Music Mood Classification Using Audio Power and Audio Harmonicity Based on MPEG-7 Audio Features and Support Vector Machine

Johanes Andre Ridoean, Riyanarto Sarno, Dwi Sunaryo
Department of Informatics
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
johanes.andre13@mhs.if.its.ac.id, riyanarto@if.its.ac.id, dwi@if.its.ac.id

Dedy Rahman Wijaya
Department of Informatics, Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
School of Applied Science, Telkom University
Bandung, Indonesia
dedyrw@tass.telkomuniversity.ac.id

Abstract*—Music can affect a person's mood. Music psychologists agree that music has a significant impact on a person's mood that determines their behavior. Therefore, our research examines the audio features that affect mood. Our method is to perform feature extraction based on MPEG-7 Low-Level Descriptors. MPEG-7 is international standardized multimedia metadata in ISO/IEC 15938. In this paper, we have made a researched about music mood classification using Audio Power and Audio Harmonicity features. The result of the extraction of the MPEG-7 obtained 17 features low-level descriptors. These features are classified using Support Vector Machine (SVM). There are two stages of SVM: training and prediction phase. Traning phase is when the machine learns to recognize the characteristics of the signal on a label while in prediction phase, it gives the predicted outcome of a label on a new signal characteristic pattern. The success rate of this experiment was 74.28% using Audio Power and Audio Harmonicity, 37.14% using Audio Spectrum Projection, and 28.57% using Audio Power, Audio Harmonicity and Audio Spectrum Projection.

Keywords—music mood classification; MPEG-7; SVM

I. Introduction

Music became popular through purchases and downloads from the internet. Most people do not know their favorite musical mood. Nowadays, users expect more semantic metadata to archive music, such as similarity, style, and mood.

Mood greatly affects a person's behavior. When someone is not in a good mood, usually they tend too lazy to do anything. On the other hand, when they are in a good mood, they will produce the best quality of work, behavior, and decision making. For example, nowadays a lot of young people likes to listen to their favorite music while they are learning something. They say they will concentrate more on doing the task if they listen to music. Background music in the classroom can have a positive effect on student focus [1]. So, a music can act affect a person's mood. Music psychologists also agree that music could be a good stimulus to improve someone's mood. That is why a music which being heard by someone can affects his/her mood.

Mood and emotion have some minor differences. Emotion is the result intense feelings directed at someone or something while the mood is a growing feeling and less intense due to lack of stimulus [2]. The mood lasted longer than emotion. Music is a good stimulus to stimulate a person's mood. Several experiments have been performed to measure the effect of music on mood.

Many experiments have been performed, but no one has used the MPEG-7 feature yet. One of them is Music Mood for Bollywood Music [3]. For mood label, they use happy, sad, exciting, and silent. JAudio is used for feature extraction method. The features are timbre features, intensity features, and rhythm feature. Features are generated in the form of a value of calculation. The result of features extraction becomes input for Random Forest classification and the K-Means algorithm is used to group music into four groups: happy, sad, exciting, and silent. The classification accuracy was 70%.

In this paper, we develop mood classifier model using Audio Power and Audio Harmonicity based on MPEG-7 Low Audio Descriptors. MPEG-7 is a multimedia metadata which is standardized metadata in ISO/IEC 15938 [4].

The discussion in this paper is organized as follows: Section II discusses the materials and methods that used the experiments. Section III discusses the results of experiments and various comparative experiments. Finally, we give the conclusion and future work in Section IV.

II. Material and Methods

From feature extraction, audio gets 17 Low-Level Descriptors [5]. Each feature describes the characteristics of a respective signal. Based on these descriptors, it is possible to identical or similar audio content including mood detection for characteristics signal.

