# Music Tempo Classification Using Audio Spectrum Centroid, Audio Spectrum Flatness, and Audio Spectrum Spread based on MPEG-7 Audio Features

Alvin Lazaro, Riyanarto Sarno, Johanes Andre R., Muhammad Nezar Mahardika
Department of Informatics Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
lazaro.alvin13@mhs.if.its.ac.id, riyanarto@if.its.ac.id, johanes.andre13@mhs.if.its.ac.id, nezarmahardika1@mail.com

***Abstract*—Music has become an integral part in human life. Recent studies show that music can affect human's mood. For example, music with slow tempo will cause the listener feel relaxed. Meanwhile, music with fast tempo will cause the listener feel excited. This paper discusses about music tempo classification using features from MPEG-7 based on Support Vector Machine (SVM). MPEG-7 is international standardized multimedia metadata in ISO/IEC 15938. The audio features used in this experiment are Audio Spectrum Centroid, Audio Spectrum Flatness, and Audio Spectrum Spread. Features are classified using SVM. The result of this operation is a classification of music based on its beats-per-minute (BPM). The classification rate of the experiment is 80%.***

***Keywords—tempo classification; MPEG-7; SVM; BPM***

## I. INTRODUCTION

Music has become an integral part in human life. Music can affect human's mood. For example, slow music will cause the listener feel relaxed. Meanwhile, fast music will cause the listener feel excited. This kind of feelings are the examples of how music affect the listeners' mood.

Today, people might want to classify their music playlist based on the tempo of the music. There are three types of music tempo slow, medium, and fast. These are based on the beats-per-minute (BPM) value of the music. In this paper, we are demonstrating a method to classify a music tempo based on its BPM value.

We are using music features from MPEG-7 to classify the tempo of the music. MPEG-7 is a multimedia metadata which are standardized metadata in ISO/IEC 15938 [1]. From that, an audio can get 17 Low-Level Descriptors [2]. For tempo classification, we will use three features from MPEG-7. They are audio spectrum centroid (ASC), audio spectrum flatness (ASF), and audio spectrum spread (ASS) [2].

Actually, a lot of people were doing experiments with music tempo classification. In this paper, we decided to take three previous works in music tempo classification as the examples. The first previous work is classification of music tempo by Yu-Yao Chang and Yao-Chung Lin in 2005 [3]. The goal of Chang and Lin's work is training a system which could make a prediction and learning the speed classes of songs. Also, this system will have different speed sensitiveness. The tempo is classified as slow, medium, and fast. The result of this project can be used for retrieving music contents and giving recommendation of music. The term "tempo" in this work represents the number of BPM of a song. On the other hand, the term "speed" represents how humans feel the speed of a song. In order to bring out the soft tempo information, Chang and Lin used Inter-Onset Interval (IOI) feature in their SVM process. Each IOI candidates will be evaluated using onset detection algorithm. Then, SVM is applied to the detection result for data training. The system successfully classifies songs with slow and medium tempo with high accuracy (85%). On the other hand, the model can't classify songs with fast tempo accurately (only 59%).

The second work is analyzing tempo and beat from acoustic music signal by Eric D. Scheirer in 1998 [4]. Scheirer analyzed the musical signal using filter bank and comb filter [4]. For beat detection, Scheirer used beat resonator and tempo analysis [4]. The algorithm's performance is identical with human's performance in many musical aspects. Though this is quite good, this work still has some errors. It is unable to understand the beat relationships in music at diverse tempo. For example, a human listener understands how eighth-note patterns group in order to form quarter-note and half-note patterns. In the system, this process is done implicitly in the resonators because of the phase-locking at harmonic ratios. To make the algorithm more robust, the writer suggests to have an explicit model of rhythmic grouping of beat relationships.

The third work is detecting onset in music audio signal by Stephen Hainsworth and Malcolm Macleod in 2003 [5]. This work adopted onset detection methods. In this work, Hainsworth and Macleod discussed several steps for differencing spectral values several steps are discussed for spectral recognition. Hainsworth and Macleod use methods such as Euclidean distance, Kullback-Liebler distance, and Foote distance to find the change point from the unsmoothed measurement functions. These methods were used to choose the highest point from the unsmoothed measurement functions. At the end, Hainsworth and Macleod concluded that this technique performed very well with low computational cost and algorithmic complexity.

The purpose of our paper is to introduce our experiment in music tempo classification based on low-level descriptor from MPEG-7. We're using three features from MPEG-7. They are

978-1-5090-5864-8/17/$31.00 ©2017 IEEE 

ASC, ASF, and ASS. We hope that our method will give high accuracy in classifying music tempo.

## II. Feature Extraction

In MPEG-7, there are three features which reflect the tempo of a music. They are ASC, ASF, and ASS. Therefore, we are using these features as the indicator to classify the tempo of music.

### A. Audio Spectrum Centroid (ASC)

Audio spectrum centroid (ASC) explains log-frequency's center of force of a power spectrum [2]. The shape of the power spectrum is described as spectrum centroid. ASC gives an indication of the domination of low or high frequencies in a power spectrum.

$$C = \sum_n log_2\,(f(n)/1000)P'_x(n) \;/\; \sum_n P'_x(n) \qquad (1)$$

Fig. 1, is an example of signal plot of audio spectrum centroid.

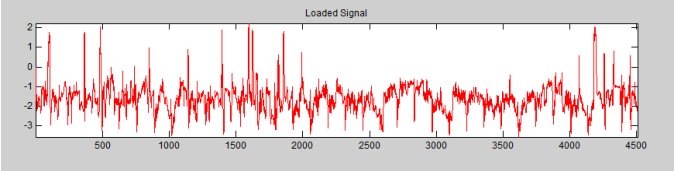


Fig. 1. Plot of audio spectrum centroid feature.

### B. Audio Spectrum Flatness (ASF)

Audio spectrum flatness (ASF) defines the planeness properties from an audio signal's spectrum [2]. The planeness properties are defined by a given number of frequency bands [2]. ASF shows how the power spectrum of a signal deviates from a frequency of a flat shape. The term "flat shape" describe the noise or impulse in a signal.

$$SFM_b = \frac{\sqrt[ih(b)-il(b)+1]{\prod_{i=il(b)}^{ih(b)} c(i)}}{\frac{1}{ih(b)-il(b)+1}\sum_{i=il(b)}^{ih(b)} c(i)} \qquad (2)$$

Fig. 2, is an example of signal plot of audio spectrum flatness.

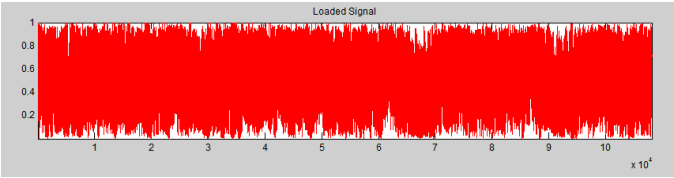


Fig. 2. Plot of audio spectrum flatness feature.

### C. Audio Spectrum Spread (ASS)

Audio spectrum spread (ASS) describes how the log-frequency spread from a power spectrum [2]. ASS grants its user to differentiate noise-like and tone-like of a signal.

$$S = \sqrt{\sum_n ((log_2\left(\frac{f(n)}{1000}\right) - c)^2\, P'_x(n)) \;/\; \sum_n P'_x(n)} \qquad (3)$$

Fig. 3, is an example of signal plot of audio spectrum spread.

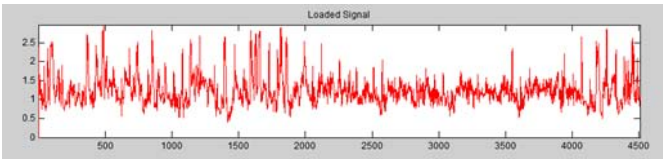


Fig. 3. Plot of audio spectrum spread feature.

### D. Beats-per-minute (BPM)

Beats-per-minute (BPM) is a unit to scale the tempo in music and heart rate [6].

### E. Wavelet

Wavelet can be used for analyzing different frequency elements from a signal (in this case, music signal). In our experiment, we use wavelets for noise reduction on the music signal. In this experiment, we are using bior 2.8 wavelet. We analyzed and calculate using Information Quality Ratio (IQR) [7].

$$wt(s,\tau) = < x, \psi_{s,\tau} > = \frac{1}{\sqrt{s}} \int_{\infty}^{\infty} x(t)\, \psi * \left(\frac{t-\tau}{s}\right) dt \qquad (4)$$

The scaling parameter is represented by s > 0. It determines the resolutions of time and frequency from the scaled base wavelet. We use a low pass on wavelet (approximation coefficients) for processing.

### F. Fast Fourier Transform (FFT)

Fast Fourier Transform (FFT) is a method to change the domain of a signal. It will change the signal from time domain into frequency domain [8]. FFT obtains the information contained when the signal rise or fall. It is also used for searching the value and a maximum value at each frequency of a signal.

### G. Support Vector Machine (SVM)

SVM is a machine-learning method that works by looking for the best hyperlane (for classification) that separates the different labels. Optimum hyperlane can be found by measuring the margin / distance between hyperlane with data closest to each label.

Mostly, all classification process use SVM. For example, recognizing a person's hand-writing, identifying the sound of a person, recognizing objects, detecting human face, and oral classification. It is a classifier for multi-dimensional data which is used to determine boundary curve between two domains effectively.

SVM will find the criterions of the judgment function through the training examples. Then, the training examples are classified into two classes. SVM will maximize the margin between these classes during its learning phase. After that, SVM is able to recognize the information patterns from the subject.

We are using SVM in this experiment because the previous experiments employed SVM for sound recognition

[10] using MPEG-7 features and gave good results. For the graph, we are using ASF as the Y-axis. Meanwhile, ASC and ASS are used as the X-axis. The reason of using these two features in the same axis is because both of them are almost similar. On the other hand, the audio spectrum flatness is very different from these two features.

### *H. Music Tempo Classification*

The tempo of music is classified by its BPM value [6]. For example, 60 BPM means one beat every second. Usually, the tempo of a music is written at the beginning of the music. This is called as a metronome marking [6]. Classical music use Italian terminology to indicate its tempo and stylistic feel.

## III. MUSIC TEMPO CLASSIFICATION

### *A. Datasets*

For audio datasets, we are using 1000 songs database. Those songs are labelled with valence and arousal score. For classifying the tempo, we are using three types of tempo, slow, medium, and fast. In the experiment, we are using 65 data training and 30 samples. Each samples contain 10 slow music,

10 medium music, and 10 fast music. The BPM value is obtained from MixMeister BPM Analyzer, a special software to obtain the BPM value of a music.

### *B. Define BPM Threshold*

At the beginning, we need to obtain the BPM value of the music. We are using a software, MixMeister BPM Analyzer, to obtain the BPM value. After that, we set the minimum BPM value for each type of tempo.

TABLE I. BPM THRESHOLD

| BPM Value | Tempo |
|---|---|
| 0 – 100 | Slow |
| 101 – 135 | Medium |
| Above 135 | Fast |

### *C. Decompose Level Wavelets*

From the extraction of ASC, ASF, and ASS, we perform Discrete Wavelet Transform (DWT) [11]. The goal of this activity is to eliminate noise in the signal without changing the original information signal [12-15]. Then we make an analysis to find the best wavelet decomposition level.

The first step is converting the domain of the signal by using FFT [11]. The purpose of this activity is to determine the information contained in the signal frequency. Then do the calculations on equation (5) to get the maximum value of the index and a signal feature.

$$[maxvalue, indexmax] = \max(abs(FFT(Feature - mean(Feature)))) \quad (5)$$

After both results are obtained, then the value is used to find the frequency range according to Table III-2 [7]. This reference table for sampling only with 1024 Hz. Different sampling values will lead to differences in the frequency range.

For searching frequency range, we perform a calculation using the Formula (6), where Fs is 1024 (sample frequency) and L is the length of the signal.

$$Fh = indexmax * Fs/L \quad (6)$$

The results of the calculation Fh is the frequency range for Table III. Then, we can find the best wavelet decomposition level for a feature.

TABLE II. DECOMPOSITION RANGE

| Decomposition Level (L) | Scope of Frequency (Hz) |
|---|---|
| 1 | 256-512 |
| 2 | 128-256 |
| 3 | 64-128 |
| 4 | 32-64 |
| 5 | 16-32 |
| 6 | 8-16 |
| 7 | 4-8 |
| 8 | 2-4 |
| 9 | 1-2 |
| 10 | 0.5-1 |
| 11 | 0.25-0.5 |
| 12 | 0.125-0.25 |
| 13 | 0.0625-0.125 |

For the type of wavelet, we use bior 2.8. Wavelet types have been based on the calculation of the best Information Quality Ratio (IQR). Based on information theory, the relationship between two variables can be measured by mutual information (MI). Mutual information quantifies how good DWT with particular MWT can reconstruct original signal x(t) [7]. The IQR value of (x(t)) and reestablish signal (y(t)) can be declared as expected value of mutual information (7).

$$IQR(x(t), y(t)) = \frac{\sum_{xi \in x(t)} \sum_{yj \in y(t)} p(x_i, y_j) log_2(p(x_i)p(y_j))}{\sum_{xi \in x(t)} \sum_{yj \in y(t)} p(x_i, y_j) log_2(p(x_i, y_j))} - 1 \quad (7)$$

The xi and yj are particular value of x(t) and y(t) respectively. Meanwhile, p(xi) and p(yj) are the marginal probability. The P(xi,yi) is the combination of probability value from xi and yj. Naturally, the range of this ratio is $0 \leq IQR \leq 1$. The biggest value (IQR=1) can be reached if DWT can perfectly reconstruct a signal without loss of information.

We can see how the wavelet remove the noises in the signal in Fig. 3. In this example, we are using Audio Power feature plot. The plot is processed through wavelet to gain the best level of

decomposition. The above (red graph) is the raw signal (Frequency range (Hz)) and the below (blue graph) is the signal after being processed through wavelet. Fig. 4, Fig. 5, Fig. 6, Fig. 7, Fig. 8, Fig. 9, and Fig. 10 are the examples.

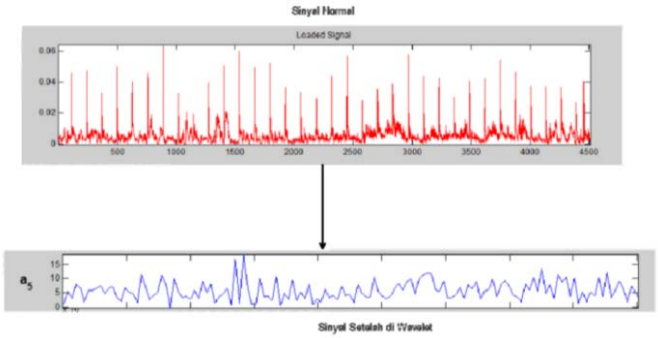


Fig. 4. Example of wave process in audio power feature.

Here are the features of ASC, ASF, and ASS after being processed through wavelet

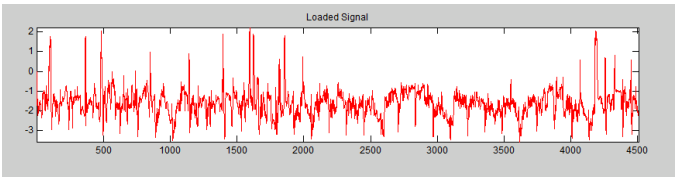


Fig. 5. Audio spectrum centroid feature.

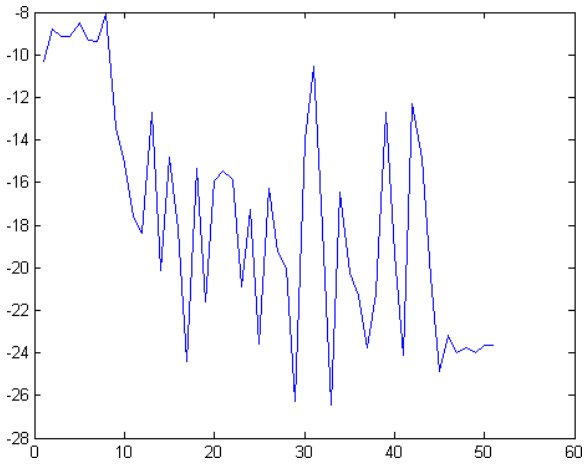

Fig. 6. Audio spectrum centroid feature after wavelet process.

