Journal of Applied Sciences

ISSN 1812-5654

Journal of Applied Sciences 13 (3): 465-471, 2013
ISSN 1812-5654 / DOI: 10.3923/jas.2013.465.471

Vowel Recognition using Discrete Tchebichef Transform

[1]Ferda Ernawan, [2]Nur Azman Abu and [2]Nanna Suryana
[1]Faculty of Information and Communication Technology, Universitas Dian Nuswantoro
Imam Bonjol No. 206, Semarang 50131, Indonesia
[2]Faculty of Information and Communication Technology, Universiti Teknikal Malaysia Melaka,
Hang Tuah Jaya, Melaka, 76100, Malaysia

Abstract: Spectrum analysis has become an elementary operation in vowel recognition. Fast Fourier Transform (FFT) has been used as a famous technique to analyze frequency spectrum of the signal in vowel recognition. Traditionally, vowel recognition required large FFT computation on each window. This study has proposed the Discrete Tchebichef Transform (DTT) as a possible alternative to the popular FFT. DTT has had lower computational complexity and it did not require complex transform with imaginary numbers. This study has proposed an approach based on 256 DTT for efficient vowel recognition. The method used a simplify set of recurrence relation matrix to compute within each window. Unlike the FFT, DTT has provided a simpler matrix setting which involves real coefficient numbers only. The experiment on vowel recognition using 256 DTT, 1024 DTT and 1024 FFT has been conducted to recognize five vowels. The experimental results have indicated the practical advantage of 256 DTT in terms of spectral frequency and time taken for vowel recognition performance. 256 DTT has been produced frequency formants that were relatively similar output of 1024 DTT and 1024 FFT in terms of vowel recognition. The 256 DTT has become potential to be a competitive candidate for computationally efficient dynamic vowel recognition.

Key words: Vowel recognition, fast fourier transforms, Discrete Tchebichef Transform

INTRODUCTION

Vowel recognition techniques typically utilize FFT to transform speech signal at time domain into frequency domain which carries out spectral transformation of speech signal. Spectral analysis requires large computation to simple speech measurement but characterizes sound more precisely. Mostly, FFT compute the speech signal 1024 sample data of speech signal for each window (Vite-Frias *et al.*, 2005). The FFT is often used to compute numerical approximations to continuous Fourier. However, a straight forward application of the FFT to computation often requires a large FFT to be performed even though most of the input data to the FFT may be zero (Bailey and Swarztrauber, 1994). In addition, FFT algorithm requires a special algorithm on imaginary numbers to compute a speech signals. FFT is an efficient algorithm that can perform Discrete Fourier Transform (DFT). The FFT takes advantage of the symmetry and periodicity properties of the Fourier Transform to reduce computation time. This study presents an alternative method to replace FFT on vowel recognition. Discrete Tchebichef Transform (DTT) is proposed instead of the popular FFT in spectral analysis.

DTT is a transform method based on orthonormal Tchebichef polynomials (Mukundan, 2004) which provide simple basis matrix. DTT is an orthonormal transform which has relatively few coefficients transform. DTT has a set of algebraic recurrence relations algorithm that involves real coefficient numbers. DTT has been recently applied in speech recognition (Ernawan and Abu, 2011), image analysis, image reconstruction (Mukundan, 2003), image projection and image compression (Abu *et al.*, 2010).

This study proposes an approach based on 256 discrete orthonormal Tchebichef polynomials as presented in Fig. 1. The smaller matrix of DTT is chosen to get smaller computation in the vowel recognition process. This study analyzes power spectral density, frequency formants and vowel recognition performance for five vowels using 256 discrete orthonormal Tchebichef polynomials.

In Fig. 1, x-axis presents the size of kernel matrix Tchebichef moment function order n and y-axis presents the polynomials of degree k.

