Efficient Discrete Tchebichef on Spectrum Analysis of Speech Recognition

Ferda Ernawan and Nur Azman Abu

Abstract—**Speech recognition is still a growing field of importance. The growth in computing power will open its strong potentials for full use in the near future. Spectrum analysis is an elementary operation in speech recognition. Fast Fourier Transform (FFT) has been a traditional technique to analyze frequency spectrum of the signals in speech recognition. FFT is computationally complex especially with imaginary numbers. The Discrete Tchebichef Transform (DTT) is proposed instead of the popular FFT. DTT has lower computational complexity and it does not require complex transform dealing with imaginary numbers. This paper proposes a novel approach based on 256 discrete orthonormal Tchebichef polynomials as efficient technique to analyze a vowel and a consonant in spectral frequency of speech recognition. The comparison between 1024 discrete orthonormal Tchebichef transform and 256 discrete orthonormal Tchebichef transform has been done. The preliminary experimental results show that 256 DTT has the potential to be more efficient to transform time domain into frequency domain for speech recognition. 256 DTT produces simpler output than 1024 DTT in frequency spectrum. At the same time, 256 Discrete Tchebichef Transform can produce concurrently four formants F_1, F_2, F_3 and F_4.**

Index Terms—**Speech recognition, spectrum analysis, Fast Fourier Transforms and Discrete Tchebichef Transform.**

I. Introduction

Spectrum analysis method using Fourier transform is widely used for digital signal processing applicable for spectrum analysis of speech recognition. Speech recognition requires heavy processing on large sample windowed data. Typically, each window consumes 1024 sample data. Since the window is large, an efficient FFT has been employed to speed up the process. 1024 sample data FFT computation is considered the main basic algorithm for several digital signals processing [1]. FFT is a traditional technique to analyze frequency spectrum of speech recognition. The FFT is often used to compute numerical approximations to continuous Fourier. However, a straightforward application of the FFT often requires a large FFT to be performed even though most of the input data may be zero [2]. FFT is a complex transform which requires operating on imaginary numbers. It is a complex exponential function that defines a complex sinusoid for a given frequency.

The Discrete Tchebichef Transform (DTT) is another transform method based on discrete Tchebichef polynomials [3][4]. This paper proposes an approach based on 256 discrete orthonormal Tchebichef polynomials to analyze spectral frequency for speech recognition. DTT is used to avoid the complex computation of FFT. DTT has a potential next candidate to transform time domain into frequency domain in speech recognition. DTT has a lower computational complexity and it does not require complex transform unlike continuous orthonormal transforms [5]. DTT does not involve any numerical approximation on friendly domain. The Tchebichef polynomials have unit weight and algebraic recurrence relations involving real coefficients. These factors in effect make DTT suitable for transforming the signal from time domain into frequency domain for speech recognition.

DTT has been applied in several computer vision and image processing application in previous work. For examples, DTT is used in image analysis [6], texture segmentation [7], image watermarking [8], image reconstruction [3][9], image compression [10] and spectrum analysis of speech recognition [5][11].

The organization of the paper is as follows. The next section gives a brief description on FFT and DTT. The matrix implementation of orthonormal tchebichef polynomials are presented in section III. Section IV shows the experiment results of speech signal coefficient of Discrete Tchebichef Transform, spectrum analysis, frequency formants and time taken performance. Section V presents the comparison of spectrum analysis, frequency formants and time taken performance between 1024 DTT and 256 DTT. Finally, section VI concludes the comparison of spectrum analysis using 1024 DTT and 256 DTT in terms of speech recognition.

II. Transformation Domain

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 $X(k)$. Let the sequence of N complex numbers $x_0, \ldots, x_{N-1}$ represent a given time domain windowed signal. The following equation defines the Fast Fourier Transform of $x(j)$:

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

Manuscript received March 22, 2011. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Mohamed Othman.

Ferda Ernawan is with the Faculty of Information and Communication Technology, Universitas Dian Nuswantoro (UDINUS), Semarang, Indonesia (e-mail: ferda1902@gmail.com).

