A MUSIC GENRE CLASSIFICATION USING MUSIC FEATURES AND NEURAL NETWORK

Ivanna K. Timotius [1], Matias H.W. Budhiantho [2], and Patrice Dwi Aryani [3]
[1,2,3] Department of Electronic Engineering, Satya Wacana Christian University
Diponegoro Street 52 – 60, Salatiga, Indonesia
ivanna_timotius@yahoo.com [1], mbudhiantho@yahoo.com [2], and miz_trice@yahoo.com [3]

ABSTRACT

Genre is a conventional way to classify music. This paper aims to implement an algorithm to classify music files into four different genre, which are classic, rock, pop, and dangdut. Dangdut is Indian influenced Indonesian popular music. An automatic music genre classification system will help music collectors to classify their music collection.

The genre classification algorithm is divided into music feature extractor and classifier. The feature extraction of the system is constructed by timbre feature extraction and rhythm feature extraction. The timbre feature extraction is done by mel frequency cepstral coefficients (MFCC). The rhythm feature is constructed by full wave rectification, low pass filtering, down sampling, mean removal, and autocorrelation. The classifier is formed by back propagation neural network (BPNN).

By using this algorithm, the average of accuracy is 80%, that is of 100% accuracy for classic, 70% accuracy for rock, 90% accuracy for pop, and 60% accuracy for dangdut.

Keywords: Genre, music, music features, MFCC, autocorrelation, BPNN.

1 INTRODUCTION

The word genre is taken from a Latin word genus, which means kind or class. A genre can be described as a type or category based on structure, theme, or functional criteria [1]. Music genre can be defined as music category having similar style or element. The similarity can be characterized from the music instruments that are used and the rhythm structure in the music [2]. An automatic music genre classification will help music collectors in classifying and searching their music collection.

The genre classification algorithm is constructed by music feature extractor and a classifier. The feature extraction of the system is constructed by timbre feature extraction and rhythm feature extraction. The timbre feature extraction is done by mel frequency cepstral coefficients (MFCC). The rhythm feature is done by full wave rectification, low pass filtering, down sampling, mean removal, and autocorrelation. The classifier is form by back propagation neural network (BPNN). This work is limited to four different music genres, those are classic, rock, pop, and dangdut.

2 MUSIC FEATURES

Music features that we use in this system are timbre features and rhythm features. These features are arranged into a music feature vector.

2.1 Timbre Features

Every music genre has its unique main musical instruments. Classical music is generally dominated by piano and violins. Pop music is generally dominated by electric guitars, drum, and keyboard. Rock music is generally dominated by electric guitars and drum. Dangdut music is dominated by a special drum. MFCC is used because of its ability in imitating human perception in hearing sound [3].

Cepstral coefficients are the result of Fourier transform inverse of the spectrum logarithm [4]. The cepstral coefficients are used since the cepstral analysis is good in separating the source signal and the filter impulse response.

Mel is a measure of pitch received by human hearing [5]. The human hearing perception followed a logarithmic scale. The mel scale is approximated using Eq. 1.

$$mel(f) = 1127.\log\left(1+\frac{f}{700}\right) \quad (1)$$

where f is frequency (Hz). Figure 1 shows the nonlinear relation between mel and frequency. The curve approximately linear at the frequency below 1000 Hz and logarithmic at frequency above 1000 Hz. The curve implied that the human perception for the frequency below 1000 Hz is linearly related to the actual frequency.

ISSN 2085-1944

Figure 2 explains the steps to calculate the MFCC. It start with a Hamming windowing, discrete Fourier transform (DFT), mel-filter bank, natural logarithm, and discrete cosine transform (DCT).

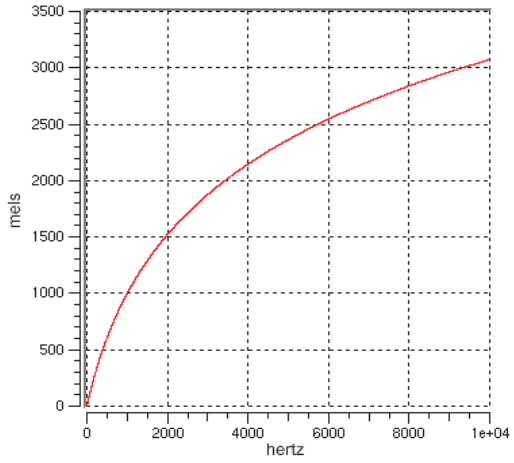


Figure 1. Mel scale

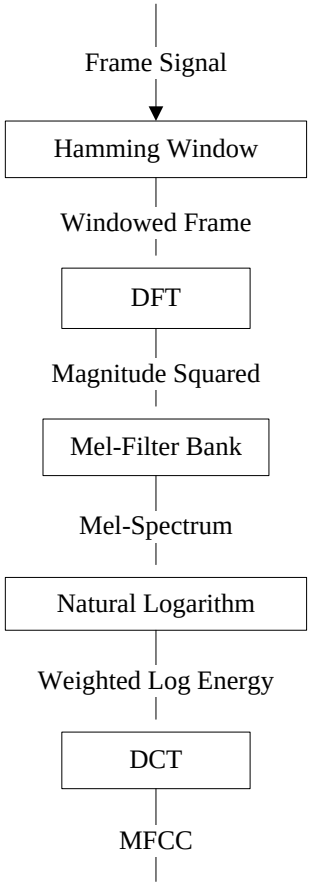


Figure 2. MFCC flowchart

The music signal taken from an audio file is a non-stationer signal [2][6]. Since the frequency analysis needs a stationer signal, the music signal input is divided into several frames. In dividing these frames, a windowing process is needed to minimized the discontinuity at the frame boundary [2]. We need overlapping Hamming windows since it has sufficient attenuation at the frame boundaries and simple equation [7]:

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

where N is the number of samples in a frame.

The human ears recognize timbre by a frequency domain analysis. For that reason, this system is developed in a frequency domain analysis using DFT. The phase information of the spectrum is not important for a music genre classification. Therefore, this system uses the energy information:

$$\left|X[k]\right|^2 = \left|DFT\left\{y[n]\right\}\right|^2 \qquad (3)$$

where $y[n]$ is the windowed signal and $|X[k]|^2$ is the spectral energy density of the signal.

We implemented mel filter bank of 40 overlapped triangle filters. The first 13 filters have an equal bandwidth, while the last 27 filters bandwidths increase logarithmically. The central frequency distance of the 13 first filters is 133.33 Hz. The central frequency distances of the 27 last filters are increased by the factor 1.0718 [2][8]. Mel spectrum is the spectrum that is filtered according to the human hearing perception:

$$Y_i = \sum_{k=0}^{N/2} \left|X[k]\right|^2 \cdot H_i(k) \qquad 1 \le n \le N \qquad (4)$$

where Y_i is the i^{th} mel spectrum, $H_i[k]$ is the i^{th} mel filter response.

Since the sound loudness received by human ear is logarithmic then:

$$\tilde{Y}_i = \ln(Y_i) \qquad i = 1, 2, \ldots, 40 \qquad (5)$$

where $\tilde{Y}_i$ is the weighted log energy.

DCT is performed to get the real value of the cepstrum:

$$c[n] = \sum_{i=1}^{40} \tilde{Y}_i \cdot \cos\left[n \cdot (i - 0.5) \cdot \frac{\pi}{40}\right]$$
$$n = 1, 2, \ldots, M \qquad (6)$$

where $c[n]$ is the n^{th} cepstral coefficient and M is the number of cepstral coefficient.

2.2 Rhythm Features

Rhythm is related to the appearance of regular accent in the music. Classic have a weak accent, pop have a medium and regular accent, rock

have a strong accent, and dangdut have a unique accent that sound like ‘dang’ and ‘dut’.

The music accent can be recognized by the signal magnitude [9]. Therefore, a full wave rectification of the framed signal, low pass filtering, down sampling, mean removal, and autocorrelation is needed:

$$x[f] = |x[n]| \tag{7}$$

where $x[n]$ is the time domain music signal in a frame and $x[f]$ is the full wave rectified signal.

The music rhythm is usually form by a relatively low frequency. Therefore, by a low pas filter we catched the signal that contained rhythm information:

$$x[l] = 0.01 \cdot x[f] + 0.99 \cdot x[l-1] \tag{8}$$

where $x[l]$ is low pas filtered output signal.

We then down sampling the signal to decrease the computation time. However, this down sampling process should not eliminate the global trend of the signal:

$$x[d] = x[k \times l] \tag{9}$$

where k is the down sampling scale. We removed the mean of the signal to make the detected rhythm more obvious:

$$x[m] = x[d] - E\{x[d]\} \tag{10}$$

where $E\{x[d]\}$ is the mean of $x[d]$.

Rhythm is a regular and repeated pattern in music. We applied autocorrelation to detect this regular pattern:

$$y[k] = \frac{1}{N}\sum_{m=1}^{N-k} x[m] \cdot x[m+k] \tag{11}$$

where $y[k]$ is the auto correlated signal.

The rhythm can be predicted from the auto correlated signal combination from each frame. The maximum peak of the combined signal shows the strong rhythm [9]. The important information from the maximum peak are the time information of the rhythm (showed by the period), the strength of the rhythm (showed by the magnitude), and the period ratio.

3 NEURAL NETWORK

A neural network is a mathematical model or computational model based on biological neural networks. Neural network attempt to use some organizational principles in a network of weighted directed graphs in which the nodes are artificial neurons and directed edges (with weights) are connections between neuron outputs and neuron inputs. Neural networks have the ability to learn complex nonlinear input-output relationships, use sequential training procedures, and adapt themselves to the data [10]. The discriminant function of a feed forward neural network with one hidden layer is [11]

$$g(\mathbf{x}) = f_o\left\{\sum_{j=1}^{n_H} w_{kj} f_h\left(\sum_{i=1}^{d} w_{ji} x_i + w_{j0}\right) + w_{k0}\right\} \tag{12}$$

where $\mathbf{x}$ is the feature vector, w_{kj} and w_{ij} are the weight factors, d is the length of feature vector, n_H is the number of hidden layer, f_h is the function related to the hidden layer, and f_o is the function related to the output layer. The function related to the hidden layer used in our designed system is a linear function:

$$f_h(x) = x \tag{13}$$

The function related to the output layer used in the classification phase is a hardlim function:

$$f_o(x) = \begin{cases} 1, & x > x_{th} \\ 0, & x < x_{th} \end{cases} \tag{14}$$

where x_{th} is the threshold value. In our designed system x_{th} is chosen 0.

We obtained the values of weight factors by conducting training phases. For our training phase we implemented a back propagation neural network. The function related to the output layer used in the training phase is sigmoid function:

$$f_o(x) = \frac{1}{1+\exp(-x)} \tag{15}$$

The training process starts with weight factors initialization. We could reformulate Eq. (12) into:

$$z\_in_j = \sum_{i=1}^{d} w_{ji} x_i + w_{j0} \tag{16}$$

$$z_j = f_h(z\_in_k) \tag{17}$$

$$y\_in_k = \sum_{j=1}^{n_H} w_{kj} z_j + w_{k0} \tag{18}$$

$$y_k = f_o\{y\_in_k\} \tag{19}$$

The output layer weight factor renewal is done by

$$\delta_k = (t_k - y_k) f_o'(y\_in_k) \tag{20}$$

$$\Delta w_{kj} = \alpha \delta_k z_j \tag{21}$$

$$w_{kj}(new) = w_{kj}(old) + \Delta w_{kj} \tag{22}$$

where t_k is the k^{th} target, and α is the learning acceleration value. The hidden layer weight factor renewal is done by

$$\delta\_in_j = \sum_{k=1}^{m} w_{jk} \delta_k \tag{23}$$

$$\delta_j = \delta\_in_j f_h'(z\_in_j) \tag{24}$$

$$\Delta w_{ji} = \alpha \delta_j x_i \tag{25}$$

$$w_{ji}(new) = w_{ji}(old) + \Delta w_{ji} \tag{26}$$

4 DESIGN, IMPLEMENTATION AND RESULT

The input of the system are mono music wav files of 44.1 kHz sampling period. The output of the system is the genre classification result. The training phase is done by a back propagation neural network (BPNN). Whereas, the classification phase is done by a feed forward neural network, using the same network obtained from the training phase.

The system takes only 3 seconds music wav file from the 60th second until the 63rd second. These timing is chosen since generally music files are already in their main part after 60 seconds that show their timbre and rhythm. The three seconds duration are assumed to be enough to obtain information on music features and can save the computational cost [9].

A music signal is considered stationer in 20-30 ms long [2]. In the system implementation, we used a 25 ms frame size. This frame size implies that the lowest frequency of interest is 40 Hz and the number of samples is 1102. Therefore, the Hamming overlapped low pass filter is implemented in 1102 samples.

