Local Linear Wavelet Neural Network and RLS for Usable Speech Classification

Suchismita Sahoo[1], Sushree Sangita Sahoo[2], M.R.Senapati[3] and P.K.Dash (SMIEE)[4]

[1] Department of Computer Science and Engineering
M.Tech Scholar, Koustuv Institute of Self Domain
Biju Patnaik University of Technology

[2] Department of Information Technology
M.Tech Scholar, Koustuv Institute of Self Domain
Biju Patnaik University of Technology

[3] Department of Computer Science and Engineering
Gandhi Engineering College
Biju Patnaik University of Technology

[4] Director Research
S'O'A University

Abstract

While operating in a co -channel environment, the accuracy of the speech processing technique degrades. When more than one person is talking at same time, then there occurs the co-channel speech. The objective of usable speech segmentation is identification and extraction of those portions of co-channel speech that are degraded in a negligible range but still needed for various speech processing application like speaker identification. Some features like usable speech measures are extracted from the co-channel signal to differentiate between usable and unusable types of speech. The features are extracted recursively by this new method and variable length segmentation is carried out by making sequential decision on class assignment of LLWNN pattern classifier. The correct classification using this technique is 84.5% whereas the false classification is 15.5%. The result shows that the proposed classifier gives better classification and is robust.

Keywords: *Co-channel Speech; Usable Speech, Sequential Detection; LLWNN; RLS; Speaker Identification.*

1. Introduction

Studies till date has revealed that speech has enough information about the speaker to perform reliable speaker identification even when the overall target speech energy is equal to that of the interferer speech energy. The performance of speaker identification system degrades under some co-channel conditions, i.e., at a particular instant of time, when two persons are talking, a signal of speech from the prominent speaker is termed as target speech. The signal from the interfering speaker is referred as interferer speech. It was found that, the Target-to- Interferer Ratio (TIR) (i.e. energy ratio of the target and the interfering speech signals) is a good metric to identify portions of the speech data usable for speaker identification [1]. Speech signals are first transformed away from the time domain, which is a p rocess of feature extraction or preprocessing. The inevitable loss of possibly useful information for speech discrimination tasks is due to pre-processing.

For the speaker identification system, studies have revealed that about 40% of the co-channel speech has all frames having a TIR (computed over a 10 m sec frame) greater than 20dB is considered to be "usable. It was also found that the accuracy of speaker identification system can be improved when extracted usable speech segment is used as an input to the speaker identification system instead of co-channel speech [2]. The application of usable speech processing for speaker identification system is represented in Figure 1. Extracted usable speech segments and the co-channel data considered in this research and recorded over a single microphone are used by the speaker identification system. Hence the co-channel data cannot make it to compute target speech energy and interference speech energy. So usable speech measures which have high correlation with the TIR is required to determine the usability in co-channel speech.

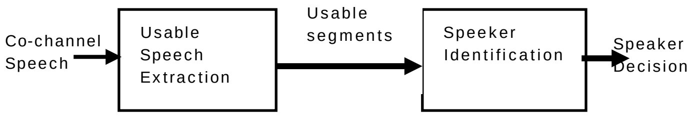


Figure 1: Application of Usable Speech Extraction System for Speaker Identification System.

Several usable speech measures have been proposed [3, 4, 5, 6, 7, 8] with mediocre performance in usable speech identification. Usable speech segments are identified measures to perform a frame-by-frame (of a fixed frame length) and consider periodicity or structure of the speech frame. In this paper, we are considering a different approach to identify the usable speech segments. The speech data is studied at a sample-by-sample level and the fixed size frame analysis is eliminated. The coefficients of the all-pole speech model is obtained by the recursive least squares (RLS) algorithm. These coefficients are used in a *LLWNN* pattern classifier to obtain sample-by-sample class associations. The Speech segmentation is achieved by grouping samples belonging to the same class together by the sequential probability ratio (SPR) test. The SPR is a class ratio computed for each class and segment class associations are made comparing by the SPR with a fixed threshold. The lenience to be given for the segmentation scheme determines the threshold. The term LLWNN-RLS of the algorithm denotes the segmentation scheme being developed.

The system was tested with clean data recorded in the studio environment as well as in various noisy environments to test the performance and accuracy of the system in different environments.

