A Robust Harmony Structure Modeling Scheme for Classical Music Opus Identification

Samuel Kim, Panayiotis G. Georgiou, and Shrikanth Narayanan

Signal Anlaysis and Interpretation Lab. (SAIL)
University of Southern California, Los Angeles, USA.
kimsamue, georgiou, shri@sipi.usc.edu

***Abstract*—A robust algorithm to model the harmony structure of a music piece is proposed. The harmony structure is extracted directly from a music audio signal using a second-order statistic of chroma feature vectors. The method is experimentally shown to be robust against the degradation of chroma feature vectors due to noisy pitch estimation in our classical music opus identification evaluation. To analyze the effects of the noisy pitch estimation, we propose a noise model that describes difference between the oracle chroma feature vectors as obtained from a symbolic representation and those extracted from the rendered audio signal. The results suggest that the harmony structure modeling scheme employing the covariance matrix is more robust than the alternative investigated second-order statistics. The results also show that the proposed method obtains 84.3% accuracy with the symbolic representations and 72.0% with the synthesized audio data, which suggest that the proposed harmony structure modeling method has room for further improvement by addressing the signal processing challenges of pitch extraction, or through employing more robust features.**

***Index Terms*—music information retrieval, music fingerprint, multipitch analysis, polyphony music signal.**

I. Introduction

An important aspect of music information retrieval (MIR) systems is the extraction and processing of relevant musical descriptions according to target applications. For some applications, such as music transcription, melody extraction, and rhythm detection, retrieving the musical attributes is the main goal, while other applications use the acquired musical attributes for their ultimate goals; genre classification, artist identification, and mood classification are good examples. Consequently, these applications which utilize the acquired musical attributes as features of a given piece of music should devise algorithms both to extract useful musical attributes and to compare the derived features as needed by the application.

In this work, we aim to model the *harmony structure* of a music piece, and utilize it as a discriminant feature in applications that employ music similarity; specifically we focus on classical music opus identification. The goal of the classical music opus identification is to identify the same classical music recorded under different conditions, such as different players, tempo, instruments, and even with different orchestrations. It is very similar to cover song identification which is one of the evaluation categories in *music information retrieval evaluation exchange* (MIREX), an annual evaluation at the *International Music Information Retrieval System Evaluation Laboratory* (IMIRSEL) [1]. While the term 'cover song' is used in a broad sense including pop music, the classical music opus identification in this work is restricted to classical music. This specific application is motivated by the fact that there are many recordings of the same classical music piece.

Excellent cover song identification efforts have been introduced at the MIREX evaluation,that also enabled a fair comparison between various systems (see [2] for an overview). Most studies, however, require considerable computation to build sophisticated models and to compute the similarities. Although Jensen proposed a time trajectory filtering scheme to mitigate the complexity of the direct cross-correlation method [3], designing the filter bank is still a heuristic process. Recently, we proposed a novel music fingerprint extraction algorithm that captures various musical attributes, such as harmony structure and temporal dynamics [4], [5]. The metric is musically meaningful, as well as it provides a simple and powerful similarity measure; the experimental results showed that the proposed music fingerprint is efficient in terms of both accuracy and complexity. The rationale behind the main idea is that different recordings of the same music have similar harmony structures.

Besides the computation problems, the signal processing challenges in extracting pitch information have not yet been addressed fully in published studies. Although it is usually difficult to exctract accurate pitch information due to various reasons such as spectral overlap of overtones especially in polyphonic and multi-timbre music audio signals ([6], [7] for a good overview of related work), many systems utilize the pitch information in various forms, such as melodies, chroma features, and chord representations.

The present paper addresses the effects of inaccurate pitch information presented in chroma features on deriving a robust harmony structure model based on the music fingerprint framework [4], [5]. We propose a noise model to provide an analytic approach to this problem, and investigate the implications of noise on music similarity measure especially in the context of classical music opus identification. To evaluate this study, it is crucial to have the relevant ground truth information for the music audio which is difficult to establish easily in practice. As an intermediate step in that direction, therefore, we utilize MIDI data and their corresponding synthesized audio signals, which enables us to have the ground truth information of the audio signals.

II. Chroma Feature Vector

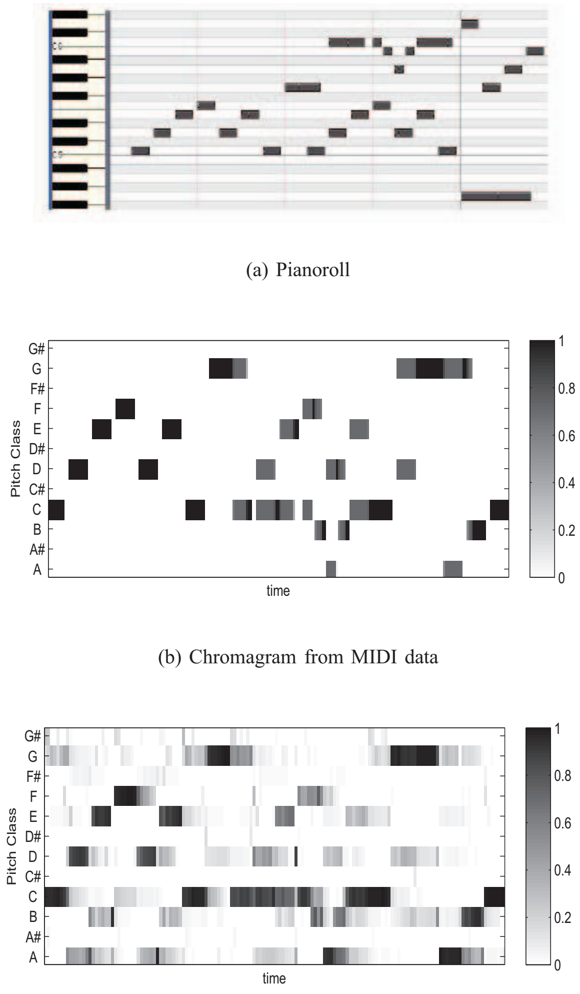
The chroma feature describes an energy distribution on the Western chromatic scale, and it is based on Shepard's helix model which factorizes the perception of frequency into *tone height* and *chroma* [8].

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

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,

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

where $s\left(t, 2^{c+k}\right)$ represents a short time power spectrum at time t. Quantizing the chroma c 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



(a) Pianoroll

(b) Chromagram from MIDI data

(c) Chromagram from synthesized audio

Fig. 1. An example of chromagram from MIDI data and synthesized audio signal compared with pianoroll (BWV772). The representation from audio is noisier compared with that obtained from MIDI.

called chroma feature vectors, and are widely used in music audio processing. Each element of the vector represents the energy for the corresponding pitch class at the time instance t.

However, as we described earlier, it is usually difficult to exctract accurate pitch information due to various reasons such as spectral overlap of overtones especially in polyphonic and multi-timbre music audio signals. To analyze the effects of the challenges, we extract the chroma feature vector from MIDI data assuming that MIDI data provide the ground truth information. We compute the time duration of the notes that correspond to the chromatic pitch classes at a time instant, i.e.,

$$\bar{x}_c(t) = \sum_e E_d \cdot I_e(c,t) \quad , \tag{3}$$

where E_d represents the duration of the e-th MIDI event within a short-time analysis window at time t. $I_e(c,t)$ is an indicator function for whether the e-th event in MIDI data corresponds with the chroma pitch class c at time t, i.e.,

$$I_e(c,t) = \begin{cases} 1 & E_s \leq t < E_e,\, E_c = c \\ 0 & \text{otherwise} \end{cases} \quad , \tag{4}$$

where E_s, E_e, and E_c represent the starting time, ending time, and the chroma pitch class value of the event, respectively. Since the chroma c is a discrete value in MIDI data, it can form a twelve dimensional vector $\bar{\mathbf{x}}(t)$ without quantization.

In practice, the analysis is performed on a short-time segment, and hence the discrete short-time segment index number n is used instead of continuous time t. Therefore, $\mathbf{x}[n]$ and $\bar{\mathbf{x}}[n]$ represent the chroma feature vector at the n-th segment for audio signal and MIDI data, respectively. In our previous work, we adopted a beat-synchronous segmentation method from [9] to investigate dynamics between consecutive beats. However, in this work, we utilize a fixed length segment since the beat detection itself may introduce artifacts toward similarity measure. We use a 64 ms analysis window along with a 32 ms overlap. Since the chroma feature vectors from audio signals and MIDI data are in different metrics, we normalize the chroma vectors to have unit norms.

Fig. 1 shows an example of chroma feature vectors from MIDI data and synthesized audio signal compared with the pianoroll of the MIDI data. In the figure, one can easily observe that the chroma feature vectors obtained from the synthesized audio signal introduce noise. For example, there are several non-zero quantities in the chroma vectors from the audio signal when the corresponding pitch class is not played (e.g. G, A, B, and C in the first 4 notes). This might be caused by the characteristics of overtones, which impose considerable amount of energy on perfect 5-th (7 chromatic interval) notes. Noise due to release-time differences can be also observed. Residual energy beyond MIDI events and vanishing energy during MIDI events are also evident. This might be caused by the fact that the release-time is dependent on the musical instruments. The chroma feature vectors from polyphonous and multiple-instrument audio signals would be even noisier than the given example which is nearly monophonic with one musical instrument.

III. Harmony structure

A. Harmony structure

In Western music, the term *harmony* represents the simultaneous use of different pitch classes. The way of building the harmony, i.e. *harmony structure*, is often governed by common practice period of western music, genre, and characteristics of composers. We hypothesize that the harmony structure is a unique feature characterizing a piece of music and that different recordings of the same music have similar harmony structures. To represent the relationship between individual pitch classes quantitatively, we proposed a covariance matrix of chroma feature vectors as a music fingerprint, i.e.,

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

where T represents the matrix transpose. This simple covariance matrix provides the relationship between individual pitch classes in terms of second-order statistics. From the music fingerprint, we could extract the various musical attributes, such as usage of pitch classes and individual harmony structures. In our previous work [4], we employed the covariance matrix among various second-order statistics, such as mean matrix and correlation matrix, to capture the harmony structure. The reason using the covariance matrix instead of other second-order statistics, however, has not been explicitly presented. In this section, we investigate how the alternatives fare in terms of how the noise due to the choice of specific signal

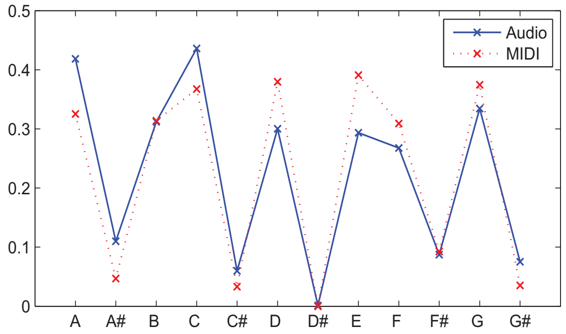


Fig. 2. An example of usage of pitch classes information; a comparison between music fingerprints from audio signal and MIDI data (BWV 772).

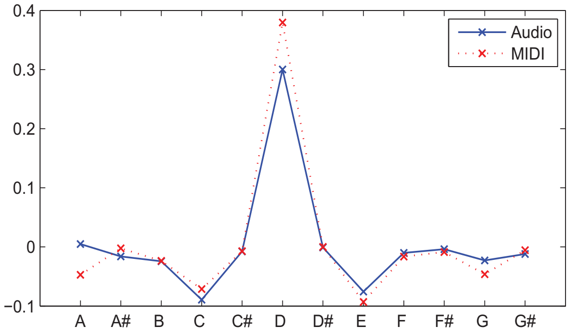


Fig. 3. An example of harmony structure information; a comparison between music fingerprints from audio signal and MIDI data (BWV 772).

representation affects the final decision toward classical music opus identification.

Fig. 2 and Fig. 3 explore the above ideas further by comparing the audio and the MIDI data with which the music audio is synthesized. They illustrate the global usage of pitch classes and the individual harmony structure, respectively. The dotted lines represent the values from MIDI data, and the solid lines represent the values from the audio signal. Although they both provide insights on the distribution prevalence of the various pitch classes and the individual harmony structure, the gap between the two lines shows the noise introduced in estimating chroma features directly from the audio.

Then, we empoly 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.

$$s_{ij} = \sum_k \sum_l \phi_{kl}^{(i)} \phi_{kl}^{(j)} \quad , \tag{6}$$

where ϕ_{kl} represents the k-th row and l-th row element of the music fingerprint $\mathbf{\Phi}$. Greater value represents higher similarity between two pieces of music.

B. Noise model

We propose a noise model describing how the noise introduced in the feature extraction procedure in dealing with audio signals affects the final decision of the opus identification. Although minimizing the noise from the signal processing challenges is outside of the scope of this work, the proposed noise model can provide a rigorous approach to analyze the effects of the noise toward the opus identification decision.

