Speaker Identification using Row Mean of DCT and Walsh Hadamard Transform

Dr. H B Kekre[1], Vaishali Kulkarni[2],

[1]Senior Professor, Computer Dept., MPSTME, NMIMS University.
[2]Associate Professor, EXTC Dept., MPSTME, NMIMS University.

Sunil Venkatraman[3], Anshu Priya[4], Sujatha Narasimhan[5]

[3, 4, 5], B-Tech EXTC students, EXTC Dept., MPSTME, NMIMS University.

***Abstract*—** In t**his paper we propose a unique approach to text dependent speaker identification using transformation techniques such as DCT (Discrete Cosine Transform) and WHT (Walsh and Hadamard Transform). The feature vectors for identification are extracted using two different techniques using the transforms, one without overlap and the other with overlap. The results show that accuracy increases as the feature vector size is increased from 64 onwards. But for feature vector size of more than 512 the accuracy again starts decreasing. The maximum accuracy without overlap is more than with overlap for both the transforms. Also the results show that DCT performs better than WHT. The maximum accuracy obtained for DCT is 94.28% for a feature vector size of 512.**

Keywords- Speaker identification, Speaker Recognition, DCT, WHT, Row Mean

I. INTRODUCTION

Security poses the biggest threat in today’s world due to extensive use of internet technology as well as due to multi-user applications. The solution to this problem is attained by granting access to authorized users. Speaker Recognition technology can be used for restricting services to authorized people. This technique makes use of the speaker’s voice to substantiate their identity for the purpose of controlled access to information and reservation service, control of financial transactions, entrance into safe or reserved areas, buildings, voice mail and voice dialing [1 – 2]. Speaker recognition is the process of automatically recognizing the speaker on the basis of individual information subsumed in speech signals. Speaker recognition can be classified into two types: speaker identification and speaker verification. Speaker identification is the process of ascertaining a speech utterance from an unknown speaker by comparing it with speech models of known speakers. Speaker verification is the process of accepting or rejecting the identity claimed by a speaker by comparing it with a model for the speaker whose identity is being claimed [3].

Speaker identification dilemma is categorized into Text-Dependent and Text-Independent systems. The Text-Dependent systems require the speaker to provide utterances of specific predefined key words or sentences, whereas the latter do not rely on a specific text. Compared to text dependent Speaker Identification, text independent Speaker Identification is more convenient because the user can speak freely to the system. However it requires longer training and testing utterances to achieve good performance [4 – 6]. Research and development on speaker recognition methods and techniques has been undertaken for well over four decades and it continues to be an active area. Approaches have spanned from human aural and spectrogram comparisons, to simple template matching, to dynamic time-warping approaches, to more modern statistical pattern recognition approaches, such as neural networks and Hidden Markov Models (HMMs) [7 – 12]. It is interesting to note that, we are still striving to extract and recognize different information from the speech signal. Many of the same features and techniques successfully applied to speech recognition have also been used for speaker recognition. Over this same time, research and development corpora have evolved from small, private corpora (5-10 speaker) under laboratory clean, controlled conditions (single session, read speech) to large, publicly available corpora (500+ speakers) reflecting more realistic and challenging conditions (extemporaneous speech from landline and cellular telephone channels). Benchmark evaluations using common corpora and paradigms have been conducted for several years (e.g. YOHO, CAVE project, NIST) allowing comparison of technical approaches and focusing effort on common challenges. The field has matured to the point that commercial applications of speaker recognition have been steadily increasing since the mid-1980s, with a large number of companies currently offering this technology [13 – 16], but the results are still not satisfactory.

We have proposed speaker recognition using vector quantization in time domain by using LBG (Linde Buzo Gray), KFCG (Kekre’s Fast Codebook Generation) and KMCG (Kekre’s Median Codebook Generation) algorithms [17 - 19] and in transform domain using DFT, DCT and DST [20].

The concept of row mean of the transform techniques has been used for content based image retrieval (CBIR) [21 – 24]. This technique also has been applied on speaker identification by first converting the speech signal into a spectrogram [25]. We have proposed speaker identification using row mean of different transform techniques [26].

In this paper we propose an algorithm for speaker Identification using the DCT and WHT by two methods, namely without overlap and with overlap. The performance of both the transform is analyzed and the results have been shown. The transform techniques have been explained in section II and the two methods of feature extraction are described in section III. Section IV explains the feature matching technique, section V comprises of the results and the conclusion is given in section VI.