The rest of the paper is organized as follows. A detail discussion of *LLWNN* as pattern classifier as given in [9, 10] and concepts of sequential detection is presented in section 2. Discussion about RLS-LLWNN algorithm is enumerated in section 3. Experimental setup, testing and result validation of the new classification scheme is carried out in section 4. Finally, section 5 draws the conclusions.

2. Background

The RLS-*LLWNN* algorithm consists of two main parts. First, the feature extraction steps using the recursive least squares [11] and second, a *LLWNN* classifier performs sequential classification.

2.1. Local linear wavelet neural network

In terms of wavelet transformation theory, wavelets in the following form:

$$\psi=\{\psi_i=|a_i|\,\psi\left(\frac{x-b_i}{a_i}\right): a_i,b_i \in R^n, i \in Z\}, \quad (1)$$

X=(x1,x2,......xn),

X=(x1,x2,......xn),

bi=(bi1,bi2,....bin),

are a family of functions generated from one single function Ψ(x) by the operation of dilation and translation. Ψ(x), which is localized in both the time space and the frequency space, is called a mother wavelet and the parameters ai and bi are named the scale and translation parameters, respectively. The x represents inputs to the WNN model.

In the standard form of WNN, the output of a WNN is given by

$$f(x)=\sum_{i=1}^{M}\omega_i\psi_i(x)=\sum_{i=1}^{M}\omega_i\,|a_i|^{-1/2}\,\psi\left(\frac{x-b_i}{a_i}\right), \quad (2)$$

Where Ψi is the wavelet activation function of i^{th} unit of the hidden layer and ωi is the weight connecting the i^{th} unit of the hidden layer to the output layer unit. Note that for the n-dimensional input space, the multivariate wavelet basis function can be calculated by the tensor product of n single wavelet basis functions as follows.

$$\psi(x)=\prod_{i=1}^{n}\psi(x_i) \quad (3)$$

Obviously, the localization of the i^{th} units of the hidden layer is determined by the scale parameter ai and the translation parameter bi. According to the previous researches, the two parameters can either be predetermined based upon the wavelet transformation theory or be determined by a training algorithm. Note that the above WNN is a kind of basis function neural network in the sense of that the wavelets consists of the basis functions. Note that an intrinsic feature of the basis function networks is the localized activation of the hidden layer units, so that the connection weights associated with the units can be viewed as locally accurate piecewise constant models whose validity for a given input is indicated by the activation functions. Compared to the multilayer perceptron neural network, this local capacity provides some advantages such as the learning efficiency and the structure transparency. However, the problem of basis function networks is also led by it. Due to the crudeness of the local approximation, a large number of basis function units have to be employed to approximate a given system. A shortcoming of the WNN is that for higher dimensional problems many hidden layer units are needed.

Local Linear wavelet network in fact is a modification of WNN. The architecture of the proposed LLWNN is shown in Fig. 1. Its output in the output layer is given by

$$y=\sum_{i=1}^{M}(\omega_{i0}+\omega_{i1}x_1+\ldots+\omega_{in}x_n)\psi_i(x)$$

$$=\sum_{i=1}^{M}(\omega_{i0}+\omega_{i1}x_1+\ldots+\omega_{in}x_n)\,|a_i|^{-1/2}\,\psi\left(\frac{x-b_i}{a_i}\right), \quad (4)$$

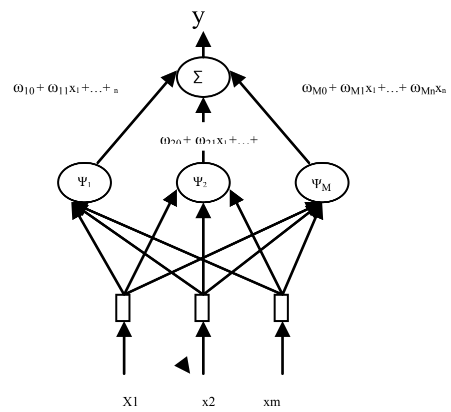


Figure 2: A local linear wavelet neural network.

Where X = [x1,x2,….,xn] Instead of the straight forward weight ω_i (piecewise constant model), a linear model

vi = $\omega_{i0} + \omega_{i1}x_1 + \ldots + \omega_{in}x_n$ (5)

is introduced. The activities of the linear models vi (I = 1,2,….,M) are determined by the associated locally active wavelet functions $\psi_i(x)$ (I = 1,2,….,M) thus vi is only locally significant. The motivations for introducing the local linear models into a WNN are as follows: (1) local linear models have been studied in some neuro-fuzzy systems and shown good performances [12, 13]; and (2) local linear models should provide a more parsimonious interpolation in high-dimension spaces when modeling samples are sparse. The scale and translation parameters and local linear model parameters are randomly initialized at the beginning and are optimized by recursive least square algorithm.

2.2 Sequential Detection

Wald [14] introduced the concept of sequential test and formulated *sequential probability ratio test* (SPRT). The test was designed to decide between two simple hypotheses sequentially. Given two constant as the upper and the lower stopping thresholds and the hypotheses H_0 and H_1 , by observing the data and computing the accumulated log likelihood ratio sequentially, SPRT can make a decision on either continuing observation or stopping the testing accepting H_0 and H_1. This algorithm needs pre-determined threshold value.

Sequential detection scheme over the *LLWNN* class association is made and hence automatically segmenting the speech data into two classes. Hypothesis H_0 corresponds to declaring the segment as usable speech and H_1 corresponds to declaring the segment as unusable speech. On every in- coming sample of speech, the SPRT is done and one of the three possible decisions is made.

1. Decide H_0
2. Decide H_1
3. Not enough information to decide either H_0 or H_1 .

If decision 1) or 2) is made, the hypothesis testing procedure stops. Otherwise, an additional observation is taken, and the test is performed again. This process continues until a decision is made either in favor of H_0 or H_1 . Note that the number of observations taken to obtain a decision of H_0 or H_1 is not fixed but a random variable.

3. RLS-LLWNN Algorithm

The RLS-LLWNN algorithm performs segmental classification of speech data. The segmental classification is accomplished by classifying on a sample-by-sample basis. It is easy to realize that a sample-by-sample classification would need large amounts of computational power. Hence in this algorithm we extract features recursively and simultaneously perform classification.

Recursive-least squares was used to obtain the auto-regressive model coefficients. A step-size of 0.4 and forgetting-factor of 0.99 was used in the computation of the weight vector $\tilde{w}[n]$ for sample point n:

$$\tilde{w}[n] = \tilde{w}[n-1] + k[n]e[n] \tag{6}$$

Where $e[n]$ is the error in prediction and $k[n]$ is the update factor which is computed using the step-size, forgetting factor and inverse correlation matrix. The new weight vector computed at every recursion is used to determine the *a posteriori* class probabilities $P(\omega_i\,|\tilde{w})$ for the *LLWNN* classifier using equation 1. The class probabilities are then mapped as labels $\Phi[n]$: "1" representing usable speech class and "0" represents the unusable speech class:

$$\phi[n] = \begin{cases} 1, & \text{if } P(\omega_1\,|\,) \geq P(\omega_0\,|\,\tilde{w}) \\ 0, & \text{otherwise} \end{cases} \tag{7}$$

In the next step the segmental classification is done, by considering a sequence of labels obtained at each recursion. We define the SPR j for segment j and classes 0 and 1 as:

$$\varphi = \frac{(n - n_s^j) - \sum_{k=n_s^j}^{n} \phi[k]}{n - n_s^j} \tag{8}$$

and $$\varphi = \frac{\sum_{k=n_s^j}^{n} \phi[k]}{n - n_s^j} \tag{9}$$

where n_s^j denotes the index of the beginning sample point and we will use n_e^j to denote the segment end sample point of the segment j. The class ratios φ_i^j are compared to a fixed threshold ρ . A valid range for this threshold is $0.5 \leq \rho \leq 1$. If $\varphi_i^j \geq \rho$ for any $\{i : 0, 1\}$, then the segment j between indices n_s^j and $n_e^j = n$ is assigned to class ω_i and a new segment with start point index $n_s^{j+1} = n+1$