Suppose the chroma feature vector from the MIDI synthesized audio signal is corrupted by additive noise, i.e.,

$$\mathbf{x}[n] = \overline{\mathbf{x}}[n] + \varepsilon[n] \quad , \tag{7}$$

where $\overline{\mathbf{x}}$ and ε represent a chroma feature vector from MIDI data and a noise vector which can be observed in Fig. 1, respectively. Then, the music fingerprint can be represented as

$$\Phi = \overline{\Phi} + \Delta \tag{8}$$

where $\overline{\Phi}$ and Δ represent the music fingerprint from MIDI data and the noise matrix, respectively. Examples of the noise matrix are depicted in Fig. 2 and Fig. 3. In the proposed covariance matrix framework, the noise matrix can be written as

$$\Delta = 2\left\{E\left[\varepsilon\overline{\mathbf{x}}^T\right] - E\left[\varepsilon\right]E\left[\overline{\mathbf{x}}\right]^T\right\} + \left\{E\left[\varepsilon\varepsilon^T\right] - E\left[\varepsilon\right]E\left[\varepsilon\right]^T\right\}. \tag{9}$$

If other second-order statistics are utilized to model the harmony structure, those can be also easily derived: the correlation matrix can be written as

$$\Delta = 2\left\{E\left[\varepsilon\overline{\mathbf{x}}^T\right]\right\} + \left\{E\left[\varepsilon\varepsilon^T\right]\right\}. \tag{10}$$

And the mean matrix is

$$\Delta = 2\left\{E\left[\varepsilon\right]E\left[\overline{\mathbf{x}}\right]^T\right\} + \left\{E\left[\varepsilon\right]E\left[\varepsilon\right]^T\right\}. \tag{11}$$

Consequently, the similarity measure is written as

$$s_{ij} = \overline{s}_{ij} + \zeta \tag{12}$$

where $\overline{s}_{ij}$ and ζ denote the similarity measure using MIDI data and the noise factor, respectively. In other words, the noise due to the inaccurate chroma extraction is summarized in the similarity noise, and directly affects the classification decision ζ.

IV. Experiments and Results

A. Database

Recall that the goal of this work is to investigate the implications of the challenges in estimating accurate pitch information directly from music audio signals. For the purposes of evaluation, therefore, we utilize MIDI data and their corresponding synthesized audio signals, which enables us to have the ground truth pitch information of the audio signals, something that is difficult to obtain in practice otherwise.

In our database, there are about 2000 pieces of various classical music composers: Bach, Beethoven, Brahms, Chopin, Debussy, Handel, Haydn, Mozart, Schubert, Tchaikovsky, and Vivaldi (about 1000 songs and 2 variations of each song). They were originally recorded in the MIDI format [10], and the audio signal for each was generated using Timidity++ toolkit [11] at 16kHz sampling rate. The types of variations in recording the same music may vary; some pieces were recorded in different keys, some in different instruments or tempos, and others in different arrangements. The range of recording length is from 1 minute to 10 minutes, and the songs whose length exceed 10 minutes were truncated to 10 minutes for simplicity. We use one of the two versions as a query, and the other as a reference. For classification, we make a decision by the maximum similarity score among the reference data set.

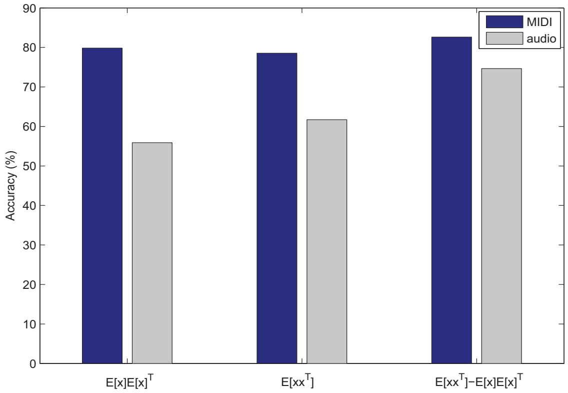


Fig. 4. Performance comparison between various second-order statistics using MIDI data and audio signal.

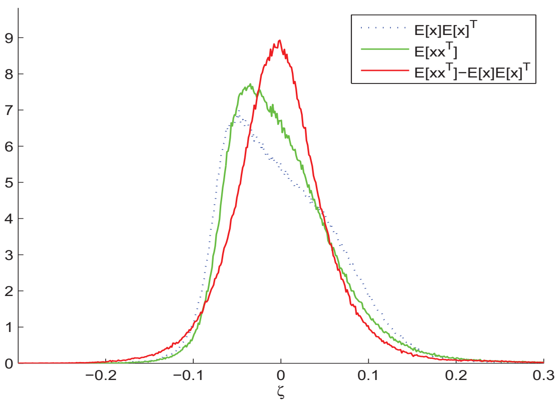


Fig. 5. Distribution of noise in similarity measure.

B. *Results and Discussion*

Fig. 4 illustrates the performance gaps in the various second-order statistics used to generate the music fingerprint: mean matrix and correlation matrix as well as the proposed covariance matrix. It also shows the performance gap between features extracted from MIDI data and their synthesized audio signals. This implicates the limitations of the signal processing involved. While the accuracy using MIDI data does not vary significantly, the accuracy using the audio signal highly depends on how the music fingerprints are generated. This implies that the effects of the noise in dealing with audio signals toward the opus identification vary depending on the modeling methods used for harmony structure.

According to the proposed noise model, the performance gaps are directly captured by ζ, which summarizes the noise in processing audio. Fig. 5 depicts the histogram-based probability density of ζ for each music fingerprint generation scheme (for better illustration, the average value is subtracted to make zero-mean distribution). Having very close standard deviation values, the distribution of the noise in the proposed method seems symmetric while the distribution of the others are skewed to negative values. This introduces more classification errors by adding biased noise to similarity measure.

The skewness is from the wrong assumptions embedded in the noise matrix model in (10) and (11). Compared with (9), the method using correlation matrix assumes $E[\varepsilon]E[\overline{\mathbf{x}}]^T = \mathbf{0}$ and $E[\varepsilon]E[\varepsilon]^T = \mathbf{0}$ which are equivalent to zero-mean signal processing noise. This assumption is not necessarily true in practical cases. The assumption embedded in the method using mean matrix is even more stronger and unrealistic. It assumes $E[\varepsilon\overline{\mathbf{x}}^T] = \mathbf{0}$ which is equivalent to saying $\overline{\mathbf{x}}$ and ε are orthogonal. It also assumes that the signal processing noise is uncorrelated with itself. As it is shown earlier in Fig. 1, however, the assumption is not valid in the given chroma feature extraction algorithm. The noise seems highly correlated with the corresponding pitch class (e.g. considerable amount of energy on the perfect 5-th pitch class of the played pitch class).

Although the proposed algorithm is shown to be fairly robust in dealing with the audio signal (72.0% accuracy), it also provides an idea that one can at least achieve 84.3% accuracy with the proposed harmony structure modeling scheme by solving signal processing challenges related to representation uncertainty.

V. Conclusion

A robust algorithm to model the harmony structure of a given music piece was proposed toward classical music opus identification. To analyze the challenges in estimating pitch information directly from audio signal, we introduce a noise model to describe the difference between the chroma feature vectors from MIDI data and from synthesized audio signals. In the proposed noise model framework, we experimentally showed that the harmony structure modeling scheme with covariance matrix is more robust against the noise in chroma features than other second-order statistics. The results also suggested that the proposed harmony structure modeling scheme has more room to improve by equipping a robust pitch information extracting algorithm.

Our future goal is to extend our analysis on the proposed algorithm with real recordings including pop music as well as classical music. Devising algorithms that reduce the noise in dealing with audio signal and model various musical attributes for improved performance will be studied as well.

References

[1] J. S. Downie, "The music information retrieval evaluation exchange (mirex)," *D-Lib Magazine*, vol. 12, no. 12, 2006.

[2] M. Casey, R. Veltkamp, M. Goto, M. Leman, C. Rhodes, and M. Slaney, "Content-based music information retirval: Current directions and future challenges," *Procedings of the IEEE*, vol. 96, no. 4, 2008.

[3] J. Jensen, M. Chistensen, D. Ellis, and S. Jensen, "A tempo-insensitive distance measure for cover song identification based on chroma features," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Processingl*, 2008.

[4] S. Kim, E. Unal, and S. Narayanan, "Music fingerprint extraction for classical music cover song identification," in *International Conference of Multimedia and Expo*, 2008.

[5] S. Kim and S. Narayanan, "Dynamic chroma feature vectors with applications to cover song identification," in *IEEE International Workshop on Multimedia Signal Processing*, 2008.

[6] S. Saito, H. Kameoka, K. Takahashi, T. Nishimoto, and S. Sagayama, "Specmurt analysis of polyphonic music signal," *IEEE transactions on Audio, Speech, and Language Processing*, vol. 16, no. 3, 2008.

[7] A. Klapuri, "Multipitch analysis of polyphonic music and speech signals using an auditory model," *Special Issue of the IEEE transaction on Audio, Speech and Language Processing on Music Information Retrieval*, vol. 16, no. 2, 2008.

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

[9] D. Ellis, "Beat tracking with dynamic programming," in *International Symposium on Music Information Retrieval*, 2006.

[10] "http://www.classicalarchives.com/."

[11] "http://timidity.sourceforge.net/."