DYNAMIC CHROMA FEATURE VECTORS
WITH APPLICATIONS TO COVER SONG IDENTIFICATION

Samuel Kim and Shrikanth Narayanan

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

ABSTRACT

A new chroma-based dynamic feature vector is proposed inspired by psychophysical observations that the human auditory system detects relative pitch changes rather than absolute pitch values. The proposed chroma-based dynamic feature vector describes the relative pitch change intervals. The utility of the proposed feature vector incorporated with a music fingerprint extraction algorithm is experimentally explored within a music cover song identification framework. The results with a Classical music database suggest that the proposed biologically plausible dynamic chroma feature vector can be successfully added to the conventional chroma feature vector as a complementary feature; it provides a 5.8% relative performance improvement.

1. INTRODUCTION

The chroma-based feature is one of the most popular features in the music information retrieval (MIR) domain. Based on Shepard's helix model, which factorizes the perception of frequency into pitch height and pitch chroma [1], this feature only considers chroma information. It represents an energy distribution over the Western pitch classes from A to G♯ for a given audio segment. Examples of applications that utilize the chroma-based feature include chord recognition, music segmentation, and similarity measure.

In our previous work, to capture dynamic information we proposed a delta chroma feature vector inspired by the delta mel frequency cepstral coefficient in speech recognition systems [2]. By capturing the dynamic information between the adjacent chroma feature vectors, we demonstrated that we could improve Classical music cover song identification accuracy where the goal is to identify the different performances of the same song. We offered a simple way to model the dynamic information of the chroma feature vectors at the feature level. The implications of the proposed dynamic feature, however, was not fully addressed in that work.

In the present paper, we introduce a dynamic chroma feature vector motivated by psychophysical observations. It is a well known fact that humans perceive or produce relative pitch changes with greater ease than absolute pitch values, and this characteristic has been utilized in several music information retrieval systems (e.g. query-by-humming systems [3]). This argument can be partially supported by results in neuroscience. In [4], Warren *et. al.* used a functional magnetic resonance imaging system to show the psychophysical effects of the pitch changes in the human brain by manipulating the pitch of a given signal. The results showed specific brain regions of activation attributed to pitch chroma changes: the pitch chroma change was represented in the anterior to primary auditory cortex, while the pitch height change was represented in the posterior to primary auditory cortex. These observations inspired us to explore the usefulness of dynamic chroma information.

In this work, the proposed algorithm attempts to model relative chroma changes in conjunction with the conventional chroma feature vectors. The utility of the proposed algorithm is experimentally explored within a Classical music cover song identification application. The rest of the paper will describe the baseline system, the proposed algorithm, and the experimental results.

2. BASELINE SYSTEM

In our previous work [2], we proposed a simple dynamic feature vector, i.e.,

$$\Delta\mathbf{c}\left[n\right] = \mathbf{c}\left[n+1\right] - \mathbf{c}\left[n\right] \ , \tag{1}$$

where $\mathbf{c}\left[n\right]$ represents a chroma feature vector extracted from music audio signal segment n, and the segmentation is done in a beat synchronous way [5]. By considering only one adjacent feature vector rather than applying a several-tap long FIR filter, we could obtain the dynamic information between the adjacent time segments, a popular method in feature representations for automatic speech recognition. Although it successfully models the intensity change in each pitch class (an example is shown in Fig. 1(b)), the relative pitch change is not modeled in this dynamic feature vector.