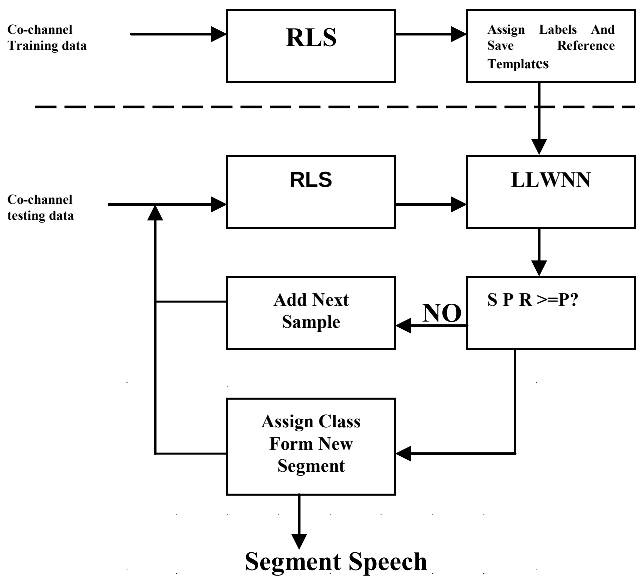


Figure 3: Usable Speech Segment Classification Using Sequential LLWNN Classifier.

The RLS-*LLWNN* algorithm performing classification of co-channel data is illustrated with the block diagram shown in figure 2. Co-channel speech is the input to the system and the output is speech segmented into usable and unusable classes. The speech segment with usable or unusable labels. The steps shown above the dashed line represents the training process and the steps below the dashed line represent the testing process.

4. Experimental setup and Results

To evaluate the proposed segmentation algorithm, two separate schemes for training and testing were designed. The training scheme requires *a priori* knowledge of the class associations of the training data.

4.1 Training

The training process involves the RLS coefficient computations and assigning labels based on the TIR values and the TIR threshold of 20dB. Speech data was taken from TIMIT database was used for all the experiments. Subsets of 42 files from speakers spanning the entire dialect regions were chosen. The original speech was sampled at 16 kHz and re sampled to 8 kHz after low-pass filtering to 3 kHz. Two utterances were read at a time (hence making a total number 861 co-channel utterances) and their amplitudes were scaled such that the energy of the two utterances are equal over the entire utterance. The scaled speech data was added to simulate co-channel data recorded over a single microphone. The TIR values were computed over fixed frame sizes of 10 m sec. The RLS coefficients $\tilde{w}$ *[nT]* were extracted to make the coefficients synchronous with the TIR values. *T* is the number of samples corresponding to the frame size of 10 m sec.

4.2 Testing

A two fold testing was carried out on the co-channel utterances. Out of 861 co-channel utterances, 431 utterances were used for the training process and the remaining 430 utterances were used for testing and evaluation. The first step in the testing stage is to perform RLS and obtain the coefficients. At every recursion the *LLWNN* classifier performs classification and assigns class labels as described in section 3.

In this experiment, the value of *k* was chosen as 9, and SPR of 0.65 was chosen. These numbers gave the best performance in the experiments. A Correct detection (hit) is said to occur when both the RLS-LLWNN classifier and TIR identifies a segment of speech belonging to the same class and a false hit when RLS-LLWNN classifier and TIR declares a speech segment to belong in different classes. Figure 3 shows the comparison of detection of usable and unusable speech segment between proposed algorithm and using TIR with 20dB threshold.

The best network model was selected based on the classification performance of the sequential *LLWNN* pattern classifier based on the following criterion. Accordingly we have drawn the confusion matrix. In addition, we draw two graphs to reach the conclusion: One for the mean of CCR versus its standard deviation and the other for the mean of the ASCE versus mean of MDL. These graphs help us to decide which classifier is better in its performance. In both plots, each classifier is represented by a symbol.

In the graph of the average of CCR versus its standard deviation, a good classifier should appear in the lower right corner of the graph. In the graph of average of MDL versus average of ASCE, a good classifier should appear in the bottom left of the plot. In addition, corresponding to these graphs, we summarize the results in Tables. In these Tables, the highest CCR's are given in boldface.

1. Correct Classification Rate (CCR) and Average Squared Classification Error (ASCE):

$$CCR = \frac{\sum_{k=0}^{c-1} CC_k}{n}; \quad ASCE = \frac{\sum_{k=0}^{c-1} [nk-cck]^2}{n} \qquad (10)$$

where nk is the number of observations in class k, and CCR is the number of correctly classified observations in the class k. The best functional network is the one with both highest CCR and smallest ASCE. We construct the confusion matrix, which is a c x c matrix, its diagonal contains the number of correctly classified observations, CCR, and the off-diagonal elements are the number of misclassified observations, mck, for k = 0,…, c - 1

2. *Computational cost (Time of execution):* It is the time needed to execute the classifier till obtaining the best model in both calibration and validation. The less computation cost is the better classifier.