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

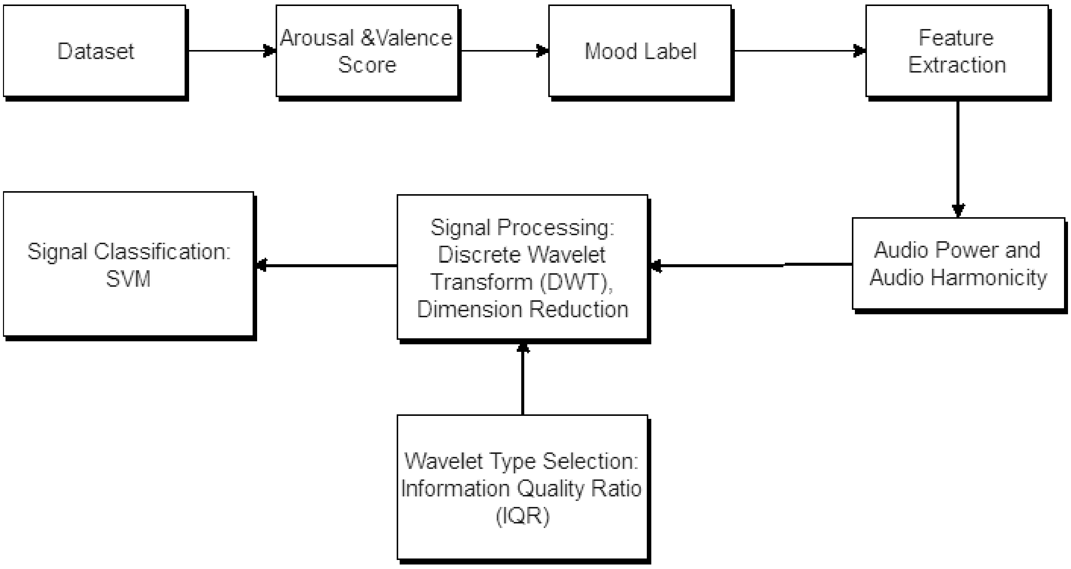
Dataset
Arousal &Valence Score
Mood Label
Feature Extraction
Audio Power and Audio Harmonicity
Signal Processing: Discrete Wavelet Transform (DWT), Dimension Reduction
Signal Classification: SVM
Wavelet Type Selection: Information Quality Ratio (IQR)

Fig. 1. General step by step music mood classification.

The mood on a music influenced by the harmony of tone and rhythm of the music [6]. In MPEG-7, features that reflect the rhythm and harmony of tone are Audio Power and Audio Harmonicity. Therefore, general step by step music mood classification Fig. 1, we use Audio Power and Audio Harmonicity as an indicator of the determination of the following:

A. Audio Power (AP) [5]

Audio Power feature represents the temporally smoothed instantaneous power (square of waveform values). To get the value of the AP can use (1).

$$AP(l) = \frac{1}{N_{hop}} \sum_{n=0}^{N_{hop}-1} \left|s(n+lN_{hop})\right|^2 \; (0 \leq l \leq L-1) \quad (1)$$

Where L is total number of time frames, *s(n)* is the average square waveform, l is the index frame and N_{hop} is number of time samples between two successive frames. Fig. 2, is an example of a signal plot Audio Power.

B. Audio Harmonicity(AH) [5]

Audio Harmonicity feature represents the degree of harmonicity of an audio signal. The purpose is to use size harmonicity to make the distinction between voice has a harmonic spectrum and non-harmonic spectrum. Fig. 3, is an example of a signal plot Audio Harmonicity.

C. Discrete Wavelets Transform

From the extraction of Audio Power and Audio Harmonicity, we perform Discrete Wavelet Transform (DWT) [7]. The goal is to eliminate noise which contained in the signal without changing the original information signal [8-10]. Then we have done an analysis to find the best wavelet decomposition level.

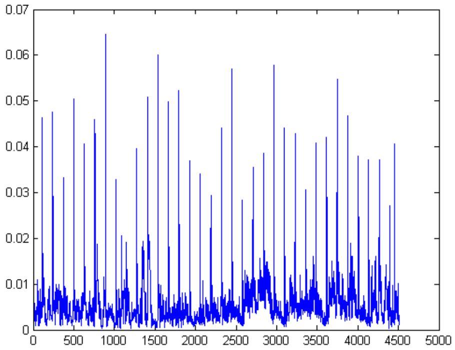

Fig. 2. Plot audio power feature.

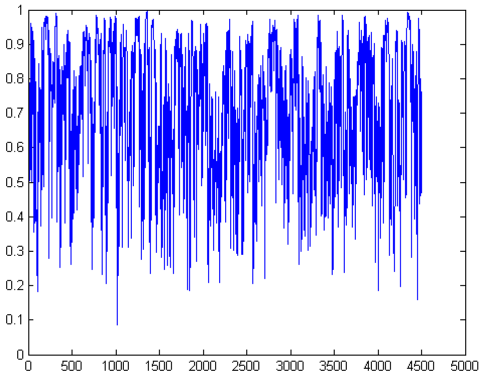

Fig. 3. Plot audio harmonicity feature.