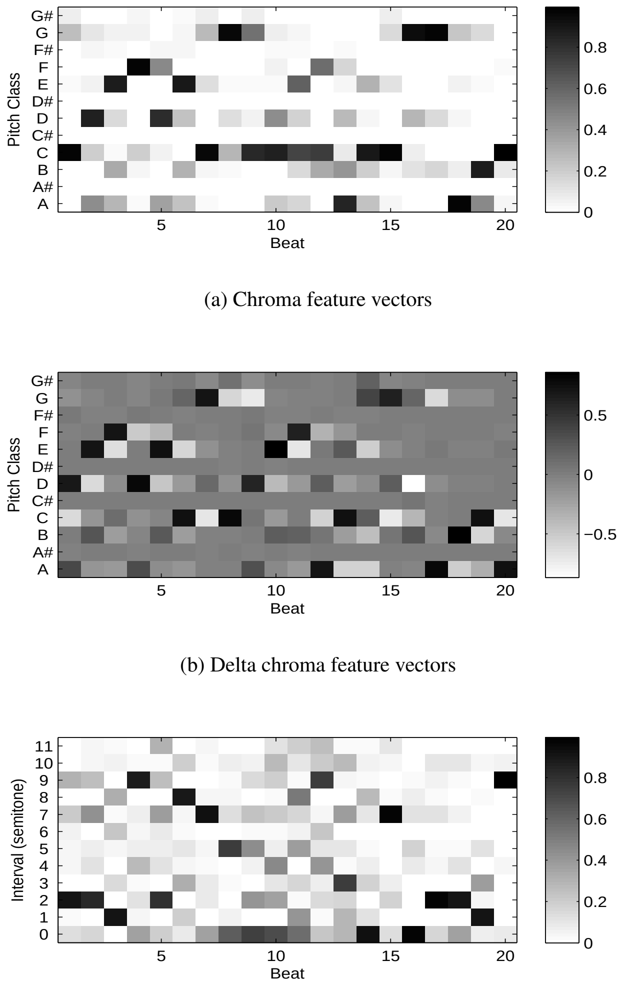
In [2], we also proposed a method to extract a music fingerprint that models the harmony structure of a given music piece. Specifically, we proposed a covariance matrix of chroma feature vectors and it was experimentally shown to encapsulate unique musical attributes successfully. When applied in the context of cover song identification (i.e., recognizing a music piece regardless of its performance variants), the method outperformed a conventional state-of-the-art system in terms of both accuracy and speed.

3. PROPOSED ALGORITHM

3.1. Proposed Delta Chroma Feature

We define chroma change as a relative interval between the pitch classes that are played sequentially in terms of semitone. For example, if the pitch class "D" is played after "C" is played, the relative chroma change interval is $+2$ semitones. A scalar value would

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



(a) Chroma feature vectors

(b) Delta chroma feature vectors

(c) Proposed delta chroma feature vectors

Fig. 1. Examples of chroma feature vectors, delta chroma feature vectors, and proposed delta chroma feature vectors extracted from music audio (BWV772).

represent the chroma change in case of a monophonic melody signal. In most cases, however, the music audio signal is polyphonic representing a mixture of multiple pitches from various instruments. It leads to multiple chroma changes at the same time. For example, if the pitch classes “C” and “G” are played simultaneously after “D” is played, the relative chroma change interval can be both -2 semitones and $+5$ semitones. To deal with the simultaneous multiple chroma changes, a vector representation is required. Here, we propose a new vector representation to describe the degree of chroma changes on all possible intervals.

Note that the magnitude of the delta chroma feature in (1), i.e. $\|\Delta \mathbf{c}[n]\|$, represents the Euclidean distance between two adjacent chroma feature vectors. It can be also interpreted as the likelihood of not sustaining the same pitch classes (zero interval chroma change); smaller the value it represents, the more likely the pitch classes move toward the zero interval chroma change (no change). In other words, if the value is close to zero, it is likely for the pitch classes to be retained as they are.

We can get similar quantities considering any chroma change interval i by circularly rotating the latter chroma feature vector, i.e.,

$$\left\|\Delta \mathbf{c}^i[n]\right\| = \left\|\mathbf{c}^i[n+1] - \mathbf{c}[n]\right\| \quad ;0 \le i \le 11\ , \tag{2}$$

where $\mathbf{c}^i$ represents the rotated vector $\mathbf{c}$ whose elements are circularly moved by i semitones. The value represents the unlikelihood of moving toward i chroma change interval. Similar to the zero chroma change interval case shown above, the smaller value it represents, the more likely the pitch classes move toward the i chroma change interval. For simplicity, we define the range of i as in the above equation (2). One should note that i is modulus of 12 so that a -2 interval can be interpreted as $+10$ interval and vice versa.