3. *The Minimum Description Length (MDL) criterion:* As explained in [15], and then the best model is the one with the smallest MDL value. The form of the description length for the classification problem using the functional network is defined as

$$L(\Theta_k) = \frac{m \log(nk)}{2} + \frac{nk}{2} \log\left(\frac{1}{nk}\sum_{i=1}^{n} \varepsilon_i^2(\Theta_k)\right) \quad (11)$$

For all k = 0… c-1, where m and k are the number of elements in the family and the category levels, respectively. We note that the principle $L(\Theta_k)$ is the code length of the estimated parameters Θk, ∀k = 0, 1, 2,…,c-1

1. We note that the description length has two terms:

(a) The first term $\frac{m\log(nk)}{2}$ is a penalty for including too many parameters in the functional network model.

(b) The second term $\frac{nk}{2} \log\left(\frac{1}{nk}\sum_{i=1}^{n} \varepsilon_i^2(\Theta_k)\right)$ measures the quality of the functional network model fitted to the training set. Therefore, the best model is the model with the smallest value of its description length. MDL is the best model performance. In addition, we draw two graphs to reach the conclusion regarding the performance of RLS-KNN and RLS-LLWNN classifiers. The graphs are drawn to represent two ideas, one for the mean of CCR versus its standard deviation and the other for the mean of the ASCE versus mean of MDL. These graphs help us to decide which classifier is better in its performance. In both plots, each classifier is represented by a symbol. In the graph of the average of CCR versus its standard deviation, a good classifier should appear in the lower right corner of the graph. In the graph of average of MDL versus average of ASCE, a good classifier should appear in the bottom left of the plot.

Table 1: Confusion Matrix

Class	Class1	Class2
Class1	86	14
Class2	17	83

The rows of the confusion matrix represent the performance of actual classes and the columns represent the identified classes. The first row represents performance of classifying usable speech and the second row represents the performance of classifying unusable speech. The percentage of correct identifying usable speech is 86% and unusable is 83%. The false alarms are 14% and 17% respectively. This gives the overall identification rate of 84.5%.

The proposed algorithm was compared with the best performing usable speech measure: Adjacent Pitch Period Comparison (APPC) and with RLS-KNN as given in [16] and the results are presented in figure 4. It should be noted that the sequential *LLWNN* pattern classifier was able to increase correct detection of usable speech by 11.5 relatively with respect to the performance of APPC and 5.5 % relative to KNN classifier.

Only 48 sample points on an average were required by the proposed sequential LLWNN technique to make a class decision i.e., at least one of the SPR, exceeding 0.65. Fixed 320 samples (40ms) frames of speech for usable speech were used by other usable speech measure. Thus it can be inferred that less data are required by the new methods to derive statistical decision.

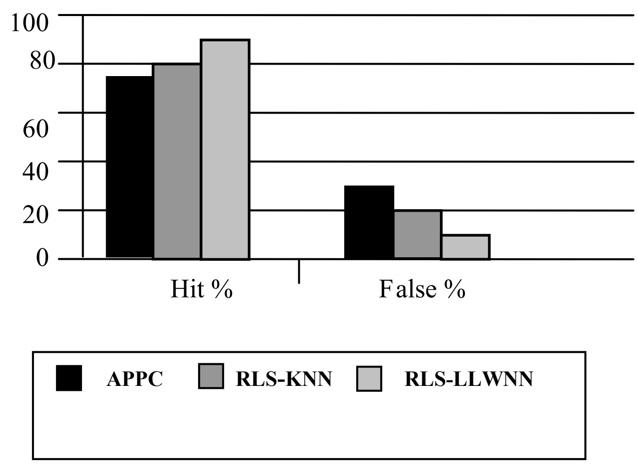


Figure 4: Comparison of Results of Detection of Usable Speech: Black bars represent performance using APPC measure, gray bars using sequential k- NN classifier and light gray representing LLWNN classifier.

