VOL. 11, NO. 9, MAY 2016 ISSN 1819-6608
ARPN Journal of Engineering and Applied Sciences

www.arpnjournals.com

ON THE PERFORMANCE OF SEGMENT AVERAGING OF DISCRETE COSINE TRANSFORM COEFFICIENTS ON MUSICAL INSTRUMENTS TONE RECOGNITION

Linggo Sumarno
Electrical Engineering Study Program, Sanata Dharma University, Yogyakarta, Indonesia
E-Mail: lingsum@usd.ac.id

ABSTRACT

In the Discrete Cosine Transform (DCT) domain, the tones of musical instruments can be divided into two groups. The first one with the single significant local peaks and the second one with the multiple significant local peaks. The second one can be divided into two sub groups, which have many and a few significant local peaks. This research deal with multiple significant local peaks. In this research, segment averaging was used to reduce the number of DCT coefficients, in the DCT domain. In this case, the reduced number of DCT coefficients called feature extraction coefficients. Based on the experiment, when the segment averaging of DCT coefficients was used optimally for the tones which had many (i.e. thirteen) and a few (i.e. three) significant local peaks, it could give 8 and 16 feature extraction coefficients respectively. So, in order that segment averaging of DCT coefficients could be used optimally, either for the tones which have many or a few significant local peaks in the DCT domain, it could use segment length 4 points and DCT length 64 points. By using it, it could give 16 feature extraction coefficients.

Keywords: tone recognition, segment averaging, DCT, feature extraction.

1. INTRODUCTION

Based on the human perception, musical instruments have two aural characteristics: a particular kind of tuning (scale) and the particular kind of sound (timbre) [1]. The scale is an aural characteristic of high and low tones in a musical instrument. Thus, if the scale is getting higher, the higher the tone. Conversely, if the scale is getting lower, the lower the tone.

The timbre is aural characteristics of instrument type. Based on the timbre, in the DCT domain, the tone of musical instruments can be divided into two groups. The first one with single significant local peaks (which also called monophonic), and the second one with multiple significant local peaks (which also called polyphonic). For the multiple significant local peaks, it can be divided into two subgroups. The first one with many significant local peaks, and the second one with a few significant local peaks. Figure-2 shows examples of timbre representation in the DCT domain, for pianica and soprano recorder musical instruments that shown in Figure-1.

(a) Pianica

(b) Soprano recorder

Figure-1. Pianica dan soprano recorder.

In Figure-2, pianica is an example of a musical instrument with many significant local peaks in the DCT domain, while the soprano recorder is an example of a musical instrument with a few significant local peaks in the DCT domain.

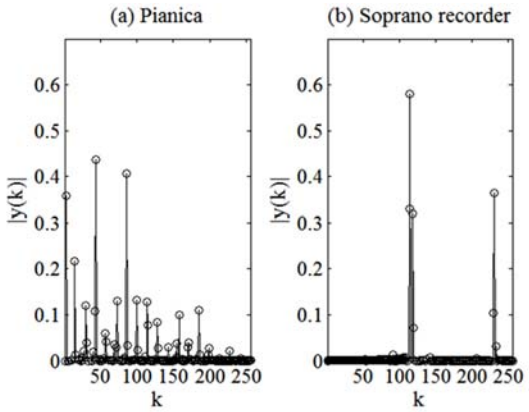

Figure-2. Timbre representation in the DCT domain $y(k)$ (see equation 2) of the tone 'C', for pianica and soprano recorder, and the use of 256 points DCT.

Previous researches associated with the tone recognition, essentially made use using time domain or transformed domain approaches. Tone recognition researches using time domain approach [2] [3], were basically based on autocorrelation. This time domain approach, deal less well to polyphonic tones. Previous researches using the transformed domain approach,

basically based on Fast Fourier Transform (FFT) [4] [5] [6] [7] or DCT [8]. Moreover, in this approach, there were basically based on a fundamental frequency [4] [5], and the other were not based on a fundamental frequency [6] [7] [8]. This transformed domain approach deal well with polyphonic tones. Furthermore, in this paper, it will be discussed the recognition of polyphonic tones which is not based on a fundamental frequency.

In the previous research, Surya [6] developed a pianica tone recognition using Kaiser Window, FFT, and the correlation function. In order to get a recognition rate of 100%, at least 128 feature extraction coefficients were. Sumarno [8] developed further a pianica tone recognition using Gaussian window, DCT, and the cosine distance function. In order to get the recognition rate of 100%, at least required a number of 32 feature extraction coefficients. Sumarno [7] developed further a pianica tone recognition using Blackman window, FFT, and the Euclidean distance function, as well as FFT windowing coefficient. For the recognition rate of 100%, at least a number of 12 feature extraction coefficients was required. By looking at the number of coefficients required for feature extraction, it can be seen that the tone recognition research to reduce the number of feature extraction coefficients is still wide open.