The first step is to convert the signal from the time domain to the frequency domain by Fast Fourier Transform (FFT) [7].

The goal is to determine the clarity of the information contained in the signal frequency. Then do the calculations on (2) to get the maximum value of the index a signal.

$$[\max value, index\ max] = max\left(abs\left(FFT\left(S - mean(S)\right)\right)\right) \qquad (2)$$

Where S is a feature of audio power and audio harmonicity. After both the results obtained, then the value is used to find the frequency range according to Table I [11]. The rule for determining level of decomposition in Table I can be expressed by the following (3) [7]:

$$\frac{f_q}{2^N + 1} \leq f_{char} \leq \frac{f_q}{2^N} \qquad (3)$$

Where f_q is the sampling frequency, f_{char} is the dominant frequency, and N is decomposition level. This reference table for sampling only with 1.024 Hz. Different sampling values will lead to differences in the frequency range. To search for the frequency range, perform calculations using (4).

TABLE I. DECOMPOSITION LEVEL TABLE

Decomposition level (L)	Frequency range (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

$$Fh = [index\ max] * Fs/L \qquad (4)$$

Where Fs is 1.024 (sample frequency) and L is the length of the signal. The results of the calculation Fh is the frequency range for Table I. Then, we have found the best wavelet decomposition level for a feature.

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 Mother Wavelet (MWT) can reconstruct original signal $x(t)$ [11]. IQR of original signal $(x(t))$ and reconstructed signal $(y(t))$ can be expressed as expected value of mutual information. For the calculation IQR can use (5).

$$IQR\left(x(t), y(t)\right) = \frac{\sum_{x_i \in x(t)} \sum_{y_j \in y(t)} p\left(x_i, y_j\right) log_2\left(p(x_i) p(y_j)\right)}{\sum_{x_i \in x(t)} \sum_{y_j \in y(t)} p\left(x_i, y_j\right) log_2\left(p\left(x_i, y_j\right)\right)} - 1 \qquad (5)$$

Where x_i and y_j are particular value of $x(t)$ and $y(t)$ respectively. $p(x_i)$ and $p(y_j)$ are the marginal probability and $P(x_i, y_i)$ is joint probability of x_i and y_j. 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.

Fig. 4, and Fig. 5, is an example approximate value plot Audio Power has been done wavelet bior 2.8 with the best level decomposition. From the plot demonstrates that the signals are decomposed with the best level does not eliminate the authenticity of the information.

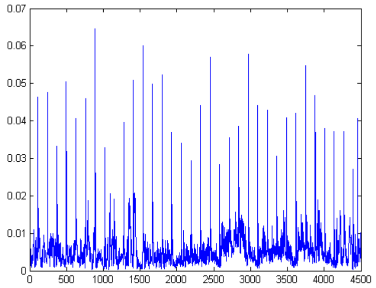

Fig. 4. Audio power after dwt.

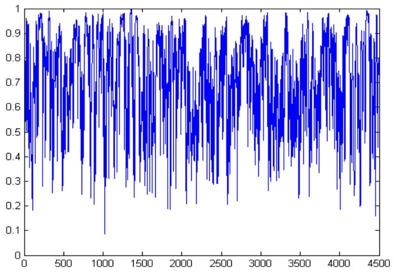

Fig. 5. Audio harmonicity after dwt.

D. Datasets

Audio datasets have been obtained from 1000 songs database which labeled with valence and arousal score [12-13]. The dataset is composed of pieces in 45 seconds with the format .mp3. To obtain a good result, we converted from .mp3 into a .wav before feature extraction. Each music has a score of valence and arousal. The valence and arousal score between 1 to 9.

For mood's label, there are happy, relaxed, sad, and angry (Russel's Diagram) [14]. We redraw again and divided based on arousal and valence score. Fig. 6, is the distribution curve mood.