Nur Azman Abu is with the Faculty of Information and Communication Technology, Universiti Teknikal Malaysia Melaka (UTeM), Melaka, Malaysia (e-mail: nura@utem.edu.my).

where $k = 0, \ldots, N-1$, $x(j)$ is the sample at time index j and i is the imaginary number$\sqrt{-1}$. Then, $X(k)$ 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. It reduced the time complexity from $O(n^2)$ to $O(n\log n)$. In this process, the transform is partitioned into a sequence of reduced-length transforms that is collectively performed with reduced computation [12]. The FFT technique also has performance limitation as the method. FFT is operating a complex field computationally dealing with imaginary numbers. FFT has not been changed and upgraded unlike Number Theoretic Transform (NTT).

Discrete orthonormal Tchebichef transform is an approach based on Tchebichef polynomials. 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 $t_k(n)$, $n = 1, 2, \ldots, N-1$ are defined using the following recurrence relation [9]:

$$t_0(n) = \frac{1}{\sqrt{N}}, \tag{3}$$

$$t_k(0) = \sqrt{\frac{N-k}{N+k}}\sqrt{\frac{2k+1}{2k-1}}\,t_{k-1}(0), \tag{4}$$

$$t_k(1) = \left\{1 + \frac{k(1+k)}{1-N}\right\} t_k(0), \tag{5}$$

$$t_k(n) = \gamma_1 t_k(n-1) + \gamma_2 t_k(n-2), \tag{6}$$

$$k = 1, 2, \ldots, N-1,\ n = 2, 3, \ldots, \left(\frac{N}{2}-1\right),$$

where

$$\gamma_1 = \frac{-k(k+1) - (2n-1)(n-N-1) - n}{n(N-n)}, \tag{7}$$

$$\gamma_2 = \frac{(n+1)(n-N-1)}{n(N-n)}, \tag{8}$$

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

$$X(k) = \sum_{n=0}^{N-1} x(n)t_k(n), \tag{9}$$

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

where $X(k)$ denotes the coefficient of orthonormal Tchebichef polynomials. $x(n)$ is the sample of speech signal at time index n. The inverse DTT is given in the following equation:

$$x(n) = \sum_{k=0}^{N-1} X(k)t_k(n), \tag{10}$$

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

The orthonormal Tchebichef polynomials are proper especially when Tchebichef polynomials of large degree are required to be evaluated. The orthonormal Tchebichef polynomials matches for signal processing which have large sample data represents speech signal. The Tchebichef transform involves only algebraic expressions and it can be compute easily using a set of recurrence relations (3)-(8) above.

III. The Implementation of Discrete Orthonormal Tchebichef Transform

A. Sample Speech Signal

The voice used in this experiment is the vowel ‘O’ and the consonant ‘RA’ from the International Phonetic Alphabet [13]. A speech signal has a sampling rate frequency component of about 11 KHz. The sample sounds of vowel ‘O’ is shown in the Fig. 1.

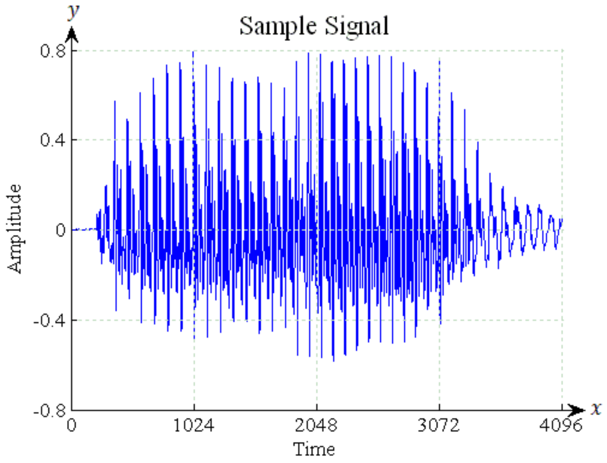


Figure 1. The sample sound of the vowel ‘O’.

B. 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 preset to be 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.

C. 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:

$$x(n) = S(n) - \alpha S[n-1] \tag{11}$$

where α is the pre-emphasis coefficient, the value used for α is typically around 0.9 to 0.95. $S(n)$ is the sample data which represents speech signal with n is $0 \le n \le N-1$, where N is the sample size which represent speech signal. The speech signals after pre-emphasis of the vowel ‘O’ [13] is shown in Fig. 2.

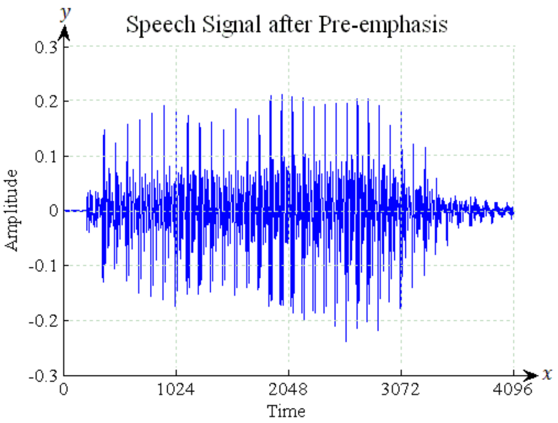


Figure 2. Speech Signal after Pre-emphasis.

D. Speech Signal Windowed

The sample of vowel ‘O’ and consonant ‘RA’ have 4096 sample data which representing speech signal. The sample of speech signal is windowed into several frames. On one hand, speech signal of the vowel 'O' and consonant ‘RA’ are windowed into four frames. Each window consumes 1024 sample data as illustrated in Fig. 3.

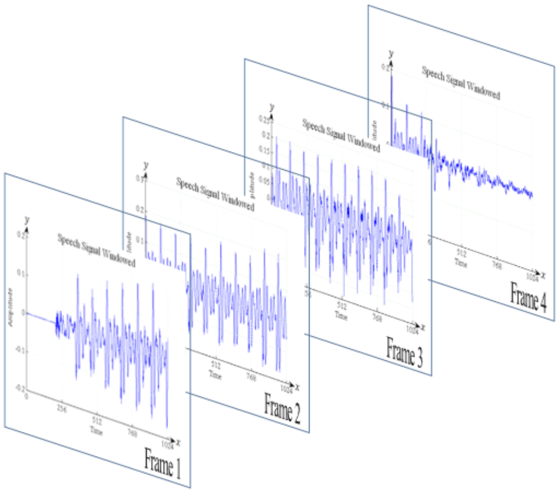


Figure 3. Speech Signal Windowed into Four Frames.

In this experiment, the fourth frame for 3073-4096 sample data is being used for analysis and evaluated using 1024 discrete orthonormal Tchebichef polynomials and FFT. The 1024 discrete orthonormal Tchebichef polynomials are shown in the Fig. 4.

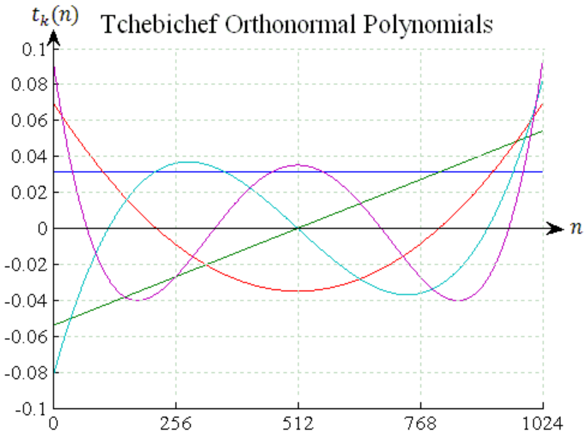


Figure 4. The First Five 1024 Discrete Orthonormal Tchebichef Polynomials $t_k(n)$ for $k = 0, 1, 2, 3$ and 4.

On the other hand, the sample speech signal of the vowel 'O' and consonant ‘RA’ are windowed into sixteen frames. Each window consists of 256 sample data which representing speech signal. The sample speech signals windowed into sixteen frames are shown in Fig. 5. In this study, the third frame for 513-768 sample data is used to compute using 256 discrete orthonormal Tchebichef polynomials as shown in Fig. 6.

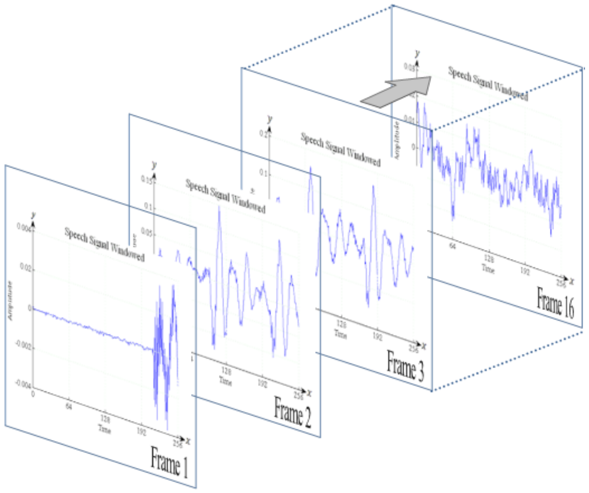


Figure 5. Speech Signal Windowed into Sixteen Frames.