This paper describes a research about the tone recognition of pianica and soprano recorder tones, which each of them represent tones with many and a few significant local peaks in the DCT domain respectively (see Figure-2). It will be investigated the performance of segment averaging that inspired by Setiawan [9] to be used on a tone recognition system using DCT. In more detail, it will be investigated the influence of segment length and also the number of significant local peaks in the DCT domain, to the recognition rate and the number of feature extraction coefficients. The tone recognition used in this research adopted the template matching approach [11].

2. RESEARCH METHODOLOGY

2.1 Material and tools

The research materials were isolated tones of pianica and soprano recorder, in wav format. It was obtained by recording the tones of pianica and soprano recorder, by using sampling rates 4800 Hz and 2400Hz respectively. The magnitude of sampling rates were chosen according to the Nyquist criterion, namely, the amount of the minimum sampling rate is twice the highest analog frequency of 2400 Hz for tone 'B' on pianica, and 995Hz for the tone 'B' on the soprano recorder. Based on the evaluation, recording duration for 2 seconds was adequate, because the sound produced was already in the steady state condition, especially in the middle part of the data, which was chosen for the purpose of frame blocking.

Research tools were a pianica (Brother brand) and a soprano recorder (Yamaha brand). A microphone Genius MIC-01A was used to capture the sound signals. A computer with an Intel Core i3 3220 processor and 4GB of RAM, was used to process the captured sound signals.

2.2 Design of the Tone Recognition System

The tone recognition system, shown in block diagram in Figure-3. The input is wav file and the output is a number that indicates the recognized tone.

Frame blocking is the process of taking a frame signal from a long signal series of signal [12]. The purpose of the frame blocking is to reduce the number of data signals to be processed. The effect of this reduction is a reduction in computing time. In this research, a frame signal is captured from the middle part of the signal, by assuming that in the middle part, the signal has reached its steady state time. The length of a frame signal was evaluated by using a number of lengths namely 32, 64, 128, and 256 points. The length of a frame signal has the same length with the length of DCT in the next process.

Normalization is the process of setting the maximum value to one, in a signal data. Normalization aimed to eliminate differences in maximum values from a number of signal data that came from recording results.

Windowing is a process of reducing discontinuities at the edges of the signal. At a recorded signal, usually found discontinuities at the edges of the signal. This case would give a number of harmonic signals in the transformed signal. The appearance of these harmonic signals would affect the accuracy of the feature extraction using DCT [8]. In order to reduce the appearance of harmonic signals, so the edges of the input signal needs to be reduced by using windowing [12].

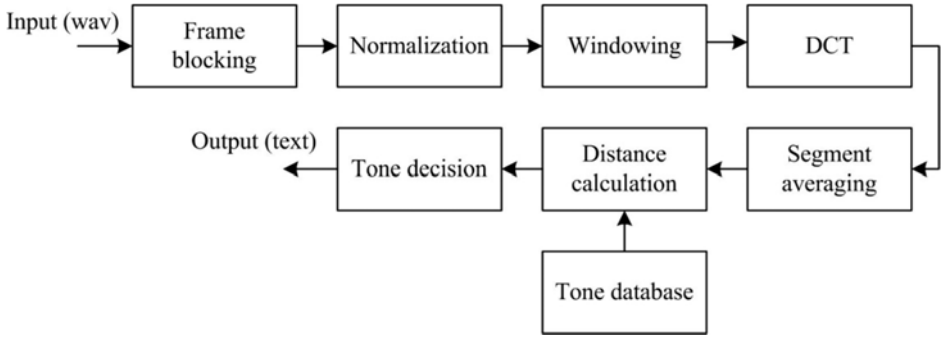

Figure-3. Block diagram of the tone recognition system.

