SPEECH CLASSIFICATION USING ZERNIKE MOMENTS

Manisha Pacharne[1] and Vidyavati S Nayak[2]

[1]Department of Computer Engineering, DIAT (DU), Pune-411025 (India)
[1]ce9manisha@diat.ac.in
[2]Armament Research & Development Establishment, Pashan, Pune-411021 (India)
[2]vidyavatinayak@diat.ac.in

ABSTRACT

Speech recognition is very popular field of research and speech classification improves the performance for speech recognition. Different patterns are identified using various characteristics or features of speech to do there classification. Typical speech features set consist of many parameters like standard deviation, magnitude, zero crossing representing speech signal. By considering all these parameters, system computation load and time will increase a lot, so there is need to minimize these parameters by selecting important features. Feature selection aims to get an optimal subset of features from given space, leading to high classification performance. Thus feature selection methods should derive features that should reduce the amount of data used for classification. High recognition accuracy is in demand for speech recognition system. In this paper Zernike moments of speech signal are extracted and used as features of speech signal. Zernike moments are the shape descriptor generally used to describe the shape of region. To extract Zernike moments, one dimensional audio signal is converted into two dimensional image file. Then various feature selection and ranking algorithms like t-Test, Chi Square, Fisher Score, ReliefF, Gini Index and Information Gain are used to select important feature of speech signal. Performances of the algorithms are evaluated using accuracy of classifier. Support Vector Machine (SVM) is used as the learning algorithm of classifier and it is observed that accuracy is improved a lot after removing unwanted features.

KEYWORDS

Feature Extraction, Feature Selection, Zernike Moment, SVM, Speech Classification.

1. INTRODUCTION

Speech processing requires careful attention to issues [1] like various types of speech classes, speech representation, feature extraction techniques and speech classifiers. There are many features associated with the speech like Mel-Frequency Cepstral Coefficients (MFCC), Wavelet, and Principal Component Analysis (PCA) which describe the speech and helps to classify the speech signal into various categories. In this paper Zernike moments are used to classify the speech signal based on the shape of spectral region. In this approach Zernike moments are computed and norm of these Zernike moments are taken as features for each audio file and then they are ranked using various feature ranking algorithms. We are selecting important features based on the ranking to improve the classification accuracy of classifier. Unwanted features can increase computation time and can have impact on the accuracy of the speech recognition.

Zernike moments have been used by many recognition systems like human face recognition [2], fingerprint recognition [3], visual speech recognition [4] in the past. In visual speech recognition different moments of video data of the speaker's mouth are extracted and used for recognition. In these recognition systems Zernike moments of images are taken and used while in our approach Zernike moments of audio data are taken and used for classification. There are applications [5] of

D.C. Wyld, et al. (Eds): CCSEA 2011, CS & IT 02, pp. 294–303, 2011.

DOI: 10.5121/csit.2011.1227

Zernike moments used as features with various classifiers for classification but we use Support Vector Machine (SVM) classifiers for classification.

This paper is organized as follows. First, introduction section, next section gives a brief on the feature extraction of .wav file to generate the dataset for our experiment. Next section presents the different algorithms used in our feature selection approach and describes the classification using SVM. Dataset used, experimental methodology and results are reported in section experimental result. Finally, we conclude our work in the last section.

2. Feature Extraction

Feature extraction is a process where a segment of audio is characterized into a compact numerical representation. One dimensional speech signal is converted into two dimensional images for extracting its Zernike moments. The Zernike moments are useful in image analysis and pattern recognition due to their orthogonality and rotation invariance property. They are also used in wide range of applications [6] on image analysis, reconstruction and recognition due to minimal redundancy, rotation invariance and robustness to noise.

2.1. Computation of Zernike Moments

Zernike moments [6] are defined to be the projection of image function on the orthogonal basis functions [7]. The basis functions $V_{n,m}$ (x,y) are given by

$$V_{n,\,m}(x,y) = V_{n,m}(\rho,\theta) = R_{n,m}(\rho)e^{j\,m\,\theta} \qquad (1)$$

Where n is non-negative integer, m is non-zero integer subject to the constraints n-|m| is even and |m| < n, ρ is the length of vector from origin to (x,y), θ is angle between vector ρ and the x axis in a counter clockwise direction and $R_{n,m}(\rho)$ is the Zernike radial polynomial.