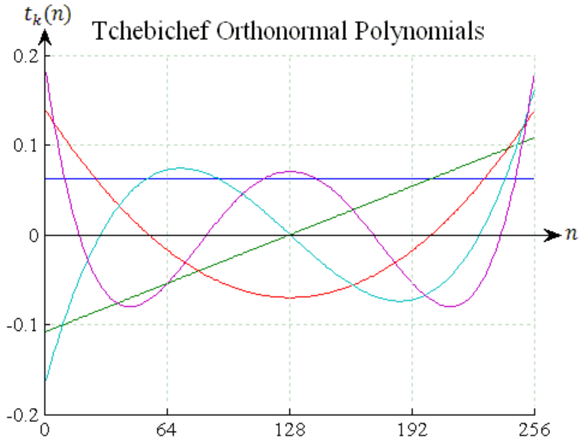


Figure 6. The First Five 256 Discrete Orthonormal Tchebichef Polynomials $t_k(n)$ for $k = 0, 1, 2, 3$ and 4.

E. Coefficient of Discrete Tchebichef Transform

Coefficients of DTT of order $n = 256$ sample data are given as follow formula:

$$\boldsymbol{TC} = \boldsymbol{S} \qquad (12)$$

$$\begin{bmatrix} t_0(0) & t_0(1) & t_0(2) & \cdots & t_0(n-1) \\ t_1(0) & t_1(1) & t_1(2) & \cdots & t_1(n-1) \\ t_2(0) & t_2(1) & t_2(2) & \cdots & t_2(n-1) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ t_{n-1}(0) & t_{n-1}(1) & t_{n-1}(2) & \cdots & t_{n-1}(n-1) \end{bmatrix} \begin{bmatrix} c_0 \\ c_1 \\ c_2 \\ \vdots \\ c_{n-1} \end{bmatrix} = \begin{bmatrix} x_0 \\ x_1 \\ x_2 \\ \vdots \\ x_{n-1} \end{bmatrix}$$

where $\boldsymbol{C}$ is the coefficient of Discrete Tchebichef Transform, which represents $c_0, c_1, c_2, \ldots, c_{n-1}$. $\boldsymbol{T}$ is matrix computation of Discrete Orthonormal Tchebichef Polynomials $t_k(n)$ for $k = 0, 1, 2, \ldots, N-1$. $\boldsymbol{S}$ is the sample of speech signal window which is given by $x(0), x(1), x(2), \ldots, x(n-1)$. The coefficient of DTT is given in as follows;

$$C = T^{-1}S \qquad (13)$$

IV. Experiment Results

A. *Speech Signal Coefficient of DTT*

The speech signal of vowel ‘O’ and consonant ‘RA’ using 1024 discrete orthonormal Tchebichef polynomials are shown on the left of Fig. 7 and Fig. 8. Next, speech signal of vowel ‘O’ and consonant ‘RA’ using 256 discrete orthonormal Tchebichef polynomials are shown on the right of Fig. 7 and Fig. 8.