Corresponding Author: Ferda Ernawan, Faculty of Information and Communication Technology, Universitas Dian Nuswantoro, Imam Bonjol No. 206, Semarang 50131, Indonesia

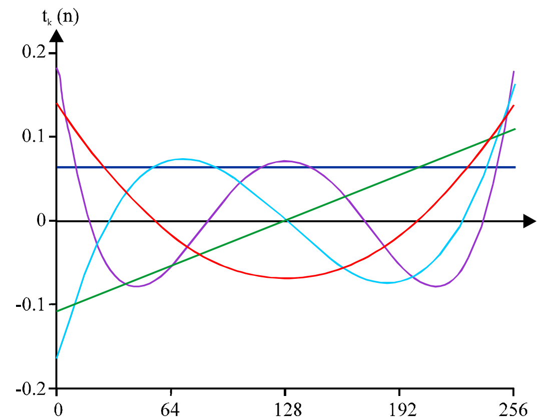

Fig. 1: First five discrete orthonormal Tchebichef polynomials $t_k(n)$ for k = 0,1,2,3 and 4

DISCRETE TCHEBICHEF TRANSFORM

The orthonormal Tchebichef polynomials use the set recurrence relation to approximate the speech signals. For a given positive integer N (the vector size) and a value n in the range [1, N-1], the orthonormal version of the one dimensional Tchebichef function is given by following recurrence relations in polynomials $t_k(n)$ of moment order k in polynomials $t_k(n)$ (Mukundan, 2004):

$$t_k(n) = a_1 n t_{k-1}(n) + a_2 t_{k-1}(n) + a_3 t_{k-2}(n) \tag{1}$$

for k = 2, 3, ..., N-1 and n = 0, 1,..., n-1.
Where:

$$a_1 = \frac{2}{k}\sqrt{\frac{4k^2-1}{N^2-k^2}} \tag{2}$$

$$a_2 = \frac{(1-N)}{k}\sqrt{\frac{4k^2-1}{N^2-k^2}} \tag{3}$$

$$a_3 = \frac{(k-1)}{k}\sqrt{\frac{2k+1}{2k-3}}\sqrt{\frac{N^2-(k-1)^2}{N^2-k^2}} \tag{4}$$

The starting values for the above recursion can be obtained from the following equations:

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

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

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

$$t_k(n) = \gamma_1 t_k(n-1) + \gamma_2 t_k(n-2) \tag{8}$$

k = 1, 2,..., N-1 and n = 2, 3,...,(N/2-1).
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 orthonormal Tchebichef polynomials set $t_k(n)$ of order N is defined as:

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

k = 0, 1,..., N-1.
where, X(k) denotes the coefficient of orthonormal Tchebichef polynomials n = 0, 1,..., N-1. x(n) is the sample of speech signal at a time index of n.

EXPERIMENTAL ANALYSIS

Sample sounds: The sample sounds of five vowels used here are male voices. The sample sounds of vowels have a sampling component at a frequency rate about of 11 kHz. As vowel data, there are three classifying events in speech, which are silence, unvoiced and voiced. By removing the silence part, the speech sound provides useful information of each utterance. One important threshold is required to remove the silence part. In this experiment, the threshold is 0.1. This means that any zero-crossings that start and end within the range of t_α, where, $-0.1 < t_\alpha < 0.1$ are to be discarded.

Speech signal windowed: The samples of five vowels have 4096 sample data. On one hand, the samples of speech signal of vowels are windowed into four frames. Each frame consumes 1024 sample data which represents speech signal. In this study, the sample speech signal for 1-1024, 1025-2048, 2049-3072, 3073-4096 sample data is represented on frames 1, 2, 3 and 4, respectively. In this experiment, a sample speech signal on the third frame is chosen as a sample to evaluate and analyze using 1024 DTT and 1024 FFT. On the other hand, the sample speech signals of the vowels are windowed into sixteen frames. Each window consists of 256 sample data which represents speech signals. In this study, the speech signals of five vowels on the tenth as a sample in the middle frame are used to analyze the use of 256 DTT. In the middle frame of speech signal consists significant

