MUSIC FINGERPRINT EXTRACTION FOR CLASSICAL MUSIC COVER SONG IDENTIFICATION

Samuel Kim, Erdem Unal, and Shrikanth Narayanan

Speech Analysis and Interpretation Lab. (SAIL)
University of Southern California, Los Angeles, USA.
{kimsamue, unal, shri}@sipi.usc.edu

ABSTRACT

An algorithm for extracting music fingerprints directly from an audio signal is proposed in this paper. The proposed music fingerprint aims to encapsulate various aspects of musical information, such as overall note distribution, harmony structure, and their temporal changes, all in a compact representation. The utility of the proposed music fingerprint to the task of automatic classical music cover song identification is explored through experimental studies; specifically, the goal here is to identify the different versions of the same music through similarity comparisons of the music fingerprints. The results show an improved performance over the state-of-the-art cover song identification systems in terms of both accuracy and speed: the accuracy improved by approximately 40% while the search speed is about 60 times faster than the conventional system.

1. INTRODUCTION

Music information retrieval (MIR) technologies are becoming increasingly more important with the growing need for mining and searching of vast amounts of music archives, both personal libraries and those available in public repositories. Recent advances in storage and networking capabilities have accelerated the multimedia data explosion, and have fueled this need for intuitive and efficient data mining/search schemes. There has been several music information retrieval applications that have been recently proposed and developed, and it is particularly notable that an annual evaluation has been introduced through the international music information retrieval system evaluation laboratory's (IMIRSEL) "music information retrieval evaluation exchange" (MIREX), to facilitate comparisons among different systems (see [1] for details).

Building a music information retrieval system, however, is challenging due to several factors. These can be grouped into two major categories: signal processing issues and music theoretical issues. Since the music audio typically represents a mixture of several instruments or voices, from a signal processing point of view, it is necessary to handle multiple pitches (polyphonic) and multiple timbres (a domain with several instruments). Similarly, from a musical point of view, mathematical modeling of the complex dynamics and the interaction between various aspects of music, such as rhythm, harmony structure, and chord progression, also has several open problems. To tackle these complex challenges from a practical point of view, in MIR application development, researchers tend to focus on handling specific aspects describing the underlying music depending on the target application (e.g., genre classification, note transcription) simplifying the complexity to some extent. In this work, we focus on the cover song identification task.

Assuming that attributes of musical information, such as note distribution, harmony structures, and note change tendency, capture key features describing the music especially in cover song identification, we propose a novel music fingerprint representation encapsulating relevant musical information. The use of such music fingerprint idea, however, is not new and studies on extracting descriptive features from music data have been carried out in a variety of MIR contexts. For example, Unal *et al.* devised a fingerprint based on the relative pitch movement for query-by-humming systems, where the users can hum to create a query input to retrieve their song of interest [2]. In [3], Haitsma *et al.* proposed several parameters that should be considered in extracting music fingerprints. The memory requirement and the searching speed, as well as the accuracy of the system are included in the parameters. They focused on discovering the exact same song with different types of audio compression algorithms, so that it can be utilized in copyright protection applications over the internet or broadcasting.

A key requirement in designing music fingerprints is the ability to perform similarity measurements that are meaningful to the target application. It should be noted that measuring similarity between two different pieces of music audio is central to many applications such as automatic classification of genre, artist, and mood expressed in the music. In seeking appropriate similarity measures, researchers have been pursuing a variety of different approaches. Mandel *et al.* utilized a timbre-related feature to measure the similarity in terms of the artist of music [4]. They computed the overall distribution of mel frequency cepstral coefficients (MFCC) for each music audio. Similarity measurements via chord recognition are also studied. Lee [5] and Unal [6] have used chord recognition results, using hidden markov model (HMM) based and rule-based approaches, respectively, to compute similarity of the different pieces of music, for automatic identification of the cover song. Recently, Ellis *et al.* proposed a cross-correlation based cover song identification system, which won the first place in the MIREX 2006 evaluation [7]. Among the various requirements for the music fingerprint design proposed in [3], the focus of numerous prior efforts, however, has been primarily on improving accuracy. In contrast, the music fingerprint proposed in this work is expected to be memory efficient and fast, as well as to provide a highly accurate similarity measurement.

This research was supported in part by the fund from the National Science Foundation (NSF).

The performance of the proposed music fingerprint, in terms of both accuracy and speed, will be evaluated within a classical music cover song identification application. The specific application is motivated by the need to identify a song in the presence of a number of different versions of the same song, especially in classical music. They might be recorded in different key or tempo, with different players or conductors, and perhaps even with different orchestration. The goal of the proposed music fingerprint is to capture and model the essential musical features that are robust to these variations. The rest of the paper will describe background features, the proposed music fingerprint, and evaluation experiments and results.

2. CHROMA-BASED FEATURE

In this work, we use the chroma features based on Shepard's helix model, which factorize the perception of frequency into *tone height* and *chroma* as follows [8].

$$f = 2^{h+c} \qquad h \in \mathbb{Z}, \;\; c \in [0, 1) \tag{1}$$

where h, c, and f represent tone height, chroma, and frequency, respectively. We can compute the *chromagram* by first performing a short-time power spectrum analysis,

$$x_c(t) = \sum_k s\left(t, 2^{c+k}\right) \tag{2}$$

where $s\left(t, 2^{c+k}\right)$ represents a short time power spectrum at time t. Appropriately quantizing the chroma into twelve levels yields a twelve dimensional vector $\mathbf{x}(t)$ that can closely match the Western chromatic pitch classes (A to G#). These quantized quantities are usually called chroma feature vectors, and are widely used in the music audio processing [2, 5, 7]. Each element of the vector represents the energy for the corresponding pitch class at the time instance t. In practice, since the power spectral analysis is performed on a short-time segment, the discrete short-time segment index number n is used instead of continuous time t. Therefore, $\mathbf{x}(n)$ represents the chroma feature vector at the n-th segment. Recently, Ellis and Poliner proposed a modified version of the chroma feature using a beat-synchronous analysis window instead of fixed length of analysis window [7]. The beat-synchronous analysis provides a flexible length analysis window at the beat level, an analysis which is expected to be robust to tempo variation. This is based on their previous work on beat detection algorithm [9]. We adopt their beat synchronous approach to generate the music fingerprint proposed in this paper.

3. MUSIC FINGERPRINT

As pointed out earlier, in the design of music fingerprints, it is desirable to have less memory and complexity as well as higher accuracy in capturing the unique characteristics of the music for the target application. Toward the goal, we propose a simple covariance matrix of beat-synchronous chroma feature vectors as a music fingerprint, i.e.

$$\mathbf{\Phi} = E\left[\left(\mathbf{x} - E\left[\mathbf{x}\right]\right)\left(\mathbf{x} - E\left[\mathbf{x}\right]\right)^T\right] \tag{3}$$

where T represents the matrix transpose.

Fig. 1 shows an example of the fingerprint. Since it is the covariance matrix of the chroma feature vectors, each element of the

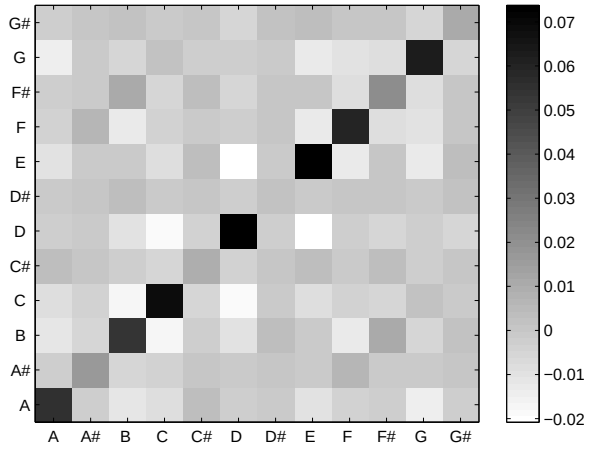

Fig. 1. An example of a music fingerprint using chroma feature vectors (BWV 772)

matrix is an energy-related quantity. Consequently, the diagonal elements of the covariance matrix represent the degree of presence of each pitch class in terms of energy. Furthermore, each column of the covariance matrix denotes the degree of co-presence of each pitch class with a given pitch class. Since the co-presence of two or more pitch classes represents the harmony information, it reveals the harmony structure of the music.