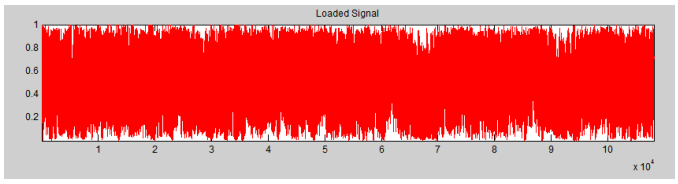


Fig. 7. Audio spectrum flatness feature.

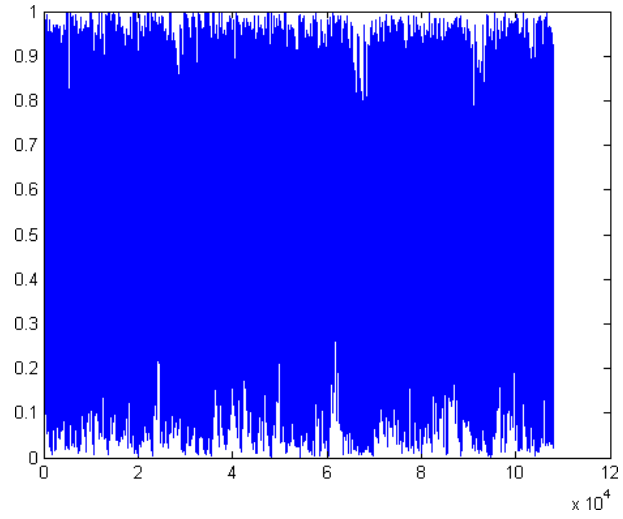

Fig. 8. Audio spectrum flatness feature after wavelet process.

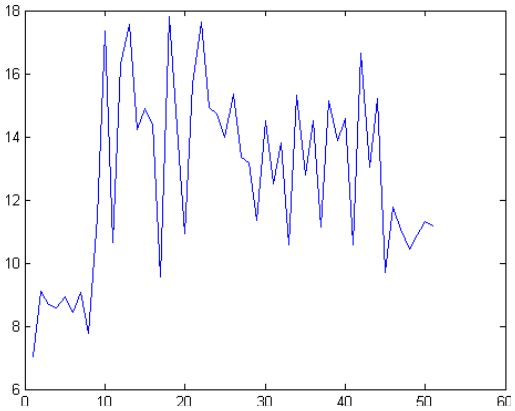

Fig. 9. Audio spectrum spread feature.

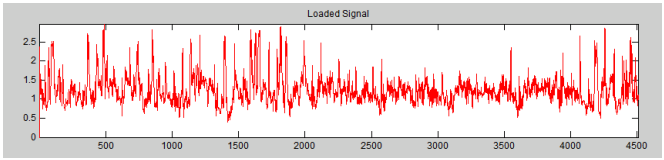


Fig. 10. Audio spectrum spread feature after wavelet process.

## D. *Data Training*

In this activity, we are doing data training before proceeding into testing stage. There are 65 data involved in this activity. Each song is labelled with valence and arousal score. The data training consists of three types of tempo, slow, medium, and fast.

## E. *Combining Features*

After the wavelet stage, we are combining these three features to form a new feature. The new feature, example A, is a combination of ASC, ASF, and ASS [16]. The result will be used for data training. As shown in Fig. 11.

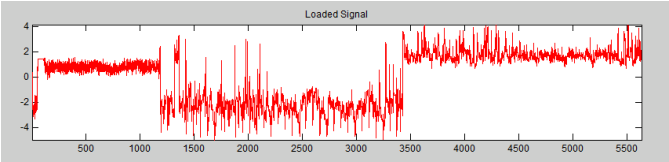


Fig. 11. Signal plot audio spectrum centroid, audio spectrum flatness, and audio spectrum spread.

### *F. System Architecture*

This is the system architecture for our experiment. There is the scheme of feature extraction process, perform wavelet, combine features until training data. Scheme below is the flow of data from the testing. The result is a tempo label (slow / medium / fast) of a song. As shown in Fig. 12.