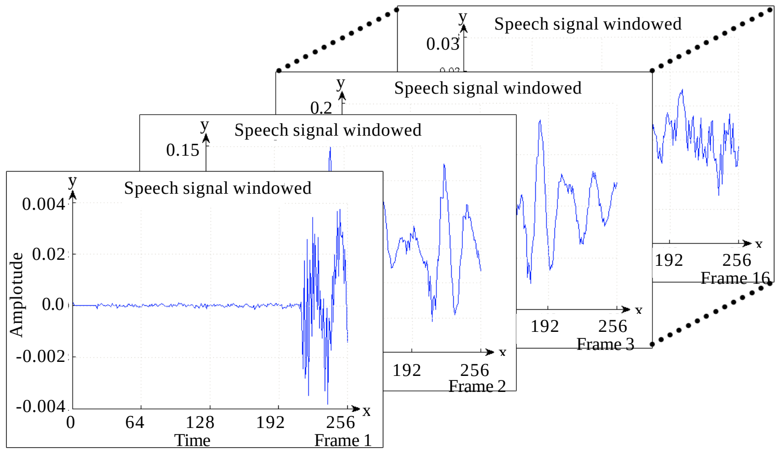

Fig. 2: Speech signal windowed into sixteen frames

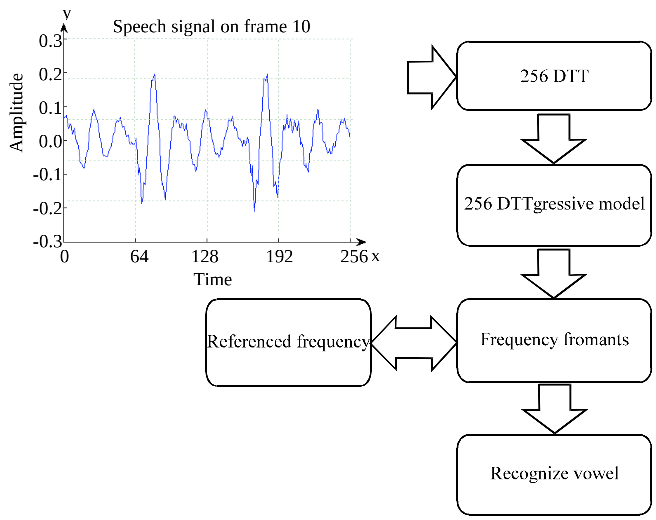

Fig. 3: Visualization of vowel recognition using DTT

important data of speech signal. Therefore, the sample in the middle has been chosen to analyze and evaluate. The sample of speech signal is presented in Fig. 2.

Since, we are administering vowel recognition in English, the speech signal shall be analyzed are on the middle vowels. Typically an English word has significant data on the middle speech signal of the vowel. It is also critical to provide a dynamic recognition module on the vowel that is immediately recognized. The visual representation of vowel recognition using DTT is given in Fig. 3.

Next, autoregression is used to generate formants or detect the peaks of the frequency signal. These formants are used to determine the characteristics of the vocal by comparing them to referenced formants. The referenced formants comparison is defined base on the classic study of vowels (Peterson and Barney, 1952). Then, the comparison of these formants is to decide on the output of the vowel.

Coefficients of discrete tchebichef transform: This section provides a representation of DTT coefficient formula. Consider the discrete orthonormal Tchebichef polynomials definition (2)-(8) above, a set kernel matrix of 256 orthonormal polynomials are computed with speech signals on each window. The coefficients of DTT of order n = 256 sample data for each window are given as in the following formula:

$$TC = S \qquad (12)$$

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

where, C is the coefficient of discrete orthonormal Tchebichef polynomials which represents c_0, c_1, c_2,...,c_{n-1} T is matrix computation of discrete orthonormal Tchebichef polynomials $t_k(n)$ for k = 0, 1,...,N-1. S is the sample of speech signal window which is given by x(0), x(1), x(2),..., x(n-1). The coefficient of DTT is given in as follows:

$$C = T^{-1}S \tag{13}$$

Spectrum analysis: Spectrum analysis is used to analyze the spectrum picked up and recording system (Schubert, 2005). The spectrum analysis using DTT can be defined in the following equation:

$$p(k) = |c(n)|^2 \tag{14}$$

$$c(n) = \frac{x(n)}{t_k(n)} \tag{15}$$

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 using 256 DTT of the vowel 'O' for 256 sample data is shown in Fig. 4. The spectrum analysis via FFT can be generated as follows:

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

where, X(k) is FFT coefficients of the speech signal. The spectrum analysis using FFT of vowel 'O' is shown in Fig. 5.

Where the x-axis show the frequency of speech signals and y-axis represent the power spectrum of the speech signals. Refer to Fig. 4 and 5, spectrum analysis of vowel 'O' using FFT produces simpler output than DTT.

