iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 7 **Issue:** X **Month of publication:** October 2019

DOI: http://doi.org/10.22214/ijraset.2019.10130

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: 7.177
Volume 7 Issue X, Oct 2019- Available at www.ijraset.com

Emotion Recognition from Audio Signals using SVM and Naive Bayes : Review

Miss. Pranali Bendale[1], Dr. Kanchan Bhagat[2], Dr. Jitendra Chaudhari[3]
[1]*Dept of Electronics and Telecommuication, J.T.M.College of Engg. Faizpur, Mharashtra,*
[2]*P.G. Co-ordinator, Dept of Electronics and Telecommuication, J.T.M.College of Engg. Faizpur, Mharashtra.,*
3Associate Professor, CHARUSAT Space Research and Technology Center, Charotar Universityof Science and Technology, Changa. Anand, Gujrat

Abstract: ***Emotions are expressed in many ways by human being. Vocal communication is more convenient medium of communication having roughly two types: speech and song that are closely related to each other. Speech and song has many important parameters which are most important in emotion recognition process such as short time energy, ZCR, Mel Frequency Cepstrum Coefficient. Database is the vital part of the system to recognize audio using several classifiers. In this paper, We are reviewing concepts related to the emotion recognition such as Feature extraction, RAVDESS Data Bases, different Classifiers: Support Vector Machine (SVM), Naive Bayes, HMM, GMM, ANN for accurate and specific knowledge of audio emotion recognition system.***

Keywords: RAVDEES, audio, SVM, Naïve Bayes, HMM, GMM, ANN, Classifiers, Prosodic features, spectral features

I. INTRODUCTION

Researchers are trying to develop a system which is able to recognize emotions from audio with various techniques like pattern recognition, speech signal processing, artificial intelligence, human computer interaction, human psychology and many more[1]. Emotions are very important to give proper meaning to the speech. Emotions like fearful, sad, happy, disgust, anger, surprise and neutral. It is very easy for human brain to recognize emotions from audio whichever it be a speech or song, but it is challenging to the computer or machine to recognize emotion of audio. Speech processing processes the audio by several methods to make machine to recognize emotion from audio.

In this paper we are reviewing emotion expressed in speech and song using various models and analyze the commonalities and differences present in emotion expression across these two communication domains. Finally, we study different classification techniques, SVM, HMM, GMM, ANN and Naïve Bayes separately for speech and song. Classifiers come under supervised learning. Some may be either for regression purpose or classification purpose. The novelty of this paper includes some steps and procedures to emotion recognition from different domains such as song and speech as well as male and female actors.

The pattern recognition system offers Speech emotion recognition as an application in which patterns of derived speech features such as Pitch, Energy, MFCC are mapped using classifier like ANN, SVM, HMM etc.

Emotions play important role in human interaction. Human possess and express emotions in everyday interactions with others. There may be different types of signs and codes that indicate emotions. One of the most important aspects of human-computer interaction is to train the system to understand human emotions through voice. People can use their voice to order commands to many electrical devices. Hence it is important to make the devices understand human emotions and give a better experience of interaction. Speech recognition is the ability of a machine or program to identify words and phrases in spoken language and determine the emotions of the speaker such as normal, anger, happiness, excitement and sadness.

A. Emotion recognition from audio has many more applications in our daily life. Some of these applications include:
B. Enhancing the human and machine interaction.
C. In psychiatric diagnosis, lies detection.
D. Analyzing the behavioral study in call centre conversation between customer and employee.
E. In aircraft cockpits, for the better performance speech emotion recognition system trained for stress speech.
F. For understanding the criminal's behavior that would help for analysis in criminal department.
G. Emotion recognition in robotic.

II. REVIEW WORK

A. Authors study the dependency of the physiology with human emotions in [2, 3]. In the approach, physiological characteristics like temperature and electro dermal activity are used as an input for emotion recognition. The results indicate that an emotional relationship between humans and computers is developed which enables the development of a personal human-friendly robot [2, 3].

B. In [4], a method of measuring the heart rate of a patient sitting on a chair is proposed. In this method, electro-mechanical film and traditional ear lobe photo-plethysmo graph (PPG) was embedded into the chair. Twenty four participants participated in the experiment in order to demonstrate whether human emotions can be directed to computer in the same way as to society.

C. In [5], it is argued that emotional state of person is dependent on heart rate variation (HRV). The results of the study revealed that death can be caused by the emotional stress due to heart disorders like acute myocardial infarction. It can also make prediction on risk of developing the hypertension. However, it was proved that positive emotions may be beneficial in treatment of that hypertension by causing alterations in HRV.

D. Speech parameters for detection and analysis of the vocal fold pathology were used in the idea proposed in [6]. They developed algorithms for speech signal processing to create characterizing healthy and pathology conditions of human vocal folds model. The methodology behind this idea was based on extraction of the separate speech signal components from both healthy and assumed pathology conditions. They applied iterative maximum likelihood (ML) estimation for solving that problem.

E. In [7], they developed an algorithm for the detection of hypernasal resonance. It provides noninvasive contactless interaction with patient, hence maximizing the accurate detection of speech due to the naturalness of speaking in comfort conditions without extra devices on the face and body.

F. In [8], speech therapy through telemedicine technology with a group of patients was conducted in an experiment. Interaction among patients through speech improves their recovery and positively influence on the quality of life [8]. The same was then improved by monitoring heart rate detected from the speech of participants and evaluate the progress of recovering and prevent any risks related to the heart failure.

G. In [9], authors introduced a remote detection of the Body Mass Index from the speech signal. They designed novel approach for remote monitoring the patients' weight in order to control the risks of diseases and death which are resulted from underweighting or in opposite from overweighting. The importance of the telemedicine was also discussed in this study, showing its benefits from different aspects such as comfort, low cost, time efficiency [9].

H. The work in [10] proposed an innovative approach for measuring the heart rate continuously of using mobile phones. The design consisted of these sub-systems: the first one records the signals and performs an offline analysis of the heart rate; the second system provides support for remote real time monitoring of the detected electrocardiogram (ECG) signal by sending the data to the medical centre or doctor through certain communication media; the third system performs a local real time classification of collected data. The main advantage of this design is, it supports for mobility of both patient and doctor [10]. This design can be improved by applying speech signal recording for evaluating the heart rate instead of using ECG sensor that sends the signal to the mobile phone via Bluetooth connection.

III. SPEECH EMOTION RECOGNITION SYSTEM

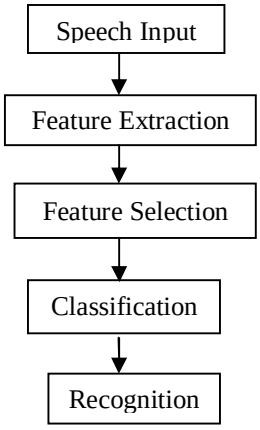


Fig. 1: Speech Emotion Recognition System

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.177
Volume 7 Issue X, Oct 2019- Available at www.ijraset.com

The system has these major modules: speech input database, feature extraction, feature selection, classifier & recognized output as illustrated in figure – 1 above.

Overall, the system is relay on deep analysis of the generation mechanism of speech signal, extracting some of features which contain information about speaker's emotion & taking appropriate pattern recognition model to identify states of emotion the most common way to recognize speech emotion is to first extract important features that are related to different emotion states from the voice signal then feed those features to the input end of a classifier and obtain different emotions at the output. At the end, performance of several systems is also calculated.

IV. SPEECH INPUT DATABASE

In this experimental study, audio samples are collected in the form of .wav files. Collection of these input data files is called datasets. Now a day's, many datasets are available due to innovations in the speech recognition field like RAVDESS dataset for audio database [11], iris for flower databases and many more. After data acquisition data training is carried out.

Preprocessing: However the data acquisition there is some noise in the input data. This noise should be processed before carrying out next steps of feature extraction and feature selection [12]. Preprocessing is used to simplify operation of the machine to classify features.

V. FEATURES EXTRACTION FROM AUDIO SIGNALS

Feature extraction is the most important step in audio classification tasks. Feature extraction is the measure of competing a compact numerical representation that can be used to characterize a segment of audio.

1) *Maximum and Minimum Value of Signal:* Highest value of signal amplitude is maximum value of signal and lowest value of signal amplitude is minimum value of signal are the basic features in feature extraction method. These are expressed as (x_{max}) and (x_{min}) respectively.
2) *Mean*: Mean is the average value,

$$m= (1/n)\sum_{i=1}^{n} x_i$$

3) *Standard Deviation*: It is the distribution of data from mean.

$$STD= \sqrt{(1/N-1)\sum_{i=1}^{N}(x[n]- (1/N)\sum_{i=1}^{N}X[n])^2}$$

4) *Dynamic Range:*

$$DR= \log_{10}(L_{max}/L_{min})$$

5) *Crest Factor:* Crest factor is ratio of peak value of signal.
6) *Autocorrelation Time:* It is defined as highest peak in short time autocorrelation sequence. It is used to evaluate how close audio signal is to a periodic one.
7) *Energy Entropy*: It is the value of change in energy
8) *Short time Energy*: Short time energy of a frame id defined as the sum of square of signal samples normalized by the frame length and converted to decibles.

$$E= 10\log_{10}((1/N)\sum_{n=0}^{N-1} x^2[n])$$

9) *ZCR:* Zero Crossing Rate is a measure of number of time the signal value crosses the zero axes. This feature is majorly used to separate noise.

$$Z= n_c (f/n)$$

10) *Spectral Roll Off:* it is defined as boundary frequency fr, such that a certain percent P and the spectral energy for a given audio frame is concentrated below fr.

$$\sum_{K=0}^{fr}|X(k)| = P(\sum_{k=0}^{k-1}|X(k)|)$$

11) *Spectral Centroid*: Spectrum Centroid is defined asss the center of gravity (COG) of the spectrum for given audio from and is computed as,

$$S_c = \frac{\sum_{k=0}^{k-1} k\,|X(k)|}{\sum_{k=0}^{k-1} |X(k)|}$$

12) *Spectral Flux:* the spectral flux measures the spectrum fluctuations between two consecutive audio frames. It is defined as[13],

$$S_f = \sum_{k=0}^{k-1} (|X_m(k)| - |X_{m-1}(k)|)^2$$

VI. FEATURE SELECTION

Feature selection is the process of selecting proper features from large set of available extracted features to maximize the performance of learning algorithm. In this process feature reduction is done. Reduction of features improves the quality of classifier capacity of prediction. Feature selection also reduces computer storage and processing time required for classification of data [14].

VII. CLASSIFIERS

1) *SVM:* Support vector machine Classifier is the supervised learning method to train the data. In SVM emotional speech data is converted into train data and test data sets. Then the feature extraction is performed on both training data and test data sets. It predicts presence and absence of specified features are related or not related to other features. SVM can be used for both either classification or regression purpose; that is by linear or nonlinear manner. In this classifier machine is trained to predict the output. Kerenel functions are used for classification in SVM classifier. Fig2 shows pictorial representation of SVM classifier algoritthm [15] The advantage of support vector classifier is that it can be extended to nonlinear boundaries by the kernel trick[1].

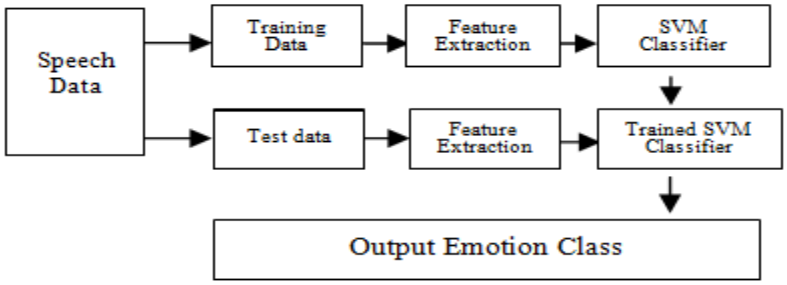


Fig2. SVM classifier working algorithm

2) *Naive Bayes:* Naive Bayes machine learning algorithms are based on unrealistic independent assumptions but still they are attractive due to their simple framework and reasonable performance. Naive Bayes text classifier technique is more efficient and simple to implement as compared to other classifier technique. Pure naïve bayes Classifer model is designed as Multivariate Benoulli Model. To improve performances on rare categories where the model parameters are unreliable, we can introduce multivariate Poisson model for naive Bayes text classification and a weight enhancing method to improve performances on rare categories where the model parameters are unreliable[3].
3) *HMM:* Hidden Markov Models are performed by many ways like traditional HMM, single channel HMM and multi channel HMM. Single channel on the additional benefit of a sub-phoneme speech model at the emotional state level instead of the phoneme level. Multi-channel HMM increases flexibility of single channel HMM. In the multi channel HMM, there are two steps in training phase. The first step requires training of each single-channel HMM to an emotional state while the second step combines the emotion- dependent single-channel HMMs into a multi-channel HMM. HMM priorities stress classification while classifying emotions of the speech followed by speech classification [1].
4) *GMM:* Gaussion Mixture Model or Mixture of Gaussion Model estimates probability density function (PDF) of given emotional Speech data. GMM discovers relation between clusters and speakers. GMM estimates EM algorithm [1].

5) ANN: The short- time features are used as an input to an ANN. Short time features classify utterances into emotional states. In the Artificial Neural Network classifier a phoneme group, such as fricatives (FR), vowels (VL), semi-vowels (SV). An utterance is a speech segment corresponding to a word or a phrase and in ANN classifier, the utterance is partitioned into several bins containing some frames each. ANN approach based classifiers are used for emotion classification due to their ability to find nonlinear boundaries which separate the emotional states. The mostly feed forward class is used that of ANNs, in which the input feature values propagate through the network in a forward direction on a layer-by-layer basis the ANN-based classifiers can be used in two ways: 1.An ANN is trained to all emotional states. 2.A number of ANNs is used, where each ANN is trained to a specific emotional state. In the first case, the number of output nodes of the ANN equals the number of emotional states and in the second case, ANN has one output node [1].

VIII CONCLUSION

In this review different classifiers like Support Vector Machines (SVM), Naïve Bayes classifiers, GMM, ANN, HMM are provided by Machine learning to recognize human emotions from audio signals. We have reviewed to develop a multi task system for classification of male and female actors emotions and emotions from speech and song audio signals using the same set of features. From the audio clips we can calculate various features like maximum and minimum value of signal, mean, standard deviation, dynamic range, crest factor, autocorrelation time, energy entropy, short time energy, ZCR, spectral roll off, spectral centroid and spectral flux. After extraction of audio features we are going to detect the class of emotion by using SVM or Naïve Bayes classifiers. We are also looking for the best technique to be proven very effective for the calculation of human emotions with good recognition rate as well as useful for psychiatrics to determine the mental status of the user as well as it is also effective for E learning to identify the mood of listener. We are going to implement SVM and Naive Bayes techniques using MATLAB software.

The next step will be the identification of features indicating whether the speech contains which kind of emotion; so this can be the challenge for future work which can be done by integrating various techniques from machine learning. Data classification methods and machine learning methods can be combined to improve the accuracy. Future work in expanding existing techniques to handle more linguistic and semantic patterns will surely be an attractive opportunity for researchers and business people alike.

REFERENCES

[1] Speech Emotion Recognition Based on SVM and ANNXianxin Ke, Yujiao Zhu, Lei Wen, and Wenzhen Zhang, vol:8 No:3

[2] K. H. Kim, S. Bang, S. Kim, Emotion recognition system using short-term monitoring of physiological signals, Medical and biological engineering and computing 42 (3) (2004) 419–427.

[3] S.-B. Kim, K.-S. Han, H.-C. Rim, S. H. Myaeng, Some effective techniques for naive bayes text classification, Knowledge and Data Engineering, IEEE Transactions on 18 (11) (2006) 1457–1466.

[4] J. Anttonen, V. Surakka, Emotions and heart rate while sitting on a chair, in: Proceedings of the SIGCHI conference on Human factors in computing systems, ACM, 2005, pp. 491–499.

[5] R. McCraty, M. Atkinson, W. A. Tiller, G. Rein, A. D. Watkins, The effects of emotions on short-term power spectrum analysis of heart rate variability, The American journal of cardiology 76 (14) (1995) 1089–1093.

[6] L. Gavidia-Ceballos, J. H. Hansen, Direct speech feature estimation using an iterative em algorithm for vocal fold pathology detection, Biomedical Engineering, IEEE Transactions on 43 (4) (1996) 373–383.

[7] D. A. Cairns, J. H. Hansen, J. E. Riski, A noninvasive technique for detecting hypernasal speech using a nonlinear operator, Biomedical Engineering, IEEE Transactions on 43 (1) (1996) 35.

[8] C. Pierrakeas, V. Georgopoulos, G. Malandraki, Online collaboration environments in telemedicine applications of speech therapy, in: Engineering in Medicine and Biology Society, 2005. IEEE-EMBS 2005. 27thAnnualInternational Conference of the, IEEE, 2006, pp. 2183–2186.

[9] B. J. Lee, B. Ku, J.-S. Jang, J. Y. Kim, A novel method for classifying body mass index on the basis of speech signals for future clinical applications: A pilot study, Evidence-Based Complementary and Alternative Medicine 2013.

[10] R. Gamasu, Ecg based integrated mobile tele medicine system for emergency health tribulations, Int J Biosci Biotechnol 6 (1) (2014) 83–94.

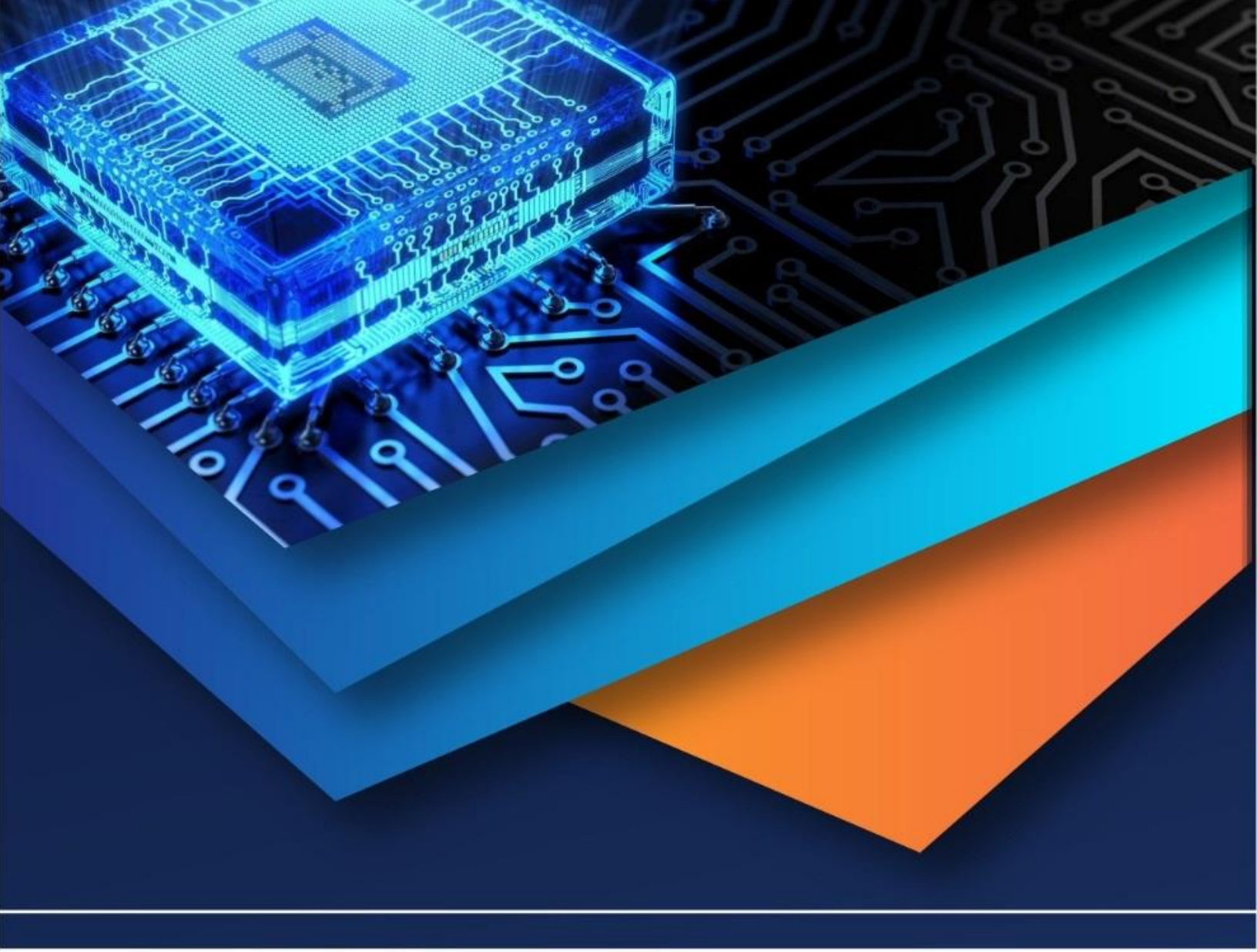
[11] Livingstone SR, Russo FA (2018) The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. PLoS ONE 13(5): e0196391.https://doi.org/10.1371/journal.pone.0196391

[12] Rajni, Dr. Nripendra Narayan Das, Emotion recognition from audio signal, (IJARCET) Vol 5, Issue 6, June 2016

[13] Yizhar Lavner and Dima Ruinskiy,A Decision-Tree-Based Algorithm for Speech/Music Classification and Segmentation, EURASIP Journal on Audio, Speech, and Music ProcessingVolume 2009, Article ID 239892

[14] Jasleen, Dawood Dilber, feature selection and extraction of Audio signal, IJIRSET,Vol. 5, Issue 3, March 2016

[15] Ritu D.Shah, Dr. Anil.C.Suthar, Speech Emotion Recognition Based on SVM Using MATLAB, IJIRCCE, Vol. 4, Issue 3, March 2016



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY