The performance of DST-Wavelet feature extraction for guitar chord recognition

Linggo Sumarno[1]*

[1]Electrical Engineering Study Program, Sanata Dharma University, Yogyakarta, Indonesia

Abstract. Small systems can be designed to be more energy-efficient compared to larger systems. On small systems, the need for data processing with small data sizes becomes a necessity. In the context of small systems for guitar chord recognition, there are indications that further efforts can be made to reduce the size of feature extraction data. This paper introduces DST (Discrete Sine Transform)-Wavelet feature extraction to achieve this reduction. Basically, this work evaluated the frame blocking length, the number of DST cutting factors, and the type of wavelet filters (Daubechies and biorthogonal families) to obtain the optimal number of feature extraction data. Based on the evaluation, the optimal result obtained was a number of four feature extraction data. This optimal result was obtained by using a frame blocking length of 512 points, a DST cutting factor of 0.5, and a biorthogonal 3.3 wavelet filter. Testing with 140 test chords using these four feature extraction data could give an accuracy of up to 92.86%.

1 Introduction

The increase of the primary energy consumption, in the long run, could increase the environmental deterioration [1]. Therefore, efforts are needed to reduce energy consumption. One way to achieve this reduction is by using low-power devices. Small systems are particularly suitable for low-power devices, as they can be designed to be more energy-efficient compared to larger systems. However, one limitation of small systems is the small data size in data processing.

A guitar chord recognition system can be developed on a small system using an FPGA (Field Programmable Gate Array) [2-3]. This small system benefits from the small data size in the data processing. One of the data components that can be reduced in chord recognition is feature extraction data. One of the methods for this feature extraction is based on PCP (Pitch Class Profile) [4], and its derivatives [5-7], which can give 12 feature extraction data. Another method for this feature extraction is based on segment averaging [8-9], which could give eight [8] and six [9] feature extraction data respectively. Yet another method for this feature extraction is based on MFCC (Mel Frequency Cepstral Coefficients) [10-11], which can give 13 feature extraction data. Meanwhile, the recent method that is based on MFCC [12] can give four feature data extraction. However, by using this four feature extraction data, the performance accuracy of this MFCC based feature extraction is still below 90%.

This paper introduces a combination of DST and wavelet for feature extraction, in order to further reduce the size of feature extraction data. Even though common DST and wavelet are used in this paper, this combination of DST and wavelet has never been used before.

2 Methodology

2.1 Development of recognition system

A recognition system has been developed in order to perform the research, as shown in Fig. 1. This system was implemented using Python software. A more detailed description of the recognition system is written below.

* Corresponding author: lingsum@usd.ac.id

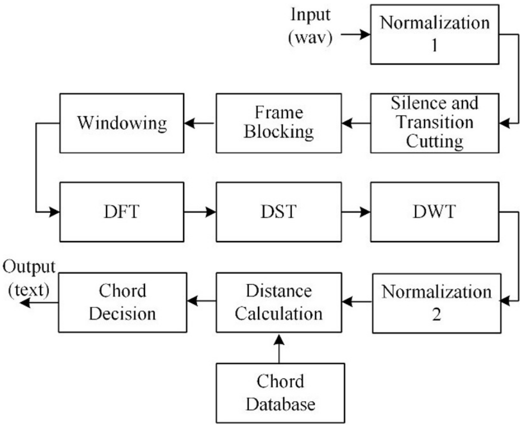

Fig. 1. The recognition system in this work

2.1.1 Input

The system input is two-second signal data recorded in WAV format from seven guitar chords (C, D, E, F, G, A, and B). Visual observation of the signal data, showed that a two-second recording is sufficient to obtain the steady state part of the signal data. In this case, the chord information is basically in this steady state part. In addition, recording was carried out at a sampling frequency of 5 kHz. This sampling frequency meets Shannon's sampling theorem [13], which state that the sampling frequency must be equal to or greater than twice the highest frequency component of the signal. In this work, the highest frequency component is 392 Hz from the G4 note, which is a subset of the G chord. As a note, the guitar used in this recording is a Yamaha CPX-500-II.

2.1.2 Normalization 1

Normalization 1 is the first normalization in the recognition system. This normalization is an attempt to adjust the maximum value of the input signal data to either 1 or -1. This adjustment is necessary because the maximum value can vary due to the chord recording process.

2.1.3 Silence and transition cutting

Silence and transition cutting is process to remove the silence and transition parts of the input signal data. This cutting is carried out because there is no chord information in the silence and transition parts. Based on the observation results, the silence part can be cut using a threshold value of 0.5. Subsequently, the transition part can be cut by removing the first 200 milliseconds of the signal data.

2.1.4 Frame blocking

Frame blocking is a process to cut a short signal data from the longer one [14]. This cutting is done at the left part of the signal data. Basically, this short signal data length (frame blocking length) will affect the resolution of the signal data resulting from the DFT process. Signal resolution that is too low or too high will have a negative effect on the discrimination level of feature extraction, which will finally reduce the accuracy. For this reason, in this work, frame blocking lengths of 128, 256, 512, 1024, and 2048 points will be evaluated.

2.1.5 Windowing

Windowing is a process used to minimize the left and right edges of the signal data sequence. If these edges are not minimized, it will cause spectral leakage to appear in the signal data resulting from the DFT process. This spectral leakage will give rise to other frequencies that are not related to the frequencies in the chord. For this reason, the appearance of spectral leakage needs to be minimized using a window. This work made use of the Hamming window, which is a window that is widely utilized in the field of signal processing [15].

2.1.6 DFT

DFT (Discrete Fourier Transform) is a transformation that converts a finite sequence of data signal into a finite sequence of complex values of data signal in the complex frequency domain. DFT of N length data signal can be expressed as follows.

$$Y_k = \sum_{n=0}^{N-1} y_n e^{-\frac{2\pi ikn}{N}} \text{ ,for } 0 \le k \le N-1 \qquad (1)$$

This work used the magnitude of the complex values of data signal. Since the sequence of the magnitude of the complex values of data signal is symmetric, only half of the left side was used.

2.1.7 DST

DST is a transformation related to the Fourier transform. This transformation is similar to the DFT above. DST uses the sine function, while DFT uses the sine and cosine functions (which are expressed in complex exponential form). DST of N length data can be expressed as follows.

$$Y_k = 2\sum_{n=0}^{N-1} y_n \sin\left[\frac{\pi(k+1)(2n+1)}{2N}\right] \text{, for } k = 0,\dots,N-1 \quad (2)$$

In this work DST was used for data compression. For this reason the output of DST process will be evaluated by using a number of cutting factors 0.25, 0.5, and 1. In this case the output of DST process will be taken from the leftmost data to 0.25 x N, 0.5 x N, and 1 x N as the right most data.

2.1.8 DWT

DWT (Discrete Wavelet Transform) is a transformation that decomposes signal data into a number of signal data that have a certain resolution and frequency bandwith. As an illustration, a decomposition method used in this work is shown in Fig. 2.

As shown in Fig. 2, H_0 and H_1 respectively indicate filtering with Low Pass Filter and High Pass Filter which comes from the wavelet filter used. In this work, the wavelet filter used will be evaluated from a number of existing wavelet filters, namely, Daubechies 1, 2, 3, 4, 5, and 6, as well as biorthogonal 1.1, 1.3, 1.5, 2.2, 2.4, 2.6, 3.1, 3.3, and 3.5.

For the filtering mentioned above, the filtering mode with periodization is used. This filtering mode will produce the same output signal data length as the input signal data length. The ↓2 sign in Fig. 2 indicates downsampling by a factor of 2. Additionally in Fig. 2 also, c_3 indicates the input signal data, while c_0, d_0, d_1, and d_2 indicate the DWT output signal data with a certain resolution and frequency field.

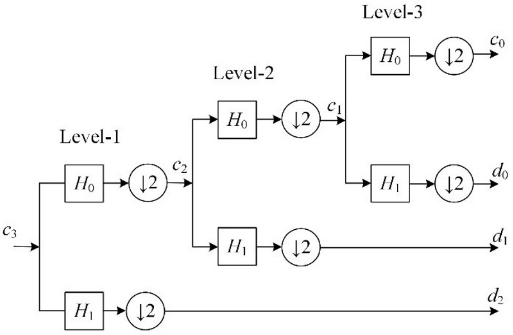

Fig. 2. Three-level decomposition of DWT.

This work used a decomposition level that depends on the length of the input signal data. For a signal data of length *N*, the decomposition level (L_d) can be expressed as follows.

$$L_d = log_2(N) \qquad (3)$$

As an illustration, as shown in Fig. 1, decomposition level up to level-3. Thus the L_d value is 3, which is related to the length of the signal data from the 8 point DWT input. In this work, the length of the DWT input signal data will vary, depending on the length of the blocking frame described above.

In this work, not all of the output from the DWT process would be used. For example, from Fig. 2 above, with a DWT input signal data length of 8 points, the decomposition results c_0, d_0, d_1, and d_2 will be obtained, which are illustrated in Fig. 3.

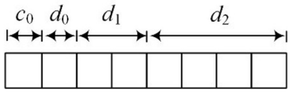

Fig. 3. Decomposition results from DWT in Fig. 2, for 8 point DWT input.

From Fig. 3, not all decomposition data c_0, d_0, d_1, and d_2 are taken as output of the DWT process. In this case, what is taken is the data on the far left (c_0) then the data is scanned to the right until the amount of data that will be used.

In this work, the number of feature extraction data to be evaluated was only from 1-8. This is related to the aim of this work, which is to find the minimum possible number of feature extraction data, less than eight, which is still acceptable.

2.1.9 Normalization 2

Normalization 2 is the second normalization in the recognition system. This normalization is an attempt to adjust the maximum value of the DWT output signal data to either 1 or -1. The classification algorithms will benefit from this normalization [16]. As a note, the output data from the Normalization 2 process is called feature extraction from the input signal data.

2.1.10 Distance calculation, chord database, and chord decision

Distance calculation and chord database indicate a classification method that uses the template matching method [17-19]. The chord database contains a set of chord feature extraction references from the chords used in this work. In more detail, in the chord database there are seven chord feature extraction references, each of which represents the seven chords used in this work.

To create the chord database above, feature extraction was carried out from 10 samples for each chord. As a note, this feature extraction is taken from the output of the Normalization 2 process above. Then the average of the feature extraction from the 10 samples is calculated. The average results become the chord feature extraction reference. Because in this work there are seven chords, in the chord database there are seven chord feature extraction references.

Distance calculations in this work used the cosine distance function. This distance function is a form of distance expression from cosine similarity. This similarity has been widely used in calculating similarity values [20-21]. The results of this distance calculation are seven distance values. These values are the result of distance calculations from feature extraction of the input signal data and the seven chord feature extraction references in the chord database.

Chord decision is determining the output chord related to the input signal data. In this case, from the seven distance values above, the smallest distance of the seven distance values is then sought. A chord associated with the smallest distance is then determined as the output chord.

2.2 Testing

To test the system, 20 other samples were taken for each chord. Because in this work there are seven chords, a total of 140 chords are used to test the system.

3 Performance testing and results

Performance testing was performed by simultaneously evaluating variations in frame blocking length parameters, DST cutting factor, wavelet filter, and the number of feature extraction data. In more detail, the variations in these parameters have been described in the Research Methodology section above.

In general, the results of performance testing are optimal results obtained when using frame blocking length: 512 points, DST cutting factor: 0.5, wavelet filter: biorthogonal 3.3, and the number of feature extraction data: 4. Then, in more detail, the results of Performance testing is shown partially in Tables 1, 2, and 3.

Table 1. Testing results of the introduced feature extraction using a DST cutting factor of 0.5 and a biorthogonal 3.3 wavelet filter. Results shown: Accuracy (%).

Frame blocking length (points)	**Number of feature extraction data**							
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
128	14.29	46.43	40.71	52.86	62.86	61.43	73.57	78.57
256	14.29	42.14	67.14	73.57	76.43	82.14	88.57	87.14
512	14.29	55.00	83.57	**92.86**	92.86	91.43	93.57	97.14
1024	14.29	32.14	45.00	58.57	71.43	79.29	81.43	84.29
2048	14.29	36.43	52.86	52.86	57.14	62.86	71.43	75.71

Table 2. Testing results of the introduced feature extraction using a frame blocking length of 512 points and a biorthogonal 3.3 wavelet filter. Results shown: Accuracy (%).

DST cutting factor	**Number of feature extraction data**							
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
1.0	14.29	45.00	60.00	75.71	81.43	90.00	90.00	90.00
0.5	14.29	55.00	83.57	**92.86**	92.86	91.43	93.57	97.14
0.25	14.29	53.57	77.14	80.00	82.86	91.43	92.86	91.43

Table 3. Testing results of the introduced feature extraction using a frame blocking length of 512 points, and a DST cutting factor of 0.5. Results shown: Accuracy (%).

