# SPECTRUM ANALYSIS OF SPEECH RECOGNITION VIA DISCRETE TCHEBICHEF TRANSFORM

Ferda Ernawan [1] and Nur Azman Abu, Nanna Suryana[2]
[1]Faculty of Information and Communication Technology Universitas Dian Nuswantoro (UDINUS)
Semarang, Indonesia
[2]Faculty of Information and Communication Technology Universiti Teknikal Malaysia Melaka (UTeM)
Melaka, Malaysia
[1]e-mail: ferda1902@gmail.com ,[2]e-mail: nura@utem.edu.my, nsuryana@utem.edu.my

## ABSTRACT

Speech recognition is still a growing field. It carries strong potential in the near future as computing power grows. Spectrum analysis is an elementary operation in speech recognition. Fast Fourier Transform (FFT) is the traditional technique to analyze frequency spectrum of the signal in speech recognition. Speech recognition operation requires heavy computation due to large samples per window. In addition, FFT consists of complex field computing. This paper proposes an approach based on discrete orthonormal Tchebichef polynomials to analyze a vowel and a consonant in spectral frequency for speech recognition. The Discrete Tchebichef Transform (DTT) is used instead of popular FFT. The preliminary experimental results show that DTT has the potential to be a simpler and faster transformation for speech recognition.

**Keyword-**Speech recognition, Fast Fourier Transforms, Discrete Cosine Transform and Discrete Tchebichef Transform.

## 1. INTRODUCTION

Speech signal methods using Fourier transform are commonly used in speech recognition. One of the most widely used speech signal methods is the Fast Fourier Transform (FFT). FFT is a basic technique for digital signal processing applicable for spectrum analysis. The FFT is often used to compute numerical approximations to continuous Fourier. However, a straightforward application of the FFT to computationally often requires a large FFT to be performed even though most of the input data to the FFT may be zero [1].

Another transformation is Discrete Cosine Transform (DCT). DCT is a discrete transform whose kernel is defined by the cosine function. It is not popular to use in speech recognition, although it produces a clear speech signal representation and spectrum analysis. DCT does not produce clear efficient third formant $F_3$ in speech recognition.

The Discrete Tchebichef Transform (DTT) is another transform method based on discrete Tchebichef polynomials [2][3]. DTT has a lower computational complexity and it does not require complex transform unlike continuous orthonormal transforms. DTT does not involve any numerical approximation. DTT has been applied in several computer vision and image processing application in previous work. For example, DTT is used in image analysis [4][5], texture segmentation [6], multispectral texture [7], pattern recognition [8], image watermarking [9], monitoring crowds [10], image reconstruction [2][11][12], image projection [13] and image compression [14]-[16]. However, DTT has not been used in audio processing.

A brief description on FFT, DCT and DTT is given in Section II. Section III presents the experimental results of spectrum analysis on speech recognition via FFT, DCT and DTT. Section IV emphasizes on the importance of third formant $F_3$ in speech recognition, comparative speech signal and spectrum analysis among FFT, DCT and DTT. Lastly, section V will conclude the comparison of spectrum analysis via FFT, DCT and DTT.

International Conference on Graphic and Image Processing (ICGIP 2011), edited by Yi Xie, Yanjun Zheng,
Proc. of SPIE Vol. 8285, 82856L · © 2011 SPIE · CCC code: 0277-786X/11/$18 · doi: 10.1117/12.913491

## 2. TRANSFORMATION DOMAIN

### 2.1 Fast Fourier Transform

FFT is an efficient algorithm that can perform Discrete Fourier Transform (DFT). FFT is applied in order to convert time domain signals $x(j)$ into the frequency domain $X(k)$. The sequence of $N$ complex numbers $x_0, \dots, x_{N-1}$ represents a given time domain signal. The following equation defines the Fast Fourier Transform of $x(j)$:

$$X(k) = \sum_{j=1}^{N} x(j) e^{\frac{-2\pi i}{N}(j-1)(k-1)} \tag{1}$$

where $k = 0, \dots, N-1$, $x(j)$ is the sample at time index $j$ and $i$ is the imaginary number $\sqrt{-1}$. $X(k)$ is a vector of $N$ values at frequency index $k$ corresponding to the magnitude of the sine waves resulting from the decomposition of the time indexed signal.The inverse FFT is given in the following equation:

$$x(j) = \frac{1}{N} \sum_{k=1}^{N} X(k) e^{\frac{(-2\pi i)}{N}-(j-1)(k-1)} \tag{2}$$

The FFT takes advantage of the symmetry and periodicity properties of the Fourier Transform to reduce computation time. In this process, the transform is partitioned into a sequence of reduced-length transforms that is collectively performed with reduced computation [17].

The FFT technique also has performance limitation as the method. FFT is a complex transform which operates on an imaginary number and especial algorithm. It is a complex exponential that defines a complex sinusoid with frequency and it has not changed or upgraded.

### 2.2 Discrete Cosine Transform

The Discrete Cosine Transform has been used in frequency spectrum analysis, data compression, convolution computation and image processing [18]. For example, let$x = [x(0), x(1), \dots, x(N-1)]^T$, with $T$denoting column vector, $x$ represents a frame of speech samples applied as an input to a speech coder. $x$ is transformed into a vector $X = [X(0), X(1), \dots, X(N-1)]^T$, where $N$ denotes the number of coefficients.

$$X(k) = b(k) \sum_{j=0}^{N-1} x(j) \cos \frac{\pi(2j-1)(k-1)}{2N} \tag{3}$$
$$k = 0, 1, \dots, N-1$$

where all the coefficients are real numbers and

$$b(k) = \begin{cases} \frac{1}{\sqrt{N}} & k = 0 \\ \sqrt{\frac{2}{N}} & k = 1, 2, \dots, N-1 \end{cases}$$

The inverse of DCT (IDCT) is given in the following equation:

$$x(j) = \sum_{k=0}^{N-1} b(k) X(k) cos\left(\frac{\pi(2j+1)k}{2N}\right) \tag{4}$$
$$j = 0, 1, \dots, N-1$$

### 2.3 Discrete Tchebichef Transform

For a given positive integer $N$ (the vector size) and a value $n$ in the range $[1,\ N-1]$, the $N$ order orthonormal Tchebichef polynomials$t_k(n),\ n = 1,2, \dots, N-1$is defined using the following recurrence relation [11]:

$$t_0(n) = \frac{1}{\sqrt{N}}, \tag{5}$$

$$t_k(0) = \sqrt{\frac{N-k}{N+k}} \sqrt{\frac{2k+1}{2k-1}} t_{k-1}(0), \tag{6}$$

$$t_k(1) = \left\{1 + \frac{k(1+k)}{1-N}\right\} t_k(0), \tag{7}$$

$$t_k(n) = \gamma_1 t_k(n-1) + \gamma_2 t_k(n-2), \tag{8}$$
$$k = 1, 2, \dots, N-1,\ n = 2, 3, \dots, \left(\frac{N}{2} - 1\right),$$

where

$$\gamma_1 = \frac{-k(k+1) - (2n-1)(n-N-1) - n}{n(N-n)}, \tag{9}$$

$$\gamma_2 = \frac{(n+1)(n-N-1)}{n(N-n)}, \tag{10}$$

The forward Discrete Tchebichef Transform (DTT) of order $N$ is defined as:

$$X(k) = \sum_{n=0}^{N-1} x(n) t_k(n), \tag{11}$$

$$k = 0, 1, \ldots, N-1,$$

where $X(k)$ denotes the coefficient of orthonormal Tchebichef polynomials.

The inverse DTT is given in the following equation:

$$x(n) = \sum_{k=0}^{N-1} X(k) t_k(n), \tag{12}$$

$$n = 0, 1, \ldots, N-1,$$

The first few discrete orthonormal Tchebichef polynomials are shown in Fig. 1.

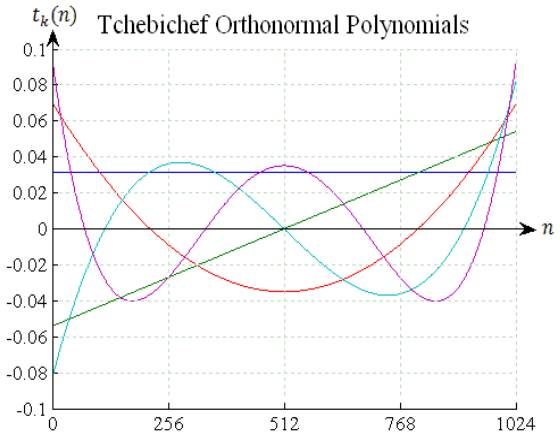


Figure 1. The discrete orthonormal tchebichef polynomial $t_k(n)$ for $k = 0, 1, 2, 3$ and 4.

## 3. EXPERIMENTAL RESULT

The voice used is a male voice based on standard voice of vowel. The sounds of the vowel 'O' and the consonant 'RA' are used from the International Phonetic Alphabet [19]. A speech signal has a sampling rate frequency component at about 11 KHz. The sample sound of the vowel 'O' is shown in Fig. 2.

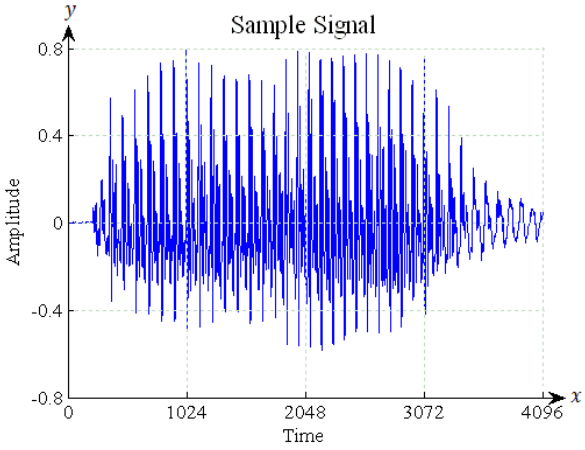


Figure 2. The sample sound of the vowel 'O'.

### 3.1 Silence detector

Speech signals are highly redundant and contain a variety of background noise. At some level of the background noise which interferes with the speech, it means that silence regions have quite a height zero-crossings rate as the signal changes from one side of the zero amplitude to the other and back again. For this reason, the threshold is included to remove any zero-crossings. In this experiment, the threshold is 0.1. This means that any zero-crossings that start and end within the range of $t_a$, where $-0.1 < t_a < 0.1$, are not included in the total number of zero-crossings in that window.

### 3.2 Pre-emphasis

Pre-emphasis is a technique used in speech processing to enhance high frequencies of the signal. It reduces the high spectral dynamic range. Therefore, by applying pre-emphasis, the spectrum is flattened, consisting of formants of similar heights. Pre-emphasis is implemented as a first-order Finite Impulse Response (FIR) filter defined as:

$$S_n = E(n) - \alpha E[n-1] \tag{13}$$

where $\alpha$ is the pre-emphasis coefficient, the value used for $\alpha$ is typically around 0.9 to 0.95. $E(n)$ is the sample data which represents speech signal with $n$ is $0 \leq n \leq N-1$, where $N$ is the sample size which represent speech signal. The speech signals after pre-emphasis of the vowel 'O' [19] is shown in Fig. 3.

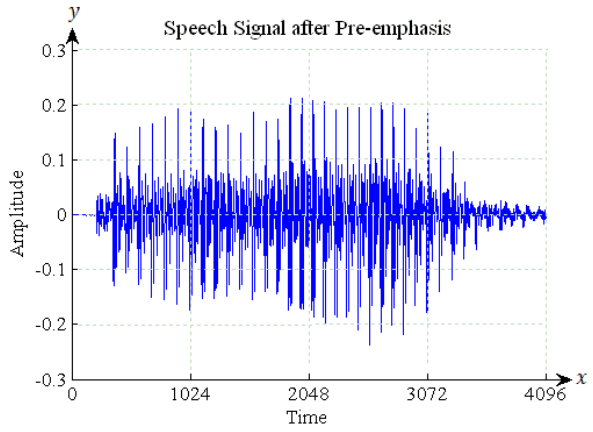


Figure 3. Speech signals after pre-emphasis of the vowel 'O'.

### 3.3 Windowing

Speech Recognition via FFT uses windowing function. A windowing function is used on each frame to smooth the signal and make it more amendable for spectral analysis.

Hamming window is a window functions used commonly in speech analysis to reduce the sudden changes and undesirable frequencies occurring in the framed speech. Hamming window is defined as:

$$w(k) = 0.54 - 0.46\cos\left[\frac{2\pi k}{L-1}\right] \tag{14}$$

where $L$ represents the width of $S_n$ and $k$ is an integer, with values $0 \leq k \leq L-1$. The resulting windowed segment is defined as:

$$x(k) = S_n w(k) \tag{15}$$

where $S_n$ is the signal function and $w(k)$is the window function. Whereas, DTT consists coefficient of DTT, therefore the window is inefficient when the sample data are multiplied by a value close to zero. Any transition occurring during this part of the window will be lost so that the spectrum is no longer true real time.

In this study, a sample of speech signal is windowed into four frames. Each window consists of 1024 sample data which represent speech signal. In this experiment, the fourth frame for 3073-4096 sample data is used. The speech signals via FFT, DCT and DTT of the vowel 'O' and the consonant 'RA' are shown on the left, middle and right of Fig. 4 and Fig. 5.

### 3.4 Spectrum analysis

The spectrum analysis via FFT and DCT can be generated as follows:

$$p(k) = |X(k)|^2 \tag{16}$$

The spectrum analysis via FFT and DCT of the vowel 'O' and the consonant 'RA' [19] is shown on the left and middle of Fig. 6 and Fig. 7. The spectrum analysis via DTT can be defined as:

$$p(k) = |c(n)|^2 \tag{17}$$

$$c(n) = \frac{x(n)}{t_k(n)} \tag{18}$$

where $c(n)$ is the coefficient of DTT, $x(n)$ is the sample data at time index $n$ and $t_k(n)$ is the computation matrix of orthonormal Tchebichef polynomials. The spectrum analysis via DTT of the vowel 'O' and the consonant 'RA' is shown on the right of Fig. 6 and Fig. 7. The frequency formants of the vowel 'O' and the consonant 'RA' [19] via FFT, DCT and DTT as numerically are shown in Table I and Table II respectively.

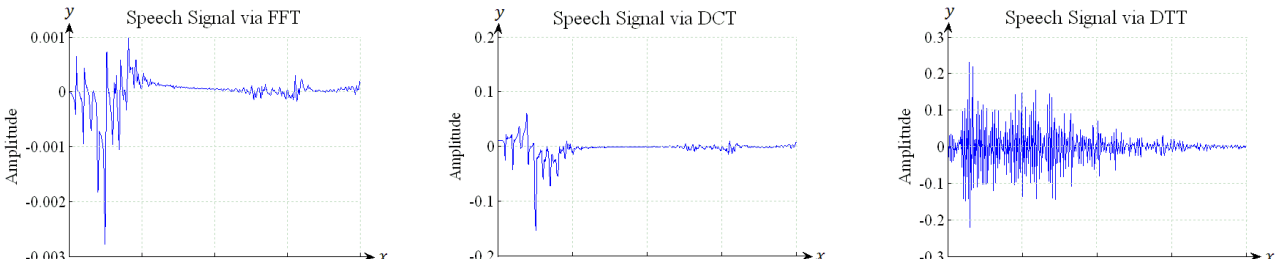


Figure 4. Imaginary part of FFT (left), coefficient of DCT (middle) and coefficient of DTT (right) for speech signal of the vowel 'O'.

