Vowel Recognition using Discrete Tchebichef Transform

1Ferda Ernawan, 2Nur Azman Abu and 3Nanna Suryana
1Faculty of Information and Communication Technology, Universitas Dian Nuswantoro
Imam Bonjol No. 206, Semarang 50131, Indonesia
2Faculty 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 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 straightforward 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):

\[
t_k(n) = a_0 t_{k-1}(n) + a_1 t_{k-2}(n) + \cdots + a_{k-1} t_1(n)
\]

for \( k = 2, 3, \ldots, N-1 \) and \( n = 0, 1, \ldots, n-1 \).

Where:

\[
a_0 = \frac{2}{k} \sqrt{\frac{k^2 - 1}{N^2 - k^2}}
\]

\[
a_1 = \frac{(N - k)}{k} \sqrt{\frac{k^2 - 1}{N^2 - k^2}}
\]

\[
a_i = \frac{(k - i)}{k} \sqrt{\frac{2k + 1}{2k - 3}} \sqrt{\frac{N^2 - (i - 1)^2}{N^2 - k^2}}
\]

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

\[
t_0(n) = \frac{1}{\sqrt{N}}
\]

\[
t_1(n) = \sqrt{\frac{N-k}{N+k}} t_1(n)
\]

\[
t_{0}(0) = t_0(0)
\]

\[
t_{1}(0) = \left[ \frac{1}{1-N} \right] t_1(0)
\]

\[
t_k(n) = \gamma_1 t_k(n+1) + \gamma_2 t_k(n-2)
\]

\( k = 1, 2, \ldots, N-1 \) and \( n = 2, 3, \ldots, (N/2-1) \).

Where:

\[
\gamma_1 = \frac{-k(k+1) - 2(n-1)(n-N-1) - n}{n(N-n)}
\]

\[
\gamma_2 = \frac{(n+1)(n-N-1)}{n(N-n)}
\]

The forward discrete orthonormal Tchebichef polynomials set \( t_k(n) \) of order \( N \) is defined as:

\[
x(k) = \sum_{n=0}^{N-1} x(n) t_k(n)
\]

where, \( x(k) \) denotes the coefficient of orthonormal Tchebichef polynomials \( n = 0, 1, \ldots, N-1 \). \( x(n) \) 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 \( t_{0} \) where \(-0.1 < t_{0} < 0.1 \) 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
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 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 chebichef transform:** This section provides a representation of DTT coefficient formula. Consider the discrete orthonormal Chebichef 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:

\[
TC = S
\]  

(12)
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:

$$C = T^{-1}S$$

(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:

$$p(k) = |x(k)|^2$$

(14)

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

(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:

$$p(k) = |X(k)|^2$$

(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:

$$pw(k) = \frac{|c(n)|^2}{(t_k - t_{k-1})}$$

(17)

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 density using FFT.
**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 \( y_i \) using DTT of order \( v \) can be expressed in the following equation:

\[
y_i = \sum_{k=1}^{v} a_k y_{i-k} + e_i
\]

where, \( a_k \) are real value autoregression coefficients, \( e_i \) is the coefficient of DTT at a frequency index \( j \), \( v \) is 12 and \( e_i \) 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 \( y_i \) using FFT of order \( v \) is given in the following equation:

\[
y_i = -\sum_{k=1}^{v} a_k e_{i-k} + e_i
\]

where, \( a_k \) are real value autoregression coefficients, \( e_i \) 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
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 (F₁), second frequency formant (F₂) and third frequency formant (F₃) 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 (F₁), second frequency formant (F₂) and third frequency formant (F₃) 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 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>
<thead>
<tr>
<th>Vowels</th>
<th>Formants</th>
<th>256 DTT</th>
<th>1024 DTT</th>
<th>1024 FFT</th>
</tr>
</thead>
<tbody>
<tr>
<td>i</td>
<td>F₁</td>
<td>236</td>
<td>226</td>
<td>214</td>
</tr>
<tr>
<td></td>
<td>F₂</td>
<td>2411</td>
<td>2411</td>
<td>2444</td>
</tr>
<tr>
<td></td>
<td>F₃</td>
<td>3468</td>
<td>3466</td>
<td>3434</td>
</tr>
<tr>
<td>c</td>
<td>F₁</td>
<td>301</td>
<td>301</td>
<td>322</td>
</tr>
<tr>
<td></td>
<td>F₂</td>
<td>1485</td>
<td>1485</td>
<td>1453</td>
</tr>
<tr>
<td></td>
<td>F₃</td>
<td>2347</td>
<td>2357</td>
<td>2401</td>
</tr>
<tr>
<td>a</td>
<td>F₁</td>
<td>624</td>
<td>581</td>
<td>667</td>
</tr>
<tr>
<td></td>
<td>F₂</td>
<td>1012</td>
<td>979</td>
<td>1055</td>
</tr>
<tr>
<td></td>
<td>F₃</td>
<td>2648</td>
<td>2670</td>
<td>2637</td>
</tr>
<tr>
<td>o</td>
<td>F₁</td>
<td>452</td>
<td>452</td>
<td>462</td>
</tr>
<tr>
<td></td>
<td>F₂</td>
<td>710</td>
<td>710</td>
<td>689</td>
</tr>
<tr>
<td></td>
<td>F₃</td>
<td>3208</td>
<td>3219</td>
<td>3208</td>
</tr>
<tr>
<td>u</td>
<td>F₁</td>
<td>301</td>
<td>301</td>
<td>247</td>
</tr>
<tr>
<td></td>
<td>F₂</td>
<td>710</td>
<td>699</td>
<td>689</td>
</tr>
<tr>
<td></td>
<td>F₃</td>
<td>3380</td>
<td>3380</td>
<td>3413</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Vowels</th>
<th>256 DTT (sec)</th>
<th>1024 DTT (sec)</th>
<th>1024 FFT (sec)</th>
</tr>
</thead>
<tbody>
<tr>
<td>i</td>
<td>0.577231</td>
<td>0.982382</td>
<td>0.648941</td>
</tr>
<tr>
<td>e</td>
<td>0.584104</td>
<td>0.998914</td>
<td>0.648901</td>
</tr>
<tr>
<td>a</td>
<td>0.589120</td>
<td>0.963208</td>
<td>0.758664</td>
</tr>
<tr>
<td>o</td>
<td>0.574317</td>
<td>0.953711</td>
<td>0.662206</td>
</tr>
<tr>
<td>u</td>
<td>0.579469</td>
<td>0.978917</td>
<td>0.703741</td>
</tr>
</tbody>
</table>

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. Alternately, 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 (F1), second frequency formant (F2) and third frequency formant (F3) 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