Wavelet filter	**Number of feature extraction data**							
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Daubechies 1	14.29	72.14	80.71	87.14	89.29	89.29	92.86	94.29
Daubechies 2	14.29	52.14	51.43	65.00	84.29	95.00	95.71	96.43
Daubechies 3	14.29	45.00	30.00	50.00	68.57	75.71	94.29	94.29
Daubechies 4	14.29	46.43	67.14	63.57	54.29	75.00	87.14	87.14
Daubechies 5	14.29	40.71	35.71	35.71	56.43	91.43	95.71	95.71
Daubechies 6	14.29	42.86	71.43	72.14	54.29	67.14	67.14	67.14
Biorthogonal 1.1	14.29	72.14	80.71	87.14	89.29	89.29	92.86	94.29
Biorthogonal 1.3	14.29	39.29	72.86	74.29	76.43	74.29	85.71	87.86
Biorthogonal 1.5	14.29	30.00	70.71	63.57	68.57	75.71	87.14	90.71
Biorthogonal 2.2	14.29	47.14	54.29	79.29	77.14	78.57	88.57	92.14
Biorthogonal 2.4	14.29	51.43	53.57	72.86	72.86	79.29	87.86	91.43
Biorthogonal 2.6	14.29	47.14	51.43	65.71	71.43	67.14	84.29	86.43
Biorthogonal 3.1	14.29	47.14	62.14	72.86	84.29	87.14	92.14	90.71
Biorthogonal 3.3	14.29	55	83.57	**92.86**	92.86	91.43	93.57	97.14
Biorthogonal 3.5	14.29	51.43	78.57	84.29	82.14	85.71	88.57	93.57

4 Discussion

Looking at the length of the frame blocking and accuracy, Table 1 indicates that if the length of the fram blocking is too short or too long it will reduce accuracy. This is because, if the frame blocking length is too short, it will cause the signal data resulting from the DFT process to be less detailed (because the signal resolution is too low). On the other hand, if the frame blocking length is too long, the signal data resulting from the DFT process will have too much detail (because the signal resolution is too high). Signal data that lacks detail or too much detail will reduce the discrimination level of feature extraction, which will finally reduce accuracy.

Viewed from the perspective of DST cutting factor and accuracy, Table 2 indicates that if the DST cutting factor is too small or large it will reduce accuracy. This is because, if the DST cutting factor is too small, it will cause the signal data resulting from the DST process to have a frequency bandwith that is too narrow. On the other hand, if the DST cutting factor is too large, the signal data resulting from the DST process will have a frequency bandwith that is too wide. Signal data whose frequency bandwith is too narrow or too wide will reduce the discrimination level of feature extraction, which will finally reduce accuracy.

Table 4. Performace comparison of feature extraction methods

Feature Extraction Methods	**Number of Feature Extraction Data**	**Accuracy (%)**	**Number of Test Chords**
Improved PCP [3]	12	95.83	192
CRP Enhanced PCP [2]	12	99.96	4608
Segment averaging with SHPS and logarithmic scaling [8]	8	100	140
Segment averaging and subsampling [9]	6	91.43	140
MFCC [12]	4	89.29	140
DST-Wavelet (this work)	4	92.86	140

Note: The table above only displays the minimal number of feature extraction data required to achieve an accuracy above 89%.

Viewed from the wavelet filter and accuracy side, Table 3 indicates that if you use a suitable wavelet function, it will provide optimal accuracy. This is related to how well the combination of LPF and HPF of the wavelet function used can separate low frequency and high frequency fields, in multi-resolution conditions. If the separation is appropriate, it will produce an optimal level of feature extraction discrimination, which will also produce optimal accuracy.

Looking at the number of feature extraction data, Tables 1, 2, and 3 indicate that in general, if the amount of feature extraction data increases (from 1 to 8), the tendency of accuracy is increasing. This means, up to 8 feature extraction data, the data is still classified as essential features. By using these essential features we can reveal the essential information from the input data. This essential information firstly, refers to the most relevance and discriminative characteristics in the feature extraction data. Secondly, this essential information enables the separation of classes [22]. In this work, if the essential features increase (from 1 to 8), it will further increase the discrimination level of feature extraction, which in turn will increase accuracy.

5 Performance comparison

Performance comparison of the introduced feature extraction methods with others is shown in Table 4. As indicated in Table 4, the introduced feature extraction method can be said to be the most efficient. The recognition system only needs four feature extraction data, in order to give an accuracy of up to 92.86%.

6 Conclusion and further research

This paper introduces a combination of DST and wavelet for feature extraction, for use in guitar chord recognition. This kind of combination could give performance accuracy up to 92.86% by using the number of feature extraction data 4. This performance can be obtained by using a frame blocking length of 512 points, a DST cutting factor of 0.5, and a biorthogonal 3.3 wavelet filter. For further research, other feature extraction methods can be explored. In this case by using four or less feature extraction data, the recognition system can give a higher accuracy.

References

1. E.A. Agbede, Y. Bani, W.N.W. Azman-Saini, N.A.M Naseem, The impact of energy consumption on environmental quality: empirical evidence from the MINT countries, Environ Sci Pollut Res 28(38), ,54117-54136, (2021)
2. K. Vaca, M. M. Jefferies, X. Yang, An Open Audio Processing Platform with Zync FPGA, in Proceeding of 2019 22nd IEEE Int. Symp. Meas. Control Robot. Robot. Benefit Humanit. ISMCR 2019, D1-2-1-D1-2–6, (2019)
3. K. Vaca, A. Gajjar, X. Yang, Real-Time Automatic Music Transcription (AMT) with Zync FPGA, in Proceeding of IEEE Comput. Soc. Annu. Symp. VLSI, ISVLSI, vol. 2019-July (2019), 378–384, (2019)
4. T. Fujishima, Realtime Chord Recognition of Musical Sound: A System Using Common Lisp Music, in ICMC Proceedings, 9(6), 464–467, (1999)
5. P. Rajpurkar, B. Girardeau, T. Migimatsu, A Supervised Approach To Musical Chord Recognition, Stanford Undergrad. Res. J. 15, 36–40, (2015)
6. K. Ma, Automatic Chord Recognition, Personal Project, 2016. http://pages.cs.wisc.edu/~kma/projects.html (accessed Mar. 12, 2023).
7. E. Demirel, B. Bozkurt, X. Serra, Automatic chord-scale recognition using harmonic pitch class profiles, in Proceeding of Sound Music Computing Conference 2019, 72–79, (2019)
8. L. Sumarno, Chord recognition using segment averaging feature extraction with simplified harmonic product spectrum and logarithmic scaling, Int. J. Electr. Eng. Informatics 10(4), 753–764 (2018)
9. L. Sumarno, Chord Recognition using FFT Based Segment Averaging and Subsampling Feature Extraction, in Proceeding of 2020 8th International Conference on Information and Communication Technology, ICoICT 2020, 465–469,(2020)
10. L. Ivanov, J. Dunn, Automatic Identification of Guitar Types from Prerecorded Audio, in Proceeding of 33rd FLAIRS Conference 2020, 308–311, (2020)
11. D.A. Talavera, E.S.C. Nase, L.D. Pancho, A.L. Ilao, Transcription of guitar chords from acoustic audio, J. Adv. Inf. Technol. 11(3), 149–154 (2020)
12. L Sumarno, The Performance of MFCC Feature Extraction for Guitar Chord Recognition, in Proceedings of the International Conference on Information Technology and Digital Applications 2021 (ICITDA 2021), 020012-1 - 020012-8 (Published 2023)
13. L. Tan, J. Jiang, Digital Signal Processing Fundamentals and Applications, 3rd ed. (Elsevier Inc., Oxford , 2019).
14. O.K. Hamid, Frame Blocking and Windowing Speech Signal, J. Information, Commun. Intell. Syst. 4(5), 87–94 (2018).
15. H. Rakshit, M. A. Ullah, A comparative study on window functions for designing efficient FIR filter, in Proceeding of 2014 9th Int. Forum Strateg. Technol. IFOST 2014, July 2014, 91–96, (2014)

16. M. Voelsen, D.L. Torres, R.Q. Feitosa, F. Rottensteiner, C. Heipke, Investigations on feature similarity and the impact of training data for land cover classification, ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci. 5(3), 181–189 (2021)
17. I. Izonin, R. Tkachenko, N. Shakhovska, B. Ilchyshyn, K.K. Singh, A Two-Step Data Normalization Approach for Improving Classification Accuracy in the Medical Diagnosis Domain, Mathematics 2022, 10(11), 1942, (2022)
18. A.K. Jain, R.P.W. Duin, J. Mao, Statistical pattern recognition: A review, IEEE Trans. Pattern Anal. Mach. Intell. 22(1), 4–37 (2000)
19. A. Massari, R.W. Clayton, M. Kohler, Damage detection by template matching of scattered waves, Bull. Seismol. Soc. Am. 108(5), 2556–2564 (2018)
20. H.U. Zhi-Qiang, Z. Jia-Qi, W. Xin, L.I.U. Zi-Wei, L.I.U. Yong, Improved algorithm of DTW in speech recognition, in Proceeding of IOP Conf. Ser. Mater. Sci. Eng., 563(5), (2019),
21. S. Sohangir, D. Wang, Improved sqrt-cosine similarity measurement, J. Big Data 4, 1 (2017).
22. H.R. Shahdoosti, F. Mirzapour, Spectral–spatial feature extraction using orthogonal linear discriminant analysis for classification of hyperspectral data, Eur. J. Remote Sens., 50(1), (2017)