Power spectral density: Power Spectral Density (PSD) is the estimation of distribution of power contained in a signal over a frequency range (Khandoker *et al*., 2008). The unit of PSD is energy per frequency. PSD represents the power of amplitude modulation signals. The power spectral density using DTT is provided as follows:

$$pw(k) = 2\frac{|c(n)|^2}{(t_2 - t_1)} \tag{17}$$

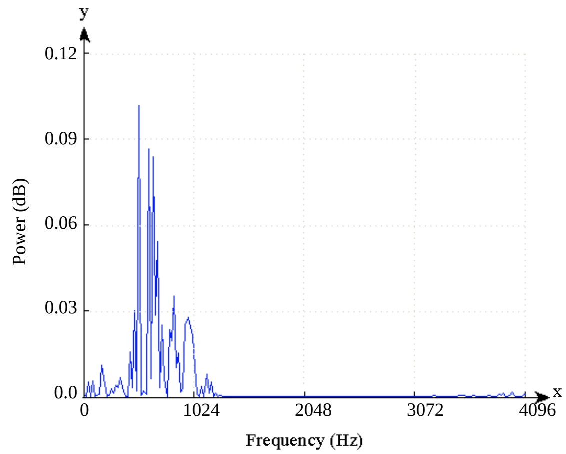

Fig. 4: Spectrum analysis using 256 DTT of vowel 'O' on frame 10

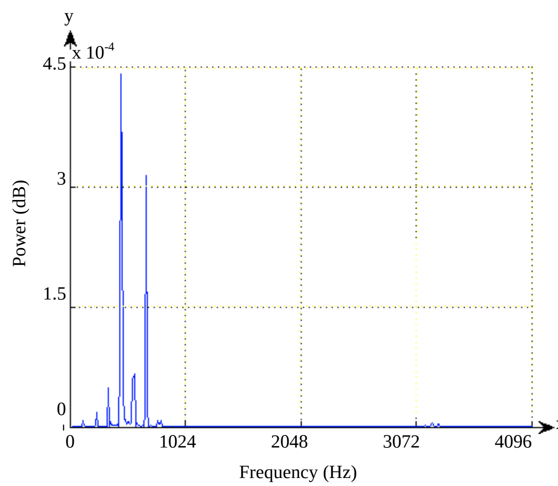

Fig. 5: Imaginary part of FFT for spectrum analysis of vowel 'O' on frame 3

where, c(n) is coefficient of dicrete Tchebichef Transform. (t_1, t_2) are precisely the average power of spectrum in the time range. The power spectral density using 256 DTT for vowel 'O' is shown in Fig. 6. The one-sided PSD using FFT can be computed as:

$$ps(k) = 2\frac{|X(k)|^2}{(t_2 - t_1)} \tag{18}$$

where, X(k) is a vector of N values at a frequency index k, the factor 2 is due to add for the contributions from positive and negative frequencies. The power spectral density using FFT for vowel 'O' on frame 3 is shown in Fig. 7, where, the x-axis show the frequency of spectral density and y-axis represent the power spectral