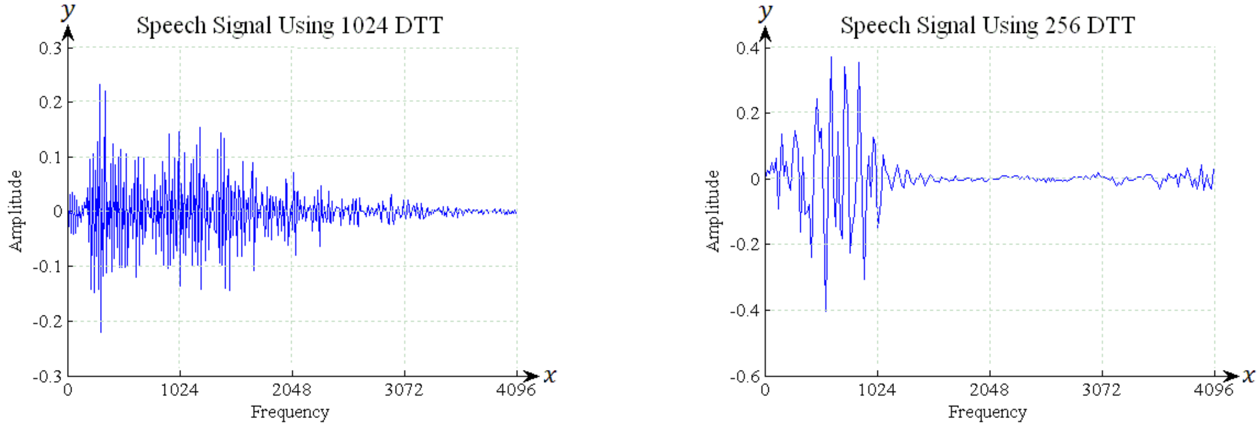


Figure 7. Coefficient of DTT for 1024 sample data (left) and coefficient of DTT for 256 sample data (right) for speech signal of vowel ‘O’.

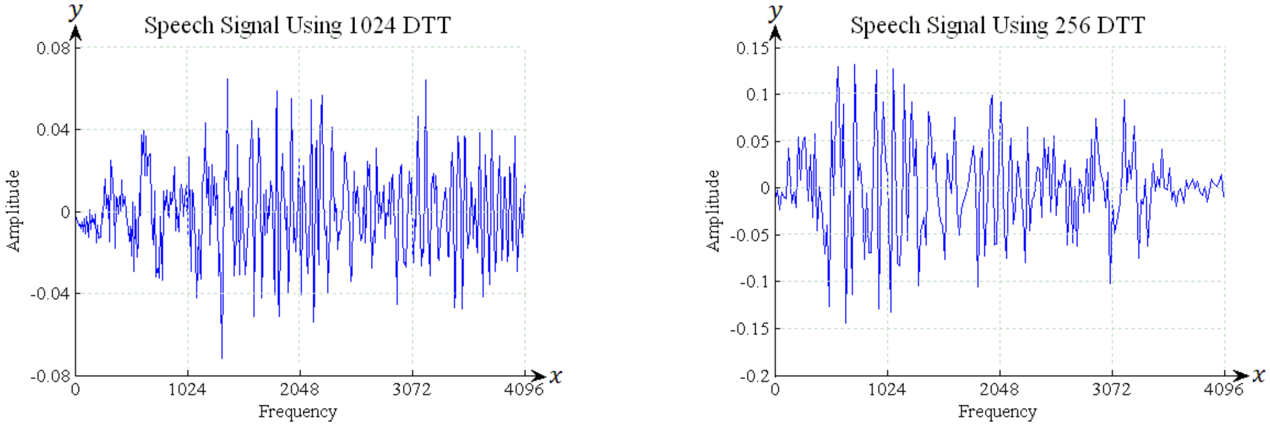


Figure 8. Coefficient of DTT for 1024 sample data (left) and coefficient of DTT for 256 sample data (right) for speech signal of consonant ‘RA’.

B. *Spectrum Analysis*

Spectrum analysis is the absolute square value of the speech signal, so the values are never negative. The spectrum analysis using DTT can be defined in the following equation:

$$p(k) = |c(n)|^2 \qquad (14)$$

where $c(n)$ is coefficient of discrete Tchebichef transform. The spectrum analysis using 1024 DTT and 256 DTT of the vowel ‘O’ and the consonant ‘RA’ is shown in Fig. 9 and Fig. 10 respectively.