Hamming window [13] is a window that is commonly used for windowing. Hamming window $w(n)$ with a width of $N + 1$ points defined below.

$$w(n) = 0.54 - 0.46\cos\left(2\pi\frac{n}{N}\right), \quad 0 \le n \le N \qquad (1)$$

DCT (Discrete Cosine Transform)

DCT is a transformation method. It is used to transform the signal from the time domain to the DCT domain. DCT of a series $u(n)$ which has N points in length, formulated as follows [10].

$$y(k) = \alpha(k)\sum_{n=0}^{N-1} u(n)\cos\left[\frac{\pi(2n+1)k}{2N}\right], \; 0 \le n \le N-1 \qquad (2)$$

where

$$\alpha(0) \triangleq \sqrt{\frac{1}{N}}, \quad \alpha(k) \triangleq \sqrt{\frac{2}{N}} \text{ for } 0 \le n \le N-1. \qquad (3)$$

Segment averaging, which inspired by Setiawan [9], is a process to reduce the size of the signal. Basically, the signal that have been reduced in size represents the basic shape of the signal pattern. In this research, segment averaging was used to reduce the size of the signal after DCT process. Results of the reduction of the signal was called the feature extraction of the signal. The algorithm of segment averaging is shown below.

Segment averaging algorithm

1. Determine a series $y(k) = \{y(0), y(1), ..., y(N-1)\}$ where $N = 2^p$ for $p \ge 0$.
2. Determine a segment length L where $L = 2^q$ for $0 \le q \le p$.
3. Divide the series $y(k)$ based on the segment length L . Therefore, it will be resulted a number of M segments as follow

$$M = \frac{N}{L}, \qquad (4)$$

and also a series $f(r) = \{f(1), f(2), ..., f(L)\}$ in each segment.

4. Compute the average value in each segment $z(v)$ as follow:

$$z(v) = \frac{1}{L}\sum_{r=1}^{L} f_v(r), \quad 1 \le v \le M \; . \qquad (5)$$

In this research, a number of the segment length was evaluated based on the DCT length. Table-1 shows that.

Distance calculation is a process of calculating the distance, between the feature extractions of an input signal with the feature extraction of a number of signals in the tone database. Euclidean distance is a distance function that commonly used [14]. Euclidean distance is defined by

Table-1. Evaluation of the segment length based on the DCT length.

DCT length (points)	**Segment length** (points)
256	1, 2, 4, 8, 16, 32, 64, 128, and 256
128	1, 2, 4, 8, 16, 32, 64, and 128
64	1, 2, 4, 8, 16, 32, and 64
32	1, 2, 4, 8, 16, and 32

$$E(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{i=1}^{m}(x_i - y_i)^2} \qquad (5)$$

where x and **y** are two vectors which have equal length (one vector is the vector of the input signal, whereas the other one is a vector that exists in the database), and m is the length of the vector **x** or **y**. Calculation of the distance is an indication of template matching [11] approach was used in this research.

Tone decision is a process to determine the tone of the input signal. Tone decision was caried out by finding the minimum value of distance calculation values. This values came from calculating the distance between a feature extraction of the input tone and a set of feature extraction of tones in a tone database. A tone which has a minimum distance value, determined as the output tone.

2.3 Tone database

Tone database was needed in the distance calculation process. To create the tone database, for each instrument (pianica and soprano recorder), the author take a number of 10 samples for each tone ('C', 'D', 'E', 'F', 'G', 'A', and 'B'). In this research it was assumed, by taking 10 samples for each tone, all variations for each tone pattern have been elaborated. Because from 7 tones, there are 10 samples for each tone, so in total there are 70 tones to create a tone database. Figure-4 shows a block diagram of the database generation process for each tone.

In this research, for each instrument (pianica and sprano recorder), a number of 30 sets of the tone database were generated. It were created based on a combination of values that were evaluated as shown in Table-1.

2.4 Test tones

Test tones were needed to evaluate the performance of the recognition system, at various tone databases that described above. In this research, for each instrument (pianica and soprano recorder), taken 10 samples for each tone ('C', 'D', 'E', 'F', 'G', 'A' and 'B'). Thus for each of these instruments there are a total of 70 test tones.

2.5 Recognition rate

Recognition rate was used to rate the performace of recognition system as descibed above. It was calculated by the following equation

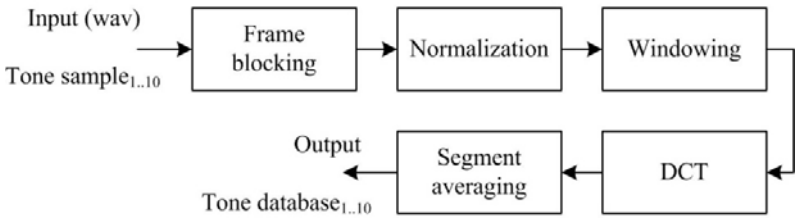

Figure-4. Block diagram of the tone database generation.

$$\text{Recognition rate} = \frac{\text{Number of recognised tones}}{\text{Number of test tones}} \times 100\% \quad (7)$$

where the number of test tones, as described above, is 70 tones for each musical instrument (pianica and soprano recorder).