The Zernike radial polynomials, $R_{n,m}(\rho)$, are defined as:

$$R_{n,m}(\rho) = \sum_{k=|m|,n-k=even}^{n} \frac{(-1)^{\frac{n-k}{2}!}\ \frac{n+k}{2}!}{\frac{n-k}{2}!\ \frac{k+m}{2}!\ \frac{k-m}{2}!}\rho^k = \sum_{k=|m|,n-k=even}^{n} \beta_{n,m,k}\,\rho^k \qquad (2)$$

Note that $R_{n,m}(\rho) = R_{n,-m}(\rho)$.The basis functions in equation (1) are orthogonal thus satisfy

$$\frac{n+1}{\pi} \int_{x^2+y^2\,\leq 1} V_{n,m}(x,y)\,V^*_{p,q}(x,y) = \delta_{n,p}\delta_{m,q}$$

Where

$$\delta_{a,b} = \begin{cases} 1 & a=b \\ 0 & otherwise \end{cases}$$

The Zernike moments of order n with repetition m for a digital image function f(x,y) is given by [7]

$$Z_{n,m} = \frac{n+1}{\pi}\sum \sum_{x^2+y^2\leq 1} f(x,y)V^*_{n,m}(x,y)$$

Where $V^*_{n,m}(x,y)$ is the complex conjugate of $V_{n,m}(x,y)$. To compute the Zernike moments of a given image, the image centre of mass is taken to be the origin. The function f(x,y) can then be reconstructed by the following truncated expansion [7].

$$\sum_{n=0}^{N} \frac{C_{n,0}}{2} R_{n,0}(\rho) + \sum_{n=1}^{N} \sum_{m>0} (C_{n,m} Cosm\theta + S_{n,m} Sinm\theta) R_{nm}(\rho)$$

Where N is the maximum order of Zernike moments we want to use, $C_{n,m}$ and $S_{n,m}$ denote the real and imaginary parts of $Z_{n,m}$ respectively.

3. FEATURE SELECTION

Many features are associated with the speech data and some of them can be redundant. There is a need to remove least important features from the available data to reduce storage requirements, training and testing time and thus improving classifier performance. The goal of feature selection [8] is driving an optimal subset of features from a given space leading to high classification performance and it should derive the features that will reduce the amount of data used for classification. However, the search for a subset of relevant features introduces an additional complexity in the modeling task.

Feature selection [9] methods also help machine learning algorithms produce faster and more accurate solutions because they reduce the input dimensionality and they can eliminate irrelevant [10] features.

3.1. Algorithms used for Feature Selection

There are two different ways for selecting important features in a feature subspace. Some of the algorithm use exhaustive search of the feature subspace which is unaffordable for all but a small initial number of features. While some of them uses heuristic search strategies which are more feasible than exhaustive ones and can give good results, although they do not guarantee finding the optimal subset. In this paper algorithms like Information Gain, Gini Index, Fisher Score, Chi Square , ReliefF, t-Test are used for feature ranking and selection.

In Information Gain, weight is assigned to the feature based on the information content of the feature. Information content of the feature is calculated using entropy. The Gini Index is a statistical measure of dispersion. It is based on another statistical phenomenon called the Lorentz curve, and is commonly used to quantify wealth distributions and has many applications. The Fisher Score [11] is a method for determining the most relevant features for classification. It uses discriminative methods and generative statistical models. Chi Square method [12] measures the relevance between the feature and the class. If the measure value is higher means that relevance between feature and class is strong. It means that the feature is having greater contribution to the category. ReliefF [13] selects the nearest neighbour samples from each category, and these nearest neighbour samples are considered as k. The t-Test is a statistical hypothesis where the statistic follows a student distribution and it is a basic test that is limited to two groups. For multiple groups, each pair of groups needs to be compared. The basic principle is to test the null hypothesis that the means of the two groups are equal.

3.2 Classification using Support Vector Machines (SVM)

In computer science, support vector machines (SVMs) are a set of related supervised learning methods that analyze data and recognize patterns, used for classification and regression. SVM constructs a hyperplane or set of hyperplanes in a high or infinite dimensional space, which can be used for classification. Support Vector Machine (SVM) is primarily a classier method that performs classification tasks by constructing hyperplanes in a multidimensional space that separates cases of different class labels. Optimal separating hyperplane is defined as the maximum-margin hyperplane in the higher dimensional feature space. Sometimes it is necessary to map the data into high dimensional space. To reduce the error in the classifications uses kernel [14] to solve quadratic optimization problem that occur while computing the hyperplane for separating data point. There are many kernel functions like polynomial, radial bias functions (RBF) etc. The experiments in this paper are done using RBF kernel. There are two reasons for using SVM [15] for performing classification. Firstly, it is having low expected probability of generalization errors and secondly, its speed and scalability.

