MAKING CHROMA FEATURES MORE ROBUST TO TIMBRE CHANGES

Meinard Müller∗
Saarland University and MPI Informatik
Campus E1 4, D-66123 Saarbrücken

Sebastian Ewert,† Sebastian Kreuzer
Universität Bonn, Informatik III
Römerstr. 164, D-53117 Bonn

ABSTRACT

Chroma-based audio features are a well-established tool for analyzing and comparing music data. By identifying spectral components that differ by a musical octave, chroma features show a high degree of invariance to variations in timbre. In this paper, we describe a novel procedure for making chroma features even more robust to changes in timbre and instrumentation while keeping their discriminative power. Our idea is based on the generally accepted observation that the lower mel-frequency cepstral coefficients (MFCCs) are closely related to timbre. Now, instead of keeping the lower coefficients, we will discard them and only keep the upper coefficients. Furthermore, using a pitch scale instead of a mel scale allows us to project the remaining coefficients onto the twelve chroma bins. Our systematic experiments show that the resulting chroma features have indeed gained a significant boost towards timbre invariance.

Index Terms— Chroma feature, MFCC, timbre-invariance, audio matching, music retrieval

1. INTRODUCTION

One main goal of content-based music analysis and retrieval is to reveal semantically meaningful relationships between different music excerpts contained in a given data collection. Here, the notion of similarity used to compare different music excerpts is a delicate issue and largely depends on the respective application. In particular, for detecting harmony-based relations, chroma features have turned out to be a powerful mid-level representation for comparing and relating music data in various realizations and formats [2, 4, 5, 7, 8]. Chroma-based audio features are obtained by pooling a signal’s spectrum into twelve bins that correspond to the twelve pitch classes or chroma of the equal-tempered scale. Identifying pitches that differ by an octave, chroma features show a high degree of robustness to variations in timbre and are well-suited for the analysis of Western music which is characterized by a prominent harmonic progression [2]. In particular, such features are useful in tasks such as cover song identification [4, 8] or audio matching [5, 7], where one often has to deal with large variations in timbre and instrumentation between different versions of a single piece of music.

In this paper, we present a method for making chroma features even more robust to changes in timbre while keeping their discriminative power as needed in matching applications. Here, our general idea is to discard timbre-related information expressed by certain mel-frequency cepstrum coefficients (MFCCs). More precisely, recall that the mel-frequency cepstrum is obtained by taking a decorrelating cosine transform of a log power spectrum on a logarithmic mel scale [6]. A generally accepted observation is that the lower MFCCs are closely related to the aspect of timbre [1, 9]. Therefore, intuitively spoken, one should achieve some degree of timbre-invariance when discarding exactly this information. As our main contribution, we combine this idea with the concept of chroma features by first replacing the nonlinear mel scale by a nonlinear pitch scale. We then apply a cosine transform on the logarithmized pitch representation and only keep the upper coefficients, which are finally projected onto the twelve chroma bins to obtain a chroma representation. The technical details of this procedure are described in Sect. 2. We report on two experiments showing that our novel chroma features indeed have gained a significant boost towards timbre invariance. We first describe an experiment based on audio data systematically synthesized by different instruments (Sect. 3). Then, using real audio data, we show how our novel features improve the matching quality between harmonically-related music excerpts contained in different versions and arrangements of the same piece of music (Sect. 4). Conclusions and prospects on future work are given in Sect. 5.

2. FEATURE DESIGN

In this section, we present the technical details for our novel audio features, see Fig. 1 for an overview. As front end transform, the audio signal is decomposed into 120 frequency bands corresponding to the MIDI pitches 1 to 120 using a suitable multirate filter bank. We then take the short-time mean-square power (local energy) for each of the 120 subbands by convolving the squared subband signals with a
The inverse DCT and projected onto the twelve chroma bins to obtain a 12-dimensional chroma vector. Finally, all chroma vectors are normalized to have unit length. The resulting audio features are referred to as CRP \((n)\) (chroma DCT-reduced log pitch) features, see Fig. 1.

As illustration, we consider the second Waltz of the Jazz Suite No. 2 by Shostakovich, which also serves as running example in the subsequent sections. The theme of this piece appears four times played in four different instrumentations (strings, clarinet, trombone, tutti). Furthermore, there are significant differences between the four themes with respect to secondary voices. Due to these differences, the resulting conventional chromagrams may strongly deviate from each other. This is illustrated by Fig. 2 (a) and (b) showing the conventional chromagrams of the theme’s beginning of the first (strings) and third (trombone) excerpt in an interpretation by Yablonsky. