iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 8 **Issue:** I **Month of publication:** January 2020

DOI: http://doi.org/10.22214/ijraset.2020.1007

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 8 Issue I, Jan 2020- Available at www.ijraset.com

Speech Emotion Recognition Systems: Review

Panay Kumar Rahi[1], Dr. Maitreyee Dutta[2]
[1]*M.E. Scholar, Department of Electronics & Communication Engineering, NITTTR, Chandigarh, India.*
[2]*Professor, Department of Computer Science & Engineering, NITTTR, Chandigarh, India.*

***Abstract:** In this paper, survey of different types of speech emotions recognition system are analyzed. In human machine interface application, emotion recognition from the speech signal has been research topic since many years. The classifiers are used to differentiate emotions such as anger, disgust, fear, joy or happiness, sadness, surprise, neutral state, etc. In speech, domain's important factor is to understand the speech. But automated speech detection is a big challenge which is solved by extracting best features of signals (speech). In Emotion Detection, domain has two types of features i.e. Important Utterance and Prosodic features. Finally, in respect of performance speech emotion recognition systems are discussed in the last section of this survey. This section also provides the possible ways of enhancing performance of Speech Emotion Recognition Systems (S.E.R.).*
***Keywords:** Classifier, Emotion Recognition, Feature extraction, SVM's.*

I. INTRODUCTION

The speech signals are more easiest and natural method of communication among persons. This fact has motivated researchers to think of speech as a fast and efficient method of interaction between human and machine [6]. There are different ways for communication but the speech signals are one of the easiest, fastest and most natural techniques of communications between two or more person. Therefore the speech can be the fast and very efficient method of interaction or communication between human and machine [1].

A. Emotions

Emotion is very difficult concepts to understand or define in psychology. There are many different definitions of emotions in the scientific literature. In everyday speech, emotion is any relatively brief conscious experience characterized by intense mental activity and a high degree of pleasure or displeasure [34]. Emotions conveyed in speech can be grouped into two main categories:

1) Consciously Expressed Emotion &
2) Unconsciously Expressed Emotion.

Consciously expressed emotions are usually more obvious compare to unconsciously expressed emotions *i.e.*, when anyone raises their voice in speaking, they are often consciously expressing that they are angry, sad, fear or any other feelings. In other cases, the only indication of a person trying to conceal their anger or any expression or feeling may be a slight terseness to their words with their actual emotions [4]. Identifying emotions in textual input presupposes the existence of some suitable taxonomy of emotional states. Emphasis has traditionally been placed on the set of six "universal" emotions: ANGER, DISGUST, FEAR, JOY, SADNESS, and Surprise [7].

B. Sensory Modalities of Emotions

There is vigorous debate about what exactly individual can express nonverbally. Persons can express their emotions via many different kinds of nonverbal communications *viz* facial expressions, quality of speech and physiological signals. Here we discuss about these categories.

1) *Facial Expressions:* The human faces are wisely expressive, can express countless emotions without saying any single words. The facial expressions are universal unlike other forms of nonverbal communication, In all cultures the facial expressions for happiness, sadness, anger, surprise, fear, and disgust are the same.
2) *Speech:* In emotional expression voices are very important modality in addition with faces. Speech is a relevant communicational channel enriched with emotions: the voice in speech not only conveys a semantic message but also the information about the emotional state of the speaker. There are some important voice feature vectors which have been chosen for research *i.e.,* Fundamental Frequency, Mel-Frequency Cepstral Coefficient (MFCC), Linear Prediction Cepstral Coefficient (LPCC), etc.

3) *Physiological Signals:* The physiological signals are closely related to autonomic nervous systems which allow assessing objectively emotions. They are such as Electro-Encephalo-Gram (EEG), Electro-Cardio-Gram (ECG), Electro-Myo-Gram (EMG), Respiration (RSP), Blood Pressure (BP), Heart Rate (HR), Skin Temperature (ST), Skin Conductance (SC), and Blood Volume Pulse (BVP) [8]. It is also useful for those people who suffer from physical or mental illness so exhibit problems with facial expressions or tone of voice by using physiological signals for recognition of their emotions [9].

II. SPEECH EMOTION RECOGNITION SYSTEM

Speech Emotion Recognition (S.E.R.) systems are actually Pattern Recognition System. The Speech Emotion Recognition (S.E.R.) systems show same stages that are also available in the pattern recognition system. The structure of Speech Emotion Recognition (S.E.R.) systems shown below, it contains five main modules emotional speech input, feature extraction, feature selection, classification, and recognized emotional output [1].

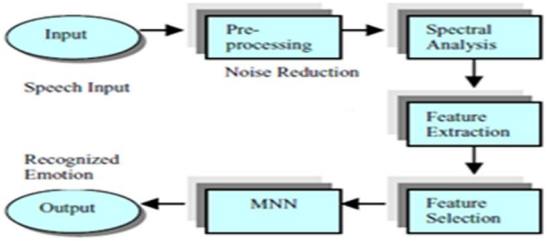


Fig. 1 Structure of Speech Emotion Recognition System

A speech emotion recognition system consists of two stages:

1) A front-end processing unit that extracts the appropriate features from the available (speech) data, and
2) A classifier that decides the underlying emotion of the speech utterance.

A. *Pre-Processing Flow*

These processes are divided into two major categories: first is Speech Processing and second is Emotion Recognition. A speech input (an utterance) is input into the speech processing module. Firstly, the speech features are calculated for those utterances. Than after, the utterances are divided into the number of speech periods. At last, the speech features are extracted for each speech period, and in feature vector, features are compiled for the utterance [4]. The transfer function of the pre- emphasis filter is usually given by:

$$H(z) = 1 - 0.97z^{-1}$$

An important issue in the design of a speech emotion recognition system is the extraction of suitable features that efficiently characterize different emotions.

1) *Prosody and Utterance Features:* In linguistics, prosody features are phonetic segments (vowels and consonants). These contribute to linguistic functions *i.e.*, intonation, tone, stress, and rhythm. Prosody may reflect various kinds of features from the speaker. On the other hand, utterances are the emotional state of the speaker which are in the form of the utterance (statement, question, or command). Prosody is the presence of irony or sarcasm, emphasis, contrast, and focus or other elements of language that may not be encoded by grammar. An Utterance is a smallest unit of speech in analysis of spoken language. It is a continuous piece of speech beginning and ending with a very clear pause. It is generally bounded by silence in the case of oral languages. Utterances do not exist in written language, it appears only in oral, but can be represented and delineated in written language in many ways.

B. *Classification*

After calculation of the features in the Speech Emotion Recognition system (S.E.R.), the best features are provided or given to the classifier. The classifier recognizes the emotion in the speaker's speech utterance. Various types of classifier have been proposed for the task of Speech Emotion Recognition (S.E.R.) [1]. Next, extraction and selection of the best features which are appropriate for the best possible differentiation of emotional states, than in final step these features are used to train the classifier and test whether it can classify different emotional states or not. From the literature, the most popular classification algorithms are described in this section. Various types of classifiers have been used for the task of speech emotion recognition HMM, GMM, SVM, Artificial Neural Networks (ANN), k-NN and many others. The main types of classifiers are as follows:

1) *Quadratic Discriminant Analysis (QDA):* Quadratic Discriminant Analysis (QDA) is very common method to use. It assumes that the likelihood of each class is normally distributed and uses the posterior distributions to estimate the class for a given test point [10]. The normal (Gaussian) parameters of each class are usually estimated from training points with Maximum Likelihood (M.L.) estimation [11].
2) *K-Nearest Neighbor (KNN):* K-Nearest Neighbor (KNN) classifies unlabeled samples or testing data by their similarity with the training data. In general, given an unlabeled sample *X*, the KNN classifier finds the K-Nearest Neighborhood samples in the training data and it labels the sample *X* with the class label that appears most frequently in the K-Nearest Neighborhood samples in the training data [12].
3) *Support Vector Machines (SVMs):* The Support Vector Machines (SVMs) are the classifier that separates or divides a set of objects into classes so that the distance between the class borders is as large as possible. The main role of SVMs is to separate two classes with a hyper-plane so that the minimal distance between elements of both classes and the hyper-plane will maximal [13].
4) *Artificial Neural Network (ANNs):* Artificial Neural Networks (ANNs) are mainly used to estimate or approximate functions which depends on a large number of inputs and that are generally unknown [14]. Basically, Artificial Neural Networks (ANNs) are computational models. They use the idea of natural neurons that receive signals and exchange messages among each other [15].

C. *Features Extraction*

Traditional speech features are typically extracted from power spectrum or amplitude spectrum of speech signals. The speech signal contains a large amount or number of parameters which reflects all type of emotional characteristics. The main point in emotion recognition is that what kind of features should be used. In context with the recent researches, many common features are extracted, *i.e.*, energy, pitch, formant, and some spectrum features such as Mel-Frequency Cepstrum Coefficients (MFCC), Linear Prediction Coefficients (LPC) and Modulation Spectral Features (MSF). The structure of feature extraction is shown below in figure.

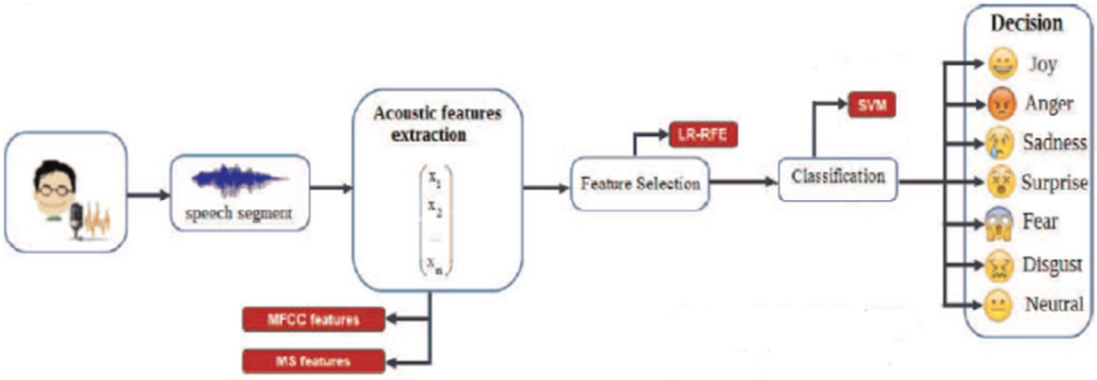


Fig. 2 Structure of Speech feature extraction

Classification of Speech Features

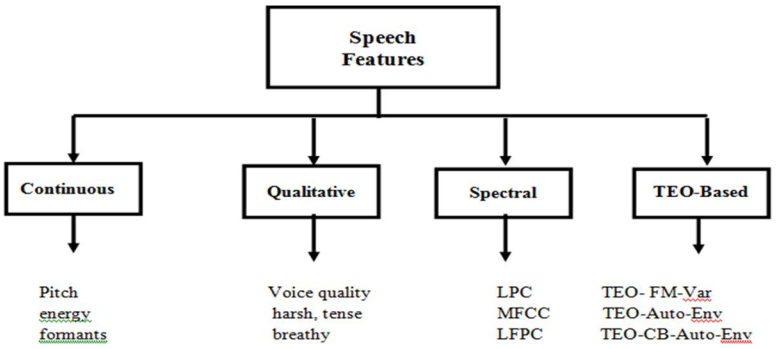


Fig. 3 Classification of Speech Features

In accordance to several studies, features can be divided into the following categories:

1) Pitch-related features,
2) Formants features,
3) Energy-related features,
4) Timing features, and
5) Articulation features [16].

III. LIERATURE REVIEW

The main objectives of the literature reviews are to develop ideas and established knowledge on a particular topic. The literature review is a piece of discursive prose. It helps in problem identification, formulation, evaluation and shapes path for conducting research in particular direction with use of efficient, appropriate scientific approach, methods or techniques or combination of techniques.

In [1] reviewed speech emotion recognition based on the previous or past technologies in which they use different classifiers for the recognition of emotions. The classifiers are used to differentiate or distinguish universal emotions such as anger, happiness, sadness, surprise, neutral state, etc. The database for the speech emotion recognition system is the emotional speech samples and the features extracted from these speech samples are the energy, pitch, Linear Prediction Cepstrum Coefficient (LPCC), Mel Frequency Cepstrum Coefficient (MFCC). The classification performance is based on extracted features. Inference about the performance and limitation of speech emotion recognition system based on the different classifiers are also discussed.

In [2] proposed a three-level speech emotion recognition model to solve the speaker independent emotion recognition problem, which classify six speech emotions, including sadness, anger, surprise, fear, happiness and disgust from coarse to fine. The appropriate features are selected from 288 candidates by using Fisher rate which is also regarded as input parameter to Support Vector Machine (SVM) for every level. In order to validate and evaluate the proposed system, Principal Component Analysis (PCA) for dimension reduction and Artificial Neural Network (ANN) for classification are adopted. It is used to design four comparative experiments, including Fisher + SVM, PCA + SVM, Fisher + ANN, PCA + ANN. The experimental results proved that Fisher is better than Principal Component Analysis (PCA) for dimension reduction and Support Vector Machine (SVM) is more expansible as compare to Artificial Neural Network (ANN) for speaker independent speech emotion recognition. The average recognition rates for each proposed level are 86.5%, 68.5% and 50.2% respectively.

In [3] introduced the basic course of speech emotion recognition, which includes processing of speech signal, speech feature extraction and speech emotion recognition. After choosing the useful features such as Mel-Frequency Cepstral Coefficients (MFCC) and its transient parameters, a better performance with the application of Back Propagation Neural Networks (BPNNs) is obtained. Furthermore, the decision trees with multi-features are used to recognize speech emotion for comparison.

In [4] discussed the significance of emotion recognition in speech and applicable research topic, and also proposed a efficient system for recognition of emotion using one class- in-one neural networks. The proposed system is speaker and context independent with the use of a large database of phoneme balanced words and also achieved a recognition rate of approximately 50% when testing on eight emotions.

In [5] improved automatic emotion recognition from speech by incorporating rhythm and temporal features. The Automatic Emotion Recognition researches are mainly based on applying features like MFCC's, pitch and energy/intensity. The idea focuses on borrowing rhythm features from linguistic and phonetic analysis and applying them to the speech signal on the basis of only acoustic knowledge. The set of temporal and loudness features are also exploit in addition to that. On different segments, a segmentation unit is employed in starting to separate the voiced/unvoiced and silence parts and features are explored. After that various classifiers are used for classification of emotions. On the Berlin Emotion Database, after selecting the top features using an IGR filter are able to achieve a recognition rate of 80.60 % for the speaker dependent framework.

In [6] proposed a novel deep neural architecture to extract the informative feature representations from the heterogeneous acoustic feature groups which may contain redundant and unrelated information leading to low emotion recognition performance. After obtaining the informative features, a fusion network is trained to jointly learn the discriminative acoustic feature representation and a Support Vector Machine (SVM) is used as the final classifier for recognition task. Experimental results on the IEMOCAP dataset demonstrate that the proposed architecture improved the recognition performance, achieving accuracy of 64% compared to existing state-of-the-art approaches.

IV. COMPARATIVE ANALYSIS

Review Details	Technology Used	Descriptions
In [1]	Reviewed	Features extracted from these speech samples are the energy, pitch, LPCC MFCC
In [2]	Four Comparative Experiments: Fisher + SVM, PCA + SVM, Fisher + ANN, PCA + ANN	The average recognition rates for each proposed level are 86.5%, 68.5% and 50.2% respectively.
In [3]	MFCC and its transient parameters	The better performance with the application of Back Propagation Neural Networks (BPNN's) is obtained.
In [4]	Uses large database of phoneme balanced words	Achieved a recognition rate of approximately 50% when testing on eight emotions.
In [5]	Applied features like MFCC's, pitch and energy/ intensity and Berlin Emotion Database	Achieved a recognition rate of 80.60 % for the speaker dependent framework
In [6]	Discriminative Acoustic feature representation and a Support Vector Machine (SVM) and IEMOCAP Dataset	Achieved accuracy 64% compared to existing state-of-the-art approaches

V. CONCLUSION

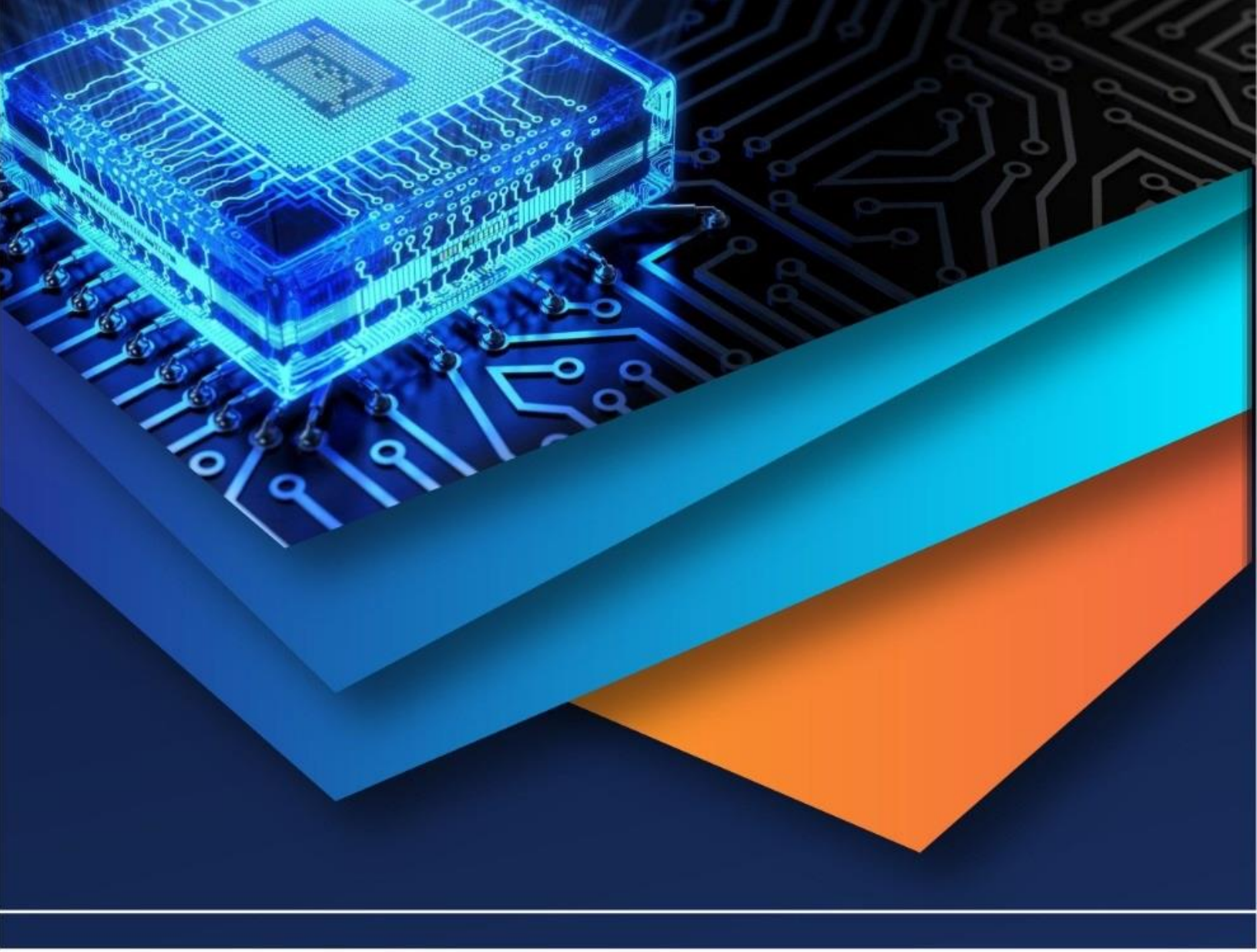
The main objective of this paper was to analyze the strengths and weaknesses of various techniques adopted in the various types or kind of Emotion Recognition System (S.E.R.). From most of literature survey it can be observed that majority of the work has been done in speech classification domain. The knowledge of speech is important for selecting the emotion, which is not represented by energy and mean of speech signal. In emotion classification, if number of emotion is increased error will also increased. It is concluded from the above that Support Vector Machine's (SVM's) gives better Recognition Rate as compare to others.

REFERENCES

[1] Ashish B. Ingale and D. S. Chaudhari, "Speech Emotion Recognition", International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-2, Issue-1, pp. 235-238, March' 2012.
[2] Lijiang Chen, Xia Maoa, Yuli Xue and Lee Lung Cheng, "Speech Emotion Recognition: Features and Classification Models", ELSEVIER, pp. 1154-1160, May' 2012.
[3] Ying Shi and Weihua Song, "Speech Emotion Recognition Based on Data Mining Technology", IEEE Sixth International Conference on Natural Computation (ICNC), pp. 615-619, 2010.
[4] J. Nicholson and K. Takahashi and R. Nakatsu, "Emotion Recognition in Speech using Neural Networks", Springer, pp. 290-296, 2000.
[5] Mayank Bhargava and Tim Polzehl, "Improving Automatic Emotion Recognition from speech using Rhythm and Temporal feature", ICECIT, pp. 139-147, ELSEVIER, 2012.
[6] Wei Jiang Zheng Wang, Jesse S. Jin, Xianfeng Han 1,and Chunguang Li, "Speech Emotion Recognition with Heterogeneous Feature Unification of Deep Neural Network", Sensors 2019.
[7] Guojun Zhou, John H.L. Hansen and James F. Kaiser, "Classification Speech under Stress based on Featured Serviced from the Nonlinear Teager Energy Operator'', IEEE Conference, Vol. 1, pp. 549-552, May 1998.
[8] He C, Yao Yj, Ye Xs, " An Emotion Recognition System based on Physiological Signals obtained by Wearable Sensors. In: Wearable Sensors and Robots, Springer, pp. 15-25, 2017.
[9] Leila Kerkeni, Youssef Serrestou, Mohamed M. Barki, Kosai Raoof and Mohamed Ali Mahjoub," Speech Emotion Recognition: Methods and Cases Study", ICAART 2018 - 10th International Conference on Agents and Artificial Intelligence, pp. 175-182, 2018
[10] Hastie, T., Tibshirani, R., Friedman, J., Hastie, T., Friedman, J., Tibshirani, R., "The Elements of Statistical Learning", Vol. 2. Springer, Berlin, 2009.
[11] Scholz, F.W.: Maximum likelihood estimation. Encyclopedia of Statistical Sciences, 1985.
[12] Peterson, L.E: K-nearest neighbor. Scholarpedia 4(2), 1883 (2009)
[13] Cortes, C., Vapnik, V.: Support-vector networks. Mach. Learn. 20(3), 273–297 (1995)
[14] Yegnanarayana, B.: Artificial Neural Networks. PHI Learning Pvt. Ltd., New Delhi (2009)
[15] Mouhannad Ali, Ahmad Haj Mosa, Fadi Al Machot and Kyandoghere Kyamakya, "Emotion Recognition Involving Physiological and Speech Signals: A Comprehensive Review", Springer International Publishing, pp. 287-302, AG 2018.
[16] Moataz El Ayadi, Mohamed S. Kamel and Fakhri Karray, "Survey on Speech Emotion Recognition: Features, Classification Schemes and Databases", Journals on Pattern Recognition (ELSEVIER), pp. 572-587, 2011.

Pranay Kumar Rahi received the Bachelors of Technology Degree in Electronics and Telecommunication Engineering from Government Engineering College, Guru Ghasidas University, Bilaspur, Chhattisgarh, India in 2004, and pursuing Masters of Engineering from National Institute of Technical Teacher's Training & Research, Punjab University, Chandigarh, India. Presently working as Assistant Professor in Department of Electrical and Electronic Engineering, Institute of Technology Korba, Chhattisgarh since 2008. He has authored more than 40 research publications and published Journal papers in the leading International and National journal.

Dr. Maitreyee Dutta has been working as Professor, Computer Science & Engineering department at NITTTR, Chandigarh. She received the Bachelor of Engineering degree from Guwahati University, Master of Technology and Ph.D from Panjab University, Chandigarh. Her research work focused on Digital Signal Processing, Image Processing, Advanced Computer Architecture and Data Warehousing & Mining. She is author and co-author of more than 50 scientific papers, published in peer-reviewed national or international journals and conferences, guided more than 100 M.Tech dissertations so far. She has guided more than 5 Ph.D students. She is a member of various professional bodies.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY