International Journal of Engineering & Technology, 7 (4.11) (2018) 29-32

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper



Genre Classification of Traditional Malay Music Using Spectrogram Correlation

S. A. Samad*, A. B. Huddin

Centre for Integrated Systems Engineering and Advanced Technologies (INTEGRA), Faculty of Engineering and Built Environment, Universiti Kebangsaan Malaysia, Bangi, Selangor, Malaysia
**Corresponding author E-mail: salinasamad@ukm.edu.my*

Abstract

A method to classify the genre of traditional Malay music using spectrogram correlation is described. The method can be divided into three distinct parts consisting of spectrogram construction that retains the most salient feature of the music, template construction that takes into account the variations in music within a genre as well as the music progresses, and template matching based on spectrogram image cross-correlation with unconstrained minimum average correlation energy filters. Experiments conducted with seven genres of traditional Malay music show that the recognition accuracy is dependent on the number of segments used to construct the filter templates, which in turn is related to the length of music segment used. Despite using a small dataset, an average recognition rate of 61.8 percent was obtained for music segments lasting 180 seconds using six relatively short excerpts.

Keywords: *genre classification; traditional Malay music; spectrogram; unconstrained minimum average correlation energy filters.*

1. Introduction

Labels that are used to categorize and describe music according to style are known as genres. Some are culturally and historically derived, while newer labels may have been invented for marketing and commercial reasons. In this sense, music genres have no strict definitions and boundaries [1]. However, even with current labels, it is fairly obvious to the listeners that certain characteristics are shared within the members of a particular genre. This observation is also true of traditional Malay music.

There are many genres of traditional Malay music, some arguably are more popular than others. For example, genres such as *Dikir Barat*, *Joget* and *Zapin* are easily recognizable by many locals as they are routinely performed at many cultural shows in Malaysia. At Malay weddings, *Ghazal* is particularly popular especially in the southern state of Johor. There are other genres such as *Keroncong*, *Dondang Sayang*, *Inang*, *Wayang Kulit*, *Caklempong*, among others that may be categorized as traditional Malay music [2].

Automatic music genre classification is important for information retrieval systems, sprouting numerous research in this field, especially for Western music genres [3-6]. Other genres have also been studied including Malaysian music genres, of which traditional Malay music is a subset [7-8]. However, as pointed out by these researchers, the main drawback for studying local genres is that there is no large dataset available such as those used for Western genres research. This is compounded by the difficulty in obtaining digital files of local genres.

Automatic music genre classification generally involves two main parts - feature extraction and classification. Popular features are similar to those used for speech recognition and extracted using similar techniques [9]. The Mel-Frequency Cepstral Coefficients (MFCC) are often used as they provide a compact representation of the spectral envelope resulting in the first coefficients ability to represent most of the energy signal. The MFCC is based on the Short-Time Fourier Transform (STFT) as do many others used for speech recognition collectively called spectral or cepstral features. In addition, time domain features such as zero-crossing rate are also utilized.

Once features are extracted from a music tract, a classifier will be employed to determine the genre. Supervised learning approaches such as K-Nearest Neighbors, Hidden Markov Models, Neural Networks, Support Vector Machines and many others similar to those used for speech recognition have been reported [3-8]. A number of studies tried to use different classifiers to try to improve performance. However, it has been shown that employing effective feature sets will have much more effect on the classification accuracy for audio signals in general [10].

In an effort to obtain more effective features, several researchers have suggested the used of spectrograms from which textural features are extracted [11-12]. A spectrogram is a visual representation of an audio signal obtained using STFT. The textual features extracted from the images are called descriptors which include entropy, homogeneity, contrast, energy and many others. It has been reported that using these features for a Latin music database with over 3000 music tracts and with an SVM classifier, good recognition rates higher than 90 percent were obtained for several genres, even though the total average recognition rate was only 60.1 for 10 Latin music genres [11].

In this paper, unlike the other techniques, the spectrograms themselves are used as features instead of extracting descriptors from the spectrograms. In order to distinguish classes using image cross-correlation, a technique based on advanced correlation filters commonly used for image biometrics is employed [13-14]. A dataset consisting of seven traditional Malay music genres are used. For each genre, a set of Unconstrained Minimum Average Correlation Energy (UMACE) filter is generated as templates and used to distinguish between the genres. Due to a limited dataset and in order to ensure good correlation, careful spectrogram construction

Copyright © 2018 Authors. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

is used so as to highlight the most salient feature of the music, while template construction using audio segmentation is used to account for the variations in music within a genre and as the music progresses.

2. Methodology

The proposed method is divided into three main parts consisting of spectrogram construction, template construction and template matching based on image cross-correlation with UMACE filters.

2.1. Spectrogram Construction

It is important that the constructed spectrogram retain the most salient feature of the music signal. To accomplish this, a method for robust audio fingerprint extraction is used for the spectrogram image construction [15].

Let an audio signal be represented by *f[n], n=0,1,....N-1*. A time-frequency transform decomposes *f* over a family of time-frequency segments $\{g_{l,k}\}_{l,k}$, where *l* and *k* are the time and frequency localization indices. With $^+$ denoting conjugate, the resulting coefficients can be written as

$$F[l,k] = \langle f, g_{l,k} \rangle = \sum_{n=0}^{N-1} f[n] g_{l,k}^{+}[n] \quad (1)$$

The STFT operation can be written as

$$g_{l,k}[n] = w[n - lu] e^{\left(\frac{i2\pi kn}{K}\right)} \quad (2)$$

where *w[n]* is a window of size *K*, which is shifted with a step u ≤ K with $0 \le l \le N/u$ and $0 \le k \le K$.

The linear spectrogram is then quantized into frequency sub-bands that are logarithmically spaced to cover a large frequency range for more features. The resulting spectrogram is then transformed to a log-magnitude representation as

$$S[l,k] = log|F[l,k]| \quad (3)$$

It is only after this operation than the spectrogram is converted into a gray scale image in order to construct the UMACE filter.

The audio format is used in this paper is the uncompressed (WAV) with a sample rate of 44.1 KHz. The audio is divided into frames using a Hanning window of length 8192 with an overlap rate of 0.75. A long frame length resulted in the spectrogram having low time resolution, which makes the representation insensitive to time variations [15]. In addition, large overlap is an advantage when dealing with using short query music against the long original signal [16-17].

2.2. Template Construction

Using the concept of time decomposition, an audio signal can be decomposed into different overlapping or non-overlapping sub-signals. Each sub-signal may be extracted from the original signal to obtain a spectrogram representing the segment. In this paper, 10-second segments separated by a gap of 20 seconds as shown in Figure 1 are employed to account for the variations in music within a genre and as the music progresses. These spectrograms are used to construct the UMACE filter templates.

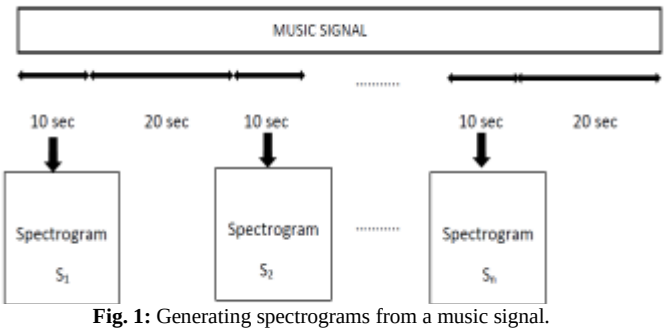

Fig. 1: Generating spectrograms from a music signal.

UMACE filters which have been successfully used for image cross-correlation in biometrics [13-14]. A UMACE filter which acts as the template are synthesized in the Fourier domain using a closed form solution. Several training images can be used to synthesize a filter template. In this case, *m* filter templates are generated from the spectrograms for each genre using the training set.

The designed filter is then used for cross-correlating with a test image in order to determine whether the test image is from the true or false class. In this process, the filter optimizes a criterion to produce a desired correlation output plane by minimizing the average correlation energy and at the same time maximizing the correlation output at the origin [15]. The optimization of the UMACE filter equation can be summarized as

$$\boldsymbol{U_{mace} = D^{-1} m} \quad (4)$$

D is a diagonal matrix with the average power spectrum of the training images placed along the diagonal elements, while *m* is a column vector containing the mean of the Fourier transforms of the training images.

2.3. Template Matching

The template matching process for a correlation filter U(u,v) and the input test image, in this case the spectrogram image S(x,y) is given by

$$c(x,y) = IFFT\{FFT\big(S(x,y)\big) * U^{+}(u,v)\} \quad (5)$$

where the test image is first transformed to frequency domain and reshaped to be in the form of a vector. It is then convolved with the conjugate of the UMACE filter, which is equivalent to cross-correlating it with the UMACE filter. The output is transformed again to the spatial domain obtaining the correlation plane.

The resulting correlation plane produces a sharp peak at the origin, while the values everywhere else are close to zero if the test image belongs to the same class as the designed filter. A simple metric called the Peak-to-Sidelobe ratio (PSR) is used to measure the sharpness of the peak where

$$PSR = \frac{peak - mean}{standard\ deviation} \quad (6)$$

where the peak is the largest value of the test image obtained from the correlation output. The mean and standard deviation are calculated from a 20x20 sidelobe region excluding a 5x5 central mask [13].

Since each genre is represented by multiple UMACE filters, a scheme is employed as illustrated in Figure 2 shown for genre 1. For each input music, the spectrograms generated as in Figure 1 are cross-correlated with the respective UMACE templates of a particular genre producing PSRs. Each value is then compared with an experimentally determined threshold in order to decide if the input music belongs to that particular genre. The threshold

values for each UMACE filter are determined in accordance to the UMACE training procedure [14]. Decision level fusion based on OR voting is used to decide if the input music belongs to that particular genre. This test is then repeated for all the genres.

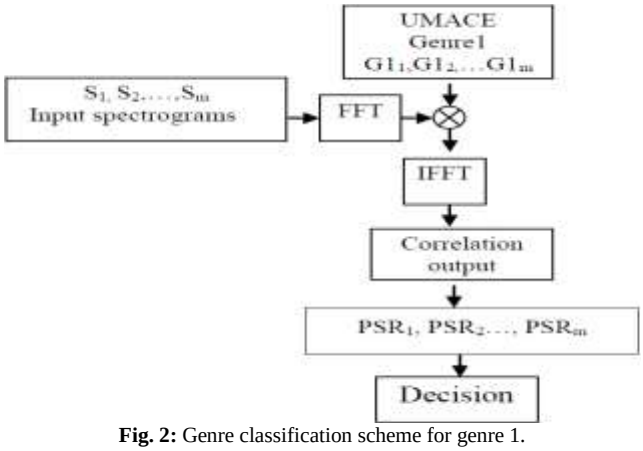

Fig. 2: Genre classification scheme for genre 1.

3. Results and Discussion

The experiments were conducted on seven Tradisional Malay music genres- *Inang, Zapin, Joget, Ghazal, Dikir Barat, Dondang Sayang* and *Keroncong*. All music samples were obtained from audio CDs. For each genre, a total of 14 music tracks were selected, equally divided between the training and testing sets.

UMACE template filters were generated for each genre from the spectrograms of the first 90 seconds of the music using three 10-second segments separated by a 20-second gap as previously explained. For comparison, the length of music is doubled to 180 seconds where six 10-second segments were extracted. UMACE filters were generated representing the extracted segments for a total of 3x7 = 21 template filters in the case of the 90 seconds music and 6x7 = 42 for the 180 seconds music. In the testing phase, each test music is divided into the corresponding segments and cross-correlated with the UMACE templates for each of the genre as previously explained. Classification accuracy is determined, defined as the ratio of correct classification to total number of test inputs. Table 1 summarizes the classification accuracy rate obtained.

Table 1: Classification accuracy rate according to genres and duration.

Genres	Classification Accuracy Rate (%)	
	90 Seconds	180 Seconds
Ghazal	46.2	75.4
Inang	30.2	38.7
Joget	38.9	58.9
Dikir Barat	51.2	80.7
Keroncong	43.7	72.9
Zapin	24.8	31.9
Dondang Sayang	48.2	73.9
Average	40.5	61.8

Table 1 shows that the average classification rate of the 180-second segments is better than that for the 90-second. This is due to the increased number of template UMACE filters used. Further doubling of the music segment is not feasible as not only will this increase the number of UMACE filters required, but this will also exceed the maximum duration of some music in the dataset. Interestingly, some genres have a much higher accuracy rate compared to others. In particular, the *Dikir Barat* has the highest classification rate at 51.2 and 80.7% for the 90- and 180-second segments respectively. This is may be due to it being a distinct genre with fast beats and with some parts, usually near the beginning, performed vocally without the accompaniment of musical instruments.

The best overall average accuracy rate of 61.8% obtained here is comparable to the spectrogram technique used for Latin music [11] with the advantage of not having to extract features from the spectrogram. The result obtained using this proposed technique is however lower than those reported by others for Malaysian music genre classification of which traditional Malay music is a subset at 88.6% [7] and 69.1% [8], where conventional techniques of extracting several features directly from audio signals were used. It may be argued that the feature used in this paper, which is the whole spectrogram image, is simpler. It is predicted that improvement in classification rates using this technique may be achieved with further research on spectrogram construction and audio partition.

4. Conclusion

Spectrogram correlation using UMACE filters has been presented for traditional Malay music genre classification. The method involved spectrogram construction retaining the most salient feature of the music, template construction catering for the variations in music within a genre and as the music progresses, and template matching based on spectrogram image cross-correlation with UMACE filters. The spectrograms themselves were used as features instead of extracting descriptors from the spectrograms. Seven traditional Malay music genres were tested- *Inang, Zapin, Joget, Ghazal, Dikir Barat, Dondang Sayang* and *Keroncong* with a pragmatically small dataset. In spite of this, an average recognition rate of 61.8 percent was obtained for 180-second music segments using six 10-seconds excerpts that start at the beginning of the music tract.

Acknowledgement

This research is supported by the Malaysian Ministry of Higher Education through the following grant: FRGS/1/2016/TK04/UKM/01/1.

References

[1] Pálmason, H., Jónsson, B.Þ., Schedl, M., Knees, P. Music genre classification revisited: An in-depth examination guided by music experts. Proceedings of the International Symposium on Computer Music Multidisciplinary Research, 2017, pp. 45-56.

[2] Nasuruddin M.G. The Malay traditional music. Dewan Bahasa dan Pustaka, 1992.

[3] Tzanetakis, G., Cook, P. Musical genre classification of audio signals. IEEE Transactions on Speech and Audio Processing, 2002, 10(5), 293-302.

[4] Lee, C.H., Shih, J.L., Yu, K.M., Lin, H.S. Automatic music genre classification based on modulation spectral analysis of spectral and cepstral features. IEEE Transactions on Multimedia, 2009, 11(4), 670-682.

[5] A. Nasridinov, Y.-H. Park. A study on music genre recognition and classification techniques. International Journal of Multimedia and Ubiquitous Engineering, 2014, 9(4), 31-42.

[6] Geng, S., Ren, G., Ogihara, M. Transforming musical signals through a genre classifying convolutional neural network. Proceedings of the International Workshop on Deep Learning and Music, 2017, pp. 48-49.

[7] Doraisamy, S., Golzari, S., Mohd, N., Sulaiman, M.N., Udzir, N.I. A study on feature selection and classification techniques for automatic genre classification of traditional Malay music. Proceedings of the International Society for Music Information Retrieval, 2008, pp. 331-336.

[8] Norowi, N.M., Doraisiamy, S., Rahmat, R.W. Traditional Malaysian musical genres classification based on the analysis of beat feature in audio. Journal of Information Technology in Asia. Journal of Information Technology in Asia, 2007, 2(1), 95-109.

[9] Yu D., Deng L. Automatic speech recognition. Springer London Limited, 2016.

[10] McKinney, M.F., Breebaart, J. Features for audio and music classification. Proceedings of the International Conference on Music Information Retrieval, 2003, pp. 151-158.
[11] Costa, Y.M., Oliveira, L.S., Koericb, A.L., Gouyon, F. Music genre recognition using spectrograms. Proceedings of the IEEE International Conference on Systems, Signals and Image Processing, 2011, pp. 1-4.
[12] Costa, Y.M., Oliveira, L.S., Silla Jr, C.N. An evaluation of convolutional neural networks for music classification using spectrograms. Applied Soft Computing, 2017, 52, 28-38.
[13] Kumar, B.V., Savvides, M., Xie, C., Venkataramani, K., Thornton, J., Mahalanobis, A. Biometric verification with correlation filters. Applied Optics, 2004, 43(2), 391-402.
[14] Samad, S.A., Ramli, D.A., Hussain, A. Lower face verification centered on lips using correlation filters. Information Technology Journal, 2007, 6(8), 1146-1151.
[15] Zhang, X., Zhu, B., Li, L., Li, W., Li, X., Wang, W., Lu, P., Zhang, W. SIFT-based local spectrogram image descriptor: A novel feature for robust music identification. EURASIP Journal on Audio, Speech, and Music Processing, 2015, 2015, 1-15.
[16] Li, W., Liu, Y., Xue, X. Robust audio identification for MP3 popular music. Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval, 2010, pp. 627-634.
[17] Yao S, Niu B, Liu J. Audio identification by sampling subfingerprints and counting matches. IEEE Transactions on Multimedia, 2017, 19(9), 1984-1995.