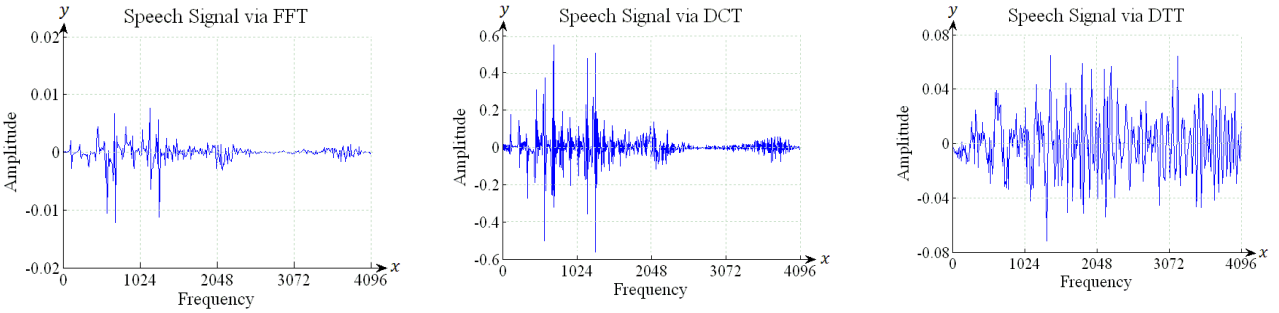


Figure 5. Imaginary part of FFT (left), coefficient of DCT (middle) and coefficient of DTT (right) for speech signal of the consonant 'RA'.

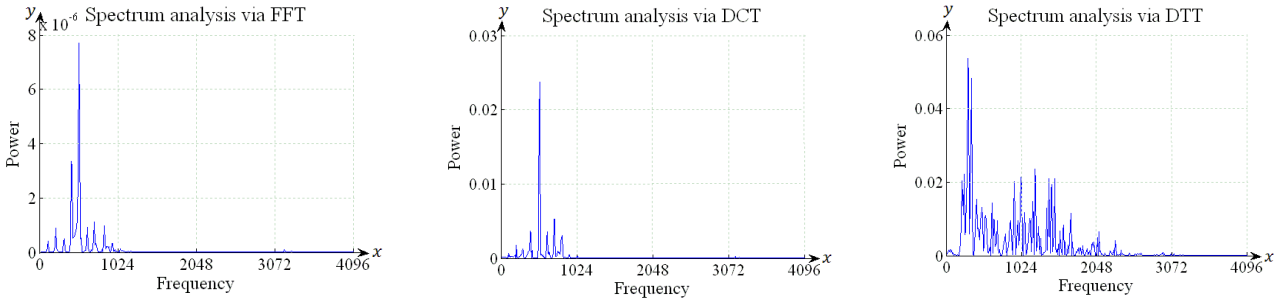


Figure 6. Imaginary part of FFT (left), coefficient of DCT (middle) and coefficient of DTT (right) for spectrum analysis of the vowel 'O'.

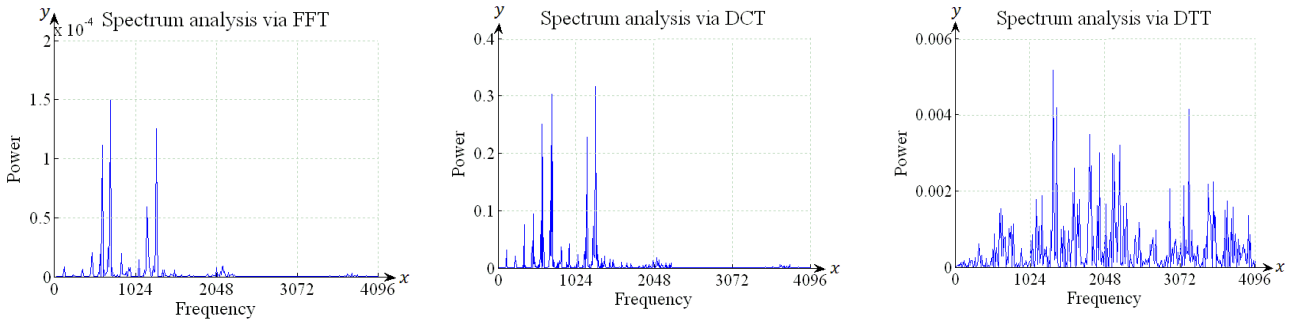


Figure 7. Imaginary part of FFT (left), coefficient of DCT (middle) and coefficient of DTT (right) for spectrum analysis of the consonant 'RA'.