4. EXPERIMENTAL RESULTS

4.1. Dataset Used

All speech files are single-channel audio data sampled at 25 kHz. All material is end pointed. (i.e. there is little or no initial or final silence). The total set consists of 1000 sentences from two different talkers. Filenames [16] consists of a sequence of 6 characters which specify the sentence spoken. For example, bbaf2n represents the sentence bin blue at F 2 now. The total data is divided into training data and testing data. Training data is used to build classifier model where as testing data is used to test the model.

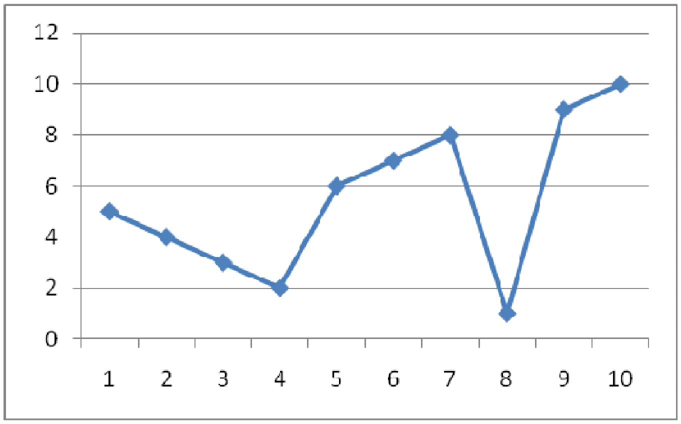

Figure 1. Feature Ranking using Information Gain

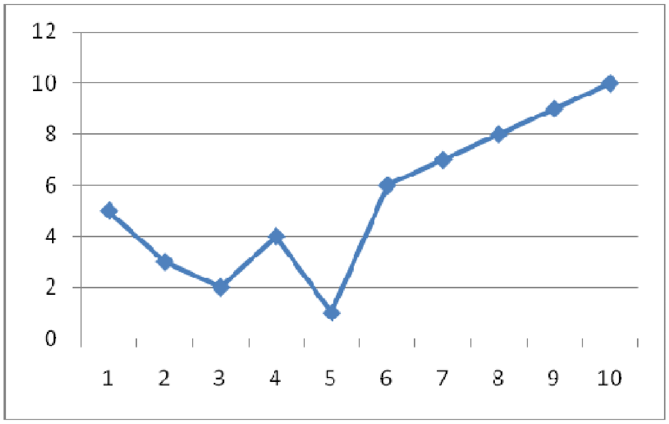

Figure 2. Feature Ranking using Gini Index

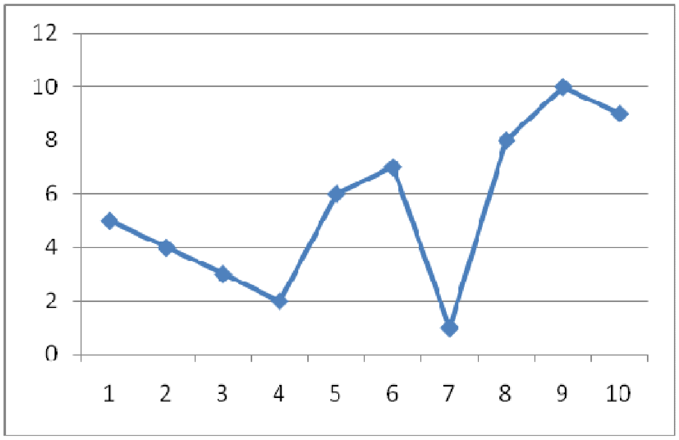

Figure 3. Feature Ranking using Fisher Score

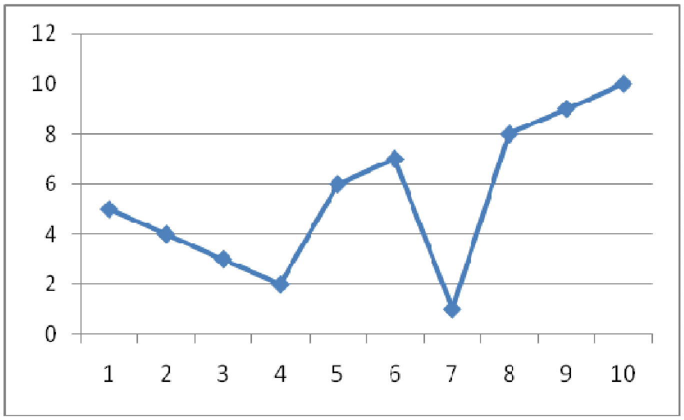

Figure 4. Feature Ranking using Chi Square

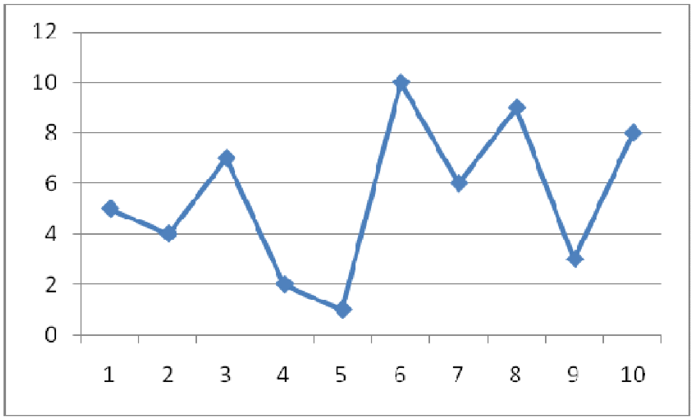

Figure 5. Feature Ranking using Relief

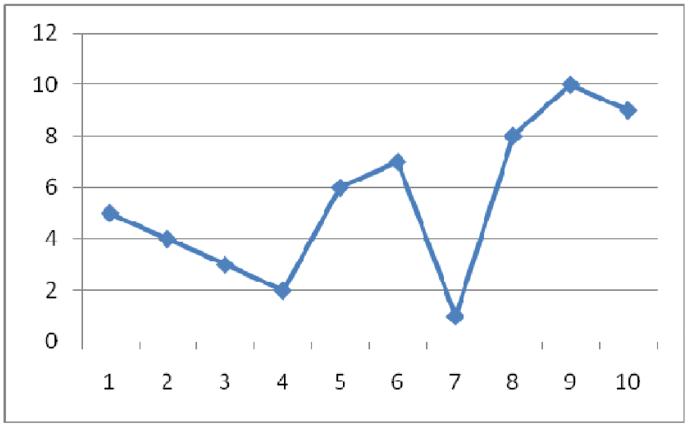

Figure 6. Feature Ranking using t-Test

4.2. Methodology

All audio files from the dataset are read and converted into two dimensional image files. Each image is then pre-processed to calculate its Zernike moments. Zernike base function for order 1, 2, .., 10 is used to calculate Zernike moments of image for different order. Zernike moments are complex number so the norm of Zernike moments for each order is considered as one feature, like wise 10 features for 10 different orders are calculated. For example consider that Zernike moments of audio file 'swwa3n.wav' is to be calculated, then waveform of the file swwa3n.wav is as shown in Figure 2.

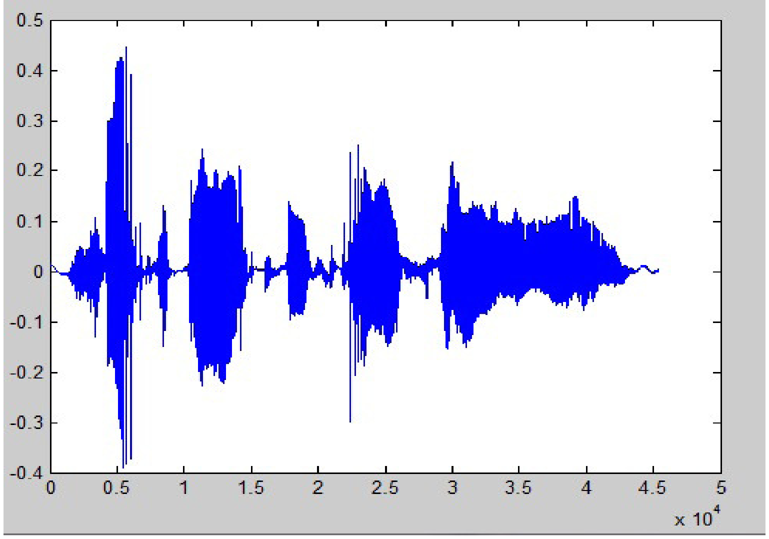

Figure 2: Signal Representation of swwa3n.wav File

This image is converted into square image of 100 pixels before calculating its Zernike moments. Figure 3 shows the image file swwa3n.jpg obtained after pre-processing which is given as input to the function calculating Zernike moments. Pre-processing is done using MATLAB.