3. RESULTS AND DISCUSSIONS

Tone recognition system test results on various combinations of DCT length and segment averaging length, for pianica and soprano recorder, are shown in Tables 2 and 3. As shown in Tables 2 and 3, in general, the longer segment length will decrease the recognition rate. This is due, as shown in Table-4, the longer segment length would further decrease the number of feature extraction coefficients.

Basically decreasing the number of feature extraction coefficients, it would decrease the pattern details. However, if the number of feature extraction coefficients are too little, too many pattern details will be lost, which causes a pattern will be more similar with the other patterns (see Figure-5 (c)). As a result, it would be increasingly difficult to distinguish between a pattern with the other patterns, thereby resulting in lower levels of recognition. Thus, it could be said, if the number of the feature extraction coefficients decrease, it would also decrease the recognition rate.

Based on Table-2 and 4, for the recognition of pianica tones that have many significant local peaks in the DCT domain (see Figure-2), the smallest number of feature extraction coefficients that gave a 100% recognition rate is 8 coefficients. This result was better than that achieved previously by Sumarno [7], which required a total of 12 coefficients. However, based on Table-3 and 4, for the recognition of soprano recorder tones that have a few significant local peaks in DCT domain (see Figure-2), the smallest number of feature extraction coefficients that gives a 100% recognition rate increased to 16 coefficients. The cause of this case was shown graphically in Figure-6.

Table-2. The test results of pianica tone recognition, in various combinations of segment length and DCT length. Results shown: Recognition rate (%).

DCT length (points)	**Segment length** (points)								
	1	**2**	**4**	**8**	**16**	**32**	**64**	**128**	**256**
256	100	100	100	100	100	100	98.6	77.1	55.7
128	100	100	100	100	100	95.7	75.7	61.4	-
64	100	100	**100**	100	88.6	75.7	57.1	-	-
32	75.7	74.3	72.9	58.6	48.6	41.4	-	-	-

Table-3. The test results of soprano recorder tone recognition, in various combinations of segment length and DCT length. Results shown: Recognition rate (%).

DCT length (points)	**Segment length** (points)								
	1	**2**	**4**	**8**	**16**	**32**	**64**	**128**	**256**
256	100	100	100	100	98.6	78.6	60	37.1	22.9
128	100	100	100	100	74.3	62.8	42.8	21.4	-
64	100	100	**100**	97.1	67.1	37.1	20	-	-
32	88.6	98.6	95.7	80	44.3	22.9	-	-	-

Table-4. The number of feature extraction coefficients in various combinations of segment length and DCT length. Results shown: The number of feature extraction coefficients.

DCT length (points)	Segment length (points)								
	1	**2**	**4**	**8**	**16**	**32**	**64**	**128**	**256**
256	256	128	64	32	16	8	4	2	1
128	128	64	32	16	8	4	2	1	-
64	64	32	**16**	8	4	2	1	-	-
32	32	16	8	4	2	1	-	-	-

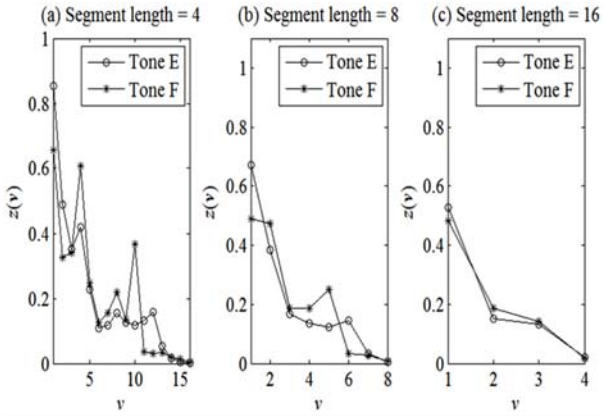

Figure-5. The influence of segment length to the calculation of segment averaging $z(v)$, for the use of 64 points DCT.

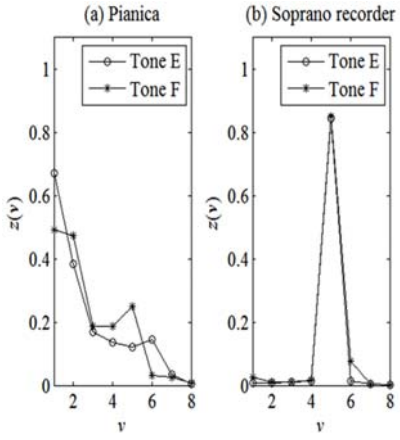

Figure-6. Examples of feature extraction difference between pianica and soprano recorder, for the use of 64 points DCT and 16 points segment length.