TABLE I. FREQUENCY FORMANTS OF THE VOWEL 'O'

| 1. Vowel 'O' | 2. FFT | 3. DCT | 4. DTT |
|---|---|---|---|
| 5. $F_1$ | 527 | 516 | 441 |
| 6. $F_2$ | 7. 764 | 8. 753 | 9. 710 |
| 10. $F_3$ | 11. 3219 | 12. 3186 | 13. 3186 |

TABLE II. FREQUENCY FORMANTS OF THE CONSONANT 'RA'

| 14. Consonant 'RA' | 15. FFT | 16. DCT | 17. DTT |
|---|---|---|---|
| 18. $F_1$ | 19. 661 | 20. 613 | 21. 624 |
| 22. $F_2$ | 23. 1301 | 24. 1259 | 25. 1248 |
| 26. $F_3$ | 27. 2160 | 28. 2121 | 29. 2131 |

## 4. COMPARATIVE ANALYSIS

The conventional method of depicting formants $F_1$ and $F_2$ only does not sufficiently represent the multi-dimensional nature of the vowel quality. Delattre [20] showed that the third formant significantly influenced listener's judgments of the vowel quality and the combination of higher formants carry a relatively significant influence on vowel perception.

More recent studies have examined the spectral features suggesting that the differences ($F_3$-$F_2$) are a more accurate way of identifying vowel frontends. Syrdal and Gopal [21] have shown that the separation between back and front vowels is more closely linked to the differences ($F_3$-$F_2$) than ($F_2$-$F_1$). However, it is important to recognize that $F_3$ and $F_4$ vary more than $F_1$ and $F_2$ as a result of the speaker characteristics. Nevertheless, they are relatively stable across vowel categories in contrast to $F_1$ and $F_2$, which vary greatly as a result of the vowel quality. The higher formants are therefore less effective carriers of phonetic information than the lower formants [22].

The speech signal of the vowel 'O' via DCT as illustrated on the middle of Fig. 4 showed that speech signal is clearer than FFT and DTT. On one hand, the speech signals of the vowel 'O' via DTT produces more noise than FFT and DCT. On the other hand, speech signals of the consonant 'RA' via FFT on the left of Fig. 5 produces a clearer from the noisy speech signal than DCT and DTT.

Spectrum analysis of the vowel 'O' via FFT on the left of Fig. 6 produces a lower power spectrum than DCT and DTT. On one hand, power spectrum via DTT on the right of Fig. 6 is higher than FFT and DCT. On the other hand, spectrum analysis of the consonant 'RA' via DCT on the middle of Fig. 7 is higher power spectrum than FFT and DTT. Spectrum analysis of the consonant 'RA' via DTT on the right of Fig. 7 produces more noise than FFT and DCT in a frequency spectrum. It is also capable to capture the third formant unlike DCT. The experimental result showed that the formants $F_1$, $F_2$ and $F_3$ among FFT, DCT and DTT were identically similar.

## 5. CONCLUSION

As a discrete orthonormal transform, DTT is a simpler and computationally more efficient than FFT. On one hand, FFT is computationally complex with the imaginary part. DTT consumes simpler and faster computation with real coefficient. It is an ideal candidate for discrete transform in speech recognition to transform time domain into frequency domain. On the other hand, DCT produces a simpler output in the frequency spectrum and it is occasionally unable to capture the third formant $F_3$. DTT is able to capture all three formants concurrently, $F_1$, $F_2$, and $F_3$. The frequency

formants via FFT, DCT, and DTT are compared. They have produced relatively identical outputs in term of speech recognitions.

## REFERENCES


[1] D.H. Bailey and P.N. Swarztrauber, "A Fast Method for Numerical Evaluation of Continuous Fourier and Laplace Transform," Journal on Scientific Computing, vol. 15, no. 5, Sep. 1994, pp. 1105-1110.

[2] R. Mukundan, "Improving Image Reconstruction Accuracy Using Discrete Orthonormal Moments," Proceedings of International Conference on Imaging Systems, Science and Technology, Jun. 2003, pp. 287-293.

[3] R. Mukundan, S.H. Ong, and P.A. Lee, "Image Analysis by Tchebichef Moments," IEEE Transactions on Image Processing, vol. 10, no. 9, Sep. 2001, pp. 1357–1364.

[4] C.-H. Teh and R.T. Chin, "On Image Analysis by the Methods of Moments," IEEE Transactions on Pattern Analysis Machine Intelligence, vol. 10, no. 4, Jul. 1988, pp. 496-513.