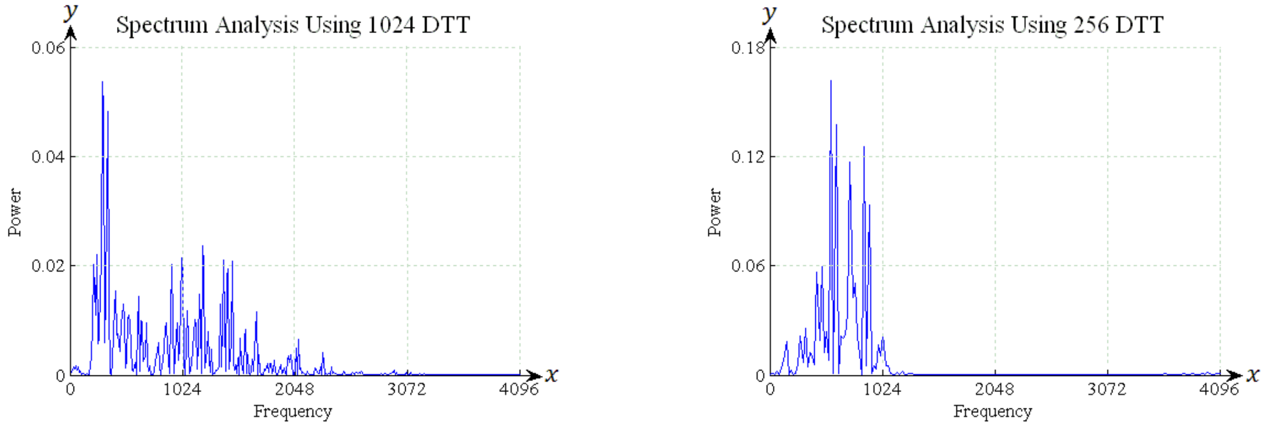
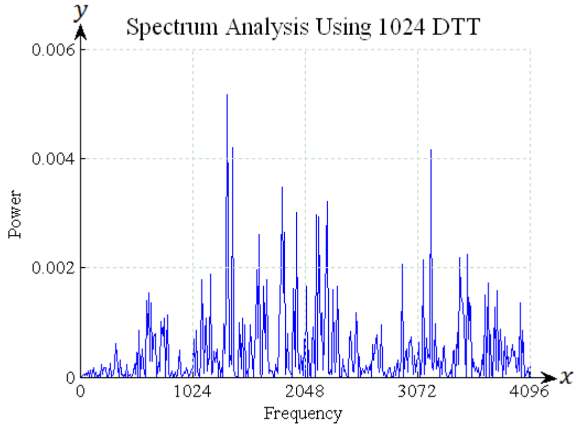


Figure 9. Coefficient of DTT for 1024 sample data (left) and coefficient of DTT for 256 sample data (right) for spectrum analysis of vowel ‘O’.



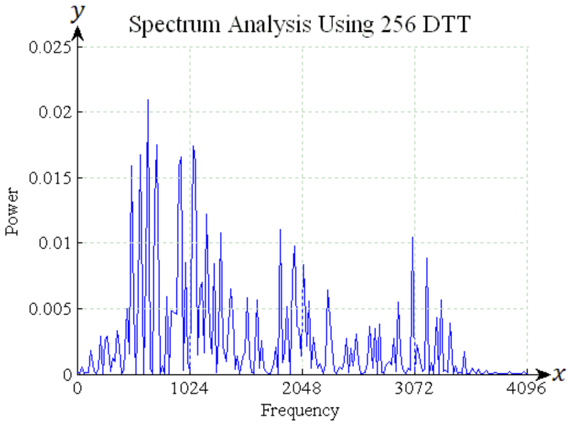


Figure 10. Coefficient of DTT for 1024 sample data (left) and coefficient of DTT for 256 sample data (right) for spectrum analysis of consonant ‘RA’.

C. *Frequency Formants*

The unique of each vowel and consonant are measured by frequency formants. Formants are a characteristic resonant region of a sound. Formants are exactly the resonant frequencies of vocal tract [14]. The frequency formants of the vowel ‘O’ and the consonant ‘RA’ have been extracted using 256 DTT as presented in Fig. 11 and Fig. 12.

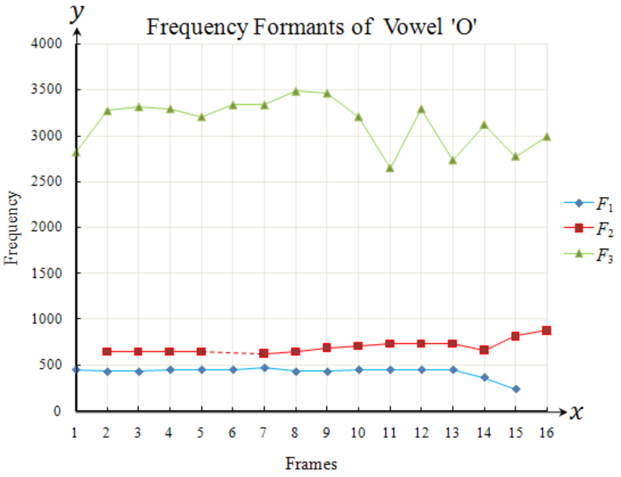


Figure 11. Frequency Formants of Vowel ‘O’ using 256 DTT.