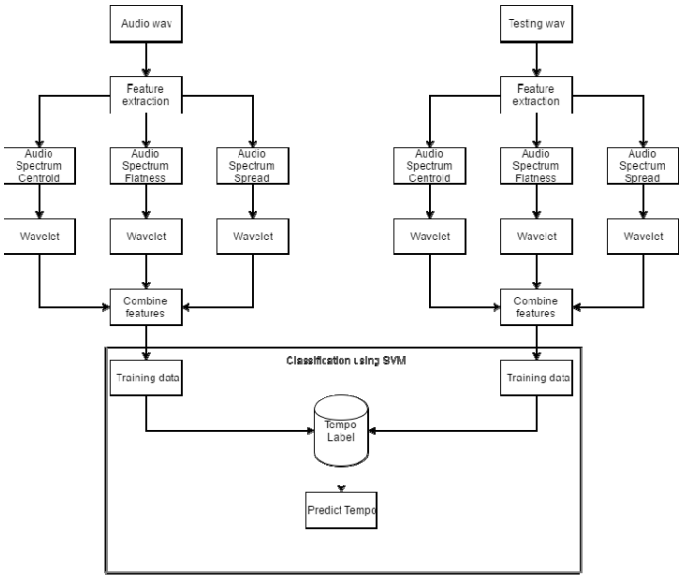


Fig. 12. System Architecture.

## IV. Experiment Result

As we mentioned in the previous chapter, we are using 30 samples for this experiment. These samples consist of 10 slow music, 10 medium music, and 10 fast music. In general, the testing result is quite accurate.

### *A. Testing Tempo Classification*

Here is the dataset used for the testing.

TABLE III. Tempo Classification Result

| | | Testing | | | Accuracy |
|---|---|---|---|---|---|
| | | ***Slow*** | ***Medium*** | ***Fast*** | |
| **Actual** | ***Slow*** | 9 | 1 | 0 | 90% |
| | ***Medium*** | 0 | 10 | 0 | 100% |
| | ***Fast*** | 0 | 5 | 5 | 50% |

There are 30 data used in the testing. These data consist of 10 slow music, 10 medium music, and 10 fast music. Some fast music (5 data) are classified as medium music. Meanwhile, there is one slow music identified as medium music. The accuracy of our system when detecting slow music is 90%. The accuracy of our system when detecting medium music is 100%. And the accuracy of our system when detecting fast music is 50%. The errors are most likely caused by the BPM value of the datasets. Some fast music has BPM value which are almost similar to the BPM value of medium music. The same thing also happened with the slow music. This problem van be fixed by re-adjusting the BPM threshold of our system. Or, we can use more data during the data training activity.

### *B. Accuracy*

Based on the experiment result, we can generate the overall accuracy of our classification. There are 30 datasets and 24 of them are identified correctly. To such a degree, the overall accuracy of our experiment is 80% as shown in Fig. 13.

| | |
|---|---|
| Datasets | = 30 |
| True | = 24 |
| Accuracy | = (True / Datasets) x 100% = 80% |

Fig. 13. Accuracy.

## V. Conculsion

This paper presents a method of classifying music by its tempo. Our method use three features from MPEG-7. They are Audio Spectrum Centroid (ASC), Audio Spectrum Flatness (ASF), and Audio Spectrum Spread (ASS). We found that BPM value holds an important role in this experiment. The goal of this experiment is to classify a music tempo based on its BPM value.

Based from the result of our experiment, the classification rate is best for slow and medium tempo. There are some flaws found during the experiment. There are some fast music identified as medium music. And also, there is one slow music identified as medium music. The classification rate of the experiment is about 80%. For future work, we are considering to improve the detection accuracy of our system. The focus is to increase the classification of music with fast tempo.

## Acknowledgement

Thank to Institut Teknologi Sepuluh Nopember (ITS) for supporting our research.

## References

[1] H.-G. Kim, N. Moreau and T. Sikora, MPEG-7 Audio and Beyond, Chichester: John Wiley & Sons, 2006. https://doi.org/10.1002/0470093366.fmatter.