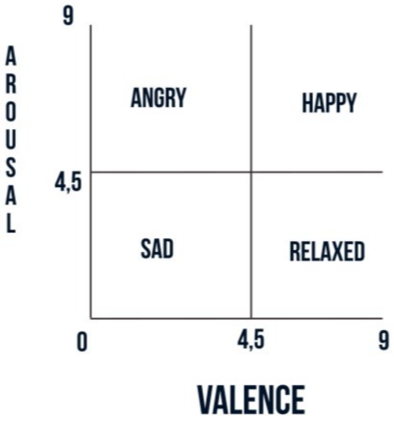

Fig. 6. Russel's Diagram.

Then we mapped mood's label based on arousal and valence score. The limit for determining high valence or high arousal is greater than or equal to 4.5.

E. Support Vector Machine (SVM)

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

SVM has been used in a variety tasks such as the classification of speaker identification, object recognition, face detection and classification of the vowel. SVM can classify multi-dimensional data which basically determines the frontier between the two classes [4].

SVM training examples define the parameters of the decision to classify the functions of two or more classes and maximize margins during the learning phase. After learning, the classification of unknown pattern can be estimated and being able to predict the data that has a new pattern.

The advantages classification using SVM are a unique solution, not sensitive to small changes of parameters, and providing increased performance compared to other algorithms.

Our experiments using SVM because the previous experiment which employs SVM for sound recognition using MPEG-7 features gave good results [15].

III. Result and Discussion

A. Dimension Reduction

Because the feature extraction results that we obtained are the form of signals with different length, we need to do a long equalization signal. The difference is due to the signal size of MPEG-7 feature extraction also consider milliseconds. Therefore, we have done an analysis for the size of a feature-length Audio Power and Audio Harmonicity after being DWT.

The result is the minimum length of features Audio Power is 4.498 and Audio Harmonicity is 4.493. Then the length of the signals are taken as that number, the rest can be ignored. It will not eliminate the information contained in the signal because the difference is only milliseconds.

B. Combining Features

The two signals extraction results will be stored in a single list. List divided by the length of data in which 0- 4498 is to feature Audio Power, and the rest is for the column Audio Harmonicity. When the classification process, a comparison is done between the same features. Fig. 7, is an example of a list that holds the characteristics that have divided the two signals per a predetermined length feature.

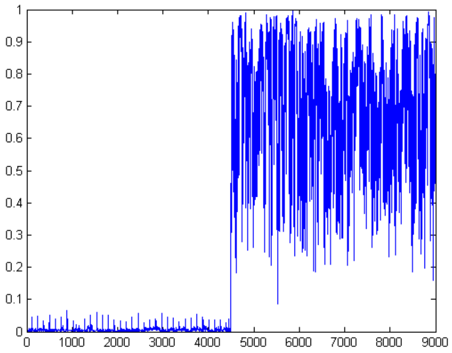

Fig. 7. List of audio power and audio harmonicity.

C. Data Traning

In this experiment, we are doing training for machine learning. There are 65 data trains for each mood. Each mood is labeled based on valence and arousal score from the dataset. Table II shows a list of the amount of training data with each label.

