Histogram to Sound Conversion: A Review

Himadri Nath Moulick[1], Moumita Ghosh[2], Poulomi Das[3], Chandan Koner[4], and Alok Kumar Roy[5]
[1] West Bengal University of Technology, India,
[2] Burdwan University, Burdwan, India
[3] Heritage Institute of Technology, Kolakata, India
[4] Dr. B.C.Roy Engineering College, Durgapur-713206, India
[5] Bankura Unnayani Institute of Engineering, Bankura, India
Email: himadri80@gmail; Mou2005be@gmail.com; poulami.das@heritageit.edu; chandan.durgapur@gmail.com; alokeroy@in.com

***Abstract*—The main goal of a voice conversion system [1]-[6] is to modify the voice of a source speaker, in order to be perceived as if it had been uttered by another specific speaker. Many approaches found in the literature convert only the features related to the vocal tract of the speaker. Our proposal is to not only convert those characteristics of the vocal tract, but also to process the signal passing through the vocal chords. Thus, the goal of this work is to obtain better scores in the voice conversion results. Also, this paper describes a method of compensating for nonlinear distortions in speech representation caused by noise. The method described here is based on the histogram equalization method often used in digital image processing. Histogram equalization is applied to each component of the feature vector in order to improve the robustness of speech recognition systems. The paper describes how the proposed method can be applied to robust speech recognition and it is compared with other compensation techniques.**

***Index Terms*—voice conversion, K-histograms, cepstral mean normalization, histogram equalization, mean and variance normalization.**

I. Introduction

Noise [7], [8] strongly degrades the performance of speech recognition systems. For this reason, robust speech recognition is one of the focus areas of speech research. Noise has two main effects on speech representation. First, it introduces a distortion in the feature space. Secondly, due to its random nature, it also causes a loss of information.

The effect of the distortion depends on the speech representation and the type of the noise, and it usually produces a nonlinear transformation of the feature space. For example, in the case of cepstral based representations, additive noise causes a nonlinear transformation that has no significant effect on high-energy frames but a strong effect on those with energy levels in the same range or below that of the noise. This distortion causes a mismatch between the training and recognition conditions. The acoustic models trained with speech acquired under clean conditions do not model speech acquired under noisy conditions accurately and this degrades the Compensation methods for robust speech recognition mainly focus on minimizing this mismatch. Some methods try to adapt the acoustic models to noisy conditions in order to allow them to represent noisy speech properly, whereas other methods try to determine the features of the clean speech from the observed noisy speech. In the former case, the noisy speech [9]-[11] is recognized using noisy models. In the latter case, a clean version of the speech is recognized using the clean models. Finally, some methods include operations in the feature extraction module in order to minimize the effect of noise superimposed on the speech representation.

Thus, for example, cepstral mean normalization is usually applied as a part of the feature extraction in order to remove the global shift of the mean affecting the cepstral vectors. This normalization compensates for the main effect of channel distortion and some of the side effects of additive noise. However, the nonlinear effects of additive noise on cepstral-based representations cannot be treated by CMN and this makes this method effective only for moderate levels of additive noise. This method is improved by mean and variance normalization because normalization of the mean and the variance yields a better compensation of the mismatch caused by additive noise.

Other methods, such as spectral subtraction or the vector Taylor series (VTS) approach yield more effective compensation of additive noise since they can deal with the nonlinear effects of noise. Robust methods based on the adaptation of acoustic models include Statistical Re-estimation and parallel model combination which apply independent corrections to each Gaussian pdf in the acoustic models. These are able to model nonlinear effects of the distortion caused by noise correctly. Most compensation methods are based on estimations of convolution and additive noise and a statistical or analytical formulation describing the effect of noise superimposed on the speech representation.

This paper describes a method of compensating for the noise affecting speech representation. The method is based on the histogram equalization (HEQ) technique, which is often used in digital image processing. This technique has been adapted here for use with speech representation. This method provides a transformation

Manuscript received July 21, 2014; revised September 21, 2014.

doi: 10.12720/lnit.2.3.226-231

mapping the histogram of each component of the feature vector onto a reference histogram. This compensates for the effect of noise processes distorting the feature space. The effectiveness of the method relies on estimating the histograms of the speech to be compensated correctly and the assumption that the effect of the noise distortion is a monotonic transformation of the representation space. The first assumption makes the method more effective as more speech frames are involved in estimating the histograms. Generally, the second assumption cannot be verified due to the random behaviour of the noise process.