Figure 3 and Figure 4 show the results of MFCC process. Figure 3 shows the mean of cepstral coefficients for each genre. The classic genre has the highest cepstral coefficient mean. The rock and dangdut genre have a high variation among their cepstral coefficient mean. The pop genre have the lowest variation among their cepstral coefficient mean. Figure 4 gives us the standard deviation of the cepstral coefficients. The rock genre have the lowest standard deviation of the cepstral coefficients. These mean and standard deviation is then combined into a music feature vector. The number of cepstral coefficients is empirically defined as 20.

The rhythm feature calculations after a autocorrelation process are shown in Figure 5. The peak shown in the figure determined the time and strength of the rhythm. Classic genre have the lowest rhythm strength. Dangdut genre have the most regular rhythm, since the time ratio between each peak is regular.

The input layer is determined according to the feature vector length, which is 60 layers. The hidden layer is determined empirically 50 layers. The output layer is 2 layers proportional to the number of genre classified. The learning acceleration value is chosen empirically 0.01.

The system is evaluated using 40 music files. Each genre is represented by 10 music files. The classification accuracy for classic genre is 100%, the classification accuracy for rock genre is 70%, the classification accuracy for pop genre is 90%, and the classification accuracy for the dangdut genre is 60%. The overall classification accuracy is 80%.

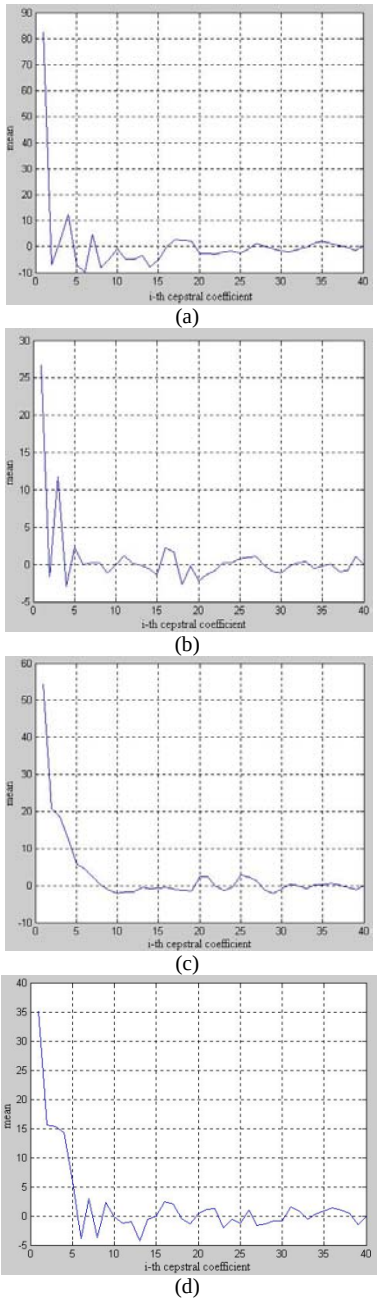
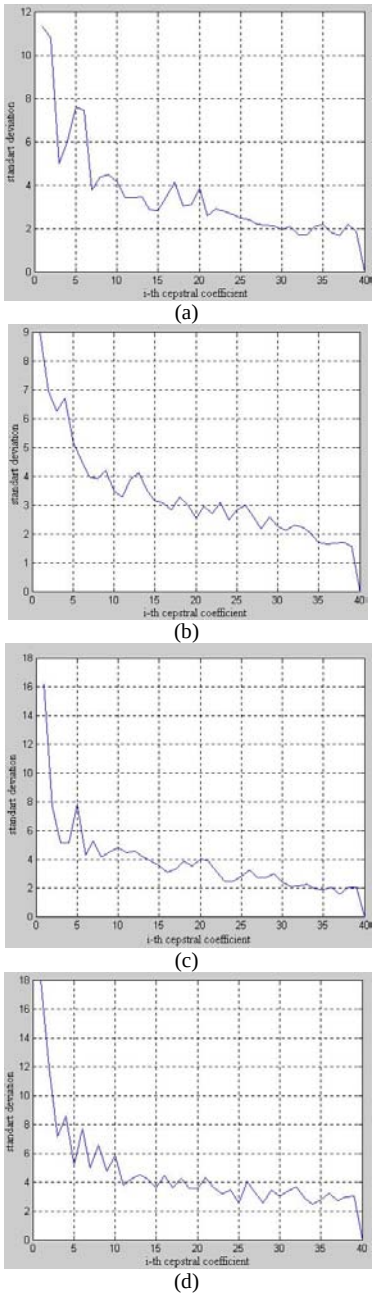
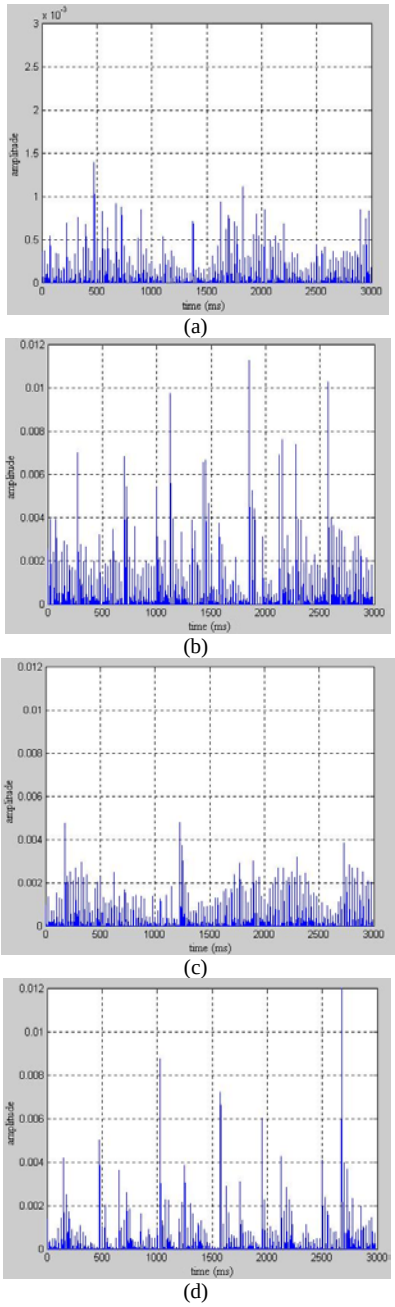