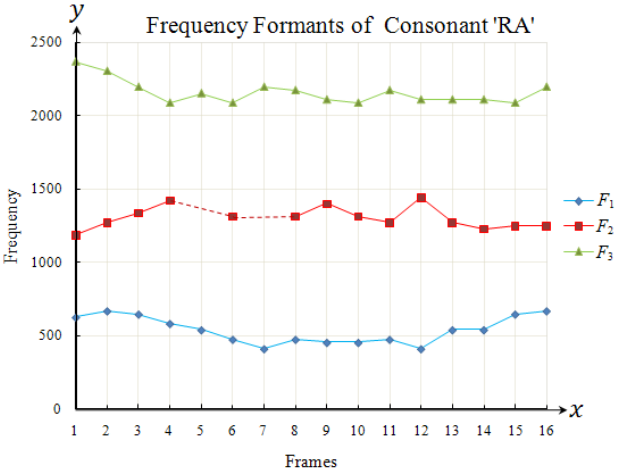


Figure 12. Frequency Formants of Consonant 'RA' using 256 DTT.

The frequency formants are detected by autoregressive model. The formants of the autoregressive curve are found at the peaks using a numerical derivative. The spectral properties of the vowels were represented through measures of first formant F_1, second formant F_2 and third formant F_3. These vector positions of the formants are used to characterize a particular vowel. The frequency formants of the vowel ‘O’ and the consonant ‘RA’ using 256 DTT, 1024 DTT and 1024 sample data FFT computation are shown in Table I and Table II respectively.

TABLE I. FREQUENCY FORMANTS OF VOWEL ‘O’

Vowel ‘O’	DTT		FFT
	256	1024	1024
F_1	430	441	527
F_2	645	710	764
F_3	3316	3186	3219

TABLE II. FREQUENCY FORMANTS OF CONSONANT ‘RA’

Consonant ‘RA’	DTT		FFT
	256	1024	1024
F_1	645	624	661
F_2	1335	1248	1301
F_3	2196	2131	2160

D. *Time Taken Performance*

The comparison time taken performance among 256 DTT coefficient, 1024 DTT coefficient and Fast Fourier Transform computation for 1024 sample data of vowel 'O' and consonant 'RA' is presented in Table III. Next, the time taken of speech recognition performance using 256 DTT, 1024 DTT and FFT is shown in Table IV.

TABLE III. TIME TAKEN OF DTT COEFFICIENT AND FFT

Vowel and Consonant	DTT		FFT
	256	1024	1024
Vowel ‘O’	0.001564 sec	0.011889 sec	0.095789 sec
Consonant ‘RA’	0.001605 sec	0.027216 sec	0.102376 sec

TABLE IV. TIME TAKEN OF SPEECH RECOGNITION PERFORMANCE USING DTT AND FFT

Vowel and Consonant	DTT		FFT
	256	1024	1024
Vowel ‘O’	0.557000 sec	0.903968 sec	0.587628 sec
Consonant ‘RA’	0.557410 sec	0.979827 sec	0.634997 sec

V. COMPARATIVE ANALYSIS

The speech signal of the vowel ‘O’ and consonant ‘RA’ using 256 DTT as shown on the right of Fig. 7 and Fig. 8 produce simpler output in spectral frequency than speech signal using 1024 DTT. Next, Spectrum analysis of the vowel ‘O’ and consonant ‘RA’ using 1024 DTT on the left of Fig. 9 and Fig. 10 produce a lower power spectrum than

256 DTT. Next, spectrum analysis of vowel ‘O’ and consonant ‘RA’ using 256 DTT on the right of Fig. 9 and Fig. 10 produce simpler output than 1024 DTT.

The frequency formants of vowel ‘O’ for sixteen frames are shown in the Fig. 11. According observation are presented in the Fig. 11, the first formant F_1 on frame 16 is not detected. The second formant F_2 on first frame and sixth frame are not captured. Next, the third formant F_3 of vowel ‘O’ for sixteen frames, the frequency formants are detected on each frame respectively. According observation frequency formants of consonant ‘RA’ are shown in the Fig. 12, the first formant F_1 and third formant F_3 are detected on sixteen frames, while the second formants F_2 on fifth and seventh frame are not captured.