[5] N.A. Abu, W.S. Lang, and S. Sahib, "Image Super-Resolution via Discrete Tchebichef Moment," Proceedings of International Conference on Computer Technology and Development (ICCTD 2009), vol. 2, Nov. 2009, pp. 315–319.

[6] M. Tuceryan, "Moment based texture segmentation," Pattern Recognition Letters, vol. 15, Jul. 1994, pp. 659-668.

[7] L. Wang and G. Healey, "Using Zernike Moments for the Illumination and Geometry Invariant Classification of Multispectral Texture," IEEE Transactions on Image Processing, vol. 7, no. 2, Feb. 1998, pp. 196-203.

[8] L. Zhang, G.B. Qian, W.W. Xiao, and Z. Ji, "Geometric invariant blind image watermarking by invariant Tchebichef moments," Optics Express Journal, vol. 15, no. 5, Mar. 2007, pp. 2251-2261.

[9] H. Zhu, H. Shu, T. Xia, L. Luo, and J.L. Coatrieux, "Translation and scale invariants of Tchebichef moments," Journal of Pattern Recognition Society, vol. 40, no. 9, Sep. 2007, pp. 2530-2542.

[10] H. Rahmalan, N. Suryana and N. A. Abu, "A general approach for measuring crowd movement," Malaysian Technical Universities Conference and Exhibition on Engineering and Technology (MUCEET2009), Jun. 2009, pp. 98-103.

[11] R. Mukundan, "Some Computational Aspects of Discrete Orthonormal Moments," IEEE Transactions on Image Processing, vol. 13, no. 8, Aug. 2004, pp. 1055-1059.

[12] N.A. Abu, N. Suryana, and R. Mukundan, "Perfect Image Reconstruction Using Discrete Orthogonal Moments," Proceedings of The 4th IASTED International Conference on Visualization, Imaging, and Image Processing (VIIP2004), Sep. 2004, pp. 903-907.

[13] N.A. Abu, W.S. Lang, and S. Sahib, "Image Projection Over The Edge," International Conference on Industrial and Intelligent Information (ICIII 2010), Proceedings 2nd International Conference on Computer and Network Technology (ICCNT2010), Apr. 2010, pp. 344-348.

[14] R. Mukundan and O. Hunt, "A comparison of discrete orthogonal basis functions for image compression," Proceedings Conference on Image and Vision Computing New Zealand (IVCNZ 04), Nov. 2004, pp. 53-58.

[15] W.S. Lang, N.A. Abu, and H. Rahmalan, "Fast 4x4 Tchebichef Moment Image Compression," Proceedings International Conference of Soft Computing and Pattern Recognition (SoCPaR2009), Dec. 2009, pp. 295–300.

[16] N.A. Abu, W.S. Lang, N. Suryana, and R. Mukundan, "An Efficient Compact Tchebichef moment for Image Compression," 10th International Conference on Information Science, Signal Processing and their applications (ISSPA2010), May 2010, pp. 448-451.

[17] S. Rapuano and F. Harris, "An introduction to FFT and time domain windows," IEEE Instrumentation and Measurement Society, vol. 10, no. 6, Dec. 2007, pp. 32-44.

[18] J. Zhou and P. Chen, "Generalized Discrete Cosine Transform," Pacific-Asia Conference on Circuits, Communications and Systems, May 2009, pp. 449-452.

[19] J.H. Esling and G.N. O'Grady, "The International Phonetic Alphabet," Linguistics Phonetics Research, Department of Linguistics, University of Victoria, Canada, 1996.

[20] P. Delattre, "Some Factors of Vowel Duration and Their Cross-Linguistic Validity," Journal of the Acoustical Society of America, vol. 34, Aug. 1962, pp. 1141-1143.

[21] K. Syrdal and H.S. Gopal, "A perceptual model of vowel recognition based on the auditory representation of American English vowels," Journal of the Acoustical Society of America, vol. 79, no. 4, Apr. 1986, pp. 1086-1100.

[22] J.H. Cassidy, “Dynamic and Target Theories of Vowel Classification: Evidence from Monophthongs and Diphthongs in Australian English,” Journal of Language and Speech, vol. 37, no. 4, Oct. 1994, pp. 357-373.