TABLE II. Number of Training Data

Mood's label	Number Of Training Data
Angry	65
Happy	65
Relaxed	65
Sad	65

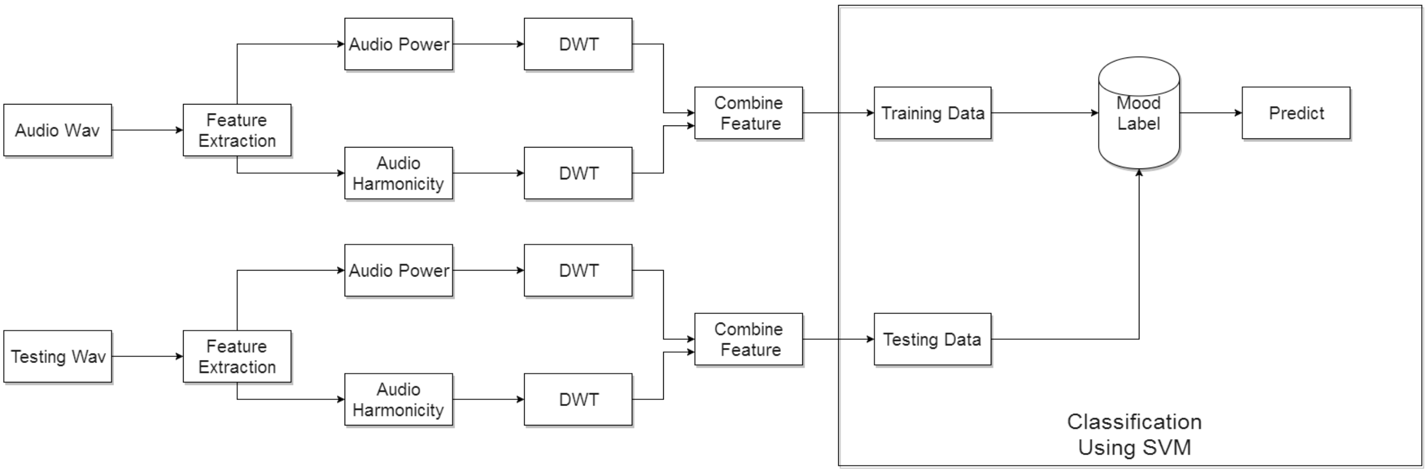

Fig. 8. Music mood classification.

D. Testing Scenario

As the descriptions above, our system architecture will be shown in Fig. 8. The details will be explained below:

1) Schemes above:

a) Audio extracted based on MPEG-7 features, got Audio Power and Audio Harmonicity.

b) Apply steps DWT for each feature.

c) Combine the two feature in a list.

d) Perform training data every mood label 65 times.

e) The machine will learn the characteristics of the signal every mood's label.

2) Schemes under:

a) Audio extracted based on MPEG-7 features, got Audio Power and Audio Harmonicity.

b) Apply steps DWT for each feature.

c) Combine the two feature in a list.

d) Perform prediction data by a trained machine.

e) The machine will give a mood predicted results.

E. Results

Totally, we are using 35 samples. These samples consist of 10 for each mood's label, but only 5 for angry because angry's mood in datasets has only 70 samples. Accuracy results in this experiment are calculated by (6).

$$Accuracy = \frac{TRUE}{TOTAL\ DATA} * 100\% \qquad (6)$$

This study evaluated the performance of the feature set which comprises Audio Power and Audio Harmonicity. The experimental result is shown in Table III. Based on the result of this experiment, three best results on the mood's label are angry, happy, and sad with successful accuracy is 100%. Total the classification success rate is about 74.28 %.

TABLE III. THE DETAIL CLASSIFICATION RESULT USING AP AND AH

Testing / Actual	Angry	Happy	Relaxed	Sad
Angry	5	0	0	0
Happy	0	10	0	0
Relaxed	0	3	1	6
Sad	0	0	0	10

F. Comparison with Audio Spectrum Projection

Now, by the same method as above we have been experimenting with Audio Spectrum Projection. In the previous experiments, these features are used to classify music [4-5]. Fig. 9, is an example of a signal Audio Spectrum Projection that has been done DWT.

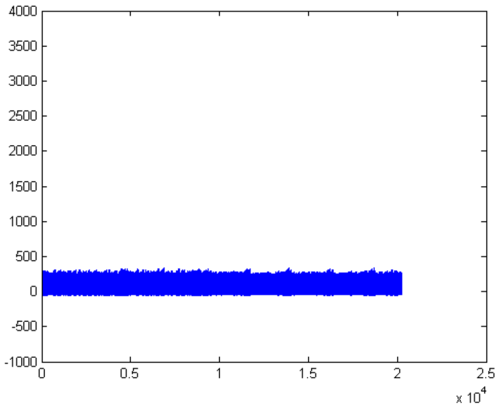

Fig. 9. Audio spectrum projection after DWT.

However, with the same method and same test data, the results using the audio spectrum projection is worse. The classification result using Audio Spectrum Projection only good for mood's relaxed. The experimental result is shown in Table

IV. On the result of this experiment, the best result on mood's label is relaxed where successfully accuracy is 70%. Total the classification success rate is about 37.14 %.

TABLE IV. THE DETAIL CLASSIFICATION RESULT USING ASP

Testing / Actual	Angry	Happy	Relaxed	Sad
Angry	1	4	0	0
Happy	0	5	4	1
Relaxed	0	2	7	1
Sad	0	4	6	0

We have tested classification using a combination of three feature: Audio Power, Audio Harmonicity, dan Audio Spectrum Projection. The experiment was performed with the same method and same test data, but the result using 3 combinations of features is getting worse. Accuracy value more decreases. Most of the music is detected as the happy's mood. The experimental result is shown in Table V. The classification success rate is about 28.57 %.