The covariance matrix of the chroma-based features, however, only captures static information at the beat level. To model the dynamic temporal information in the music, we adopt the delta feature computation idea from automatic speech recognition systems. For simplicity, we consider only one feature vector from the immediately adjacent beat to model the temporal information.

$$\Delta\mathbf{x}(n) = \mathbf{x}(n+1) - \mathbf{x}(n) \;\; . \tag{4}$$

Since the feature vectors are extracted in a beat-synchronous way, the delta feature represent the dynamics between two consecutive beats. Fig. 2 shows examples of the chroma and delta chroma feature vectors. In this figure, the bottom half represents the chroma feature vectors, and the upper half, the delta chroma feature vec-

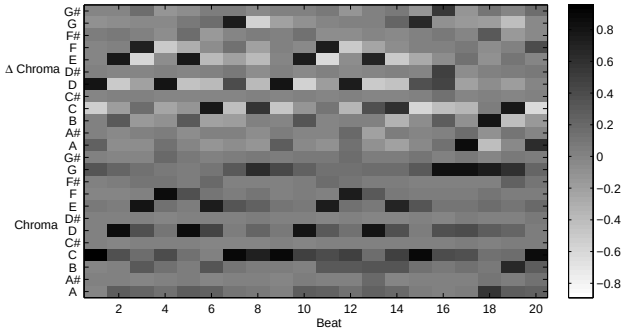

Fig. 2. An example of delta chroma feature vectors (BWV 772).

tors. For the delta chroma feature vectors, the positive and negative values denote the on-set intensity and release intensity, respectively, between two consecutive beats.

We construct a super-vector, which consists of the chroma feature vector and the delta chroma feature vector, to build the music fingerprint. Computing the covariance matrix of the super-vector is similar to the previous one:

$$\mathbf{\Phi}_{\Delta} = E\left[\left(\mathbf{x}_{\Delta} - E\left[\mathbf{x}_{\Delta}\right]\right)\left(\mathbf{x}_{\Delta} - E\left[\mathbf{x}_{\Delta}\right]\right)^{T}\right] \tag{5}$$

where

$$\mathbf{x}_{\Delta} = \begin{bmatrix} \mathbf{x} \\ \Delta\mathbf{x} \end{bmatrix}. \tag{6}$$

Fig. 3 shows an example of the music fingerprint using a super-vector of the chroma and the delta chroma features. Since the length of the vector is doubled, the size of the fingerprint is four times bigger. Note that the bottom-left third quadrant is the same with the fingerprint using only chroma feature vectors. To analyze the fingerprint in detail, let us focus on the first quadrant which represents the covariance matrix of the delta chroma feature vectors. These quantities are more related to temporal changes in intensity rather than energy distribution since delta chroma features encapsulate the temporal changes between the time segments. Diagonal elements of the quadrant denote the degree of temporal changes in intensity of individual pitch classes; The elements of each column represent the degree of temporal changes that happen with a given pitch class simultaneously. Positive values represent on-set events for the corresponding pitch class, while negative values represent release events.

When we look at the second or fourth quadrant, we can observe the cross-covariance matrix between the chroma feature vectors and the delta chroma feature vectors. Each vector describes the movements of the notes that follow with respect to a given current note. If the value is positive, there is a tendency of on-set of the note after the given note. If it is negative, there is a tendency of release of the note after the given note. If the value is close to zero, there is no crucial movement on the note after the given note.

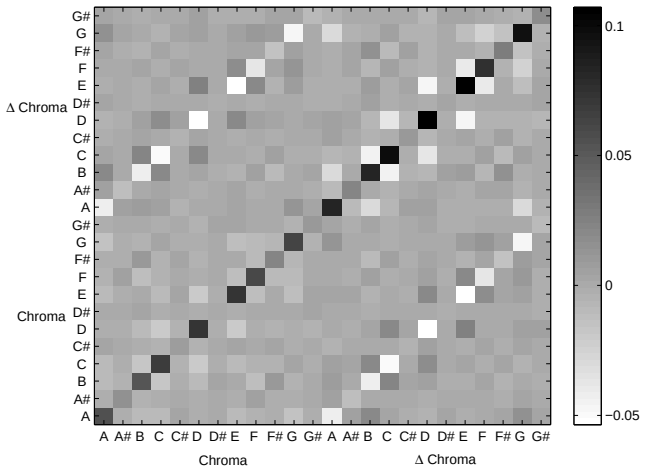

Fig. 3. An example of a music fingerprint using super-vectors of chroma and delta chroma feature vectors (BWV 772)

4. EXPERIMENTS

4.1. Experimental Setup

The performance of the proposed music fingerprint was evaluated in the context of automatic cover song identification where the goal is to determine the different versions of a given song. In our experimental database, 107 distinct pieces of classical music composed by Bach, Mozart, Brahms, Vivaldi, and Chopin were used. They were originally recorded in the MIDI format, and the audio signal for each was generated using Timidity++ toolkit [10] to have 16kHz sampling rate. Each piece of music has two different versions with possible changes of tempo, orchestration, and key. We use one of the two versions as a query, and the other as a reference.

For the baseline system, we use the system from LabRosa at Columbia University developed by Ellis *et al.* [7]. As it described earlier in Section 1, this system is based on the cross-correlation of beat-synchronous chroma feature vectors to measure the similarity between two different pieces of music and placed first at the MIREX 2006.

4.2. Similarity Measure

We use a simple template matching to measure the similarity of the two candidate music fingerprints. The similarity between music i and j is computed as follows.

$$s_{ij} = \sum_{k}\sum_{l} \phi_{kl}^{(i)}\phi_{kl}^{(j)} \quad , \tag{7}$$

where ϕ_{kl} represents the k-th row and l-th row element of the music fingerprint $\mathbf{\Phi}$. It should be noted that even in playing the same music, it is possible to transpose the key of the music. To compensate for the possible key transposition, we circularly shift one of the fingerprints in the diagonal direction by one semi-tone step to get the maximum similarity value.

$$s_{ij} = \max_{m} \sum_{k}\sum_{l} \phi_{kl}^{(i)}\phi_{kl}^{m(j)} \qquad ; 0 \le m \le 11 \ , \tag{8}$$

where

$$\phi_{kl}^{m} = \phi_{mod((k+m)/12)mod((l+m)/12)} \tag{9}$$

and $mod(\cdot)$ represents the modulus of the division. In case of using delta chroma features, the shifting process is done separately in each quadrant.

Since the music fingerprint contains energy-related quantities, it is crucial to normalize the covariance matrix appropriately to obtain a proper similarity measure. The choice of the normalization scheme itself depends on what kind of information the application emphasizes.

$$s_{ij} = \max_{m} \sum_{k}\sum_{l} N\left(\phi_{kl}^{(i)}\right) N\left(\phi_{kl}^{m(j)}\right) \ , \tag{10}$$

where $N(\cdot)$ represents the chosen normalization algorithm. In this paper, we used two different types of normalization schemes: an overall normalization (ON) and a column-wise normalization (CN). Overall normalization (ON) considers the overall energy distribution of each note and their co-presence by dividing their squared sum in the fingerprint, i.e.,

$$N\left(\phi_{kl}\right) = \frac{\phi_{kl}}{\sqrt{\sum_{m}\sum_{n}\left(\phi_{mn}\right)^{2}}} \ . \tag{11}$$

On the other hand, column-wise normalization (CN) lays emphasis on the harmony structure of the music by dividing the squared sum (energy) in the column of the fingerprint, i.e.,

$$N\left(\phi_{kl}\right) = \frac{\phi_{kl}}{\sqrt{\sum_{m}\left(\phi_{ml}\right)^{2}}}\,. \tag{12}$$

4.3. Results and Discussion

Table 1 shows accuracy and a search speed of the system compared with the baseline system. The results show that the proposed music fingerprint significantly improves the performance in terms of both complexity as well as accuracy. The proposed approach accelerates the search speed by approximately up to 60 times with a 30% relative accuracy improvement. It is also notable that a significant accuracy difference exists depending on the choice of the normalization scheme. The fingerprint method with column-wise normalization outperforms the overall normalization method, discerned more from the harmony structure than the overall note energy distribution.

Table 2 shows the performance of the music fingerprint with super-vectors of chroma feature and delta chroma feature in terms of accuracy and search speed. Although the usage of the delta feature can improve the accuracy over the conventional system, the search time is more than doubled. This is also true for the proposed fingerprint approach, but the total searching time is still affordable at the cost of some accuracy improvement. We can get about 40% relative accuracy improvement, and still offer a factor of 20 speed up compared to the conventional system. The results also show that the column-wise normalization outperforms the overall normalization even with delta chroma feature vectors.

Improvements in the storage memory requirement, another important factor in the design of music fingerprints, can be also shown. While the conventional system needs to store whole chroma feature vectors for each song, the proposed system needs to store only the music fingerprint itself whose size is 576 Byte (assuming *double* type) for each song. Even in the case of using delta chroma features, it only requires 1728 Byte.

Table 1. Performance of the fingerprint in terms of accuracy and search speed.

	[7]	Fingerprint w/ ON (11)	Fingerprint w/ CN (12)
Accuracy (%)	59.6	68.6	**80.7**
Approx. Searching Time (sec)	386	6	6

Table 2. Performance of the fingerprint with super-vectors of chroma and delta chroma feature vectors in terms of accuracy and search speed.

	[7]	[7] w/ $\mathbf{x}_{\Delta}$	FP-ON w/ $\mathbf{x}_{\Delta}$	FP-CN w/ $\mathbf{x}_{\Delta}$
Accuracy (%)	59.6	65.1	77.1	**85.3**
Approx. Searching Time (sec)	386	845	23	23

5. CONCLUSIONS AND FUTURE WORK

We proposed a music fingerprint approach for classical music cover song identification. The proposed music fingerprint not only outperforms the conventional state-of-the-art cover song identification system in terms of accuracy, but also requires very low memory and computing power. It was able to reduce the search time significantly, by up to 60 times, and improve accuracy by 40% relatively. These features of the proposed music fingerprint are promising because it can be implemented in portable devices and personal computers expending very limited computing power and memory.

From the results, we showed that the harmony structure of individual note contains more information about a song's identity than the overall note distribution. We also showed that the temporal dynamic information plays an important role in identifying the cover song. In future work, we will explore the usefulness of the proposed music fingerprint in other related applications such as automated composer, genre, and mood classification, assuming that the musical information compacted in the proposed music fingerprint can provide with those desirable discriminative information.

6. REFERENCES

[1] J. Stephen Downie, "The Music Information Retrieval Evaluation eXchange (MIREX)", *D-Lib Magazine*, Vol. 12, No. 12, 2006.

[2] E. Unal, E. Chew, P. Georgiou, S. Narayanan, "Challenging Uncertainty in Query-by-Humming Systems: A Fingerprinting Approach," *Special Issue of the IEEE transaction on Audio, Speech and Language Processing on Music Information Retrieval (MIR)*, Vol. 16, No. 2, 2008.

[3] J. Haitsma, T. Kalker, "A highly robust audio fingerprinting system," *International Symposium on Music Information Retrieval (ISMIR)*, 2002.

[4] M.I. Mandel, D.P.W. Ellis, "Song-level features and SVM for music classification," *International Symposium on Music Information Retrieval (ISMIR)*, 2006.

[5] K. Lee, "Identifying cover songs from audio using harmonic representation," *International Symposium on Music Information Retrieval (ISMIR)*, 2006.

[6] E. Unal, S. Narayanan, "Statistical modeling and retrieval of polyphonic music," *International workshop on Multimedia Signal Processing (MMSP)*, 2007.

[7] D.P.W. Ellis, G.E. Poliner "Identifying 'cover songs' with chroma features and dynamic programming beat tracking," *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2007.

[8] R.N. Shepard, "Circularity in judgments of relative pitch," *Journal of the Acoustic Society of America*, Vol. 36, No. 12, 1964.

[9] D.P.W. Ellis, "Beat tracking with dynamic programming," *International Symposium on Music Information Retrieval (ISMIR)*, 2006.

[10] http://timidity.sourceforge.net/