A Multimodal Approach to Song-Level Style Identification in Pop/Rock Using Similarity Metrics

Ching-Hua Chuan
University of North Florida
School of Computing
Jacksonville, FL, USA
c.chuan@unf.edu

Abstract—**This paper presents a multimodal approach to style identification in pop/rock music. Considering the intuitive feelings of similarity from the listener's perspective, this study focuses on features that are computed using similarity metrics for melodies, harmonies, and audio signals for style identification. Support vector machine is used as a binary classifier to determine if two songs are created by the same artist given their similarity distances in the three aspects. Experiments are conducted using songs of four well-known pop/rock bands from 6 albums. The preliminary result shows that the approach achieves the best result in correct rate of 85% using only seven similarity metrics.**

Keywords— Gaussian mixture models; melodic contour; music similarity; n-grams; style

I. Introduction

Music is a unique type of multimedia data. It not only exists in many different data formats, but it also conveys numerous musical ideas via hierarchical instrumentations and/or sounds. For example, symbolic data such as melodies and harmonies indicate the fundamental components in music compositions while acoustic data present another layer of information created by performances or sound effects.

In addition, music creation is dynamically changing. Music pieces or songs created by the same artist may not always sound the same, although they may sound more similar to each other than the other artists' work. Songs by the same artist may share certain components but not necessarily contain all of them, depending on how versatile the artist is. Therefore, in order to understand the nature of musical data, all related features should be considered while each music piece should be regarded as an individual case.

In this paper, we analyze musical data in three aspects: melodies, harmonies, and audio signals. Unlike the majority of prior work on music style that only focuses on a singular type of features [1, 2], we study how an artist's style is defined along with the three aspects by comparing the songs of one artist with those of another artist. Melodic similarity is represented by the cosine distance between two pitch class distributions and pitch interval distributions, as well as cosine distance between melodic contours in phrases. Harmonic similarity is computed by comparing chord profiles of two songs. A chord profile consists of *n*-gram chord patterns weighted by the patterns' durations. Acoustic similarity is produced by computing the Mahalanobis distance between the feature vectors extracted from two audio recordings, as well as by comparing the Gaussian mixture models built for the two recordings using Monte Carlo sampling. We conduct experiments to investigate the effectiveness of these similarity metrics in the three aspects for identifying an artist's unique style.

Since an artist's style is fluid and dynamically changing, we focus on the task of determining whether any given two songs are composed by the same artist, instead of building a fixed model based on a dataset of a particular artist. For the binary classification problem, an input instance consists of features that represent the similarity/difference between two songs using the similarity metrics for melodies, harmonies, and audio signals. Positive instances are created by comparing two songs from the same artist using these similarity metrics, while negative cases are produced by two songs from different artists. We use support vector machine as the classifier. Classification results are generated using 10-fold cross validation. Given the expeirment results, we examine how effective the proposed approach performs for recognizing the style of the artists in the dataset. We also discuss how the similarity metrics in the three aspects affect the system's performance for a particular artist.

II. Related Work

The most popular classification task for musical data is genre classificaiton. Genre classification involves classifying a given music example into one of several pre-defined classes/genres, such as classical, pop, rock, jazz. Most genre classification systems take audio recordings as input. Lower-level acoustic features such as spectral centroid, zero-crossing rates, mel-frequency cepstral coefficients [3] and features related to psychoacoustics [4] are commonly used. Commonly used classifiers include support vector machine [5], genetic algorithms [6], hiddne Markov models [7].

However, identifying an artist's signature style is a different task from genre classification. For example, an artist can be labeled with multiple genres while even artists in the same genre may have very different styles. In [1, 2], *n*-gram models are used for representing composer style in pop/rock and jazz. In these studies, *n*-grams are generated from chord

978-0-7695-5144-9/13 $31.00 © 2013 IEEE
DOI 10.1109/ICMLA.2013.143

IEEE computer society

sequences of songs by a composer, and the style of the composer is described by the extracted *n*-gram patterns with their probability distribution over the entire collection. In [8], artists' styles in pop/rock music are examined by studying the chord patterns at the end of segments as suggested by the punctuation marks in lyrics and melodic boundaries. The result shows that more than half of the chord patterns extracted at the end of segments are different from standard ones in Western music, and focusing on the non-standard patterns appears to be effective for differentiating styles among pop/rock bands.