II. TRANSFORMS TECHNIQUES

A. Discrete Cosine Transform

A discrete cosine transform (DCT) expresses a sequence of finitely many data points in terms of a sum of cosine functions oscillating at different frequencies. DCTs are important to numerous applications in science and engineering, from lossy compression of audio and images (where small high-frequency components can be discarded), to spectral methods for the numerical solution of partial differential equations.

In particular, a DCT is a Fourier-related transform similar to the discrete Fourier transform (DFT), but using only real numbers. DCTs are equivalent to DFTs of roughly twice the length, operating on real data with even symmetry (since the Fourier transform of a real and even function is real and even). There are eight standard DCT variants, of which four are common. The most common variant of discrete cosine transform is the type-II DCT, which is often called simply "the DCT". The discrete cosine transform is given by (1).

$$y(k) = w(k) \sum_{n=1}^{N} y(n) \cos \frac{\pi(2n-1)(k-1)}{2N} \qquad (1)$$

Where y(k) is the cosine transform, k=1,…, N.

$$w(k) = 1/\sqrt{N} \qquad k=1$$

$$= \sqrt{\left(\frac{2}{N}\right)} \qquad 2 \le k \le N$$

The DCT, and in particular the DCT-II, is often used in signal and image processing, especially for lossy data compression, because it has a strong "energy compaction" property [17 – 20].

B. Walsh Hadamard Transform (WHT)

The Walsh transform or Walsh–Hadamard transform is a non-sinusoidal, orthogonal transformation technique that decomposes a signal into a set of basis functions. These basis functions are Walsh functions, which are rectangular or square waves with values of +1 or –1. The Walsh–Hadamard transform returns sequency values. Sequency is a more generalized notion of frequency and is defined as one half of the average number of zero-crossings per unit time interval. Each Walsh function has a unique sequency value. You can use the returned sequency values to estimate the signal frequencies in the original signal. The Walsh–Hadamard transform is used in a number of applications, such as image processing, speech processing, filtering, and power spectrum analysis. It is very useful for reducing bandwidth storage requirements and spread-spectrum analysis. Like the FFT, the Walsh–Hadamard transform has a fast version, the fast Walsh–Hadamard transform (`fwht`). Compared to the FFT, the FWHT requires less storage space and is faster to calculate because it uses only real additions and subtractions, while the FFT requires complex values. The FWHT is able to represent signals with sharp discontinuities more accurately using fewer coefficients than the FFT. FWHT_h is a divide and conquer algorithm that recursively breaks down a WHT of size N into two smaller WHTs of size $N/2$. This implementation follows the recursive definition of the $2N \times 2N$ Hadamard matrix H_N:

$$H_N = \frac{1}{\sqrt{2}} \begin{pmatrix} H_{N-1} & H_{N-1} \\ H_{N-1} & -H_{N-1} \end{pmatrix} \qquad (2)$$

The $1/\sqrt{2}$ normalization factors for each stage may be grouped together or even omitted. The Sequency ordered, also known as Walsh ordered, fast Walsh–Hadamard transform, FWHT_w, is obtained by computing the FWHT_h as above, and then rearranging the outputs.

III. FEATURE EXTRACTION

If y (t) is the speech signal, the digitized speech signals can be represented by Y[n] where 'n' indicates the sample point which has values from 0 to N-1 and N depends on the sampling frequency and the length of the speech signal. The feature vectors are extracted from these digitized speech samples using two techniques: one without overlap and the other with overlap of the sample points.

A. *Without overlap*

The procedure for feature vector extraction without overlap is given below:

1. The digitized speech signal is divided into groups of n samples. (Where n can take values: 64, 128, 256, 512, 1024, 2048, 4096 and 8192) samples.
2. These blocks are then arranged as columns of a matrix and then transform (either DCT or WHT) is taken.
3. The mean of the absolute values of the rows of the transform matrix is then calculated.
4. These row means form a column vector (1 × n where n is the number of rows in the transform matrix).
5. This column vector forms the feature vector for the speech sample.
6. The feature vectors for all the speech samples are calculated for different values of n and stored in the database.

