Speaker Identification using Frequency Dsitribution in the Transform Domain

Dr. H B Kekre
Senior Professor, Computer Dept.,
MPSTME, NMIMS University,
Mumbai, India.

Vaishali Kulkarni
Associate Professor, Electronics and Telecommunication,
MPSTME, NMIMS University,
Mumbai, India.

Abstract— **In this paper, we propose Speaker Identification using the frequency distribution of various transforms like DFT (Discrete Fourier Transform), DCT (Discrete Cosine Transform), DST (Discrete Sine Transform), Hartley, Walsh, Haar and Kekre transforms. The speech signal spoken by a particular speaker is converted into frequency domain by applying the different transform techniques. The distribution in the transform domain is utilized to extract the feature vectors in the training and the matching phases. The results obtained by using all the seven transform techniques have been analyzed and compared. It can be seen that DFT, DCT, DST and Hartley transform give comparatively similar results (Above 96%). The results obtained by using Haar and Kekre transform are very poor. The best results are obtained by using DFT (97.19% for a feature vector of size 40).**

Keywords-Speaker Identification; DFT; DCT; DST; Hartley; Haar; Walsh; Kekre's Transform.

I. INTRODUCTION

Recently a lot of work is being carried out in the field of biometrics. There are several categories of biometrics like fingerprint, iris, face, palm, signature voice etc. Voice as a biometric has certain advantages over other biometrics like: it is easy to implement, no special hardware is required, user acceptability is more, and remote login is possible [1]. In spite of these advantages it has not been implemented to a very large extent because of the problems like security, changes in human voice etc. Human beings are able to recognize a person by hearing his voice. This process is called Speaker Identification. Speaker Identification falls under the broad category of Speaker Recognition [2 – 4], which covers Identification as well as Verification.

Speaker Identification (also known as closed set identification) is a 1: N matching process where the identity of a person must be determined from a set of known speakers [4 - 6]. Speaker Verification (also known as open set identification) serves to establish whether the speaker is who he claims to be [7]. Speaker Identification can be further classified into text-dependent and text-independent systems. In a text dependent system, the system knows what utterances to expect from the speaker. However, in a text-independent system, no assumptions about the text can be made, and the system must be more flexible than a text dependent system. Speaker Recognition systems have been developed for a wide range of applications like control access to restricted services, for example, for giving commands to computer, phone access to banking, database services, shopping or voice mail, and access to secure equipment [8 - 11]. Speaker Identification encompasses two main aspects: feature extraction and feature matching. Traditional methods of speaker recognition use MFCC (Mel Frequency Cepstral Coefficients) [13 – 16], LPC (Linear Predictive Coding) [12] for feature extraction. Feature matching has been done using Vector Quantization [17 – 21], HMM (Hidden Markov Model) [21 – 22], GMM (Gaussian Mixture Model) [23].

We have proposed Speaker Identification using row mean of DFT, DCT, DST and Walsh Transforms on the speech signal [24 – 25].We have proposed speaker recognition using the concept of row mean of the transform techniques on the spectrogram of the speech signal [26]. We have also proposed speaker identification using power distribution in the frequency domain [27 - 28].

In this paper we have extended the technique of power distribution of the frequency domain to four more transforms i.e. Hartley, Walsh, Haar and Kekre Transform. Here we have used the power distribution in the frequency domain to extract the features for the reference as well as test speech samples. The feature matching has been done using Euclidean distance. The various transform techniques have been explained in section II. In Section III, the feature vector extraction is explained. Results are discussed in section IV and conclusion ion section V.within parentheses, following the example.

II. TRANSFORM TECHNIQUES

The Transform when applied on a speech signal converts the converts it from time domain to frequency domain. In this paper seven different Transform techniques have been used. Let y(t) be the speech signal in the time domain and y0, y1, y2, yN-1 be the samples of y(t) in the time domain. The Discrete Fourier Transform of this signal is given by (1). The DFT is implemented using Fast Fourier Transform (FFT).

$$Y_k = \sum_{n=0}^{N-1} Y_n e^{-j2\pi kn/N} \tag{1}$$

Where y_n=y(nΔt) is the sampled value of continuous signal y(t); k= 0, 1, 2…, N-1.Δt is the sampling interval.

The discrete cosine transform which is closely related to the DFT has been used in compression because of its capability of reconstruction with a few coefficients.

The DCT of the signal y(t) can be given by (2) and w_k as given by (3).

$$Y_k = w_k \sum_{n=1}^{N} y_n \cos\frac{\pi(2n-1)(k-1)}{2N} \tag{2}$$

$$\begin{array}{ll} w_k = 1/\sqrt{N} & \text{For k=1} \\ \quad = \sqrt{\frac{2}{N}} & 2\le k\le N \end{array} \tag{3}$$

A discrete sine transform (DST) expresses a sequence of finitely many data points in terms of a sum of sine functions. The DST of the signal y(t) can be given by (4).

$$Y_k = \sum_{n=1}^{N} y(n)\sin(\pi\frac{kn}{N+1}) \tag{4}$$

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 is used in a number of applications, such as image processing, speech processing, filtering, and power 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 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 recursion of the definition 2N Hadamard 2N× matrix H_N *as given by (5).*

$$H_N = \frac{1}{\sqrt{2}}\begin{bmatrix} H_{N-1} & H_{N-1} \\ H_{N-1} & -H_{N-1} \end{bmatrix} \tag{5}$$

A discrete Hartley transform (DHT) is a real transform similar to the discrete Fourier transform (DFT). If the speech signal is represented by y(t) then the DHT is given by (6).

$$Y_k = \sum_{n=0}^{N-1} y_n[\cos(\frac{2\pi}{N}nk) + \sin(\frac{2\pi}{N}nk)] \tag{6}$$

The Haar transform is derived from the Haar matrix. The Haar transform is separable and can be expressed in matrix form as shown in (7).

$$[F] = [H].[f].[H]^T \tag{7}$$

Where [f] is an N×1 signal, [H] is an N×N Haar transform matrix and [F] is an N×1 transformed signal. The transformation H contains sampled version of the Haar basis function $h_k(t)$ which are defined over the continuous closed interval t Є [0, 1].

The Haar basis functions are

- When k=0, the Haar function is defined as a constant as in (8).

$$h_0(n) = 1/\sqrt{N} \tag{8}$$

- When k>0, the Haar function is defined as in (9).

$$h_k(n) = \begin{cases} 2^{p/2} & (q-1)/2^p \le t \le (q-0.5)/2^p \\ -2^{p/2} & (q-0.5)/2^p \le t \le q/2^p \\ 0 & \text{Otherwise} \end{cases} \tag{9}$$

Where 0 ≤ p < log2N and 1 ≤ q ≤ 2p

For example, when N=4, we have H_4 as given by (10).

$$H_4 = \frac{1}{2}\begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 \\ \sqrt{2} & -\sqrt{2} & 0 & 0 \\ 0 & 0 & \sqrt{2} & -\sqrt{2} \end{bmatrix} \tag{10}$$

Kekre Transform matrix can be of any size N x N, which need not have to be in powers of 2 (as is the case with most of other transforms including Haar Transform). All upper diagonal and diagonal values of Kekre transform matrix are one, while the lower diagonal part except the values just below diagonal are zero. Generalized N×N Kekre Transform Matrix can be given as in (11). The formula for generating the term K_{xy} of Kekre transform matrix is given by (12).

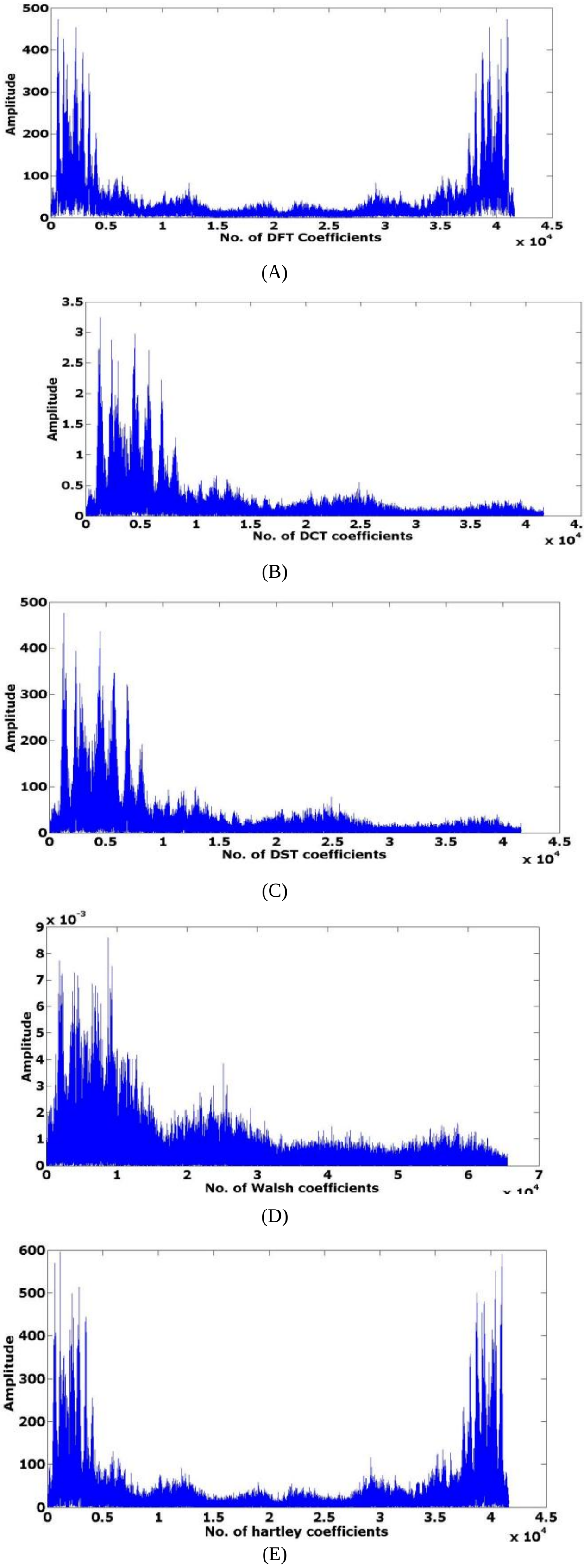
$$K_{N\times N} = \begin{bmatrix} 1 & 1 & 1 & .. & 1 & 1 \\ -N+1 & 1 & 1 & .. & 1 & 1 \\ 0 & -N+2 & 1 & .. & 1 & 1 \\ : & : & : & : & : & : \\ 0 & 0 & 0 & 0 & -N+(N-1) & 1 \end{bmatrix} \tag{11}$$

$$K_{xy} = \begin{cases} 1 & ; x \le y \\ -N+(x-1) & ; x = y+1 \\ 0 & ; x > y+1 \end{cases} \tag{12}$$

III. Feature Extraction

The feature vector extraction process is described as below.

1. The speech signal was converted into frequency domain by applying the transform techniques described in section II, for three different lengths of speech signal. (8.192 sec, 4.096 sec and 2.048 sec) as it gives 2^{16}, 2^{15} and 2^{14} samples at 8 KHz sampling rate.
2. The magnitude of the signal in the transform domain was considered for feature extraction. Figure 1 shows the magnitude plot of the various transforms for the speech signal of length 8.192 sec.



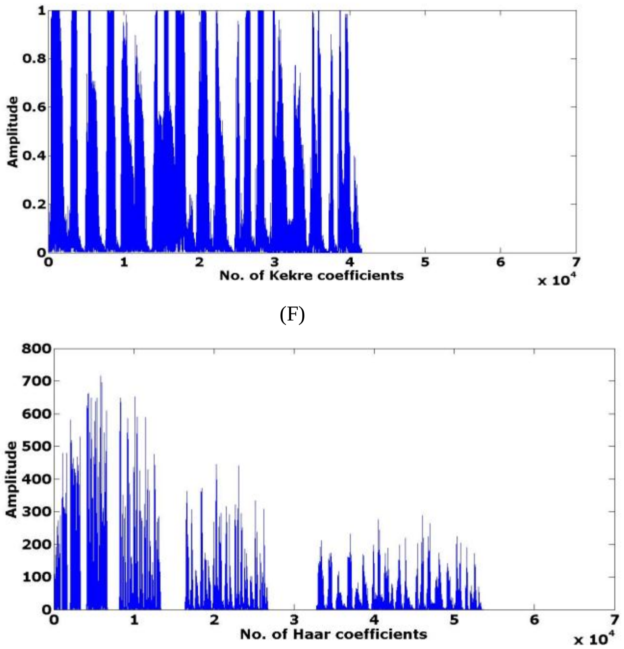
(A)

(B)

(C)

(D)

(E)



(F)

(G)

Figure 1. Frequency Spectrum of the different transforms. (A)FFT, (B) DCT (C) DST (D) Walsh (E) Hartley (F) Kekre (G) Haar

3. This was then divided into various groups and the sum of the magnitude for each group forms the feature vector.

IV. Experimental Results

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 speakers of different age group ranging from 12 to 75 years. Five iterations of four different sentences of varying lengths are recorded from each of the speakers. Twenty samples per speaker are taken. For text dependent identification, four iterations of a particular sentence are kept in the database and the remaining one iteration is used for testing. These speech signals have an amplitude range of '-1' to '+1'.