[2] “ISO/IEC 15938-4:2002(en), Information technology — Multimedia content description interface — Part 4: Audio.” [Online]. Available: https://www.iso.org/obp/ui/#iso:std:iso-iec:15938:-4:ed-1:v1:en. [Accessed: 19-Dec-2016]. https://doi.org/10.3403/bsisoiec15938.

[3] C. Yu-Yao and L. Yao-Chung, “Music Tempo (Speed) Classification”, in Stanford University. Stanford, California 94305, 2005.

[4] E.D. Scheirer, “Tempo and Beat Analysis of Acoustic Musical Signals”, in Machine Listening Group, E15-401D MIT Media Laboratory, Cambridge, Massachusetts 02139, 1998. https://doi.org/10.1121/1.421129.

[5] H. Stephen and M. Malcolm, “Onset Detection in Musical Audio Signals”, Cambridge University Engineering Department, Cambridge, CB2 1PZ, UK, 2003.

[6] BBC, “Music, Rhythm and Metre”, Available: http://www.bbc.co.uk/schools/gcsebitesize/music/elements_of_music/rhythm_metre1.shtml. [Accessed: 15-Dec-2016].

[7] D.R. Wijaya, R. Sarno and E. Zulaika, Information Quality Ratio as a novel metric for mother wavelet selection," *Chemom. Intell. Lab. Syst*. https://doi.org/10.1016/j.chemolab.2016.11.012.

[8] S. LI, H. LI and L. MA, "Music genre classification based on MPEG-7 audio features," in Proceedings of the Second International Conference on Internet Multimedia, ACM, 2010, pp. 185-188. https://doi.org/10.1145/1937728.1937772.

[9] G. Carpentier, "Information technology—Multimedia content description interface—Part 4: Audio, AMENDMENT 2: High-level descriptors. In Motion Picture Expert Group (ISO/IEC JTC 1 SC29.," July 2005. [Online]. https://doi.org/10.3403/bsisoiec15938.

[10] C.-H. Lin, M.-C. Tu, Y.-H. Chin, W.-J. Liao, C.-S. Hsu, S.-H. Lin, J.-C. Wang and J.-F. Wang, "SVM-Based Sound Classification Based on MPEG-7," International Conference on Hybrid Information Technology, pp. 536-543, August 2012.R. Nicole, “Title of paper with only first word capitalized,” J. Name Stand. Abbrev., in press. https://doi.org/10.1007/978-3-642-32692-9_67

[11] R.X. Gao and R. Yan, Wavelets: Theory and applications for manufacturing, Springer Science & Business Media, 2010. https://doi.org/10.1007/978-1-4419-1545-0

[12] M. N. Munawar, R. Sarno, D. A. Asfani, T. Igasaki, and B. T. Nugraha, “Significant Preprocesing Method In Eeg-Based Emotions Classification,” J. Theor. Appl. Inf. Technol., vol. 87, no. 2, Jun. 2016.

[13] R. Sarno, B.T. Nugraha, M.N. Munawar, R. Sarno, B.T. Nugraha, and M.N. Munawar, ‘Real Time Fatigue-Driver Detection from Electroencephalography Using Emotiv EPOC+’, *Int. Rev. Comput. Softw. IRECOS*, vol. 11, no. 3, pp. 214–223, Mar. 2016. https://doi.org/10.15866/irecos.v11i3.8562

[14] B.T. Nugraha, R. Sarno, D. Anton Asfani, T. Igasaki, and M. Nadzeri Munawar, “Classification of Driver Fatigue State Based on Eeg Using Emotiv EPOC+,” J. Theor. Appl. Inf. Technol., vol. 86, no. 3, Apr. 2016.

[15] R. Sarno, M.N. Munawar, B.T. Nugraha, R. Sarno, M.N. Munawar, and B.T. Nugraha, ‘Real-Time Electroencephalography-Based Emotion Recognition System’, *Int. Rev. Comput. Softw. IRECOS*, vol. 11, no. 5, pp. 456–465, May 2016. https://doi.org/10.15866/irecos.v11i5.9334

[16] K. Adistambha, M. Doeller, R. Tous, M. Gruhne and M. Sano, The MPEG-7 Query Format: A New Standard in Progress for Multimedia Query by Content, University of Wollongong Research Online, 2007. https://doi.org/10.1109/iscit.2007.4392066