Fig. 1 shows the row mean (feature vector) generation technique for a speech signal having 15 sample points, which is then divided into groups of 5 sample points. Figure 3 shows the row mean generation with DCT and WHT without overlap for a grouping of 64 sample points for one of the samples in the database

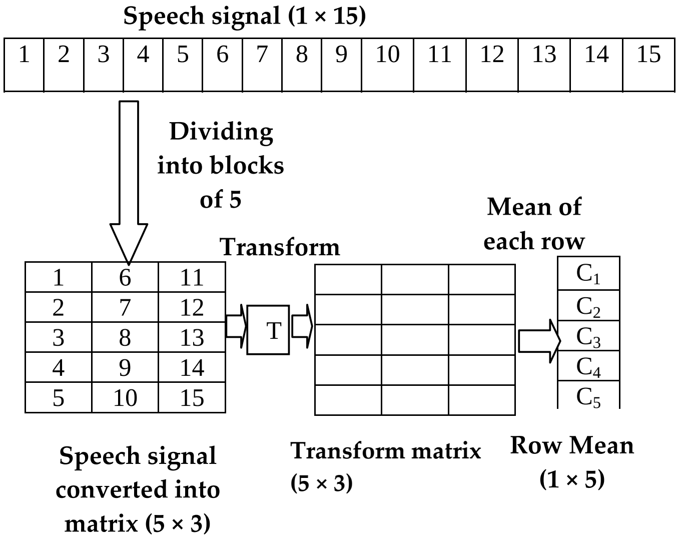

Figure 1. Row Mean Generation Technique without overlap

B. *With overlap*

The procedure for feature vector extraction without overlap is given below:

- The digitized speech signal is divided into groups of n samples. (Where n can take values: 64, 128, 256, 512, 1024, 2048, 4096 and 8192) samples, by considering an overlap of 25% between consecutive blocks.
- E.g. if the first block for a grouping of 64 is from 1 to 64, then the second block will be from 48 to 112 and so on.
- Then the procedure for feature extraction is same as that without overlap as explained in (A) above.

Fig. 2 shows the row mean (feature vector) generation technique for a speech signal having 17 sample points, which is then divided into groups of 5 sample points by considering an overlap of 2 between two consecutive blocks.

IV. FEATURE MATCHING

The matching process involves identifying the speaker. Speaker Identification or matching is done using the minimum Euclidean distance between the feature vector of the test sample and the feature vector of the samples stored in the database. The algorithm for this process is given below:

1. Read the test sample.
2. Extract the feature vector of test sample.
3. Calculate the Euclidean distance between feature vectors of the samples in the database with the test sample.
4. Select the sample which has smallest Euclidean distance with the test sample and declare the speaker corresponding to this sample from the database as the identified speaker.