Frequency formants as presented in Table 1 and Table 2 show those frequency formants of vowel ‘O’ and consonant ‘RA’ using 256 DTT produce similar identical output with frequency formants using 1024 DTT and FFT. The comparison time taken of 256 DTT coefficients, 1024 DTT coefficient, and FFT is represented in the Table 3. According to observation as presented in Table 3, the time taken of 256 DTT is computationally efficient than 1024 DTT and FFT. Next, speech recognition performance of 256 DTT produces the efficient time taken than 1024 DTT and FFT.

The Discrete Tchebichef Transform that developed in this paper is smaller computations. The experiment result have shown that the propose 256 Discrete Tchebichef Transform algorithm efficiently reduces the time taken to transform time domain into frequency domain. The efficient of 256 Discrete Tchebichef Transform is much higher than 1024 Discrete Tchebichef Transform in terms of speech recognition performance.

FFT algorithm produces the time complexity $O(nlogn)$. Next, the computation time of DTT produce time complexity $O(n^2)$. For the future research, DTT can be upgraded to reduce the time complexity from $O(n^2)$ to be $O(n \log n)$ using convolution algorithm. DTT can increase the speech recognition performance in terms getting frequency formants of a speech.

VI. Conclusion

As a discrete orthonormal transform, 256 Discrete Tchebichef Transform has a potential to perform faster and computationally more efficient than 1024 Discrete Tchebichef Transform. A 256 DTT produces simpler output in spectral frequency than DTT which takes in 1024 sample data. On one hand, FFT is computationally complex especially with imaginary numbers. On the other hand, DTT consumes simpler and faster computation with involves only algebraic expressions and the Tchebichef polynomial matrix can be constructed easily using a set of recurrence relations. Spectrum analysis using 256 DTT produces four formants F_1, F_2, F_3 and F_4 concurrently in spectrum analysis for consonant. The frequency formants using 1024 DTT and 256 DTT are compared. They have produced relatively identical outputs in terms of frequency formants for speech recognitions.

References

[1] J.A. Vite-Frias, Rd.J. Romero-Troncoso and A. Ordaz-Moreno, “VHDL Core for 1024-point radix-4 FFT Computation,” *International Conference on Reconfigurable Computing and FPGAs*, Sep. 2005, pp. 20-24.

[2] 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.

[3] R. Mukundan, “Improving Image Reconstruction Accuracy Using Discrete Orthonormal Moments,” *Proceedings of International Conference on Imaging Systems, Science and Technology*, June 2003, pp. 287-293.

[4] 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.

[5] F. Ernawan, N.A. Abu and N. Suryana, “Spectrum Analysis of Speech Recognition via Discrete Tchebichef Transform,” *Proceedings of International Conference on Signal and Information Processing*, Dec. 2010, pp. 395-399.

[6] 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.

[7] M. Tuceryan, “Moment Based Texture Segmentation,” *Pattern Recognition Letters*, Vol. 15, July 1994, pp. 659-668.

[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] R. Mukundan, “Some Computational Aspects of Discrete Orthonormal Moments,” *IEEE Transactions on Image Processing*, Vol. 13, No. 8, Aug. 2004, pp. 1055-1059.

[10] 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.

[11] F. Ernawan, N.A. Abu and N. Suryana, “The Efficient Discrete Tchebichef Transform for Spectrum Analysis of Speech Recognition,” *Proceedings 3rd International Conference on Machine Learning and Computing*, Vol. 4, Feb. 2011, pp. 50-54.

[12] 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.

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

[14] A. Patil, C. Gupta, and P. Rao, “Evaluating Vowel Pronunciation Quality: Formant Space Matching Versus ASR Confidence Scoring,” *National Conference on Communication (NCC)*, Jan. 2010, pp. 1-5.

Ferda Ernawan received Master Student Degree in Software Engineering and Intelligence from the Faculty of Information and Communication Technology, Universiti Teknikal Malaysia Melaka (UTeM). He received the Bachelor Student Degree in Information Technology from Universitas Dian Nuswantoro (UDINUS), Semarang, Indonesia in 2009. His research interests are in the areas audio processing, image processing and their applications.

Nur Azman Abu is currently serving as a senior lecturer at Faculty of ICT, Universiti Teknikal Malaysia Melaka (UTeM). He obtained his bachelor and master degree from Purdue University in 1992 and 1994 both in Mathematics. He is currently undergoing his PhD program at Universiti Teknikal Malaysia, Melaka. His current interests include cryptography, random number generation, image processing and TSP.