TABLE I. Database Description

Parameter	Sample characteristics
Language	English
No. of Speakers	107
Speech type	Read speech, microphone recorded
Recording conditions	Normal
Sampling frequency	8000 Hz
Resolution	8 bps

The simulation was done using MATLAB 7.7.0. For DFT, the FFT algorithm was used to calculate the transform coefficients. For DCT, DST and Walsh, the in-built functions in MATLAB were used. To calculate the Hartley Transform coefficients, first the FFT of the real part of speech signal was calculated and then the imaginary part of the complex transform was subtracted from its real part. This is shown in by (13).

$$\begin{aligned} Y1 &= fft(y) \\ Y2 &= real(Y1) - imaginary(Y1) \end{aligned} \qquad (13)$$

For calculating the Kekre Transform, the difficulty was to generate the Transform matrix of the order of 65536×65536, 32768×32768 and 16384×16384 which gave 'out of memory' error.

Instead of computing the transform matrix, the coefficients were calculated as given in (14).

$$\begin{aligned} S_1 &= \sum_{n=0}^{N-1} y_n & ; k = 0 \\ S_k &= S_1 - \sum_{n=0}^{N-2} y_n - (n-k) y_{N-1} & ; 0<k\leq N\text{-}1 \end{aligned} \qquad (14)$$

For calculating the Haar Transform coefficients also, the same order of Transform matrix was required. Again here also, the problem was solved by directly calculating the coefficients using the butterfly diagram approach. Thus after transforming the signal into transform domain, the magnitude plot was generated as shown in figure 1. As can be seen from the magnitude plots, the energy concentration is in the lower order coefficients. This concept was utilized and the frequency spectrum was divided into groups and the sum of the magnitude for each group formed the feature vector. The feature vectors of all the reference speech samples were calculated for the different transforms and 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 (15).

$$Accuracy(\%) = \frac{no\_of\_samples\_identified}{Total\_no.\_of\_samples\_tested} \times 100 \qquad (15)$$

The sentences in the database are of varying sizes. We have performed the simulations for three different lengths of the sentences. In the first case we considered only the first 2.048 sec (16384 samples) of the sentence for each speaker in the training as well as in the testing phase. Figure 2 shows the accuracy obtained for different Transforms for the speech signal of length 2.048 sec (16384 samples). We have begun by taking the entire spectrum as one group and then taking the sum of the magnitude as the feature vector. In this case there is only one element in the feature vector. As can be seen the accuracy is very less for all the transforms. For FFT we get an accuracy of around 6.54%. As we divide the spectrum into more number of groups and then take the sum of each group as the element of the feature vector, the accuracy goes on increasing. For FFT, the accuracy is 93.45% for a feature vector of size 56. Above a feature vector of size 56, the accuracy decreases and we an accuracy of 92.52% for a feature vector of size 88. DCT and DST also show a similar trend, with a maximum accuracy of 89.71% for a feature vector of size 40. With Walsh transform though the trend is similar, the maximum accuracy is only 79.43% for a feature vector of size 80. Hartley transform shows a behavior similar to FFT and the maximum accuracy is 93.45% for a feature vector of size 56. As can be seen from the magnitude spectrum also, the energy compaction in case of Kekre transform and Haar transform is less than other transforms. This explains the lower performance for both the transforms, Kekre transform 41.12% and Haar transform 60.74%. For the second set of simulations, the first 4.096 sec of the sentence spoken by each speaker was considered in the training as well as in the testing phase. Figure 3 shows the results obtained for this set of experiments. As can be seen from figure 3, the overall trend shown by each transform is the same as in figure 2. But here the effect of the increase in length of the speech signal considered is that the accuracy increases. With FFT, the maximum accuracy 97.19% for a feature vector of size 48. For DCT and DST, the maximum accuracy is 95.32% for a feature vector of size 48. With Walsh transform, the maximum accuracy is now around 85%. Hartley transform gives a maximum accuracy of 96.26% for a feature vector of size 48. There is no significant improvement as far as the Kekre transform and Haar transform are considered. Overall there is a gain in accuracy by increasing the length of the speech signal under consideration. Figure 4 shows the results obtained by increasing the length of the speech signal to 8.192 sec (64536 samples). If the length of