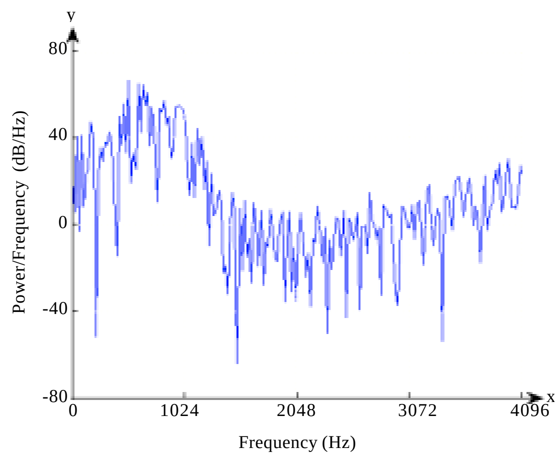

Fig. 6: Power Spectral Density of vowel 'O' using 256 DTT on frame 10

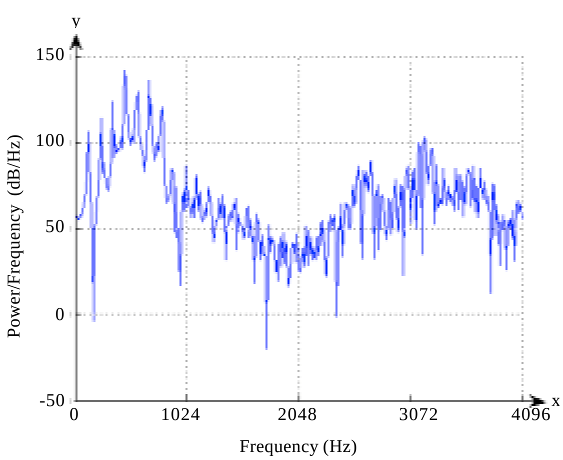

Fig. 7: Power Spectral Density using FFT for vowel 'O' on frame 3

density. The power spectral density is plotted using a decibel (dB) scale 20 log 10.

Autoregression: Speech production is modeled by an excitation filter model, where an autoregressive filter model is used to determine the vocal tract resonance property and an impulse models the excitation of voiced speech (Li and Andersen, 2006). The autoregressive process of a series y_j using DTT of order v can be expressed in the following equation:

$$y_j = -\sum_{k=1}^{v} a_k c_{j-k} + e_j \tag{19}$$

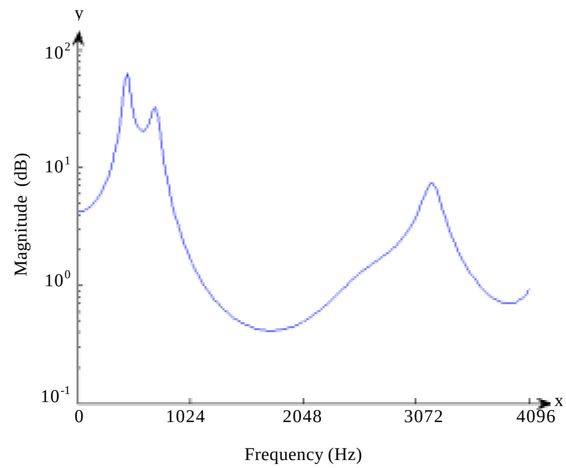

Fig. 8: Autoregressive of vowel 'O' using 256 DTT on frame 10

where, a_k are real value autoregression coefficients, c_j is the coefficient of DTT at a frequency index j, v is 12 and e_j represents the errors that are term independent of past samples. The autoregressive model using 256 DTT of vowel 'O' are shown in Fig. 8. Next, the autoregressive process of a series y_j using FFT of order v is given in the following equation:

$$y_j = -\sum_{k=1}^{v} a_k q_{j-k} + e_j \tag{20}$$

where, a_k are real value autoregression coefficients, q_j represent the inverse FFT from power spectral density, and v is 12. The peaks of frequency formants using FFT in autoregressive for vowel 'O' on frame 3 is shown in Fig. 9, where, the x-axis show the frequency formants of vowel 'O' and y-axis represent the magnitude of the formants. An autoregressive model describes the output of filtering a temporally uncorrelated excitation sequence through all pole estimate of the signal. Autoregressive models have been used in vowel recognition to represent the envelope of the power spectrum of the signal by performing the operation of linear prediction (Ganapathy *et al*., 2010). The autoregressive model is used to determine the characteristics of the vocal and to evaluate the formants. Frequency formant can be obtained from the estimated autoregressive parameters.

Frequency formants: Frequency formants are frequency resonance of vocal tracts in the spectrum of a vowel sound (Ali *et al*., 2006). The formants of the autoregressive curve are found at the peaks using a

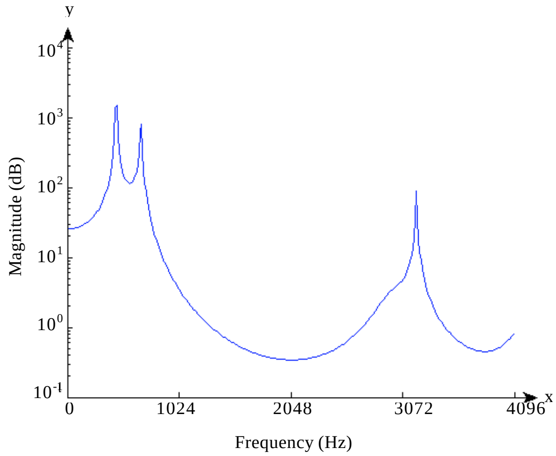

Fig. 9: Autoregressive using FFT for Vowel 'O' on frame 3

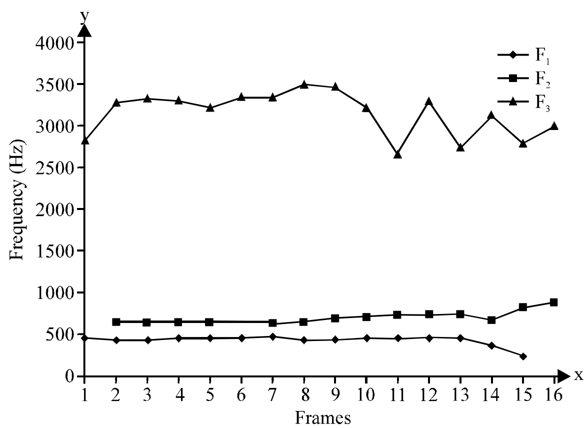

Fig. 10: Frequency Formants of Vowel 'O' using 256 DTT

numerical derivative. Formants of a vowel sound are numbered in order of their frequency like first formant (F_1), second formant (F_2), third formant (F_3) and so on. A set of frequency formants F_1, F_2 and F_3 is known to be an indicator of the phonetic identification of vowel recognition. The first three formants F_1, F_2 and F_3 contain sufficient information to recognize vowels from voice sounds. The frequency formants especially F_1 and F_2 are closely tied to the shape of the vocal tract to articulate the vowels. The third frequency formant F_3 is related to a specific sound. The frequency formants of vowel 'O' using 256 DTT are shown in Fig. 10, where, the x-axis represent speech signal frame from frame 1 to 16 and y-axis represent the frequency formants F_1, F_2 and F_3 of vowel 'O'. These vector positions of the formants are used to characterize a particular vowel. Next, the frequency peak formants of F_1, F_2 and F_3 are compared to referenced formants to decide on the output of the vowels. The comparison code for the referenced formants

Table 1: Frequency formants of five vowels

Vowels	Formants	256 DTT	1024 DTT	1024 FFT
i	F_1	236	226	215
	F_2	2411	2411	2444
	F_3	3466	3466	3434
e	F_1	301	301	322
	F_2	1485	1485	1453
	F_3	2347	2357	2401
a	F_1	624	581	667
	F_2	1012	979	1055
	F_3	2648	2670	2637
o	F_1	452	452	462
	F_2	710	710	689
	F_3	3208	3219	3208
u	F_1	301	301	247
	F_2	710	699	689
	F_3	3380	3380	3413

was written based on the classic study of vowels by Peterson and Barney (1952). The comparison of the frequency formants using 256 DTT, 1024 DTT and 1024 FFT for five vowels are shown in Table 1.

DISCUSSION

The frequency formants in vowel recognition using 256 DTT, 1024 DTT and 1024 FFT have been investigated. The speech signal was divided into different frame. As proposed a 256 forward DTT can be used in spectrum analysis in terms of vowel recognition. With reference to the experimental results as presented in Fig. 8 and 9, the peaks shape of first frequency formant (F_1), second frequency formant (F_2) and third frequency formant (F_3), respectively appear to be similar output. The frequency formants as shown in Fig. 10 show identically output among each frame. The frequency formants of vowel recognition using 256 DTT, 1024 DTT and 1024 FFT are analyzed for five vowels. Frequency formants as presented in Table 1 show that the frequency formants of vowel 'O' using DTT produce similar shape output with frequency formants using FFT. The results on Table 1 show that the peaks of first frequency formant (F_1), second frequency formant (F_2) and third frequency formant (F_3) using 256 DTT, 1024 DTT and 1024 FFT, respectively appear to be to produce output that is identically quite similar. Even though, there are missing elements of recognition, the overall the result is practically acceptable.