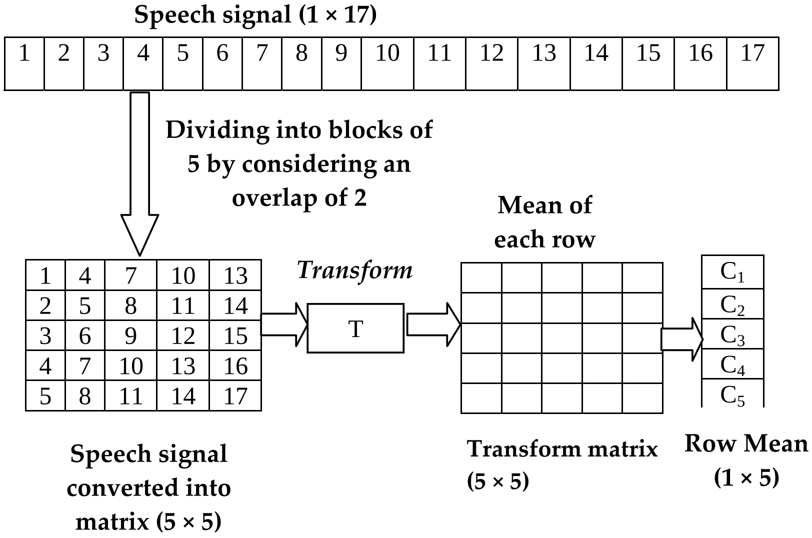

Figure 2. Row Mean Generation Technique with overlap

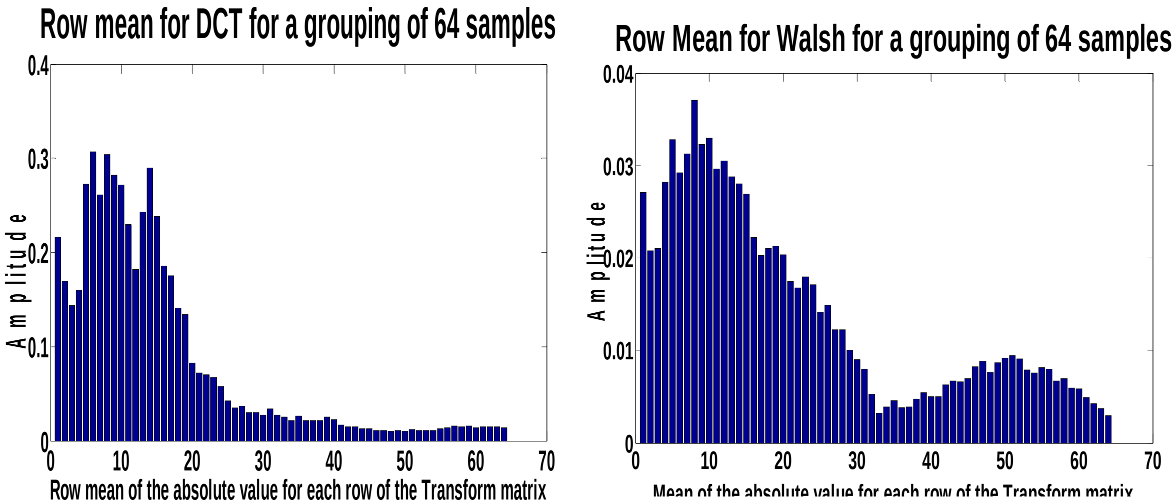

Figure 3. Row Mean Generation without overlap for a grouping of 64 sample points for one of the samples in the database

V. RESULTS

A. *Basics of speech signal*

The speech samples used in this work are recorded using Sound Forge 4.5. The sampling frequency is 8000 Hz (8 bit, mono PCM samples). Table I shows the database description. The samples are collected from different speakers. Samples are taken from each speaker in two sessions so that training model and testing data can be created. Twelve samples per speaker are taken. The samples recorded in one session are kept in database and the samples recorded in second session are used for testing.

TABLE I. DATABASE DESCRIPTION

Parameter	Sample characteristics
Language	English
No. of Speakers	105
Speech type	Read speech
Recording conditions	Normal. (A silent room)
Sampling frequency	8000 Hz
Resolution	8 bps

B. *Experimental Reults*

The feature vectors of all the reference speech samples are stored in the database in the training phase. In the matching phase, the test sample that is to be identified is taken and similarly processed as in the training phase to form the feature vector. The stored feature vector which gives the minimum Euclidean distance with the input sample feature vector is declared as the speaker identified. The accuracy of the identification system is calculated as given by equation 3.

$$Accuracy\ (\%) = \frac{number\ of\ matches}{number\ of\ samples\ tested} \qquad (3)$$

Fig. 4 shows the results obtained for DCT and WHT for different feature vector sizes without overlap. The results show that the accuracy increases as the feature vector size increases from 64 to 512 for DCT (80.95% for a feature vector of size 64 to 94.28% for a feature vector of size 512). As the feature vector size is further increased, accuracy decreases (85.7% for a feature vector of size 8192). For WHT, also, the accuracy increases as the feature vector size is increased from 64 to 1024 (72.38% for a feature vector of size 64 to 84.76% for a feature vector of size 1024). The accuracy then decreases to about 63% for a feature vector of size 8192 samples. The results show that DCT gives better results as compared to WHT. The maximum accuracy is obtained with DCT feature vector of size 512 (94.28%). Fig. 5 shows the results obtained for DCT and WHT for different feature vector sizes with an overlap of 25%. (for a feature vector of size 64, the overlap is of 16, for a feature vector of size 128, the overlap is 32 and so on). The results show a similar pattern as without overlap. The maximum

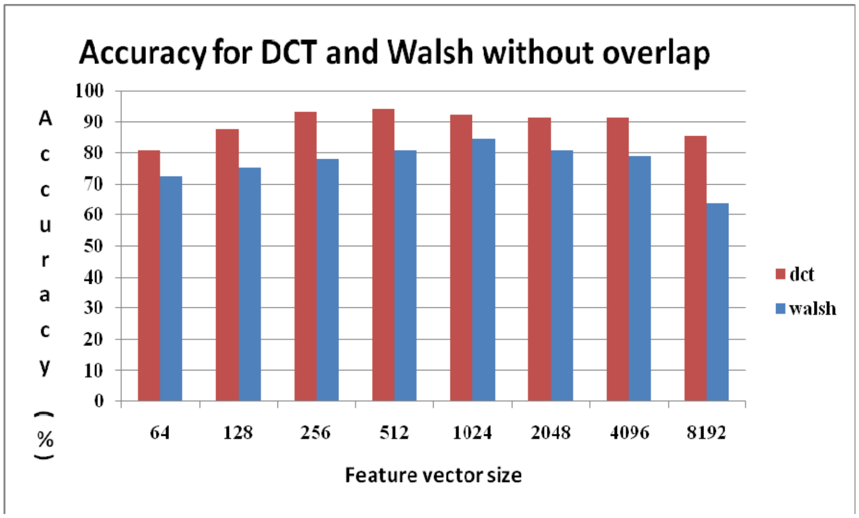

Figure 4. Accuracy of DCT and WHT without overlap for different feature vector sizes

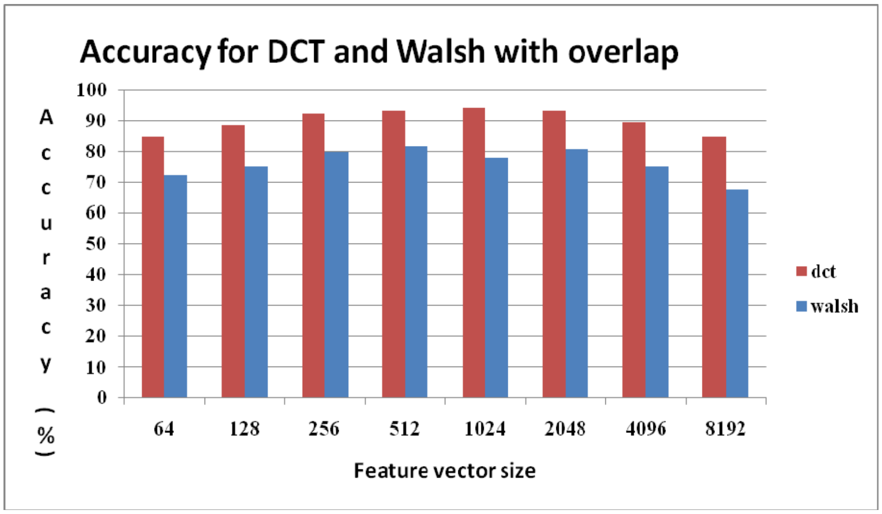

Figure 5. Accuracy of DCT and WHT with overlap for different feature vector sizes