TABLE V. THE DETAIL CLASSIFICATION RESULT USING AP, AH, AND ASP

Testing / Actual	Angry	Happy	Relaxed	Sad
Angry	1	4	0	0
Happy	0	9	0	1
Relaxed	0	9	0	1
Sad	0	10	0	0

IV. CONCLUSION

Audio Power and Audio Harmonicity are the best MPEG-7 features for music mood classification. This is due to mood's label influenced by the audio power and harmony tone of the music. When the Audio Spectrum Projection features used for classification along with Audio Power and Audio Harmonicity, the results are getting worse. The success rate obtained is 74.28%. The accuracy of classifier using Audio Power and Audio Harmonicity features is the best for mood's label like angry, happy and sad. For future research, we will use another classification methods for improving accuracy.

ACKNOWLEDGEMENT

We would like to thank Institut Teknologi Sepuluh Nopember for supporting our research.

REFERENCES

[1] D. Strachan, "The Space Between the Notes: The Effects of Background Music on Student Focus," *Masters Arts Educ. Action Res. Pap.*, May 2015.

[2] D. Hume, *Emotion and Moods*. Organizational behavior, 2012.

[3] A. Ujlambkar, O. Upadhye, A. Deshpande, and G. Suryawanshi, "Mood Based Music Categorization System for Bollywood Music," *Int. J. Adv. Comput. Res.*, vol. 4, no. 1, Mar. 2014.

[4] H.-G. Kim, N. Moreau, and T. Sikora, *MPEG-7 Audio and Beyond: Audio Content Indexing and Retrieval*. John Wiley & Sons, 2005.

[5] ISO/IEC (2001), "Information Technology — Multimedia Content Description Interface — Part 4: Audio," *FDIS 15938-4:2001(E)*, June.

[6] Z.W. Ras and A. Wieczorkowska, *Advances in Music Information Retrieval*, 1st ed. Springer Publishing Company, Incorporated, 2010.

[7] R.X. Gao and R. Yan, *Wavelets: Theory and Applications for Manufacturing*, 2011 edition. New York; London: Springer, 2010.

[8] M.N. Munawar, R. Sarno, D.A. Asfani, T. Igasaki, and B.T. Nugraha, "Significant preprocessing method in EEG-Based emotions classification," *J. Theor. Appl. Inf. Technol.*, vol. 87, no. 2, pp. 176–190, May 2016.

[9] B.T. Nugraha, R. Sarno, D.A. Asfani, T. Igasaki, and M. N. Munawar, "Classification of driver fatigue state based on EEG using Emotiv EPOC+," *J. Theor. Appl. Inf. Technol.*, vol. 86, no. 3, pp. 347–359, Apr. 2016.

[10] 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

[11] D.R. Wijaya, R. Sarno, and E. Zulaika, "Information Quality Ratio as a novel metric for mother wavelet selection," *Chemom. Intell. Lab. Syst.*, vol. 160, pp. 59–71, Jan. 2017. http://dx.doi.org/10.1016/j.chemolab.2016.11.012

[12] M. Soleymani, M.N. Caro, E.M. Schmidt, C. Ya Sha, and Y.-H. Yang, "Emotion in Music Database - MediaEval 2013 - aka 1000 songs." [Online]. Available: http://cvml.unige.ch/databases/emoMusic/. [Accessed: 19-Dec-2016].

[13] M. Soleymani, M.N. Caro, E.M. Schmidt, C.-Y. Sha, and Y.-H. Yang, "1000 Songs for Emotional Analysis of Music," *Proc. ACM Int. Multimed. Conf. Exhib.*, vol. 6, no. 1, pp. 1–14, 2015.

[14] M. Nardelli, G. Valenza, A. Greco, A. Lanata, and E.P. Scilingo, "Recognizing Emotions Induced by Affective Sounds through Heart Rate Variability," *IEEE Trans. Affect. Comput.*, vol. 6, no. 4, pp. 385–394, Oct. 2015.

[15] C. Lin, M. Tu, Y. Chin, W. Liao, C. Hsu, S. Lin, J. Wang and J. Wang, "SVM-Based Sound Classification Based on MPEG-7 Audio LLDs and Related Enhanced Features," in *Convergence and Hybrid Information Technology*, 2012, pp. 536–543.