A Performance Comparison of Approaches in the Retrieval of Music

Jan Miles Co, and Andrei Coronel

***Abstract*—**Music Information Retrieval (MIR) is a relatively new field with the main goal of developing systems that allow users to retrieve music with the use of music content as opposed to text-based searching. Currently, most of the researches in MIR systems lack a performance indicator that will enable researchers to gauge the effectiveness of a newly proposed music retrieval system. This study attempts to test the performance of a proposed retrieval system, which uses melodic features for music retrieval, against a set of existing retrieval systems, which uses acoustic features for retrieval. The proposed retrieval system was able to achieve an accuracy of 98.25%, while one of the existing retrieval systems was able to achieve an accuracy of 99.75%. Being comparable in performance, the findings indicate that the melodic features can be used for the retrieval of music.

***Keywords*—**computer music, music information retrieval, music retrieval, music retrieval accuracy.

I. Introduction

MUSIC Information Retrieval (MIR) is a relatively new field with the main goal of developing systems that allow users to retrieve music with the use of music content as opposed to text-based search [1]. The lack of a standard evaluation method in this area makes it impossible to gauge whether a Proposed MIR System is effective or not. Moreover, there is no indicator whether a previous work has improved or not. The objective of this study is to improve an MIR system that was previously proposed and to compare its performance against a set of existing MIR systems. By creating a performance indicator, which can be fairly used against other methods, the performance of the Proposed MIR System will be identified. Section 2 introduces the Proposed MIR System, discusses several retrieval systems, and stresses the lack for evaluation methods in MIR Systems. Section 3 describes the methodology for performance improvement and comparison. Section 4 presents the results of the experiment. Section 5 concludes the study.

II. Review of Related Literature

A. Retrieval of MIDI Files

J. Co and A. Coronel [2] used melodic features for the retrieval of MIDI files. In this study, several classification algorithms were used to create feature sets for the classification of music by cultural source. The feature sets with the highest classification accuracy were used as features sets for the retrieval of the MIDI files. In this study, it was identified that the training set for the identification of the feature sets for retrieval contained songs that belong to two or more categories, which caused an ambiguity to the interpretation of the results. The study also failed to compare the performance of the Proposed MIR System against existing MIR Systems, which is a common problem in numerous music retrieval system proposals due to the lack of a standard evaluation method.

B. Existing MIR Systems for Comparison

M.K. Islam et al [3] used beat and peak period features for music retrieval. 100 different music, composed of Classic, Rap, Pop, New Age, and Rock songs, were used in his study. Given a query, main beat and peak periods will be extracted. Using these features, the location of the song in the database will be identified. If the length of the query is too short, then peak features will not be used. Levenshtein distance and Euclidean distance were used as a similarity measure and an accuracy of 98% and 85% were achieved respectively.

S. Parvin and Jong Sou Park [4] also created a system that uses beat features for music retrieval. In his work entitled "An Efficient Music Retrieval Using Noise Cancellation," songs undergo two processes: Noise Cancellation and Music Retrieval. In Noise Cancellation, filtering approaches are applied to get a clean noise-free signal, which can sufficiently represent the music. Using the songs unique beat pattern, beat features are extracted from the representative music and matched against the song available on its database. 100 different music, composed of Classic, Rap, Pop, New Age, and Rock songs were used in his study. An accuracy of 94% was achieved.

P. Boonmatham et al [5] used wave features for the retrieval of Thai music. Noise reduction was also performed before feature extraction where chromagram features are extracted for the succeeding retrieval process. Two similarity measures were used: simple Euclidean distance and Cosine distance. 60 songs, coming from three types of classical music: 20 Piphat, 20 Mahori, ans 20 Khrueang Songs, were used. All of the songs were used as a query. This formula was used to compute for the retrieval accuracy:

$$Accuracy(\%) = \frac{\#correctlyretrievedsongs}{\#queries} \quad (1)$$

An accuracy of 65% was achieved when simple Euclidean distance was used as a similarity measure, while an accuracy

Jan Miles Co is with Ateneo de Manila University, Quezon City, Philippines (janmilesco@yahoo.com.ph).

Andrei Coronel is with Ateneo de Manila University, Quezon City, Philippines (docdrei@gmail.com).

of 71.66% was achieved when Cosine distance was used as a similarity measure.

D. Schnitzer [6] used MFCC features for the retrieval of music. The MFCC features used were identified to be the most effective for music retrieval in MIREX 2006. Kullback-Leibler (KL) divergence is used for similarity computation. Multiple datasets were used to test the MFCC feature sets. The first dataset contains 120 songs belonging to six genres. The second dataset contains 1,311 songs belonging to eight genres. And the last dataset contains 16,781 songs belonging to 21 genres. The Genre/Class Method described in [6] was used to measure the retrieval accuracies. An accuracy of 52.2%, 82.8%, and 52% was achieved with the datasets respectively.

C. Lack of Recommendation Systems

Existing evaluation methods were explored in order to gauge the effectiveness of MIR systems. However, not one method was identified. In fact, in 2012, it was determined by The International Society for Music Information Retrieval (ISMIR) that experiments in MIR are often initiated by graduate students, who build a test-set and make it available for the community. This setup is problematic since there is a lack of documentation with regard to the methodologies, which make it difficult to assess its reliability. And in most cases, it is impossible to perform the exact experiment in a different lab. Moreover, it was identified that there is difficulty with making datasets available due to copyright restrictions. When systems are submitted for testing, the inaccessibility of the dataset used by formal evaluators makes it difficult to improve the evaluated systems [7].

III. Methodology

A. Performance Improvement

In order to eliminate the ambiguity that was caused by the dataset in [2]. Songs that belonged to two or more categories were replaced with songs that belong to only one category. A total of 8 songs were removed from the original dataset. 17 new songs were collected to replace the songs in the original dataset. The songs were collected from [8], similar to where the original dataset was collected.

TABLE I
Songs in the Original Dataset that Belong to Two or More Categories

Song	Scottish	English	American	French
amazingg	1	1	0	0
aurelia	1	1	0	0
darwa148	1	0	1	0
deboutia	0	1	1	1
forestgrn	0	1	1	0
heartaok	0	1	1	0
jinglebe	0	0	1	1
oldhundr	1	0	1	0

TABLE II
List of New Songs and Its Categories

Song	Category	Song	Category
ojohncom	Scottish	littljen	American
oladymar	Scottish	lilyray	American
omyloves	Scottish	marine	American
ladculln	English	melind	American
ringerec	English	mineisth	American
lavans	English	myboyisc	American
fareweto	English	veuxtute	French
wareham	English	yabiendi	French
waltham	English		

The methodology described in [2] was repeated for the creation of new feature sets for music retrieval: jSymbolic was used for feature extraction and WEKA 3 was used for classification with Naive Bayes and SVM. C4.5 decision tree, MLR, and CFS with Best First Search, Greedy Stepwise, and Genetic Search were used for the reduction of the feature sets. The winning feature sets, feature sets with the highest classification accuracy with SVM, were used for retrieval similar to [2].

B. Performance Comparison

Due to the lack of standard evaluation methods, as discussed in Section 2, two methods were used to assess the quality of the Proposed MIR System. The first method compares the performance of acoustic features to melodic features. This method was conducted in order to determine the behavior of acoustic features when used for the retrieval of music specifically when working with the dataset at hand, which is specifically composed of Scottish, English, American, and French traditional MIDI files. The second method adopts the evaluation methods used by existing MIR Systems to gauge its effectiveness with regard to the retrieval of music. The existing MIR Systems used for comparison are discussed in Section 2.

1) Implementation of Acoustic Features

The main concern for the creation of an MIR system in [2] is to determine whether melodic features can be used to classify and retrieve traditional music according to cultural source, specifically: Scottish, English, American, and French. To provide an idea about the performance of the Proposed MIR System that was developed, a comparative experiment was performed with acoustic features and an MIR system was developed that implements acoustic features for music retrieval. An experiment was designed in order to test the effectiveness of melodic features against the effectiveness of acoustic features to determine which of the two is more effective in the context of classifying and retrieving traditional songs according to its cultural source.

The 400 MIDI files of the new dataset (without redundant songs) were converted to .aiff format with the following specifications: sample size = 16, sample rate = 8000, channels = mono, format = uncompressed. The OS X application called Switch was used for the conversions.

Two feature sets were used. The first feature set contains the MFCC features while the second feature set contains Beat features. The MFCC features were determined to be one of the most effective feature sets for the retrieval of music [6]. Beat

features were used by [4] for the retrieval of music. The features were extracted using jAudio, a software that extracts features from audio files. The jAudio sample rate was set to 8kHz. The datasets were loaded in jAudio and the feature sets were extracted accordingly. The datasets with the extracted features were loaded in WEKA 3 for classification with Naive Bayes and SVM with 10-fold cross-validation.

Two matrices were created for the retrieval of songs. jAudio was used for feature extraction and R was used for distance computations. The matrices were then used for the retrieval of songs, first the MFCC features, succeeded by the beat features. The results of retrieval are presented in Section 4.

2) Comparison with Existing MIR Systems

The performance of the Proposed MIR System was compared to those of [4], [5], and [3]. The Proposed MIR System with the implementation of acoustic feature sets were also included, as well as the Proposed MIR System with the original dataset and the initial feature sets. In order to apply the methods used by other MIR Systems, the Proposed MIR System was modified. Initially, the Proposed MIR System does not return the query as the most similar song if it is found on the database. Example: If the query is 4and20-scottish and the top most retrieved song is 4and20-scottish, it is removed from the list and the succeeding song will be the most similar song. After modifying the Proposed MIR System, the methods used by existing MIR Systems were implemented. This method works according to these steps:

- Use all of the songs as a query
- Check the first retrieved song (most similar song)
- Add a score if the songs are the same. The accuracy is computed by using this formula:

$$Accuracy(\%) = \frac{\#\,correctsongs}{\#\,totalsongs} \times 100 \quad (2)$$

The comparison of retrieval accuracies is presented in in Section 4.

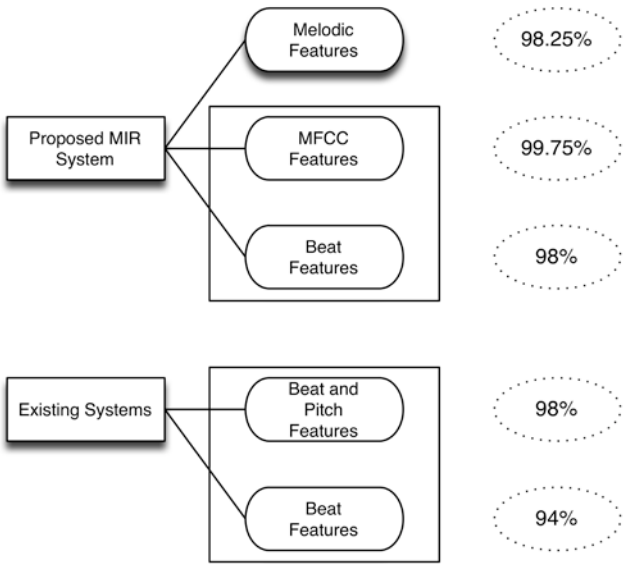

Fig. 1 Methodology for Performance Comparison

IV. Results and Discussion

The Proposed MIR System with the implementation of melodic features was compared to the Proposed MIR System with the implementation of acoustic features. The retrieval accuracies the Proposed MIR System were also compared to the retrieval accuracies of [4], [5], and [3].

TABLE III
Comparison of Retrieval Accuracies

Song	Dataset	Dataset Categories	Features Used	Accuracy
Proposed MIR System	400 (Revised Dataset)	4	MFCC	99.75%
Proposed MIR System	400 (Revised Dataset)	4	Melodic	98.25%
Proposed MIR System	400 (Revised Dataset)	4	Beat	98%
Content Based Music… [3]	100	5	Beat and Pitch	98%
Proposed MIR System	400 (Original)	4	Melodic	97.50%
An Efficient Music…[4]	100	5	Beat	94%
A Comparison of Audio…[5]	60	3	Chromagram	71.66%

When the evaluation method of existing MIR Systems were implemented in the Proposed MIR System, an accuracy of 98.25% was achieved which shows that the Proposed MIR System proved to be effective for the retrieval of music. Moreover, a higher accuracy can be achieved with the use of MFCC features, which were consistent with what was known with the feature set (to be one of the best for music retrieval) [6].

Being one of the best for music retrieval, the performance of the Proposed MIR System can be compared to the performance of MFCC features. MFCC features perform better than the Proposed MIR System by 1.5%. It can be concluded that the Proposed MIR System was comparable in performance when this evaluation method was used. This finding proves that MIDI files can be used for search and retrieval of music in the context of Scottish, English, American, and French traditional songs.

V. Conclusions and Further Studies

The findings in this study prove that the Proposed MIR System was comparable in performance to existing MIR Systems. The Proposed MIR System uses melodic features for the retrieval of music, in contrast to most MIR Systems that use acoustic features. Going back to the possible implementations of the Proposed MIR System, this can be used to improve recommender systems where acoustic files are processed in real time. The use of actual MP3 and wav files for search and retrieval corresponds to costly storage and processing requirements while MIDI files require less processing and storage requirements. There is a potential for MIDI files to be used for the online sales of music, specifically for recommendation. However, it is important to note that in this study, the use of melodic features for search and retrieval are limited to Scottish, English, American, and

French traditional songs. More experiments must be conducted whether MIDI files can truly replace acoustic files in recommender systems and the retrieval of songs of a wider genre.

ACKNOWLEDGMENT

Financial support from DOST-ERDT is greatly acknowledged.

REFERENCES

[1] Stephen J. Downie. The music information retrieval evaluation exchange (2005-2007): A window into music information retrieval research. *Acoustical Science and Technology*, 29(4):247-255, 2008. http://dx.doi.org/10.1250/ast.29.247

[2] Jan Miles Co and Andrei Coronel. Exploring melodic features for the classification and retrieval of traditional music in the context of cultural source. In *National Conference on Information Technology 2013*, http://janmilesco.files.wordpress.com/2014/02/ncite2013co.pdf.

[3] M.K. Islam, Hyung-Jin Lee, A.K. Paul, and Joong Hwan Baek. Content-based music retrieval using beat information. In *Fuzzy Systems and Knowledge Discovery, 2007. FSKD 2007. Fourth International Conference on*, volume 3, pages 317-321, 2007.

[4] S. Parvin and Jong Sou Park. An efficient music retrieval using noise cancellation. In *Future Generation Communication and Networking (FGCN 2007)*, volume 1, pages 541-546, 2007.

[5] P. Boonmatham, S. Pongpinigpinyo, and T. Soonklang. A comparison of audio features of thai classical music instrumentation. In *Computing and Convergence Technology (ICCCT), 2012 7TH International Conference on*, pages 213-218, 2012.

[6] D. Schnitzer. Mirage High-Performance Music Similarity Computation and Automatic Playlist Generation. Master's thesis, Vienna University of Technology, 2007.

[7] Geoffroy Peeters, Julian Urbano, and Gareth JF Jones. Notes from the ismir 2012 late-breaking session on evaluation in music information retrieval. 2012.

[8] Musica international, http://www.musicanet.org/robokopp/. January 2013.