Based on the above quantities, we can define a new vector representation which describes the likelihood of moving toward individual chroma change intervals. Since the above quantities are unlikelihood, we need a reciprocal function that transforms unlikelihood to likelihood values. In this work, we simply put a negative sign and add the maximum value among the elements to make a vector whose elements are non-negative. Therefore, the proposed dynamic chroma feature vector can be written as

$$\nabla \mathbf{c}[n] = \{\nabla c_0[n], \nabla c_1[n], \cdots, \nabla c_{11}[n]\}^T\ , \tag{3}$$

where

$$\nabla c_i[n] = -\left\|\Delta \mathbf{c}^i[n]\right\| + C_{\max} \tag{4}$$

and

$$C_{\max} = \max_j \left\|\Delta \mathbf{c}^j[n]\right\|\ . \tag{5}$$

Fig 1 illustrates the examples of the chroma feature vectors, the delta chroma feature vectors, and the proposed biologically plausible delta chroma feature vectors of a given music clip. As seen in the figure, the proposed delta chroma feature shows the relative chroma change interval between the adjacent time segments while the conventional delta chroma feature shows the temporal dynamic information of each pitch class.

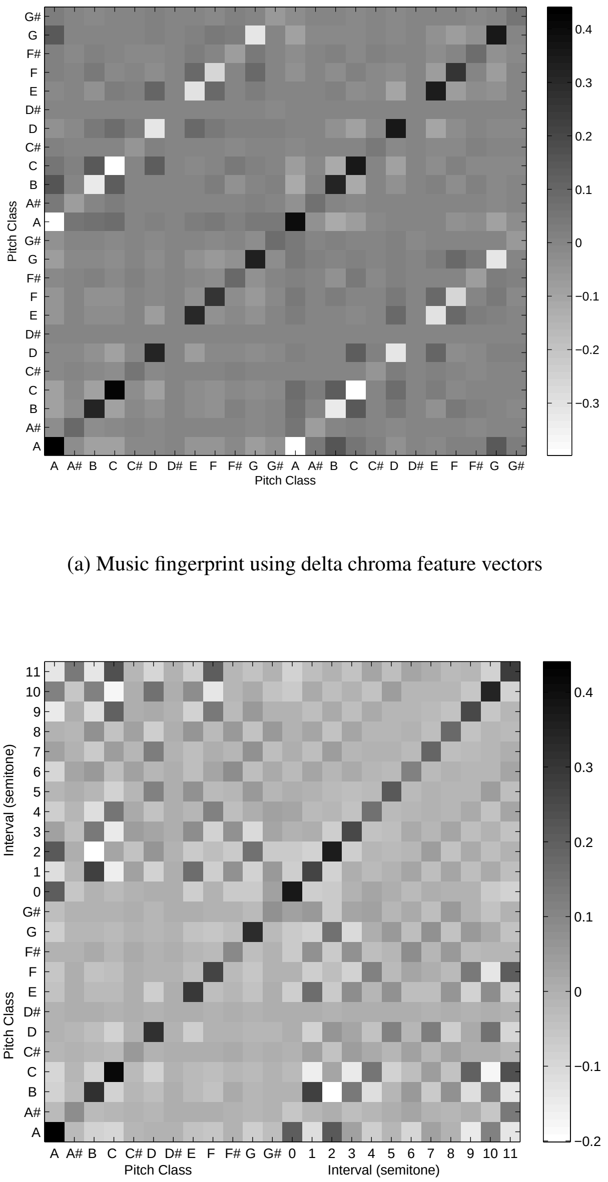
3.2. Music fingerprint

We adopt a music fingerprint extraction algorithm from our previous work [2] which utilizes a covariance matrix of the feature vectors, i.e.,

$$\mathbf{\Phi} = E\left[(\mathbf{x} - E[\mathbf{x}])(\mathbf{x} - E[\mathbf{x}])^T\right] \tag{6}$$