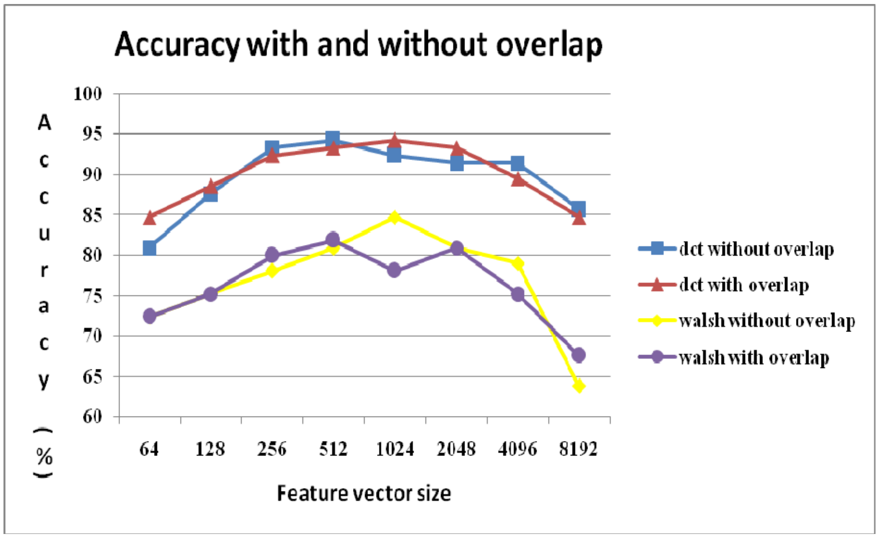

Figure 6. Comparison of results with and without overlap

accuracy for DCT is 94.28% for a feature vector of size 512. For WHT, the maximum accuracy is 81.28% for a feature vector of size 512. Fig. 6 shows the comparison of both the methods, i.e. without overlap and with overlap for both the transforms. The comparison shows that for DCT, the accuracy with overlap is more for a feature vector of size 64 and 128. But as the grouping is increased, there is not much difference and taking overlap does not increase the accuracy. For WHT also a similar trend is observed. As can be seen from the results the maximum accuracy for both the transforms is obtained without taking an overlap. Also the results of DCT are much better than WHT.

VI. CONCLUSION

In this paper we have compared the performance DCT and WHT for different feature vector size with and without overlap for speaker identification. Accuracy increases as the feature vector size is increased from 64 onwards. But for feature vector size of more than 512 the accuracy again starts decreasing. The maximum accuracy without overlap is more than with overlap for both the transforms. Also the results show that DCT performs better than WHT. The maximum accuracy obtained for DCT is 94.28% for a feature vector size of 512. The present study is ongoing and we are analyzing the performance on other transforms.

REFERENCES

[1] Evgeniy Gabrilovich, Alberto D. Berstin: “Speaker recognition: using a vector quantization approach for robust text-independent speaker identification”, Technical report DSPG-95-9-001‟, September 1995.

[2] Tridibesh Dutta, “Text dependent speaker identification based on spectrograms”, Proceedings of Image and vision computing , New Zealand 2007.

[3] Lawrence Rabiner, Biing-Hwang Juang and B.Yegnanarayana, ”Fundamental of Speech Recognition”, Prentice-Hall, Englewood Cliffs, 2009.

[4] S Furui, “50 years of progress in speech and speaker recognition research”, ECTI Transactions on Computer and Information Technology, Vol. 1, No.2, November 2005.

[5] D. A. Reynolds, “An overview of automatic speaker recognition technology,” Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP’02), 2002, pp. IV-4072–IV-4075.

[6] S. Furui. Recent advances in speaker recognition. AVBPA97, pp237--251, 1997.

[7] J. P. Campbell, ``Speaker recognition: A tutorial,'' Proceedings of the IEEE, vol. 85, pp. 1437--1462, September 1997.

[8] D. A. Reynolds, “Experimental evaluation of features for robust speaker identification,” IEEE Trans. Speech Audio Process., vol. 2, no. 4, pp. 639–643, Oct. 1994.

[9] Tomi Kinnunen, Evgeny Karpov, and Pasi Fr¨anti, “Realtime Speaker Identification”, ICSLP2004.

[10] Marco Grimaldi and Fred Cummins, “Speaker Identification using Instantaneous Frequencies”, IEEE Transactions on Audio,Speech, and Language Processing, vol., 16, no. 6, August 2008.

[11] Zhong-Xuan, Yuan & Bo-Ling, Xu & Chong-Zhi, Yu. (1999). “Binary Quantization of Feature Vectors for Robust Text-Independent Speaker Identification” in IEEE Transactions on Speech and Audio Processing, Vol. 7, No. 1, January 1999. IEEE, New York, NY, U.S.A

[12] F. Bimbot, J.-F. Bonastre, C. Fredouille, G. Gravier, I.Magrin-Chagnolleau, S. Meignier, T. Merlin, J. Ortega-García,D.Petrovska-Delacrétaz, and D. A. Reynolds, “A tutorial on text-independent speaker verification,” EURASIP J. Appl.Signal Process., vol. 2004, no. 1, pp. 430–451, 2004.

[13] Special Issue on Speaker Recognition, Digital Signal Processing, vol. 10, January 2000.

[14] B. S. Atal, “Automatic Recognition of speakers from their voices”, Proc. IEEE, vol. 64, 1976.

[15] http://www.nist.gov/speech/tests/spk/index.htm

[16] D. O’Shaughnessy, “Speech communications- Man and Machine”, New York, IEEE Press, 2nd Ed., 2000.