II. Histogram Equalization for Robust Speech Recognition

HEQ was originated as a technique for digital image processing. Its aim is to provide a transformation $x_1 = F(x_0)$ that converts the probability density function $p_0(x_0)$ of the original variable into a reference probability density function $p_1(x_1) = p_{ref}(x_1)$. The transformation therefore converts the histogram of the original variable into the reference histogram, it equalizes the histogram. The formulation of the method is described below.

Let x0 be a uni-dimensional variable following a distribution $p_0(x_0)$. A transformation $x_1 = F(x_0)$ modifies the probability distribution according to the expression in (1)

$$p_1(x_1) = p_0(G(x_1))\frac{\partial G(x_1)}{\partial x_1} \quad (1)$$

where $G(x_1)$ is the inverse transformation of $F(x_0)$. The relationship between the cumulative probabilities associated with these probability distributions is given by(2)

$$\begin{aligned} C_0(x_0) &= \int_{-\infty}^{x_0} p_0(x_0')dx_0' \\ &= \int_{-\infty}^{F(x_0)} p_0(G(x_1'))\frac{\partial G(x_1)}{\partial x_1'}dx_1' \\ &= \int_{-\infty}^{F(x_0)} p_1(x_1')dx_1' \\ &= C_1(F(x_0)) \end{aligned} \quad (2)$$

And therefore, the transformation $x_1 = F(x_0)$, which converts the distribution $p_0(x_0)$ into the reference distribution $p_1(x_1) = p_{ref}(x_1)$ (and hence converts the cumulative probability $C_0(x_0)$ into $C_1(x_1) = C_{ref}(x_1)$), is obtained from a (3)

$$x_1 = F(x_0) = C_1^{-1}[C_0(x_0)] = C_{ref}^{-1}[C_0(x_0)] \quad (3)$$

where $C_{ref}{}^{-1}[C]$ is the inverse function of the cumulative probability $C_{ref}(x_1)$, providing the value that corresponds to a certain cumulative probability. For practical implementations, a finite number of observations are considered and therefore cumulative histograms are used instead of cumulative probabilities. For this reason the procedure is referred to as histogram equalization rather than probability distribution equalization. The HEQ method is frequently applied in digital image processing as a means of improving the brightness and contrast of digital images and to optimize the dynamic range of the grey level scale. HEQ is a simple and effective method for automatically correcting images that are either too bright or too dark or that have a poor contrast.

III. Application of Histogram Equalization to the Speech Representation

HEQ allows accurate compensation of the effect of any nonlinear transformation of the feature space provided that 1) the transformation is monotonic (and hence does not cause an information loss) and 2) There are sufficient observations of the signal being compensated to allow an accurate estimate of the original probability distribution.

In the case of digital image processing, the brightness and contrast alterations are mainly due to incorrect lighting or nonlinearities in the receptors. These usually correspond to monotonic nonlinear transformations of the grey-level scale. On the other hand, an image typically contains from several thousand to several million pixels. All of them contribute to an accurate estimation of the original probability distributions. This makes HEQ very effective for image processing.

