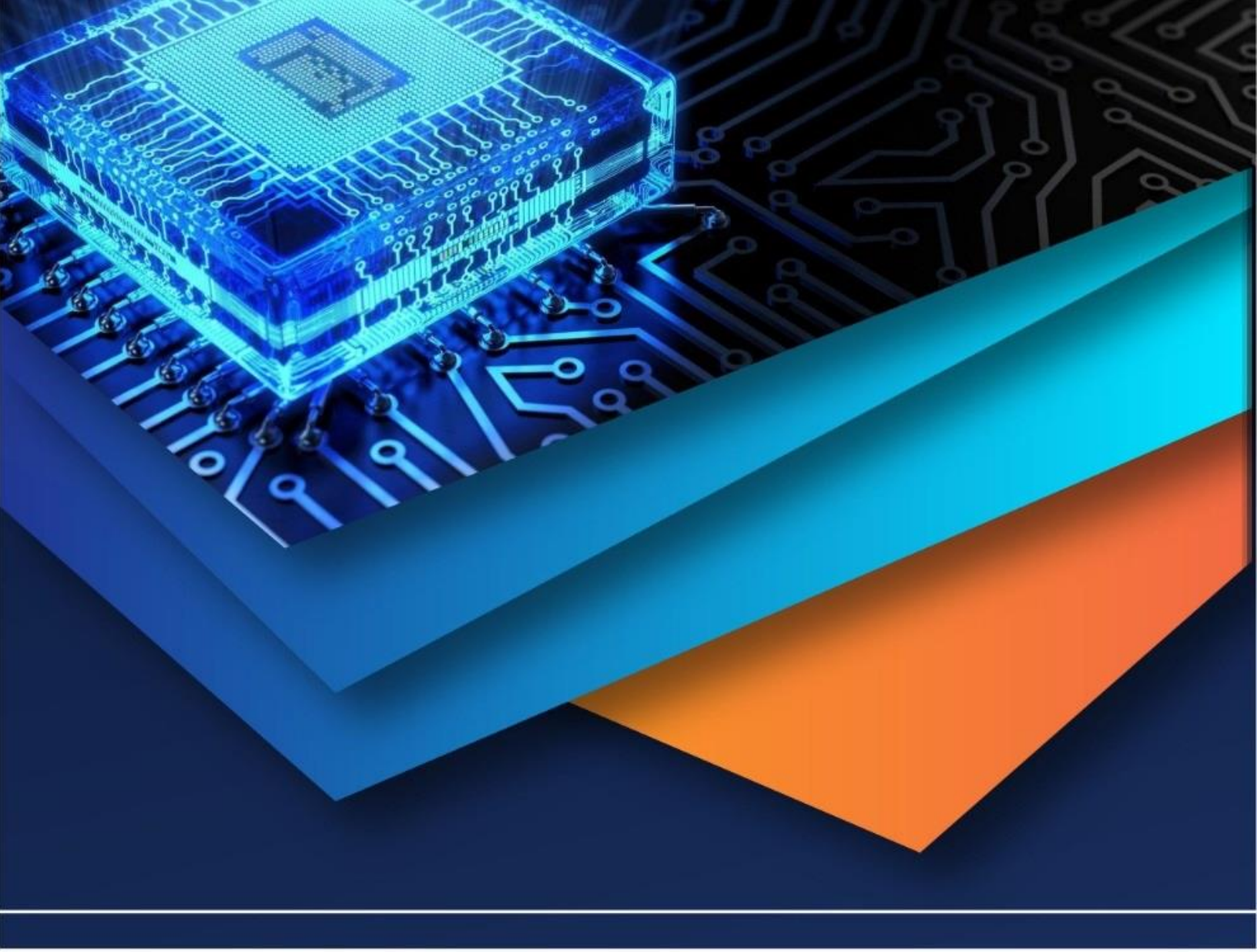
Reference No.	Features	Models	Dataset	Accuracy
[1]	MFCC, MS (modulation spectrum), FS (feature selection	Multivariate linear regression, Recurrent Neural Network, Support vector machine	Berlin Database	83%
[2]	LLD, MFCC, Graph- Based	Convolutional neural network, Graph convolutional network,	IEMO-CAP, MSP-IMPROV	65.29%
[3]	MFCC, HNR, ZCR, TEO	SVM, Auto Encoder	RML	69.3%
[4]	MFCC, OpenSMILE toolkit	Domain adversarial neural network (DANN)	IEMO-CAP	82.68%
[5]	Wav2vec-2.0	MMER, supervised contrastive learning (SCL), Augmented contrastive learning (ACL)	IEMO-CAP	81.2%
[6]	MFCC and MS	3 state SVM	Berlin	68 %
[7]	MFCC and DWT	SVM and GMM-UBM	Berlin ,German	83%
[8]	MFCC, LPCC and MEDC	SVM	Berlin	90%
[9]	2D Representation LSTM	RNN	IEMO-CAP	80%
[10]	MFCC and LPCC	HNN, SVM, RNN	IEMO-CAP	80%