[17] H B Kekre, Vaishali Kulkarni, “Speaker Identification by using Vector Quantization”, International Journal of Engineering Science and Technology, May 2010.

[18] H B Kekre, Vaishali Kulkarni, “Performance Comparison of Speaker Recognition using Vector Quantization by LBG and KFCG”, International Journal of Computer Applications, vol. 3, July 2010.

[19] H B Kekre, Vaishali Kulkarni, “Performance Comparison of Automatic Speaker Recognition using Vector Quantization by LBG KFCG and KMCG”, International Journal of Computer Science and Security, Vol: 4 Issue: 4, 2010.

[20] [20] H B Kekre, Vaishali Kulkarni, “Comparative Analysis of Automatic Speaker Recognition using Kekre’s Fast Codebook Generation Algorithm in Time Domain and Transform Domain”, International Journal of Computer Applications, Volume 7 No.1. September 2010.

[21] Dr. H.B.Kekre, Sudeep D. Thepade, Akshay Maloo “Performance Comparision of Image Retrieval using Row Mean of Transformed Column Image”, International Journal on Computer Science and Engineering Vol. 02, No. 05, 2010, 1908-1912

[22] Dr.H.B.Kekre,Sudeep Thepade “Edge Texture Based CBIR using Row Mean of Transformed Column Gradient Image”, International Journal of Computer Applications (0975 – 8887) Volume 7– No.10, October 2010

[23] Dr. H.B.Kekre, Sudeep D. Thepade, Akshay Maloo “Eigenvectors of Covariance Matrix using Row Mean and Column Mean Sequences for Face Recognition”, International Journal of Biometrics and Bioinformatics (IJBB), Volume (4): Issue (2)

[24] Dr. H.B.Kekre, Sudeep Thepade, Archana Athawale, “Grayscale Image Retrieval using DCT on Row mean, Column mean and Combination”, Journal of Sci., Engg. & Tech. Mgt. Vol 2 (1), January 2010

[25] Dr. H. B. Kekre, Dr. T. K. Sarode, Shachi J. Natu, Prachi J. Natu “Performance Comparison of Speaker Identification Using DCT, Walsh, Haar on Full and Row Mean of Spectrogram”, International Journal of Computer Applications (0975 – 8887) Volume 5– No.6, August 2010.

[26] Dr. H.B. Kekre, Vaishali Kulkarni, “Comparative Analysis of Speaker Identification using row mean of DFT, DCT, DST and Walsh Transforms”, International Journal of Computer Science and Information Security, Vol. 9,No. 1, 2011.

AUTHORS PROFILE

Dr. H. B. Kekre has received B.E. (Hons.) in Telecomm. Engg. from Jabalpur University in 1958, M.Tech (Industrial Electronics) from IIT Bombay in 1960, M.S.Engg. (Electrical Engg.) from University of Ottawa in 1965 and Ph.D. (System Identification) from IIT Bombay in 1970. He has worked Over 35 years as Faculty of Electrical Engineering and then HOD Computer Science and Engg. at IIT Bombay. For last 13 years worked as a Professor in Department of Computer Engg. at Thadomal Shahani Engineering College, Mumbai. He is currently Senior Professor working with Mukesh Patel School of Technology Management and Engineering, SVKM’s NMIMS University, Vile Parle(w), Mumbai, INDIA. He ha guided 17 Ph.D.s, 150 M.E./M.Tech Projects and several B.E./B.Tech Projects. His areas of interest are Digital Signal processing, Image Processing and Computer Networks. He has more than 300 papers in National / International Conferences / Journals to his credit. Recently twelve students working under his guidance have received best paper awards. Recently two research scholars have received Ph. D. degree from NMIMS University Currently he is guiding ten Ph.D. students. He is member of ISTE and IETE.

Vaishali Kulkarni has received B.E in Electronics Engg. from Mumbai University in 1997, M.E (Electronics and Telecom) from Mumbai University in 2006. Presently she is pursuing Ph. D from NMIMS University. She has a teaching experience of more than 8 years. She is Associate Professor in telecom Department in MPSTME, NMIMS University. Her areas of interest include Speech processing: Speech and Speaker Recognition. She more than 10 papers in National / International Conferences / Journals to her credit.