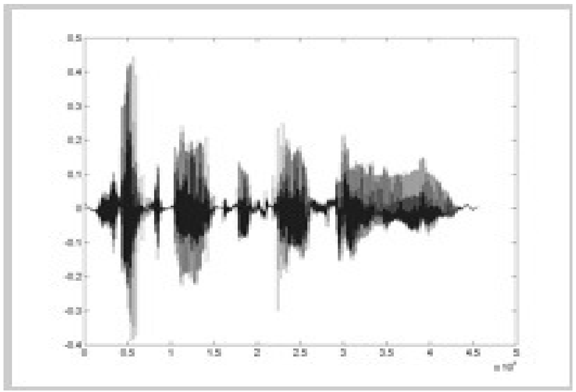

Figure 3: Input Image after Pre-processing

Then Zernike moments of the image are calculated for different order from 1 to 10. Some of them are given in Table 1.

Table 1. Zernike moments of the image for order 1 and 2.

File Name	**Order of Zernike Moments**	**Zernike Moments (ZM)**	**Feature=Norm(ZM)**
Swwa3n.jpg	1	578466.46 + 0.00i -12398.24 - 6078.90i	5.7863e+ 005
	2	578466.46 + 0.00i -12398.24 - 6078.90i 96254.98 + 0.00i -18558.13 - 1198.16i	5.8688e+ 005

The features obtained from Zernike moments are written in the comma separated file, which is then converted into a format required by LibSVM for further processing. LibSVM [17] is library for SVM, it is integrated software for classification and regression. Classification model is build using training data by LibSVM. Scaling of the input training data and testing data is done to make the model more accurate. (Lower limit and upper limit considered for the scaling is -1 to 1). Kernel parameter selection is done to improve the accuracy. Classification accuracy is predicted for test dataset based on the model build using training dataset. This will give the accuracy of dataset without feature selection. By applying various algorithms listed in the Section 3.1 for feature ranking. These algorithms calculate the weight of features and ranks are assigned based on there weight. Ranking for the various features using different algorithms is as shown in Table 2 and Figures 1, 2, 3, 4, 5 and 6 represents the graphical representation of the ranking using different algorithms. Here X-axis represents the feature number and Y-axis represents the rank assigned to the corresponding feature. Dataset for first 5, 7 and 8 ranked features are created. Scaling and kernel parameter selection is done to obtain better results. All the datasets are given as input to LibSVM to build model and predict the accuracy of test data. Accuracies of different algorithm with different number of feature combination are as shown in Table 3.

Table 2. Feature ranking using various algorithms.

Information Gain	Gini Index	Fisher Score	Chi Square	ReliefF	t-Test
5	5	5	5	5	5
4	3	4	4	4	4
3	2	3	3	7	3
2	4	2	2	2	2
6	1	6	6	1	6
7	6	7	7	10	7
8	7	1	1	6	1
1	8	8	8	9	8
9	9	10	9	3	10
10	10	9	10	8	9

Table 3. Accuracies of different algorithms with different number of feature combination for Zernike features.

Algorithm Used	Accuracy (%) without Feature Selection	Accuracy (%) with 5 Features	Accuracy (%) with 7 Features	Accuracy (%) with 8 Features
Information Gain	50	76.53	78.57	78.57
Gini Index	50	72.44	78.57	77.55
Fisher Score	50	76.53	78.57	78.57
Chi Square	50	76.53	78.57	78.57
ReliefF	50	73.46	76.53	81.63
t-Test	50	76.53	76.67	78.57

From Table 3 it observed that the accuracy of classifier for dataset for the entire features extracted (i.e. without feature selection) is 50%, while after eliminating the least important features the accuracy of classifier shows lot of improvement. Among all the algorithms ReliefF gives the best accuracy of 81.63% for 8 feature selection.

5. CONCLUSIONS

In this paper, we have used various feature selection and ranking algorithms to reduce the features of the input data to an SVM based speech recognition system. The features were ranked using Information Gain, Gini Index, Fisher Score, Chi Square, ReliefF and t-Test feature ranking algorithms. From the ranked features optimum feature subsets were identified and used for speech classification.

Classification was performed using SVM on the dataset with full features and also with the reduced feature subsets to determine the comparison in the speech classification accuracy. The effort is made to compare the performance of various subsets of features obtained from the feature selection algorithm over generated dataset. Experimentally, it is found that the reduced feature subsets give better results on the dataset than the full feature set. We are able to achieve higher speech classification accuracy on a smaller feature subset thus achieving substantial reduction of the input dataset.