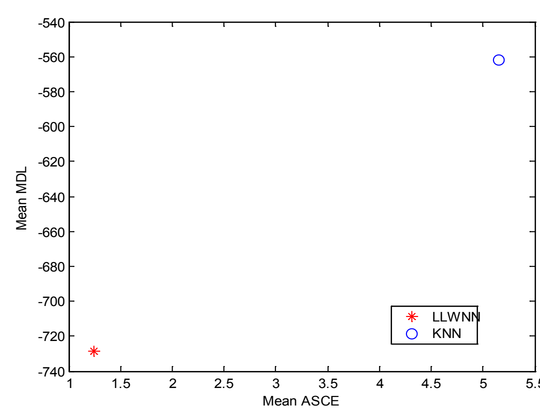
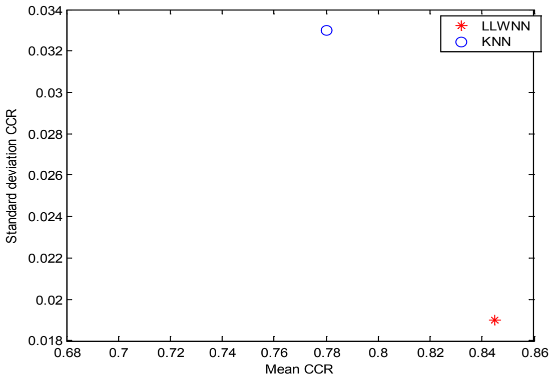


Figure 5: (a) Mean ASCE vs mean MDL



(b) Mean CCR vs σCCR

5. Conclusion

The purpose of this paper was to develop a sequential *LLWNN* classifier and evaluate it to classify the usable and unusable portions of co-channel speech in the context of speaker identification. It was found that by using [17] RLS coefficients as a feature and sequential *LLWNN* classifier; we were able to achieve identification rate of 84.5%. It was noticed that one can obtain desired performance rate by changing SPR. Also, it was observed that the proposed algorithm requires less data to decide on the class memberships.

In the proposed algorithm, we have used only one set of features; however due to the fact that speech is non- linear in nature and single a feature can not model entire system. This leads to poor classification performance. Therefore to improve the classification rate, one can think of using the usable speech measures itself as a feature for the sequential *LLWNN* classifier. In the current research, the SPR was chosen based on heuristics. It is of our next interest and we intend to use a larger data base, from medical science and/or business sector to evaluate the performance of the proposed technique.

References

[1]. R. E. Yantorno, "Co-channel speech study, final report for summer research faculty program," Tech. Rep., Air Force Office of Scientific Research, Speech Processing Lab, Rome Labs, New York, 1999.

[2]. J. M. Lovekin, R. E. Yantorno, K. R. Krishnamachari, D.B. Benincasa, and S. J. Wenndt, "D eveloping usable speech criteria for speaker identification," *IEEE, International Conference on Acousitcs and Signal Pro- cessing*, pp. 424–427, May 2001.

[3]. J. M. Lovekin, K. R. Krishnamachari, and R. E. Yan- torno, "Adjacent pitch period comparison (appc) as a usability measure of speech segments under co-channel conditions," IEEE International Symposium on Intelligent Signal Processing and Communication Systems, pp. 139–142, Nov 2001.

[4]. N. Chandra and R. E. Yantorno, "Usable speech de- tection using modified spectral autocorrelation peak to valley ration using the l pc residual," 4th IASTED Inter- national Conference Signal and Image Processing, pp. 146–150, 2002.

[5]. A. R. Kizhanatham, R. E. Yantorno, and B. Y. Smolen- ski, "Peak difference autocorrelation of wavelet transform (pdawt) algorithm based usable speech measure," IIIS Systemics, Cybernetics and Informatics, Aug 2003.

[6]. N. Sundaram, A. N. Iyer, B. Y. Smolenski, and R. E. Yantorno, "Usable speech detection using linear predictive analysis a model-based approach," *IEEE* International Symposium on Intelligent Signal Processing and Communication Systems, *ISPACS*, 2003.

[7]. A. N. Iyer, M. Gleiter, B. Y. Smolenski, and R. E. Yantorno, "Structural usable speech measure using l pc residual," IEEE International Symposium on Intelli- gent Signal Processing and Communication Systems, *ISPACS*, 2003.