In the case of automatic speech recognition, the speech signal is segmented into frames, with a frame period of about 10ms, and each frame is represented by a feature vector. The number of observations for the estimation of the histograms is much smaller than in the case of image processing (typically several hundred frames per sentence) and also an independent HEQ procedure needs to be applied to each component of the feature vector. If the method is applied for noise compensation, it should be borne in mind that the more frames that are considered when estimating histograms, the more accurate the transformation obtained for the noise compensation will be. Additionally, HEQ is intended to correct monotonic transformations but the random behaviour of the noise makes the transformation no monotonic, resulting in a loss of information in addition to the mismatch. Therefore, like other noise compensation methods, HEQ can deal with the mismatch caused by the noise but not with the loss of information caused by the random behaviour of the noise. This limits the effectiveness of HEQ based noise compensation. We applied HEQ to each component of the feature vector representing each frame of the speech signal. In order to obtain the transformation for each component, the cumulative histogram was estimated by considering 100 uniform intervals between $\mu i - 4\sigma i$ and $\mu i + 4\sigma i$ where $\mu i\ and\ \sigma i$ are the mean the standard deviation for the 8th component of the feature vector, respectively. The transformation was computed according to for the points in the center of each interval and was applied to the parameters to be compensated as a linear interpolation, as “Fig. 1” using the closet pair of points for which the transformation was computed. Original histograms were estimated using the frames of each utterance. Thus, the HEQ method was applied on a sentence-by-sentence basis. The speech representation used was based on the Mel Frequency Cepstral Coefficients (MFCC) and included the logarithm of the energy, the cepstral coefficients and the first and second associated regression coefficients. A Gaussian probability distribution with zero mean and unity variance was used as the reference probability distribution for each component. HEQ was applied as a part of the speech signal parameterization process both during training of the acoustic models and during the recognition process. Figure shows how the HEQ method compensates b for the effect of noise on the speech representation. In this case, we contaminated the speech signal with additive Gaussian white noise at SNRs ranging from 60 dB to 5 dB. The figure shows the effect of the noise and HEQ on the energy coefficient and the 3rdcepstral coefficient. The plots in the first row show the original probability distributions1 for these components and for the different noise levels. As may be seen, the noise severely affects the probability distributions of the speech causing a considerable mismatch when the training and recognition SNRs differ. The plots in the second row show how these coefficients change over time. The speech signal corresponds to the pronunciation of the Spanish digit string “8089.” Again, the mismatch caused by noise can be observed. The plots in the third row show the transformations obtained in each case in order to convert the original histograms into the reference histogram, according to the procedure described above. The histograms of the transformed speech representation are shown in the following plots and, as may be observed, they approximate to the reference Gaussian probability distribution. Finally, the last plots show how the equalized components change over time. In this case, the mismatch caused by the noise is significantly reduced. However, HEQ cannot remove completely the noise effect due to its randomness. Similar plots would be observed for the other components.

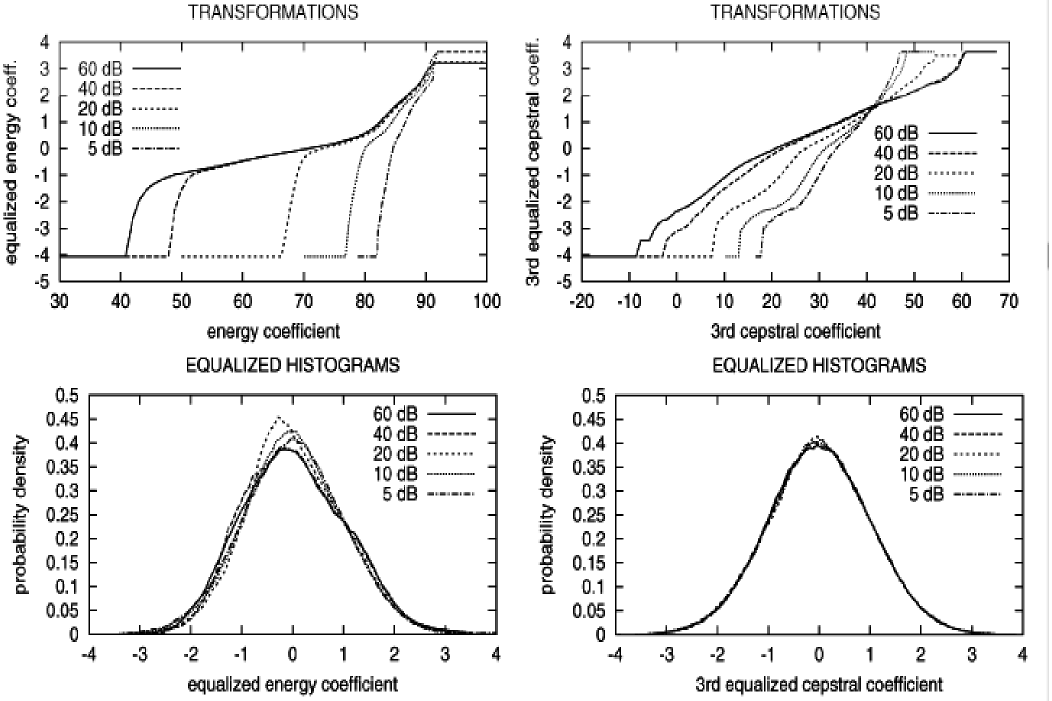

Figure 1. Original histograms estimated frames

IV. Combination of Histogram Equalization with Other Method

One of the particularities of HEQ is that its formulation does not rest on any assumptions about the speech representation or the process causing the distortion. Other methods for robust speech recognition are formulated taking into account the nature of the noise and the mechanisms affecting the speech representation in a given domain. One could expect that such methods provide a compensation of the noise effects that is more accurate than that provided by HEQ. This absence of assumptions could be considered a drawback of the proposed HEQ method.

However, because of this absence of assumptions, HEQ (Fig. 2) is able to deal with distortions coming from different processes. In particular, one could expect a compensation of the residual noise after applying other methods such as spectral subtraction, Wiener filtering or VTS. An additional improvement could be expected from compensation of this residual noise provided by HEQ.

Effect of HEQ over the speech representation for the energy co efficient (plots in the left side) and the third cepstral coefficient (in the right side)

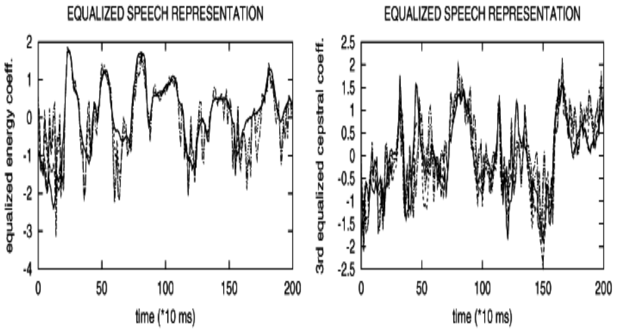

Figure 2. Residual noise HEQ.

A. Voice Conversion

The goal of voice conversion systems is to convert the voice of a source speaker, so it is perceived as being pronounced by another specific speaker, who is called the target speaker. The next subsections describe the most important aspects of the voice conversion systems.

B. Source-Filter Model

The source-filter model is a representation of the phonatory system as a filter being excited by a source signal. The filter that represents the vocal tract of the speaker is modelled by using a series of coaxial tubes. This is made by using an n degree polynomial, where n is the number of tubes used in the model. The coefficients of this polynomial are called Linear Predictive Coding (LPC). The source of the system is the air flow that comes from the lungs and passes through the vocal cords. On the other hand, the inverse model of this filter is a widely used tool to analyze speech, because it decomposes the voice into the excitation and the vocal tract model using LPC coefficients. The next subsections describe the techniques used in this work to convert each component of the vocal tract model.

C. Vocal Tract Conversion

1) Line Spectral Frequencies (LSF) transformation

In the most popular voice conversion systems pairs of source-target Line Spectral Frequencies (LSF) parameters are modelled using an approach of Gaussian mixture models. In some cases, the initialization of the parameters of the model is done by applying the k-means clustering algorithm. In this work, quantized LSF coefficients are clustered using k-histograms and source parameters are transformed into target parameters through a Non-Gaussian approach via the cumulative density function (CDF).The k-means algorithm is one of the most widely used clustering algorithms. Given a set of numeric objects Xi ε D and an integer number k, the k-means algorithm searches for a partition of D into k clusters that minimizes the within groups sum of squared errors (WGSS). This process can be formulated as the minimization of the function P (W, Q) with respect to W and Q, as shown in Equation in “Fig. 3”.

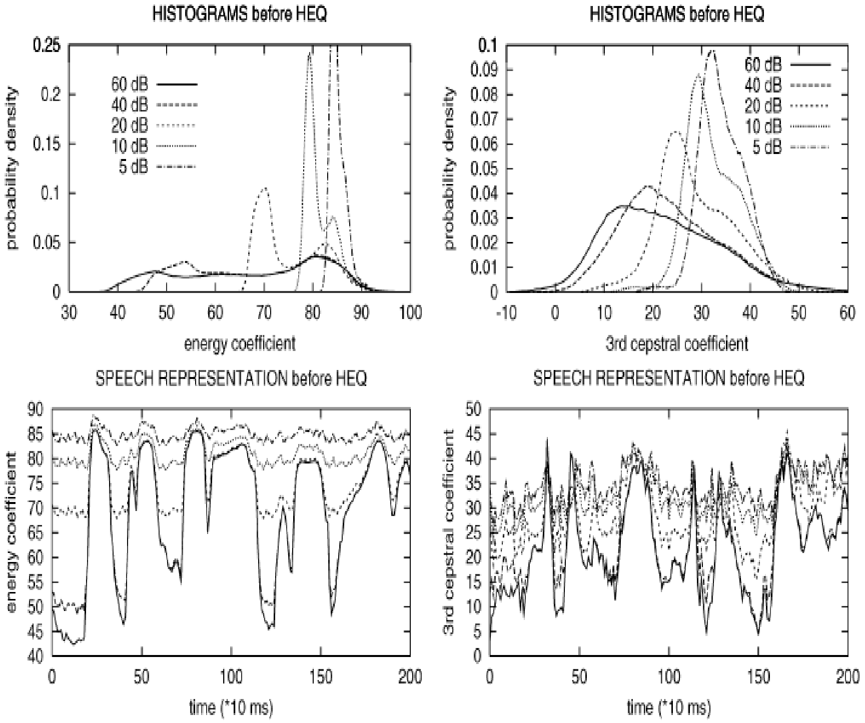

Figure 3. Vocal tract conversion noise

2) Clustering using k-histograms

K-histograms are an interesting approach to cluster categorical data. Each cluster is represented by the histograms of the elements of that cluster. Assuming that each element Xi is a vector of m categorical values xi,1...xi, m, equation can be adapted to categorical data defining a distance based on the histograms of the cluster, as shown in Equation (4).

$$Minimize\, P(W,H)=\sum_{l=1}^{k}\sum_{i=i}^{n} w_{i,l}\, d(X_i, H_L) \qquad (4)$$

©2014 Engineering and Technology Publishing where $w_{i,l}$ is the partition matrix. The distanced compares the histograms of the cluster of each element.

The clustering algorithm is explained in detail by He et al. In this paper, k-histograms are used to partition into sets the vectors of features (LSF parameters) utilized in voice conversion. The LSF parameters are discredited to estimate the counts in the histograms of each set. The source and target LSF vectors are aligned in the training set, and they are jointly partitioned using k-histograms. Then, when estimating by using histograms we make no assumptions about a particular distribution of the parameters. The conversion between source and target parameters using histograms is performed by applying a Non- Gaussian to Non-Gaussian mapping via the cumulative distribution function (CDF) coefficient by coefficient, as shown in equation (5).

$$\hat{y}_i = F_{y_j}^{-1}\left[F_{x_j}\left(x_i\right)\right] \tag{5}$$

The LSF parameter xi of the source speaker is mapped into the target LSF parameter ˆ yi using the CDF of source and target ith LSF parameter and jth set (Fxj and Fyj respectively). The different available sets are obtained using the partition of the LSF parameter space via the k-histograms clustering technique.

The decision about the set j used in the transformation of a given source feature vector x is performed calculating the joint probability of each component of the vector (of dimension K) for each possible set in the equation (6).

$$p_j = \sum_{i}^{k} \log\left(f_{x_j}\left(x_i\right)\right) \tag{6}$$

fxj is the probability that the coefficient xi belongs to set j. The vector belongs to the set j with the highest probability pj .The parameters estimated by means of Eq. 4 are used to perform the synthesis of the target speech. The next subsection explains the proposed conversion method based on the LSF transformation shown in this section.

D. Voice Conversion Using K-Histograms

The voice conversion algorithm using k-histograms can be described in four steps: windowing and parameterization, inverse filtering, parameter transformation and re-synthesis. In the first step, each utterance is divided into overlapping pitch synchronous frames with a width of two periods. An asymmetrical Hamming window is used to minimize boundary effects. The parameterization consists of a 20th order LSF vector. Then, the source excitation (the residual of LPC estimation) is calculated via inverse filtering with the LPC parameters obtained in each frame.

In the third step, the LSF parameters are transformed by using the CDF estimated for the set with the highest probability calculated as shown in Equation. The transformation includes a discretization of the LSF parameters that span from 0 to π The degree of discretization is an adjustable parameter and it is directly related to the amount of available data to estimate histogram counts. The estimated CDF is obtained by means of the training process, where source and target LSF parameter vectors are aligned to obtain the mapping function using k-histograms. The alignment information is extracted from phoneme boundaries provided by a speech recognizer. Inside the boundaries of a frame, the alignment is proportional. Finally, the transformed LSF parameters are transformed into LPC coefficients, and they are used to obtain the target converted voice by filtering the source citation. The fundamental frequency is transformed using a mean and standard deviation normalization and the signal is re-synthesized using TD-PSOLA. In Figure it is shown a scheme of a simple system that makes voice conversion only using K-Histograms, without any transformation of the source residual signal (e_s).

System based on K-Histograms, without modification of the residual signal e_s

Although K-Histograms is an approximation that uses statistical tools likewise the GMM model, Uriz *et al.* obtained a better conversion with this non-Gaussian approach, without introducing assumptions about the distribution of the LSF coefficients. The main drawback of this proposal is the discretization of LSF parameters that introduces noise in the estimation. This is reduced by using a high quantity of levels to discredited the cumulative distribution function (3140 bins), and consequently, this error is negligible.

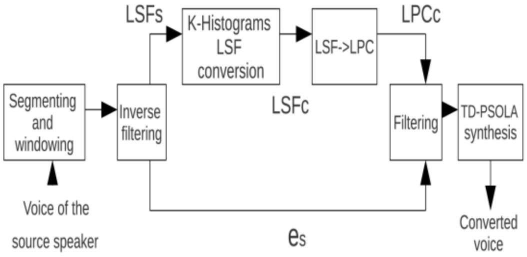

Figure 4. Feature vector or the effect that the different noise processes

V. Conclusion

In this paper, a voice conversion algorithm based on K-Histograms and averaging of residual signal stored in a database was presented. Objective and subjective experiments show that the proposed method has a higher performance in the converted voices. Since the level of similarity of the resulting audios remains the same, a better trade-off of similarity and quality than the system that re-synthesizes the voice using only one residual segment is obtained. This paper shows that the results of the system using residual averaging are slightly lower than the reference system, which uses privileged information for the excitation of the converted voice. This is important because the main limitation of the performance of this system is the size of the pre-recorded database. Thus, when increasing the size of the database, the performance of the system would increase to near the reference value. Also this paper describes an adaptation of the HEQ method to robust speech recognition. Based on an estimation of the histograms for the different components of the feature vectors in the sentence to be recognized, the method provides the transformations (one for each component) that convert the original histograms into a reference one. This method is able to compensate for the nonlinear distortions caused by noise. HEQ compensates for the effect of noise without relying on any prior assumptions about the nature of the components in the feature vector or the effect that the different noise processes affecting the speech signal produce on those components. The compensation technique put forward here has been evaluated with continuous speech recognition experiments in which the signal has been contaminated at different SNRs with different types of noise. The HEQ method has yielded significant improvements in recognition performance under noisy conditions with respect to the baseline recognizer and with respect to linear methods such as CMN and MVN. HEQ can be considered as an extension of CMN and MVN to all the moments of the pdf. This way, HEQ provides appropriate transformations to compensate for the nonlinear effects caused by noise.

REFERENCES

[1] N. Zhang, "The enhancement methods for digital image," *Popular Science & Technology*, vol. 8, pp. 27−28, 2006.

[2] C. F. Chen, C. R. Zhu, and H. Q. Song, "Image enhancement based on butter worth low pass filter," *Odern Electronics Technique*, vol. 30, pp. 163−168, 2007.

[3] Y. L. Shi, F. F. Li, and Y. D. Sun, "Image denoising based on mean filter and wavelet analysis," *Electronic Measurement Technology*, vol. 31, no. 8, pp. 140−142, 2008.

[4] M. Tang, S. D. Ma, and Q. Xiao, "Enhancing far infrared image sequences with model-based adaptive filtering," *Chinese Journal of Computers*, vol. 23, no. 8, pp. 893−896, 2000.

[5] J. L. Zhou and H. Lu, "Image enhancement based on a new genetic algorithm," *Chinese Journal of Computers*, vol. 24, no. 9, pp. 959−964, 2001.

[6] S. G. Chang, B. Yu, and M. Vattereli, "Adaptive wavelet thresholding for image denoising and compression," *IEEE Trans. on Image Processing*, vol. 9, no.9, pp. 1532−1546, 2000.

[7] S. X. Li, R. L. Wang, C. M. Li, L. Xu, and G. X. Li, "New method of image de-noising through wavelet shrinkage based on estimate of noise variance," *Application Research of Computers*, vol. 24, no. 1, pp. 220−221, 2007.

[8] S. Q. Luo and J. Han, "Adaptive template filter and its applications for medical images," *Beijing Biomedical Engineering*, vol. 21, no. 1, pp. 16−18, 2002.

[9] D. Dasgupta, "Advances in artificial immune systems," *IEEE Computational Intelligence Magazine*, vol. 1, no. 4, pp. 40−43, 2006.

[10] M. Glickman, J. Balthrop, and S. A. Forrest, "Machine learning evaluation of an artificial immune system," *Journal of Evolutionary Computation*, vol. 13, no. 2, pp. 179−212, 2005.

[11] M. Omid, A. Mahmoudi, and M. H. Omid, "An intelligent system for sorting pistachio nut varieties," *Expert Systems with Applications*, vol. 36, pp. 11528–11535, 2009.

Himadri Nath Moulick has received B.Tech and M.Tech degree from West Bengal University of Technology, India in 2007 and 2011 respectively. Currently he is an Assistant Professor of Computer Science and Engineering of Aryabhatta Institute of Engineering and Management, India. His teaching and research areas include image processing and crptography. He is an research fellow of Pacific Academic of Higher Education & Research University in Computer Science. He has published two books on Basic Computing with C and Object Oriented Language and UML. He has published more than 40 papers in International Journals and International Conferences. He has published two Indian Patents.

Dr. Chandan Koner did his Ph .D. in Computer Science & Engineering in 2012 from Jadavpur University, India. Now he is Associate Professor of Dr. B. C. Roy Engineering College, India.

Dr. Koner is a Member of the Computer Society of India, Indian Society for Technical Education, Institute of Electronics & Telecommunication Engineers, Cryptology Research Society of India, International Association of Computer Science and Information Technology, Singapore, Computer Science Teachers Association, USA, Universal Association of Computer and Electronics Engineers, Australia, International Rough Set Society, Canada and Senior Member of the Operation Research Society of India. He is reviewer of many international journals and was the reviewer of different IEEE International Conference in India and abroad like Singapore, China. He has three patents and is an author of more than fifty research papers in international journals and conference proceedings. He has delivered many invited talks and tutorials in several engineering colleges, seminars, FDPs and conferences.

Alok Kumar Ray has received B.E degree from Utkal University, India, Computer Science and Engineering department in the year 1994. He has received MBA degree in System and marketing from Utkal University, India in the year 1998. He has received M.Tech degree from West Bengal University of Technology, India in 2008. Currently he is an Associate Professor of Computer Science and Engineering of Bankura Unnayani Institute of Engineering, India. His teaching and research areas include Image Processing and computer organization. He is a research fellow of West Bengal University of Technology, India. His teaching experience in degree Engineering College is almost 13 years. Before his teaching he was engaged in Industry for a period of 3 years