III. Feature Extraction

A. Melodic Similarity

In this study, similarity between two melodies is calculated based on three melodic representations: pitch class distribution, pitch interval distribution, and melodic contour. In Western tonal music, pitches are divided into 12 pitch classes: c, $c^{\#}/d^{b}$, d, $d^{\#}/e^{b}$, e, f, $f^{\#}/g^{b}$, g, $g^{\#}/a^{b}$, a, $a^{\#}/b^{b}$, and b. The distance between two pitch classes is called a semitone distance. Pitch class distribution presents the existence of the 12 pitch classes in percentages depedining on the duration of each pitch class. Pitch interval distribution shows the distribution of distances (in terms of number of semitone distance) between adjacent pitches. Melodic contour describes the general shape of a melody by focusing on a subset of notes in the melody. Melodic contour is easier to be remembered by listeners than exact notes [9] and is commonly used for indexing and retrieval in music information retrieval [10].

Fig. 1 (a) shows an example of melodic segment. If a note in the melody is represented as a <pitch class, duration (in beats)> pair, the melodic segment in Fig. 1 (a) can be represented as <*e*, 0.5>, <*d*, 0.5>, <*b*, 0.25>, <*d*, 0.5>, <*d*, 0.25>, <*e*, 0.5>, <*d*, 0.5>, <*b*, 0.5>, <*d*, 0.25>, and <*d*, 0.25>. Therefore, the pitch class distribution for the melodic segment is [c, $c^{\#}/d^{b}$, d, $d^{\#}/e^{b}$, e, f, $f^{\#}/g^{b}$, g, $g^{\#}/a^{b}$, a, $a^{\#}/b^{b}$, b] = [0, 0, 0.5625, 0, 0.25, 0, 0, 0, 0, 0, 0, 0.1875]. The intervals between adjacent notes are [-2, -3, +3, 0, +2, -2, -3, +3, 0] in semitone distance. Therefore, the pitch interval distribution for the melodic segment is [-3, -2, -1, 0, +1, +2, +3] = [0.23, 0.27, 0, 0.2, 0, 0.1, 0.2], calculated by obtaining the sum of the duration of the two notes for an interval. The melodic contour for the melodic segment is shown in Fig. 1 (b), generated by setting the degree of resolution to 1 beat.

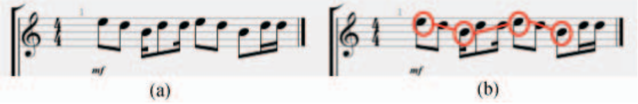


Fig. 1. A melodic segment (a) and its melodic contour (b) highlighted.

The similarity between two melodies can be calculated using cosine distance between the pitch class distributions and pitch interval distributions of the two songs. This measure uses global information, i.e., statistics on pitches in the entire melody, to provide similarity. Melodic similarity can also be computed using a local measure: by comparing the melodic contour of local phrases in two melodies. A melody is first divided into phrases, and each phrase is summarized by its melodic contour. The overall similarity is then calculated by averaging the distance between all pairs of phrases in the two melodies. More details about melodic similarity can be found in [11].

All melodies are transposed (normalized) to the key of C before producing distributions and melodic contours.

B. Harmonic Profiles

Harmony in Western tonal music is usually represented as a sequence of chords that accompany a melody. An effective way to analyze the patterns in the chord sequence is to use *n*-gram models [12, 13]. Fig. 2 shows a melodic segment harmonized by a chord sequence consisting of G major, G major, C major, C major and G major. Given the chord sequence in Fig. 2, four 3-gram patterns can be extracted: G-G-C, G-C-C, C-C-D, and C-D-G.

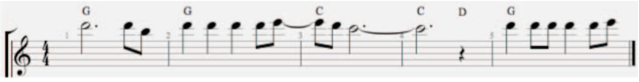


Fig. 2. A melodic segment (a) and its melodic contour (b) highlighted.

A chord profile is then constructed with the extracted *n*-gram patterns, weighted by the sum of the chord's durations in the pattern. For example, the first 3-gram pattern in Fig. 2, G-G-C, has a total duration of 12 beats. The total duration of all 3-gram patterns for this sequence is 12 + 11 + 8 + 8 = 38. Therefore, the weight for G-G-C is 12/38 ≈ 0.32.

The harmonic similarity between two songs is computed by calculating the cosine distance between the weight vectors in the chord profile of the two songs [2]. Before producing chord profiles, chords are reduced to their fundamental major and minor triads and normalized (transposed) to the key of C.

C. Acoustic Features

In addition to symbolic data such as melody and harmony, acoustic data extracted from CD or mp3 recordings provide abundant information about which instruments were used and how the notes were played. In this paper, we use discrete wavelet transform to extract low-level acoustic features from recordings. The silence in the beginning of audio recordings is removed based on signal energy using root mean square [13]. Non-music parts for recordings of live performances are also removed manually. 30-seconds of the recordings were then extracted as the excerpts for expriments. The 30-second excerpt was then divided into overlapping frames with frame duration of 0.5 seconds and hop duration around 23 milliseconds (1,024 samples for sampling rate of 44.1kHz). Each frame was then analyzed by discrete wavelet transforms to produce a compact feature vector.

As a result, acoustic data in an recording can be represented using the coefficients obtained from discrete wavelet transforms. The representation results in a matrix with rows indicating frames and columns showing the corresponding wavelet coefficients. The data in the matrix can also be further summarized by modeling each frame using Gaussian mixture models (GMMs). We applied GMMs with expectation maximization algorithm to build the most likely model for each audio excerpt. In this way, the excerpt was summarized as the parameters such as mean and covariance in the GMMs.

Acoustic similarity between two audio excerpts is calculated for both the matrix representation and Guassian mixture models. For the matrix representation, Mahalanobis distance [14] is used to assess the dissimilarity between two feature vectors $\boldsymbol{f}_i$ and $\boldsymbol{f}_j$:

$$D_M(\boldsymbol{f}_i, \boldsymbol{f}_j) = (\boldsymbol{f}_i - \boldsymbol{f}_j)^T \Sigma^{-1} (\boldsymbol{f}_i - \boldsymbol{f}_j), \qquad (1)$$

where $\boldsymbol{\Sigma}$ is the covariance matrix of the features across all audio examples in the dataset.

Given the built GMMs of two audio excerpts, we can compute similarity between the two using Monte Carlo sampling. Given two excerpts E_i and E_j, the dissimilarity between the two can be computed by first using Monte Carlo sampling to produce N_s samples using E_i's GMMs, and calculating the likelihood of these N_s samples that are generated using E_j's GMMs. The dissimilarity (distance) is calculated as a symmetric and normalized measure [15]:

$$D_G(E_i, E_j) = \sum_{n=1}^{N_s} \log p\left(\boldsymbol{S}_n^{\lambda_i} \middle| \lambda_i\right) + \sum_{n=1}^{N_s} \log p\left(\boldsymbol{S}_n^{\lambda_j} \middle| \lambda_j\right) - \sum_{n=1}^{N_s} \log p\left(\boldsymbol{S}_n^{\lambda_i} \middle| \lambda_j\right) - \sum_{n=1}^{N_s} \log p\left(\boldsymbol{S}_n^{\lambda_j} \middle| \lambda_i\right), \qquad (2)$$

where $\boldsymbol{S}$ is the set consisting of N_s samples generated based on the built GMMs of a particular example. N_s is set to 20 in this study.

In this paper we adopted the parameters, including wavelet order of five and eight Gaussian components, that were obtained for the best result in music genre classification and retrieval in [16].

IV. Datasets and Experiments

Table 1 lists the artists, albums, and number of songs collected for the experiment in this paper. In addition, Table 1 lists the top five tags obtained from last.fm to provide readers with some background information regarding these artists' musical styles. These artists were chosen because they are all labeled as rock artists but belong to different sub-genres. Acoustic data were collected from the CD recordings while melodies and chords of the songs were manually annotated based on commercial lead sheets.

TABLE I. The dataset used in this paper

Artists	Albums	No. of Songs	Top-5 Tags[a]
Beatles	The Beatles 1, The White Album	15	classic rock, rock, british, 60s, pop
Green Day	Dookie	14	punk rock, rock, punk, alternative, pop punk
Guns N' Roses	Appetite for Destruction, Greatest Hits	15	hard rock, rock, classic rock, 80s, metal
Indigo Girls	1200 Curfews	15	folk, female vocalists, singer-songwriter, acoustic, rock

[a] Tags are obtained from last.fm.

We designed a binary classification task to examine the effectiveness of the proposed similarity metrics in melody, harmony, and audio signals for artist differentiation. An input instance for the classifier consists of similarity values between two songs, and the classifier aims to report whether the given two songs are from the same artist(s).

Two sets of experiments were conducted. In the first set, all similarity metrics in melody, harmony, and audio signals were used. Positive input instances for artist A_i were produced by generating similarity values using two songs from A_i. Negative instances for artist A_i can be created by using one song of A_i and another one from A_j, j = 1, …, 4, and $j \neq i$. If the dataset only contains negative instances for A_i using a particular A_j songs, we then can observe whether the classifier can distinguish the styles between the two artists A_i and A_j.

In the second set of the experiment, similar procedures were conducted but we focused on examining the effectiveness of similarity metrics in melody, harmony, and audio signals individually.

Support vector machine was used as the classifier in this paper. For each dataset, every combination of negative instances was used. Therefore, postive instances were weighted higher in order to present the dataset with equal numbers of positive and negative cases. 10-fold cross validations were performed to generate classification results regarding overall correct rates.

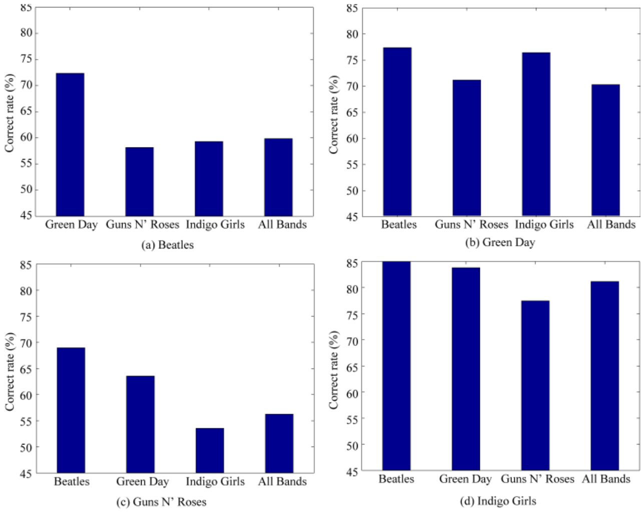
V. Results

Fig. 3 shows the classification results using all similarity metrics. Notice that the overall classification results in Fig. 3 (d) are the highest, indicating that the classifier performs the best to identify songs from Indigo Girls. It also recognizes the differences between Indigo Girls' songs from those of other artists in the dataset. The classifier achieves the second highest overall result for Green Day. One reason for achieving better classification results for Indigo Girls and Green Day can be that examples for these two bands are collected from single albums (see Table 1). The albums for Beatles and Guns N' Roses are compilations, consisting of songs over many years of the bands' music career.

Also notice that the results in Fig. 3 are not symmetric, i.e., the classification result for Indigo Girls using negative cases from Beatles is different from the Beatles' result using Indigo Girls as negative cases. Although the negative instances are the same for the two datasets, the variety in Beatles' songs may produce weaker positive cases.

Fig. 4 shows the classification results using similarity metrics individually. The results in Fig. 4 help us understand in which aspects the two bands are different. For example, the aspect that most differentiates songs of Beatles and Indigo Girls is melody, as shown in Fig. 4 (a) and (d).

Generally, the correct rates in Fig. 3 are higher than the ones in Fig. 4, indicating that considering the three similarity aspects simultaneously improves the results. But there are few cases in Beatles' and Guns N' Roses' results that using single metrics performs better. Bigger datasets and dividing songs from compilation albums into sub groups are necessary for further studies.



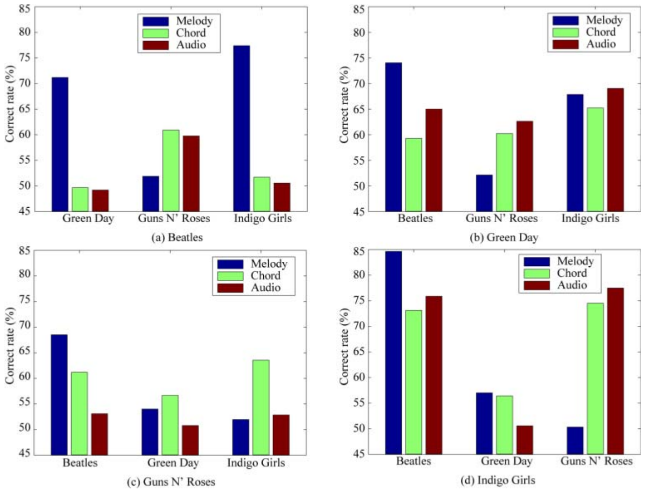
(a) Beatles

(b) Green Day

(c) Guns N' Roses

(d) Indigo Girls

Fig. 3. Classification results with all similarity metrics.



(a) Beatles

(b) Green Day

(c) Guns N' Roses

(d) Indigo Girls

Fig. 4. Classification results obtained using similarity metrics individually.

References

[1] M. Mauch, S. Dixon, C. Harte, M. Casey, and B. Fields, "Discovering chord idioms through Beatles and real book songs," in *Proceedings of the 8th International Conference on Music Information Retrieval*, 2007.

[2] M. Ogihara and T. Li, "N-gram chord profiles for composer style representation," in *Proceedings of the 9th International Conference on Music Information Retrieval*, 2008.

[3] T. Li, M. Ogihara, and Q. Li, "A comparative study on content-based music genre classification," in Proceedings of the 26th annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 282–289, 2003..

[4] M. F. McKinney and J. Breebaart, "Features for audio and music classification," in *Proceedings of the 4th International Conference on Music Information Retrieval*, 2004.

[5] S.-H. Chen and S.-H. Chen, "Content-based music genre classification using timbre feature vectors and support vector machine," in Proceedings of the 2nd International Conference on Interaction Sciences: Information Technology, Culture and Human, pp. 1095–1101, 2009.

[6] G. V. Karkavitsas, "Automatic music genre classification using hybrid genetic algorithms," *Intelligent Interactive Multimedia Systems and Services*, 11, pp. 323–335, 2011.

[7] Y. Qi, "Music analysis using hidden Markov mixture models," *IEEE Transactions on Signal Processing*, 55(11), 2007.

[8] C.-H. Chuan, "Pop/Rock musical style as defined by two-chord patterns at segmentation points in the melody and lyrics," in *Proceedings of the 11th IEEE International Symposium on Multimedia*, pp. 448–452, 2009.

[9] W. J. Dowling, "Scale and contour: Two components of a theory of memory for melodies," *Psychological Review*, 85(4), pp. 341–354, 1978.

[10] Y. E. Kim, W. Chai, R. Garcia, and B. Vercoe, "Analysis of a contour-based representation for melody," in *Proceedings of International Sympsoium on Music Information Retrieval*, 2000.

[11] T. Eerola and P. Toiviainen, "MIDI toolbox: MATLAB tools for music research," www.jyu.fi/musica/miditoolbox/, 2004.

[12] C.-H. Chuan, "Harmonic style-based song retrieval using n-grams," in *Proceedings of 1st ACM International Conference on Multimedia Retrieval*, 2011.

[13] O. Lartillot and P. Toiviainen, "MIR in Matlab (II): A toolbox for musical feature extraction from audio," in Proceedings of the 8th International Conference on Music Information Retrieval, 2007.

[14] M. I. Mandel and D. Ellis, "Song-level features and support vector machines for music classification," in *Proceedings of the 6th International Conference on Music Information Retrieval*, 2005.

[15] J.-J. Aucouturier and F. Pachet, "Improving timbre similarity: How high's the sky?" *Journal of Negative Results in Speech and Audio Sciences*, 1(1), pp. 1–13, 2004.

[16] C.-H. Chuan, "Audio classification and retrieval using wavelets and Gaussian mixture models," *International Journal of Multimedia Engineering and Management*, 4(1), 2013.

[17] C. Harte, M. Sandler, S. A. Abdallah, and E. Gomez, "Symbolic representation of musical chords: a proposed syntax for text annotations," in *Proceedings of the 6th International Conference on Music Information Retrieval*, 2005.