[8]. Y. Shao and D-L. Wang, "Co-channel speaker identifi- cation using usable speech extraction based on multi- pitch tracking", IEEE International Conference on Acoustics, Speech, and Signal Processing,, vol. 2, pp.205–208, 2003.

[9]. Yuehui Chen, Bo yang, and Jiwen Dong, "Time series prediction using a local linear wavelet neural network", *Neuro Computing, 69, (2006), pp. 449-465.*

[10]. Yuehui Chen; Jiwen Dong; Bo Yang; Yong Zhang, "Local linear wavelet neural network", Fifth World Congress on Intelligent Control and Automation (WCIA), Hangzhou, June (2004), pp.1954-1957.

[11]. S. Haykin, Adaptive Filter Theory, Pearson Education, 4 edition, September 2001.

[12]. B. Fischer, O. Nelles, R. Isermann, "Adaptive p redictive control of a heat exchanger based on a fuzzy model", Control Eng. Practice 6, (1998), pp. 259–269.

[13]. B. Foss, T.A. Johansen, "On local and fuzzy modeling, Proceedings of 3rd International Industrial Fuzzy Control and Intelligent Systems", (1993), pp. 134–139.

[14]. A. Wald, Sequential Analysis, Wiley, New York, 1947.

[15]. Emad A. El-Sebakhy, Kanaan Abed Faisal, T. Helmy, F. Azzedin, and A. Al-Suhaim, "Evaluation of Breast Cancer Tumor Classification with Unconstrained Functional Networks Classifier", IEEE International Conference on Computer Systems and applications, (2006), pp. 281-287, DOI: 10.1109/AICCSA.2006.205102.

[16]. Jashmin K Shah, Brett Y Smolenski, Robert E Yantorno, Ananth N Iyer, " S equential k-Nearest Neighbourhood Pattern Recognition for Usable Speech Classification", in the proceedings of European signal processing conference, September 6th-10th, 2004 pp. 741-744.

[17]. V. Matta S. Marano and P. Willett, "Sequential detection of almost-harmonic signals," IEEE Transaction on Signal Processing, vol. 51, 2, pp. 395–406, February 2003.

About the Authors

Suchismita Sahoo is pursuing her Masters of Technology in Computer Science and Engineering at Koustuv Institute of Self Domain under Biju Patnaik University of Technology, Rourkela. She has completed her

Bachelors of Technology in Computer Science and Engineering. Her areas of Interest are Artificial Intelligence, Soft Computing, Algorithm design and analysis and Embedded System includes her areas of interest.

Sushree Sangita Sahoo is pursuing her Masters of Technology in Information Technology at Koustuv Institute of Self Domain under Biju Patnaik University of Technology, Rourkela. She has completed her Bachelors of Technology in Computer Science and Engineering. She is working as a Lecture in Indus College of Engineering, Bhubaneswar. She is having above 02 years of teaching experience.

M.R.Senapati is pursuing his Ph.D. in Computer Science and Engineering, at Biju Patnaik University of Technology, Rourkela, in the area of Multi Cass Pattern Recognition. His major interest includes intelligent protection, digital signal processing, soft computing, FACTs. Graduated in Computer Science and Engineering, He is working as Associate professor in the department of computer science and engineering, Gandhi Engineering College. He is having 18 years of teaching experience and 01 year of industry experience.

P.K. Dash (SMIEE 1990) is working as Director Research, S'O'A University, Bhubaneswar, India. Earlier he was a professor in the Faculty of Engineering, Multimedia University, and Cyberjaya, Malaysia. He also served as a professor of electrical engineering and chairman of center for intelligent systems, National institute of Technology, Rourkela, India for more than 25 years. Dr. Dash holds D.Sc., Ph.D., M.E., and B.E. degrees in electrical engineering and had his post-doctoral education at the University of Calgary, Canada. His research interests are in the area of power quality, FACTS, soft computing, deregulation and energy markets, signal processing, and data mining. He had several visiting appointments in Canada, USA, Switzerland, and Singapore. To his credit he has published more than 150 International Journal papers and nearly 100 in International conferences. Prof. Dash is a fellow of the Indian National Academy of Engineering and senior member of the IEEE, and fellow of Institution of Engineers, India.