ACKNOWLEDGEMENTS

The authors would like to thank Valadi K. Jayaraman, Emeritus Scientist, C-DAC, Pune for valuable guidance during this research work.

REFERENCES

[1] M.A.Anusuya, S.K.Katti., (2009) “Speech Recognition by Machine: A Review”, International journal of Computer science and information security, Vol. 6, No. 3.

[2] Javad Haddadnia, Majid Ahmadi, Karim Faez, (2002) ”A Hybrid Learning RBF Neural Network For Human Face Recognition with Pseudo Zernike Moment Invariant”,IEEE,0-7803-7278-6/02.

[3] Hasan Abdel Qadar, Abdul Rahman Ramli and Syed Al-Haddad, (2007) ”Fingureprint Recognition using Zernike Moments”,The International Arab Journel of Information Technology,vol. 4, No.4.

[4] Wai C. Yau, Dinesh K. Kumar, Sridhar P. Arjunan and Sanjay Kumar, (2006) ”Visual Speech Recognition using Image Moments and Multiresolution Wavelet Images”,Proceddings of international Conference on Computer Graphics, Imaging and Visualisation, 0-7695-2606-3/06.

[5] Mostafa Nosrati, Hamidreza Amindavar, (2007) ”Applications of Zernike MomentsAs features in KNN and SVM as semi blind detectors for STBC MIMO-OFDM systems in impulsive noise environments”,IEEE,1-4244-0728-1/07.

[6] G.B.Gholam Reza Amayeh, Ali Erol and M.Nicolescu, Accurate and efficient computaion of high order Zernike moments ,Computer Vision Laboratory, University of Neveda.

[7] A. Khotanzad and Y.H. Hong., (1990) “Invariant image recognition by Zernike moments”,IEEE trans. On Patteren Anal. And Machine Intell.,Vol.12,489-498.

[8] Liu H. and Setiono R., (1995) “Chi2: Feature Selection and Discretization of Numeric Attributes”, 7th International Conference on Tools with Artificial Intelligence, pp. 338-391.

[9] Guyon I. and Elisseeff A., (2003) “An introduction to variable and feature selection”, Journal of Machine Learning Research, Vol. 3, pp. 1157-1182.

[10] G.H.John, R.Kohavi and P. fleger, (1994) “Irrelevant features and subset selection problem”, 11thInternational conference Machine Learning,pp. 121-129.

[11] X.He, D.Cai and P.Niyogi, Laplacian Score for feature selection, Department of Computer Science, University of Chicago.

[12] Yan and X. Ting, An improved $\chi 2$ statistics method for text feature selection ,College of Computer and Information Science, Southwest University, China.

[13] X.Wang, B. Wang, L.Shi, and M. Chen, (2010) “An improved combination feature selection based on relieff and genetic algorithm”, International conference on Computer science and education.

[14] Cristianini, N. and Shawe-Taylor J., (2003) “Support Vector and Kernel Methods, Intelligent Data Analysis: An Introduction Springer – Verlag”.

[15] R. Kumar, V.K. Jayaraman, B.D. Kulkarni, (2005) “An SVM classifier incorporating simultanious noise reduction and feature selection:illustrative case examples”, Journel of Pattern Recognition, Vol. No. 38, pp.41-49.

[16] Martin Cooke and Te-Won, Speech separation challenge, http://staffwww.dcs.shef.ac.uk/people/M.cooke/SpeechSeparationChallenge.htm.

[17] Chih-Chung Chang and Chih-Jen Lin., LIBSVM: A Library for Support Vector Machine 2001, http://www.csie.ntu.edu.tw/~cjlin/libsvm.

Authors

Vidyavati S Nayak was born in Karnataka, India, in 1972. She received B. E. degree in Computer Science from Gulbarga University, Gulbarga, India, in 1996 and the M Tech degree in Computer Science and Engineering from the Indian Institute of Technology, Madras, Chennai, India, in 2007.

Since February 1998, she has been working in Defence Research and Development Organization (DRDO) as scientist in the field of Computer Science. Her current research interests include computer networks, data mining, algorithms, VLSI.

Manisha Pacharne was born in Pune, India in 1986. She Received the B.E. degree in Computer Engineering from Pune University, Pune, India, in 2007 and M Tech degree in Computer Science and Engineering from Defence Institute of Advance Technology (DU), Pune, India in 2011.