Figure-6 shows that, the graphic of feature extraction for the tone 'E' and 'F', for soprano recorder is more similar, when compared with pianica. This means, the results of sporano recorder feature extraction are more difficult to be distinguished each other, when compared with the results of pianica feature extraction. Based on the discussion that described above, in order to able to be distinguished each other easily, it was the necessary to add more feature extraction coefficients. Therefore, soprano recorder requires more feature extraction coefficients when compared with pianica.

Based on Table-2, 3 and 4, it can be seen that in order that segment averaging of DCT coefficients can be used optimally, either for musical instruments that have many or a few significant local peaks in the DCT domain, it could use segment length 4 points and DCT length 64 points. By using it, it could give 16 feature extraction coefficients.

4. SUMMARY

Based on the above discussions, it can be summarized as follow:

a) If segment averaging was used for signals which have many significant local peaks in the DCT domain, it would give a fewer number of feature extraction coefficients, rather than when it was used for signals which have a few significant local peaks in the DCT domain.
b) In case of segment averaging was used for pianica signals which have many (i.e. thirteen) significant local peaks in the DCT domain, the optimal number of feature extraction coefficients was 8 coefficients. That coefficient was generated by using segment length 4 points, and DCT length 64 points.
c) In case of segment averaging of DCT coefficients was used for soprano recorder signals which have a few (i.e. three) significant local peaks in the DCT domain, the optimal number of feature extraction coefficients was 16 coefficients. That coefficient was generated by using segment length 4 points, and DCT length 64 points.
d) In order that segment averaging of DCT coefficients could be used optimally, either for musical instruments that have many or a few significant local peaks in the DCT domain, it could use segment length 4 points and DCT length 64 points. By using it, it could give 16 feature extraction coefficients.

REFERENCES

[1] Forster C. 2010. Musical Mathematics: On the Art of Science and Acoustic Instruments. Chronicle Books LLC, San Diego, California. pp. vii-viii.

[2] Cheveigné A. de and H. Kawahara. 2002. YIN, A Fundamental Frequency Estimator for Speech and Music. The Journal of the Acoustical Society of America. pp. 111-117.

[3] McLeod P. and G. Wyvill. 2005. A Smarter Way to Find Pitch. In: Proceedings of the International Computer Music Conference (ICMC'05). Barcelona.

[4] Michael Noll M. 1970. Pitch Determination of Human Speech by the Harmonic Product Spectrum, the Harmonic Sum Spectrum and a Maximum Likelihood Estimate. In: Proceedings of the Symposium on Computer Processing in Communications, Vol. XIX, Polytechnic Press: Brooklyn, New York. pp. 779-797.

[5] Mitre, A., Marcelo, Q, and Régis, F. 2006. Accurate and Efficient Fundamental Frequency Determination from Precise Partial Estimates. Proceedings of the 4th AES Brazil Conference. pp. 113-118.

[6] Surya D.E. and L. Sumarno. 2012. Recognition of Pianica Tones using Kaiser Window, FFT, and Correlation (in Indonesian). In: Proceedings of National Seminar on Engineering of Industrial and Information Technology. Yogyakarta. pp. 151-157.

[7] Sumarno L. 2014. Recognition of Pianica Tones using Blackman Window and Fast Fourier Transform Feature Extraction (in Indonesian). Media Teknika. 9(2): 84-93.

[8] Sumarno L. 2013. Recognition of Pianica Tones using Gaussian Window, DCT, and Cosine Distance (in Indonesian). Jurnal Penelitian (Research Journal). 17(1): 8-15.

[9] Setiawan Y.R. 2015. Numbers Speech Recognition using Fast Fourier Transform and Cosine Similarity (in Indonesian). Undergraduate Thesis. Sanata Dharma University. Yogyakarta. pp. 66-70.

[10] Jain A. K. 1989. Fundamentals of Digital Image Processing. Prentice-Hall International Inc. New Jersey, USA.

[11] Theodoridis S. and K. Koutroumbas. 2009. Pattern Recognition. 4th Edition. Elsevier Inc. San Diego, California. pp. 481-519.

[12] Meseguer N.A. 2009. Speech Analysis for Automatic Speech Recognition. MSc Thesis. NTNU. Trondheim.

[13] Oppenheim, A.V. and R.W. Schafer. 1989. Discrete-Time Signal Processing. Prentice-Hall. pp. 447-448.

[14] Wilson D.R. and T. R. Martinez. 1997. Improved Heterogeneous Distance Function. Journal of Artificial Intelligence Research. 6: 1-34.