where T represents the matrix transpose. If the feature vector $\mathbf{x}$ is the chroma feature vector, the music fingerprint provides the relationship between individual pitch classes which leads to model the harmony structure of a given music piece. If the feature vector $\mathbf{x}$ is a super-vector of the chroma feature vector and the delta chroma feature vector, i.e., $\mathbf{x} = \left[\mathbf{c}^T\ \Delta\mathbf{c}^T\right]^T$, the music fingerprint compacts additional musical attributes, such as temporal change information after a certain pitch class is played [2].

If we use a super-vector of the chroma feature vector and the proposed delta chroma feature vector, i.e., $\mathbf{x} = \left[\mathbf{c}^T\ \nabla\mathbf{c}^T\right]^T$, the content of the music fingerprint is somewhat different compared to the one with conventional delta chroma feature vectors (see Fig. 2



(a) Music fingerprint using delta chroma feature vectors

(b) Music fingerprint using proposed delta chroma feature vectors

Fig. 2. Examples of music fingerprints using delta chroma feature vectors and proposed delta chroma feature vectors extracted from music audio (BWV772).

(a) and (b) for the examples). Firstly, the axes of the music fingerprint with the proposed delta chroma feature vectors consist of the pitch classes and the relative intervals rather than just pitch classes. In the first quadrant of the figure, the diagonal elements represent the intensities of chroma changes. Each vector in the first quadrant describes how the chroma changes happens simultaneously with the corresponding chroma change.

In the second quadrant of the figure (it is symmetric with the fourth quadrant), each vector illustrates which direction the chroma change happens after the corresponding pitch class is played. The greater the value is, the stronger the tendency for the pitch classes that are simultaneously played with the corresponding pitch class to move toward the corresponding interval exists. It is remarkable that the temporal dynamic information is modeled as a group of the notes that are simultaneously played with the corresponding pitch class. For example, after the pitch class "A" is played, there exists a tendency for the notes that are played with the pitch class "A" to be retained as they are or move toward 2 semitones above.

One should note that the third quadrant is the same with the one of conventional music fingerprint, since they both represent the covariance matrix of the chroma feature vectors.

3.3. Similarity Measure

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

$$s_{ij} = \sum_{k=1}^{24} \sum_{l=1}^{24} \phi_{kl}^{(i)} \phi_{kl}^{(j)} \ , \tag{7}$$

where $\phi_{kl}^{(i)}$ represents the k-th row and l-th column element of the music fingerprint $\mathbf{\Phi}$ of the music piece i. A greater value represents higher similarity between two pieces of music.

It should be also noted that it is possible to transpose the key of the music even in playing the same music. To compensate for possible transposition, we take the maximum similarity value among the possible key transpositions. In the conventional delta chroma feature vectors, we circularly move the each quadrant in diagonal direction. However, with the proposed delta chroma feature vector, special care should be paid to deal with the possible transposition. Since the chroma change interval is a relative value and independent of the key, it should not be moved during the compensation process. Therefore, the first quadrant should be retained as it is and the second (or fourth) quadrant should circularly move to the right (or upper) direction to compensate for possible key difference.

4. EXPERIMENTAL RESULTS

4.1. Database

In our experimental database, there are approximately 2000 pieces of various classical music composers: Bach, Beethoven, Brahms, Chopin, Debussy, Handel, Haydn, Mozart, Schubert, Tchaikovsky, and Vivaldi (Approx. 1000 songs and 2 variations of each song). They were originally recorded in the MIDI format [6], and the audio signal for each was generated using Timidity++ toolkit [7] at 16kHz sampling rate. The length of the songs varies from 1 minute to 5 minutes, and the songs whose length exceeds 5 minutes were truncated to 5 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, i.e.,

$$\arg\max_j s_{ij}. \tag{8}$$

4.2. Results and Discussion

Table 1 shows the cover song identification accuracy according to the types of feature. It illustrates the performance improvement by using the temporal dynamic information in the chroma feature vectors. With the given database, utilizing the proposed delta chroma feature vector can boost the system accuracy by 5.8% relative improvement, while the conventional dynamic feature improves the performance by 2.5% relative improvement. It indicates that the dynamic information can offer supplementary features to capture the unique characteristics of the given music piece especially, using the proposed delta chroma feature vectors. See also the results of $\mathbf{\Phi}_{3rd}$ and $\mathbf{\Phi}_{all}$ in Table 2 which represent the same results in Table 1.

Table 2 provides an in-depth analysis by showing the system accuracies of the cases when we only use each quadrant to measure the similarity. Since individual quadrants include different aspects of the musical attributes, we perform cover song identification tasks using the individual quadrants. In the table, the subscript of the music fingerprint $\mathbf{\Phi}$ indicates that we use the corresponding quadrant of the music fingerprint for the task so that we can investigate the effects of the musical attributes embedded in the corresponding quadrant. For example, the task with $\mathbf{\Phi}_{1st}$ implies that we use the fist quadrant of the music fingerprint to analyze the effects of dynamic features only.

The results using $\mathbf{\Phi}_{1st}$ show that the dynamic chroma feature vectors by themselves are not as efficient as the chroma feature vectors in capturing the unique characteristics of a given music piece. Making a super-vector, however, with the chroma feature vector and its dynamic feature vector improves the accuracy as it was shown in Table 1. The proposed method outperforms the conventional method, although the proposed delta feature is worse than the conventional delta feature in the case of the dynamic feature vectors only.

The results using the second (or fourth) quadrant, i.e., $\mathbf{\Phi}_{2nd}$ (or $\mathbf{\Phi}_{4th}$), show the performance of the cross-covariance matrix between chroma feature vectors and their dynamic feature vectors. The task with the proposed method outperforms the case using the chroma feature vectors only as well as the case using the conventional delta feature vectors, while the performance with the conventional delta method is lower than the case using the chroma feature vectors only.

The experimental results suggest that the temporal dynamic information modeled by the proposed delta chroma feature vector can represent unique characteristics of a given music piece along with the chroma feature vector as a complementary feature.

Table 1. Cover song identification accuracy according to types of feature.

	Chroma	Delta	Proposed Delta
Accuracy (%)	74.7	76.6	**79.0**
Relative Improvement (%)	·	2.5	**5.8**

Table 2. Cover song identification accuracy according to types of feature and used quadrant.

	Delta	Proposed Delta
$\mathbf{\Phi}_{1st}$	**62.2**	51.7
$\mathbf{\Phi}_{2nd}$ or $\mathbf{\Phi}_{4th}$	69.5	**75.2**
$\mathbf{\Phi}_{3rd}$	74.7	74.7
$\mathbf{\Phi}_{all}$	76.6	**79.0**

5. CONCLUSION

We introduced a new biologically plausible dynamic chroma feature vector that models the relative chroma change intervals. Various musical attributes modeled in the music fingerprint within the proposed dynamic feature vectors were described. The experimental results with a Classical music cover song identification task indicated that the proposed dynamic chroma feature vector can be successfully adopted as a complementary feature.

In future work, we will investigate the effects of musical attributes embedded in the music fingerprint so that we can devise an algorithm to fuse those musical attributes in an optimal way. We will also explore additional music information retrieval applications, such as composer identification and music segmentation, with the proposed delta chroma feature vectors.

6. REFERENCES

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

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

[3] E. Unal, E. Chew, P. Georgiou, and 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.

[4] D. Warren, S. Uppenkamp, R. D. Patterson, and T. D. Griffiths, “Separating pitch chroma and pitch height in the human brain,” *Proc. Natl. Acad. Sci. U.S.A.*, vol. 100, no. 17, 2003.

[5] D. Ellis and G. Poliner, “Identifying ‘cover songs’ with chroma features and dynamic programming beat tracking,” in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2007.

[6] “http://www.classicalarchives.com/.”

[7] “http://timidity.sourceforge.net/.”