Figure 3. Mean of cepstral coefficients (a) classic (b) rock (c) pop (d) dangdut



(a) (b) (c) (d)

Figure 4. Standard deviation of ceptral coefficients (a) classic (b) rock (c) pop (d) dangdut



(a) (b) (c) (d)

Figure 5. Rhythm signal (a) classic (b) rock (c) pop (d) dangdut

5 CONCLUSION AND DISCUSSION

A music genre classification using MFCC, rhythm feature extraction and back propagation neural network gives an overall classification accuracy 80%. The method is good enough for real world application. In the future, a different architecture of the neural network or other classification methods might improve the classification accuracy.

REFERENCES

[1] Breure, and Leen (2001), Development of the Genre Concept [Online], Available at: http://www.cs.uu.nl/people/leen/GenreDev/GenreDevelopment.htm [Accessed: 2001].

[2] Kosina, and Karin (2002), Music Genre Recognition [Online], Available at: www.kyrah.net/mugrat/mugrat-slides.pdf [Accessed: 2002].

[3] Tzanetakis, George, and S. Lippens (2003), A comparison of Human and Automatic Musical Genre Classification.
[4] Wang, Basics of Signal Processing.
[5] Kamm, Terri, and Hynek Hermansky, Learning The Mel Scale and Optimal VTN Mapping [Online], Available at: www.clsp.jhu.edu/ws97/acoustic/reports/KHAMel.pdf.
[6] Kauchak, and Dave (2002), Automatic Musical Genre Classification of Audio Signals [Online], Available at: www.cse.ucsd.edu/classes/fa01/cse291/Audio.ppt [Accessed: 2002].
[7] DeFatta, and Lucas (1988), Digital Signal Processing: A Sistem Approach, Canada:John Wiley and Sons.
[8] Slaney, and Malcolm (1998), Auditory Toolbox Technical Report vs.2:pg29 [Online], Available at: http://rvl4.ecn.purdue.edu/~malcolm/interval/1998-010/AuditoryToolboxTechReport.pdf [Accessed: 2002].
[9] Tzanetakis, and George (2002), Musical Genre Classification of Audio Signal [Online], Available at: www.cs.uvic.ca/~gtzan/work/pubs/tsap02gtzan.pdf [Accessed: 2002].
[10] Jain A. K., Duin, R. P. W., and Mao J (2000) Statistical Pattern Recognition: A Review. IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 22, no. 1: 4-37.
[11] Duda R. O., Hart P. E., and Stork D. G. (2001), Pattern Classification, Wiley, New York.