The time taken for vowel recognition as presented in Table 2 shows that vowel recognition performance using 256 DTT requires a shorter time to recognize five vowels than 1024 DTT and 1024 FFT. The time taken of vowel recognition using 256 DTT reveals that it is faster and computationally efficient than 1024 DTT and 1024 FFT, because the 256 DTT requires a smaller matrix computation and a simpler computation field in the

Table 2: Time taken for vowel recognition performance using DTT and FFT for five vowels

Vowels	256 DTT (sec)	1024 DTT(sec)	1024 FFT (sec)
i	0.577231	0.982382	0.648941
e	0.584104	0.993814	0.643500
a	0.589120	0.963208	0.738364
o	0.574317	0.953711	0.662206
u	0.579469	0.978917	0.703741

transformation domain. The experimental results show that the proposed 256 DTT algorithm efficiently reduces the time taken to transform the time domain into the frequency domain.

CONCLUSION

FFT is a popular transformation method over the last decades. Alternatively, DTT is proposed here instead of the popular FFT. In previous research, vowel recognition using 1024 DTT has been experimented. In this paper, the simplified matrix on 256 DTT is proposed to produce vowel recognition that is a simpler and more computationally efficient than 1024 DTT. 256 DTT consumes smaller matrix which can be efficiently computed on rational domain compared to the popular 1024 FFT. The preliminary experimental results show that the peaks of first frequency formant (F_1), second frequency formant (F_2) and third frequency formant (F_3) using 256 DTT give similar output with 1024 DTT and 1024 FFT in terms of vowel recognition. Vowel recognition using a scheme of 256 DTT should perform well so as to recognize vowels. It can be the next candidate in vowel recognition.

ACKNOWLEDGMENT

The authors would like to express a very special thank to Ministry of Higher Education (MOHE), Malaysia for providing financial support on this research project by Fundamental Research Grant Scheme (FRGS/2012/FTMK/SG05/03/1/F00141).

REFERENCES

Abu, N.A., S.L. Wong, N.S. Herman and R. Mukundan, 2010. An efficient compact tchebichef moment for image compression. Proceedings of the 10th International Conference on Information Science Signal Processing and their applications, May 10-13, 2010, Kuala Lumpur, pp: 448-451.

Ali, A., S. Bhatti and M.S. Mian, 2006. Formants based analysis for speech recognition. Proceedings of International Conference on Engineering of Intelligent System, (EIS'06), Islamabad, pp: 1-3.

Bailey, D.H. and P.N. Swarztrauber, 1994. A fast method for numerical evaluation of continuous fourier and laplace transform. J. Scient. Comput., 15: 1105-1110.

Ernawan, F. and N.A. Abu, 2011. Efficient discrete tchebichef on spectrum analysis of speech recognition. Int. J. Machine Learn. Comput., 1: 001-006.

Ganapathy, S., P. Motlicek and H. Hermansky, 2010. Autoregressive models of amplitude modulations in audio compression. IEEE Trans. Audio Speech Language Process., 18: 1624-1631.

Khandoker, A.H., C.K. Karmakar and M. Palaniswami, 2008. Power spectral analysis for identifying the onset and termination of obstructive sleep apnoea events in ECG recordings. Proceeding of the 5th International Conference on Electrical and Computer Engineering, December 20-22, 2008, Dhaka, pp: 96-100.

Li, C. and S.V. Andersen, 2006. Blind identification of non-gaussian autoregressive models for efficient analysis of speech signal. Proceedings of the International Conference on Acoustic, Speech and Signal Processing, Vol. 1, May 14-19, 2006, Toulouse, pp: 1205-1208.

Mukundan, R., 2003. Improving image reconstruction accuracy using discrete orthonormal moments. Proceedings of International Conference on Imaging Systems, Science and Technology, Jun 23-26, 2003, Computer Science and Software Engineering, pp: 287-293.

Mukundan, R., 2004. Some computational aspects of discrete orthonormal moments. IEEE Trans. Image Process., 13: 1055-1059.

Peterson, G.E. and H.L. Barney, 1952. Control methods used in a study of the vowels. J. Acoustical Soc. Am., 24: 175-184.

Schubert, W.K., 2005. Micro acoustic spectrum analyzer. J. Acoust. Soc. Am., 117: 2692-2693.

Vite-Frias, J.A., Rd.J. Romero-Troncoso and A. Ordaz-Moreno, 2005. VHDL Core for 1024-point radix-4 FFT Computation. Proceedings of International Conference on Reconfigurable Computing and FPGAs, September 28-30, 2005, Puebla City, pp: 020-024.