V. CONCLUSION

After reviewing all the listed research papers, we analyzed the algorithms and models featured in the various speech emotion recognition along with their features. We compared the advantages and limitations of all the models used in different systems along with their approaches, we have determined that the Support Vector Machine (SVM) model stands out with the most favorable outcomes. In our evaluation, particularly in processing input speech from the Berlin and Spanish databases, the SVM model showcased a remarkable accuracy of 90%. This robust performance positions SVM as a leading choice for emotion recognition, surpassing other models in terms of effectiveness. The simplicity and reliability of SVM make it a compelling solution for accurately discerning emotions in speech, underscoring its potential for enhancing various systems reliant on emotion-aware technology.

REFERENCES

[1] B Leila Kerkeni, Youssef Serrestou, Mohamed Mbarki, Kosai Raoof, Mohamed Ali Mahjoub, and Catherine Cleder "Automatic Speech Emotion Recognition System using Machine Learning," in IntechOpen 2019.
[2] Amir Shirian and Tanaya Guha, "Compact Graph Architecture for Speech Emotion Recognition," in ICASSP 2021.
[3] Hadhami Aouani and Yassine Ben Ayed, "Speech Emotion Recognition with deep learning," Published by Elsevier B.V. 2019.
[4] Zheng Lian1, Jianhua Tao, Bin Liu, Jian Huang, Zhanlei Yang and Rongjun Li, "Context-Dependent Domain Adversarial Neural Network for Multimodal Emotion Recognition," in Inter-speech 2020.
[5] Sreyan Ghosh, Utkarsh Tyagi, S Ramaneswaran, Harshvardhan Srivastava and Dinesh Manocha, "MMER: Multimodal Multitask Learning for Speech Emotion Recognition," arXiv:2203.16794v5(2023).
[6] Milton A, Sharmy Roy S, Tamil Selvi S. "SVM scheme for speech emotion recognition using MFCC feature" in. International Journal of Computer Applications, 69, 2013.
[7] Divya Sree GS, Chandrasekhar P, Venkateshulu B, "SVM based speech emotion recognition compared with GMM-UBM and NN," in IJESC. 2016.
[8] Pan Y, Shen P, Shen L. Speech emotion recognition using support vector machine. International Journal of Smart Home. 2012.
[9] Lim W, Jang D, Lee T. "Speech emotion recognition using convolutional and recurrent neural networks." in Asia-Pacific. 2017.
[10] Sathit P, "Improvement of speech emotion recognition with neural network classifier by using speech spectrogram," in International Conference on Systems, Signals, and Image Processing (IWSSIP). 2015.
[11] Saaran, V., Kushwaha, V., Gupta, S., Agarwal, G. "A Literature Review on Generative Adversarial Networks with Its Applications in Healthcare," in Sharma, H., Saraswat, M., Yadav, A., Kim, J.H., Bansal, J.C. (eds) Congress on Intelligent Systems. CIS 2020. Advances in Intelligent Systems and Computing, vol 1334. Springer, Singapore.
[12] Agarwal, G., Maheshkar, V., Maheshkar, S., Gupta, S. "Recognition of Emotions of Speech and Mood of Music: A Review" in Woungang, I., Dhurandher, S. (eds) International Conference on Wireless, Intelligent, and Distributed Environment for Communication. WIDECOM 2018.
[13] Agarwal, G., Maheshkar, V., Maheshkar, S., Gupta, S. "Vocal Mood Recognition: Text Dependent Sequential and Parallel Approach", in Malik, H., Srivastava, S., Sood, Y., Ahmad, A. (eds) Applications of Artificial Intelligence Techniques in Engineering. Advances in Intelligent Systems and Computing, vol 698. Springer, Singapore.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY