Classification of Emotional Speech of Children Using Probabilistic Neural Network

H.K. Palo, Mihir Narayan Mohanty
Departement of Electronics and Communication Engineering, Institute of Technical Education and Research, Siksha 'O' Anusandhan University, Bhubaneswar, Odisha, India

Article Info

Article history:

Received Jan 2, 2015
Revised Feb 10, 2015
Accepted Feb 25, 2015

Keyword:

ANN
Hurst parameter
LP-VQC
pH feature vectors
PNN

ABSTRACT

Child emotions are highly flexible and overlapping. The recognition is a difficult task when single emotion conveys multiple informations. We analyze the relevance and importance of these features and use that information to design classifier architecture. Designing of a system for recognition of children emotions with reasonable accuracy is still a challenge specifically with reduced feature set. In this paper, Probabilistic neural network (PNN) has been designed for such task of classification. PNN has faster training ability with continuous class probability density functions. It provides better classification even with reduced feature set. LP_VQC and pH vectors are used as the features for the classifier. It has been attempted to design the PNN classifier with these features. Various emotions like angry, bore, sad and happy have been considered for this piece of work. All these emotions have been collected from children in three different languages as English, Hindi, and Odia. Result shows remarkable classification accuracy for these classes of emotions. It has been verified in standard databse EMO-DB to validate the result.

Corresponding Author:

Mihir Narayan Mohanty
Departement of Electronics and Communication Engineering,
Institute of Technical Education and Research,
Siksha 'O' Anusandhan University, Bhubaneswar, Odisha, India
Email: mihirmohanty@soauniversity.ac.in

1. INTRODUCTION

Human speech is the linguistic act that conveys information about the speaker. Although human emotions cannot change the content, but it is expressed in separate way from the normal speech. This information of a speaker can effect his/her mental states. It is challenging in extracting suitable features of human emotional speech that can best represent a particular emotion unambiguously. A comparison between different division of emotions as hard-wired vs. socially learned and primary' vs. 'secondary' emotions with its universality in containing various classes of emotions is presented in [1]. A review on different standard emotional databases used by different researchers, their accessibility and performance can be found in [2-3]. Different acoustic features as fundamental frequency (F0), energy and duration were explored in many cases as parameters of speech emotional utterances [2-3]. In [4], statistical properties of various acoustic speech emotional features like zero crossing rate (ZCR), Harmonics-to-Noise-Ratio (HNR), formant were derived to represent emotional utterances of Berlin databases. Recently researchers have focused on determining Hurst parameter and on time-frequency pH features for emotional speech [5-7]. EMO-DB database is a standard database as expressed in [8]. It has been used in most of the cases. Variouse features such as Linear predictor coefficients (LPC), Linear predictor cepstral coefficients (LPCC), Mel frequency cepstral coefficients (MFCC), Perceptual linear prediction (PLP), LP_VQC and pH vectors are used for emotional speech recognition are found in literature [9-12]. Similarly, classifiers like Multilayer perceptron (MLP), Radial basis function network (RBFN) are used for classification of different emotions by authors [10-12]. H.K. Palo et. al. emphasized on child emotional speech classification using different neural network models in [11-12].Use of K-nearest neighbourhood Linear discriminant classifier (LDC) for classification of emotions have been also used by many researchers [13-16]. A comparative study on Multilayer Percepton, Random Forest, Probabilistic Neural Networks and Support Vector Machine also has been reported for the different speech emotion classification [17-20]. Probabilistic neural network has been used in many pattern classification tasks [21-22]. Decision Trees, Artificial Neural Networks (ANN), Probabilistic neural network and random forest techniques are applied for classification of different signals [19-24]. Over the last decades different ANN classifier's performance and their comparison in classifying speech and emotions were explored by many researchers, while a little amount of work has been reported for classification of child emotions. Motivated by the flexibility incorporated in child emotions, authors have taken an attempt to classify four classes of emotions as angry, bore, sad and happy using LP_VQC and pH feature vectors with PNN classifiers.

The paper is organized as follows. Section 2 deals with the research method. In this section both the methods for classification and feature extraction are presented. The result is discussed in Section 3 and finally in section 4, it concludes this piece of work.

2. RESEARCH METHOD

2.1. Method of Classification

The choice for an appropriate classifier for emotional speech is a complex task. Popular classifiers for emotion recognition such as Linear Discriminant Classifiers (LDCs) and k-Nearest Neighbour (kNN) classifiers have been used in literature [13-16]. Although PNN is not necessarily the best classifier but it provides good statistical properties. Classification accuracies of Decision Trees such as RF, Artificial Neural Networks (ANN) and Probabilistic neural network were found to be similar. ANN required by far the highest calculation times, whereas the training and testing of RF took usually longer than PNN. Artificial Neural Network has many disadvantages, such as complex optimization, low robustness and much training time. Random Forest in contrast is easy to use, since only one variable needs to be set by the user. However, its classification accuracies can not satisfy the machine-learning methods whereas its robustness was among the best [23-24].

PNN is a special class of Radial basis function (RBF) network used for classification [17-18, 20-24]. It is beneficial to many appilications including speech because of the speed of learning. It is one of the non-parametric methods. This type of network adopts probabilistic method to classify data. The PNNs are effective discriminative classifiers with several outstanding characteristics, namely: they are having a order of magnitude much faster and accurate than multilayer perceptron networks. The networks are relatively insensitive to outliers having an inherently parallel structure and guaranteed to converge to an optimal classifier as the size of the training set increases. PNN is based on Bayes optimal classification. PNNs can generate accurate predicted target probability scores with no local minima issues. No extensive retraining is necessary if training samples are added or removed. These characteristics have made PNNs very popular and successful.

In PNN, the operations are organized into multilayered feed forward network with four layers: inputlayer, pattern layer, summation and output layer. Fundamentally it is based on Bayesian classifier technique.The first layer simply distributes the input to the neurons in the pattern layers. This layer using the given set of data points as the centers forms the Gaussian functions. It organizes the training set such that each input vector is represented by an individual processing layer called the summation layer. There are as many processing elements as the number of classes to be recognized in the summation layer. The distances from the input testing vectors to the input training vectors is computed and a vector that indicates the closeness of the training and testing inputs is produced. For each class of testing inputs the summation layer sums the contributions of previous layer output by giving a net output vector of probabilities. Basically an averaging operation of the outputs from the pattern layer for each class is performed by the summation layer A PNN uses Parzen window probabilistic density function estimator for each of the classes with a mixture of Gaussian basis functions [22]. Figure 1 shows the basic structure of probabilistic neural network.

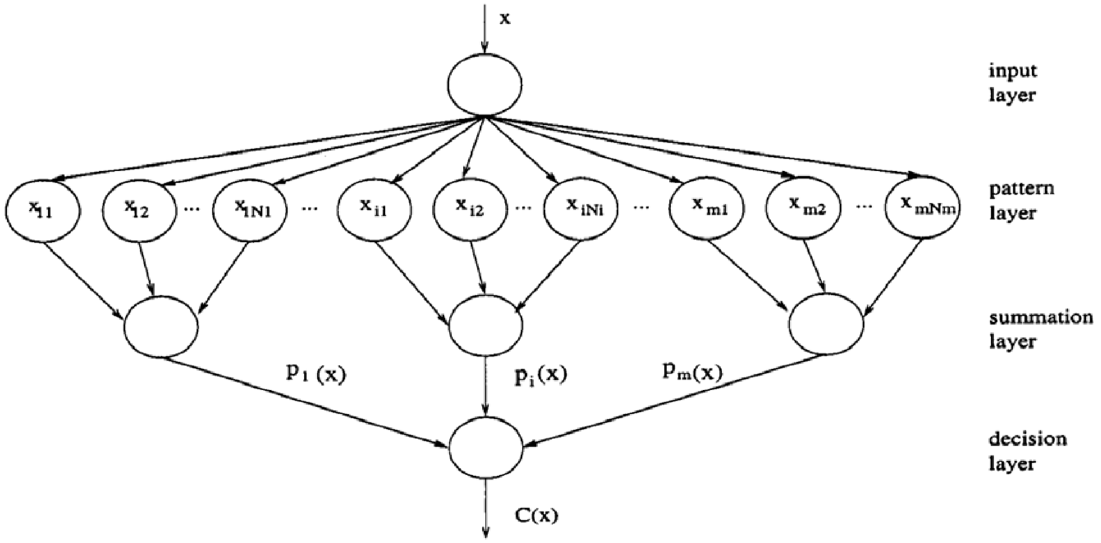

Figure 1. Basic Probabilistic Neural Network Structure

For an input pattern x_p , pattern vector dimension *d,* smoothing parameter σ, with each input training vector $x_{i,j}$ considered to be a center of a kernel function the output of the pattern layer is given by [21]

$$\emptyset_{i,j} = \frac{1}{(2\pi)^{\frac{d}{2}}\,\sigma^d} exp\left[-\frac{\left(x_p - x_{i,j}\right)^T \left(x - x_{i,j}\right)}{2\sigma^2}\right] \qquad (1)$$

For c classes of emotions the probability density function $P_i(x)$ of each class c_i is given by the equation

$$P_i\left(x_p\right) = \frac{1}{(2\pi)^{\frac{d}{2}}\,\sigma^d}\frac{1}{N_i}\sum_{j=1}^{N_i} exp\left[-\frac{\left(x_p - x_{i,j}\right)^T \left(x_p - x_{i,j}\right)}{2\sigma^2}\right] \qquad (2)$$

Where i=1,2,…, *m* and N_i is total number of training patterns for each class c_i ; x_p is the p^{th} input vector. Here *m* takes the value of four corresponding to angry, bore, sad and happy speech emotions. The variance σ should be chosen judiciously as it acts as a smoothing factor to soften the surface. For small σ the classifier may not generalize well and can create distinct modes. Larger σ allows interpolation between points. Very large σ approximate the pdf to be Gaussian resulting loss of details.

To distinguish class c_i to which input vector x_p belongs the Bayesian decision rule is applied. Here it is assumed that *a priori* probabilities and losses associated with incorrect decision is same for each class. By this rule the estimated class $\hat{C}(x)$ of the pattern x_p , based on the output of all the summation layer neurons is given by

$$\hat{C}(x) = \arg\max\{p_i(x)\},\ \ i = 1,2,\dots,m \qquad (3)$$

2.2. Feature Extraction

Selecting a suitable feature that can best represent a particular emotion is one of the foremost works in the field of emotions recognition [11]. Since child emotions are very flexible the selection becomes more tedious. In order to capture different aspects of vocal tract mechanism both time-frequency parameter and spectral parameters as features have been attempted. Further the effect of feature reduction capability of vector quantization with LPC spectral feature were compared with pH feature vectors for PNN classifiers. The two different methods for feature extraction is explained as follows.

pH time-frequency feature vectors

pH is a time-frequency vocal source feature [7-8]. It explores the time-frequency multi-resolution capability of discrete wavelet transform (DWT) to capture the higher order correlations of the speech samples thus able to classify the emotions in a better way. It consists of a vector of Hurst component H $(0 < H < 1)$ obtained from each frame of the signal concerned. Since Hurst parameter closely related to the decaying rate of autocorrelation coefficient function (ACF), $\rho(k)(-1 < \rho(k) < 1$. The relationship of the Hurst parameter as $k \to \infty$ is given as

$$\rho(k) \sim H(2H-1)k^{2(H-2)} \tag{4}$$

A value of H values of $(0 < H < 0.5)$ indicates emotions with high energy, since the dominant high frequency components have a -9dB/octave roll-off and the ACF rapidly decays to zero. Emotions with lower energy tends to have H values of $(0.5 < H < 1)$ with -15 dB/octave PSD roll-off and leads to a slowly vanishing ACF. ReScaled adjusted range (R/S) statistic, Higuchi method and Abry-Veitch (AV) estimator can be used to extract the Hurst parameter. However Abry–Veitch (AV) estimator using wavelet decomposition is less time consuming and does not assume the speech signals to be fractal as in former methods hence was a natural choice. The steps of extracting features are depicted below.

- Wavelet decomposition: By applying discrete wavelet transform (DWT) the approximation $(a(j,k))$ and detail $(d(j,k))$ coefficients of the concerned signal is obtained. Here j,k are the decomposition scale and coefficient index of each scale respectively.
- Hurst exponent computation (HC): The variance $\sigma_j^2 = (1/n_j)\sum_k d(j,k)^2$ is calculated for each scale j, n_j being number of available coefficients in each scale j. By a weighted linear regression The slope α of the plot $y_j = \log_2(\sigma_j^2)$ versus j is obtained and the Hurst parameter is computed using the relations

$$E(\sigma_j^2) = C_H j^{2H-1} \tag{5}$$

$$H = (1+\alpha)/2 \tag{6}$$

where C_H is a constant.

- Hurst parameter is computed from each segments of emotional speech signal and from the entire signal. pH vector composition is done by taking *(J+1)* values of$[H_0, H_1, \dots \dots, H_J]$.

LP_VQC features

Speech sample can be approximated as a linear combination of previous samples in LPC as described by the relation,

$$S(n) \approx a_1 s(n-1) + a_2 s(n-2) + \cdots a_p s(n-p) \tag{7}$$

Initially each utterance of emotional signal is segmented into frames to make a statistically non-stationary signal to nearly stationary followed by a Hamming window to compensate for sharp boundary discontinuities. Next to it, LPC coefficients are extracted using LP analysis. To apply vector quantization technique speech signal is divided into number of blocks from which the maximum of each block is derived to generate the code book. A code book vector is selected from similar speech samples. Based on minimum distance from the maximum signal of a block is selected to develop the code book index of that particular block. For each block above procedure is repeated. Application of vector quantization to LPC features gives the LP_VQC feature vectors.

3. RESULTS AND ANALYSIS

A database of 500 utterances from children has been developed. The database has been prepared for four emotions boredom, angry, sad and happy of five children (three boys and two girls) in the age group of six to thirteen with duration of five seconds. A cross validation approach is provided by randomly partitioning the input data into training, validation and testing sets in the ratio 0.6, 0.2 and 0.2 respectively. Testing set has been used to find the performance of the PNN classifier while adjustment of network design parameter has been performed by the validation set. The classifier performance has been studied with various spreading parameter of 0.25, 0.42 and 1.1. A spread constant of 1.1 is selected for our study as it provided maximum classification accuracy for all the four classes of emotions.

Table 1 shows the classification error in percentage for four classes of emotions with different feature set. PNN classifier indicates the error percentage of pH is lower than the LP_VQC features for all four classes of speech emotions. However the average classification error is around 04.89 percent lower in case of pH vectors as shown in Table 2. The same is also represented graphically in Figure 2. The average classification error is shown in Figure 3 graphically. The classification error found lower for child emotion with PNN classifier. It has been tested with EMO-DB data base also, though there is no availability of children database.

Table 1. Classification error in % Using PNN

Feature	Angry	Bore	Sad	Happy
LP_VQC	14.17	17.09	12.43	16.87
pH	09.17	07.19	12.35	11.25

Table 2. Average Classification error in % Using PNN

Feature	Average % Classification error
LP_VQC	15.14
pH	10.25

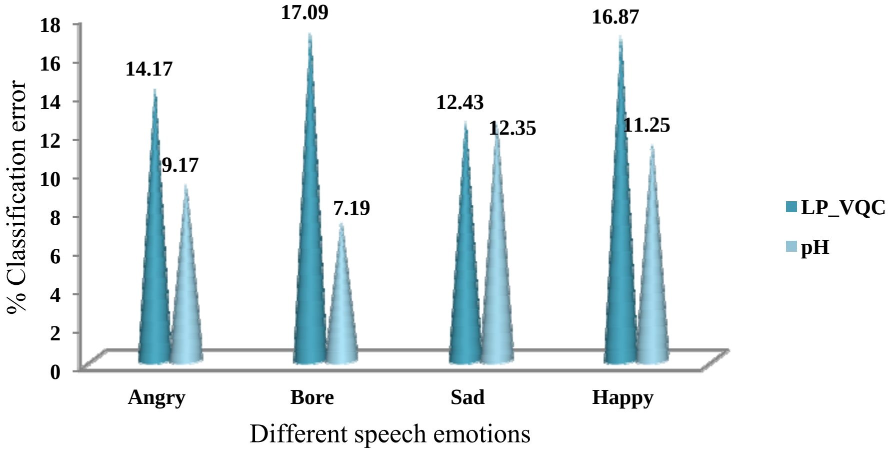

Figure 2. Percentage classification error for different emotional speech

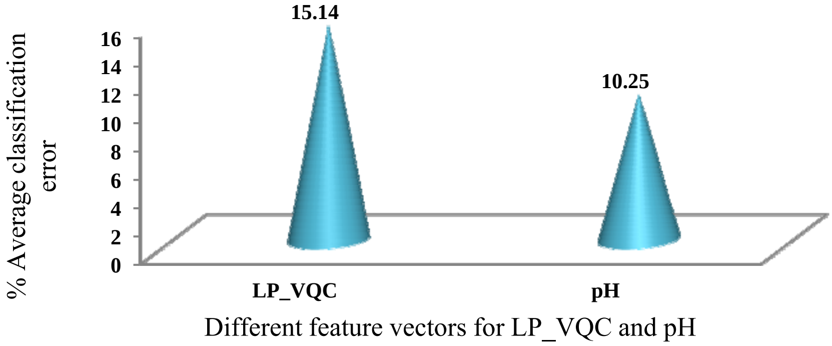

Figure 3. Percentage average classification error for various emotional speech

4. CONCLUSION

pH vector features prove to be more robust than LP_VQC spectral features. It can be applied for individual emotion for futher verification. By applying the proposed method, the importance of features could be evaluated along with the performance of the classifier. The results will be helpful for the later research on the emotion classification. Our work is intended to recognize every day activities based on emotional signals. In this paper a number of different techniques were investigated and applied as feature extractors and classification method was used and compared. Each classification step also becomes simpler and accurate as it processes a limited number of features and classes. When designing these systems, it is very important to detect situations that might affect subsequent performance. In future, audio-based classification scheme can be tested with the development of autonomous systems to recognize. The performance of newly proposed feature set was compared and found to be effective in majority of the cases. Reduction of feature set even another important technique can be tested for the proposed classified. These may be kept for future work.

REFERENCES

[1] N. Fragopanagos, J.G. Taylor. "Emotion recognition in human–computer interaction". *Neural Networks*. vol. 18, pp. 389-405, Mar. 2005.

[2] D. Ververidis, C. Kotropoulos. "Emotional speech recognition: Resources, features and methods". *Speech Communication, Elsevier*. vol. 48, no. 9, pp. 1162-1181, Apr. 2006.

[3] M.E. Ayadi, M.S. Kamel, and F. Karray. "Survey on speech recognition: Resources, features and methods". *Pattern Recognition*. vol. 44, pp. 572-587, Mar. 2011.

[4] B. Schuller et.al. "Timing Levels in Segment-Based Speech Emotion Recognition". *ICSLP, INTERSPEECH 2006*. pp. 1818-1821, 23-27 Sept. Pennsylvania 2006.

[5] E. Hurst. "Long-term storage capacity of reservoirs". *Trans. Amer. Soc.Civil Eng*. pp. 770-799, Apr. 1951.

[6] T. Higuchi. "Approach to an irregular time series on the basis of the fractal theory". *Physics D*. vol. 31, pp. 277-283, 1988.

[7] Ricardo Sant'Ana, Rosângela Coelho, and Abraham Alcaim. "Text-Independent Speaker Recognition Based on the Hurst Parameter and the Multidimensional Fractional Brownian Motion Model". *IEEE Transactions on Audio, Speech, and Language Processing*. vol. 14, no. 3, pp.931-940, May 2006.

[8] L. Zao et.al. "Time-Frequency Feature and AMS-GMM Mask for Acoustic Emotion Classification". *IEEE Signal Processing Letters*. vol. 21, no. 5, pp.620-624, May 2014.

[9] Thomas F. Quatieri. "*Discrete-Time Speech Signal Processing*". Prentice-Hall, 3rd edition. 1996.

[10] H.K. Palo, Mihir Narayan Mohanty, Mahesh Chandra. "Design of Neural Network Model for Emotional Speech Recognition". *Artificial Intelligence and Evolutionary Algorithms in Engineering Systems*. vol. 325, pp. 291-300, Springer India, Nov. 2014.

[11] H.K. Palo, Mihir Narayana Mohanty, Mahesh Chandra. "Novel Feature Extraction Technique for Child Emotion Recognition". *International Conference on Electrical, Electronics, Signals, Communication and Optimization (EESCO)-2015, IEEE*. Jan. 2015.

[12] H.K. Palo, Mihir Narayan Mohanty, Mahesh Chandra. "Use of Different Features for Emotion Recognition Using MLP Network". *Advances in Intelligent systems and computing*. vol. 332, pp.7-15, Springer India, Jan. 2015.

[13] Kwon, O.W., Chan, K., Hao, J., Lee, T.W. "*Emotion recognition by speech signals*". In: Proc. Interspeech. pp. 125–128, 2003.

[14] Batliner, A., Fischer, K., Huber, R., Spilker, J., No¨ th, E. "*Desperately seeking emotions: actors, wizards, and human beings*". In: Proc. ISCA Workshop on Speech and Emotion, Newcastle, Northern Ireland. pp. 195–200, 2000.

[15] Shami, M., Verhelst, W. "Automatic classification of expressiveness in speech: a multi-corpus study". In: Mu¨ ller, C. (Ed.), *Speaker Classification II, Lecture Notes in Computer Science/Artificial Intelligence*. vol. 4441. Springer, Heidelberg–Berlin–New York. pp. 43–56, 2007.

[16] Chuang, Z.J., Wu, C.H. "*Emotion recognition using acoustic features and textual content*". In: Proc. ICME, Taipei, Taiwan, pp. 53– 56. 2004.

[17] T. Iliou et.al. "Classification on Speech Emotion Recognition -A Comparative Study". *International Journal on Advances in Life Sciences*. vol 2 no 1 & 2, pp. 18-28 , 2010.

[18] Nermine Ahmed Hendy and Hania Farag. "Emotion Recognition Using Neural Network: A Comparative Study". *World Academy of Science, Engineering and Technology*. vol.7. pp. 1149-1155, Mar. 2013.

[19] M.M. Javidi and E.F. Roshan. "Speech Emotion Recognition by Using Combinations of C5.0, Neural Network (NN), and Support Vector Machines (SVM) Classification Methods". *Journal of mathematics and computer Science*. vol. 6, pp. 191-200, Apr. 2013.

[20] Mihir Narayan Mohanty, A Routray. "Machine Learing Approach for Emotional Speech Classification". *SEMCO-14*, Book chapter, Springer /Verlag Berlin Heidelberg, 2014.

[21] K.Z. Mao et.al. "Probabilistic Neural-Network Structure Determination for Pattern Classification". *IEEE Transactions on neural networks*. vol. 11, no. 4, pp. 1009-1016. Jul. 2000.

[22] Mihir Narayan Mohanty, V. kumar, A Routray, P. Kabisatpathy. "Classification of Power Quality Disturbances Using Parzen Kernels". *International Journal of Emerging Electric Power Systems*. vol.11, Issue 1, ISSN (Online) 1553-779X, DOI: 10.2202/1553-779X.2335, pp. 1-13, Jan. 2010.

[23] Chuang, Z.J., Wu, C.H. "*Emotion recognition using acoustic features and textual content*". In Proc. ICME, Taipei, Taiwan. pp. 53- 56, 2004.

[24] Mc Gilloway, S., Cowie, R., Doulas-Cowie, E., Gielen, S., Westerdijk, M., Stroeve, S. "*Approaching automatic recognition of emotion from voice: A rough benchmark*". In Proc. ISCA Workshop on Speech and Emotion, Newcastle, Northern Ireland. pp. 207-212, 2000.

BIOGRAPHIES OF AUTHORS

Hemanta Kumar Palo has completed his 'A.M.I.E.' from "Institute of Engineers", India in 1997 and his Master of Engineering from "Birla Institute of Technology", Mesra, Ranchi in 2011. He completed his 'Diploma in Rail Transport and Management' from " Institute of Rail Transport", India in 2003. He is having 20 years of experience in the field of Electronics and Communication Engineering from 1990 to 2010 in Indian Air Force and was an Assistant Professor in Gandhi Academy of Technology and Engineering, Odisha, in the Department of ECE from 2010 to 2012. He is the life member of IEI, India and is the memberof IEEE. Currently he is serving as an Assistant Professor in the department of Electronics and Communication Engineering of Electronics in the Institute of Technical Education and Research, Siksha „O" Anusandhan University, Bhubaneswar, Odisha, India.

Mihir Narayan Mohanty is presently working as an Associate Professor in the Department of Electronics and Communication Engineering, Institute of Technical Education and Research, Siksha „O" Anusandhan University, Bhubaneswar, Odisha. He has published over 100 papers in International/National Journals and Conferences along with approximately 20 years of teaching experience. He is the active member of many professional societies like IEEE, IET, IETE, EMC & EMI Engineers India, IE (I), ISCA, ACEEE, IAEng etc. He has received his M.Tech. degree in Communication System Engineering from the Sambalpur University, Sambalpur, Odisha. Now he has done his Ph.D. work in Applied Signal Processing. He is currently working as Associate Professor and was Head in the Department of Electronics and Instrumentation Engineering, Institute of Technical Education and Research, Siksha O" Anusandhan University, Bhubaneswar, Odisha. His area of research interests includes Applied Signal and image Processing, Digital Signal/Image Processing, Biomedical Signal Processing, Microwave Communication Engineering and Bioinformatics.