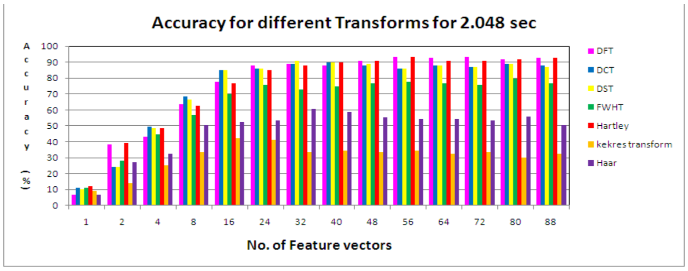


Figure 1. Accuracy for different Transforms for 2.048 sec

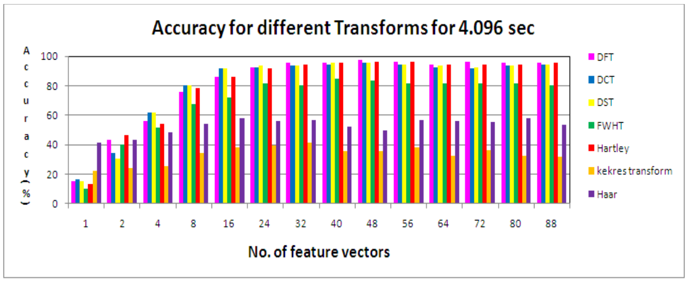


Figure 2. Accuracy for different Transforms for 4.096 sec

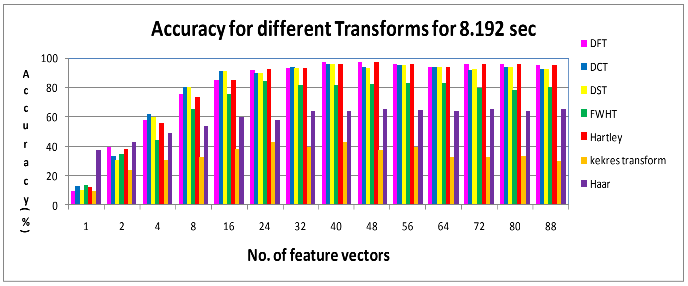


Figure 3. Accuracy of different Transforms for 8.192 sec

the speech signal is smaller than 8.192 sec, then it is padded with zeros to make them all of equal length. As can be seen from the results, there is not much gain over that obtained by considering 4.096 sec. the maximum accuracy is still 97.19% for FFT with feature vector of size 40 now. The trend shown by all the transforms remains the same.

The overall results indicate that the accuracy increases with the increase in the size of feature vector up to a certain point and then it decreases. FFT, DCT, DST and Hartley transforms give very good results. Walsh gives comparatively lower results. Haar and Kekre transform give lesser accuracy compared to all other transforms. This technique of using the magnitude spectrum is very simple to implement and gives comparable results with the traditional techniques used for speaker identification. For the present study we have not used any preprocessing techniques for the speech signal. The database is collected using different brands of locally available microphones under normal conditions. This shows that the results obtained are independent of the recording instrument specifications.

V. Conclusion and Future Scope

In this paper we have shown a comparative performance of speaker identification by using seven different transform techniques. The approach used in this work is entirely different from the studies which have been done in this area. Here we are simply using the distribution in the magnitude spectrum for feature vector extraction. Also for feature matching we are using minimum Euclidean distance as a measure. This makes the system very easy to implement. The maximum accuracy is 97.19% with FFT for a feature vector of size 48. The present study is ongoing and we are trying to analyze the transform domain still further, as it has proved to be a promising way for feature vector extraction. Different algorithms for extracting the feature vector using transforms are being developed.

References

[1] Lisa Myers, An Exploration of Voice Biometrics, GSEC Practical Assignment version 1.4b Option 1, 2004

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

[3] 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.

[4] 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.

[5] Joseph P. Campbell, Jr., Senior Member, IEEE, “Speaker Recognition: A Tutorial”, Proceedings of the IEEE, vol. 85, no. 9, pp. 1437-1462, September 1997.

[6] S. Furui. Recent advances in speaker recognition. AVBPA97, pp 237--251, 1997

[7] 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.

[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, “Real-time 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.

[12] J.Tierney,”A study of LPC Analysis ofspeech in additive noise”, IEEE Trans. Acoust., Speech Signal Processing, vol. ASSP-28, pp 389-397, Aug 1980.

[13] Sandipan Chakroborty and Goutam Saha, Improved Text-Independent Speaker Identification using Fused MFCC & IMFCC Feature Sets based on Gaussian Filter., International Journal of Signal Processing 5, Winter 2009.

[14] Speaker recognition using MFCC by S. Khan, Mohd Rafibul lslam, M. Faizul, D. Doll, presented in IJCSES (International Journal of Computer Science and Engineering System) 2(1): 2008.

[15] Speaker identification using MFCC coefficients –Mohd Rasheedur Hassan, Mustafa Zamil, Mohd Bolam Khabsani, Mohd Saifur Rehman, 3rd international conference on electrical and computer engineering (ICECE), (2004).

[16] Molau, S, Pitz, M, Schluter, R, and Ney, H., Computing Mel-frequency coefficients on Power Spectrum, Proceedings of IEEE ICASSP-2001, 1: 73-76 (2001).

[17] C.D. Bei and R.M. Gray.An improvement of the minimum distortion encoding algorithm for vector quantization, IEEE Transactions on Communications, October (1998).

[18] F. Soong, A. Rosenberg, L. Rabiner, and B-H. Juang, “A Vector Quantization Approach to Speaker Recognition,” In International Conference on Acoustics, Speech, and Signal Processing in Florida, IEEE, pp. 387-390,1985.

[19] F. K. Soong, A. E. Rosenberg, L. R. Rabiner, and B. H. Juang, “A Vector Quantization Approach to Speaker Recognition.” AT&T Technical Journal, Vol. 66, No. 2, pp. 14-26, 1987.

[20] Burton D.K. “Text-dependent Speaker verification using VQ source coding”, IEEE Transactions on Acoustics, Speech and Signal processing, vol. ASSP 35 No. 2 February 1987, pp 133-143.

[21] T Matsui and S Furui, “Comparison of Text Independent speaker recognition methods using VQ-distortion and discrete/continuous HMMs”, in Proc. IEEE ICASSP, Mar. 1992, pp. II. 157 – II.164.

[22] N. Z. Tishby, “On the Application of Mixture AR Hidden Markov Models to Text Independent Speaker Recognition,” IEEE Transactions on Acoustics, Speech, and Signal Processing, Vol. 39, No. 3, pp. 563 – 570, 1991

[23] D A Reynolds, “A Gaussian mixture modeling approach to text independent speaker identification, PhD Thesis, Georgia Inst. of Technology, Sept. 1992.

[24] 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, January 2011.

[25] Dr. H B Kekre, Vaishali Kulkarni, Sunil Venkatraman, Anshu Priya, Sujatha Narashiman, “Speaker Identification using Row Mean of DCT and Walsh Hadamard Transform”, International Journal on Computer Science and Engineering, Vol. 3, No.1, March 2011.

[26] Dr. H B Kekre, Vaishali Kulkarni, “Speaker Identification using Row Mean of Haar and Kekre’s Transform on Spectrograms of Different Frame Sizes”, (IJACSA) International Journal of Advanced Computer Science and Applications, Special Issue on Artificial Intelligence.

[27] Dr. H B Kekre, Vaishali Kulkarni,”Speaker Identification using Power Distribution in Frequency Spectrum”, Technopath, Journal of Science, Engineering & Technology Management, Vol. 02, No.1, January 2010.

[28] Dr. H B Kekre, Vaishali Kulkarni, “Speaker Identification by using Power Distribution in Frequency Spectrum”, ThinkQuest - 2010 International Conference on Contours of Computing Technology”, BGIT, Mumbai,13th -14th March 2010.

AUTHORS PROFILE

Dr. H. B. Kekre has received B.E. (Hons.) in Telecomm. Engineering, 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 has 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 450 papers in National / International Conferences / Journals to his credit. Recently twelve students working under his guidance have received best paper awards. Recently five research scholars have received Ph. D. degree from NMIMS University Currently he is guiding eight Ph.D. students. He is member of ISTE and IETE.

Vaishali Kulkarni. Author 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 around 10 years. She is Associate Professor in telecom Department in MPSTME, NMIMS University. Her areas of interest include networking, Signal processing, Speech processing: Speech and Speaker Recognition. She has 17 papers in National / International Conferences / Journals to her credit.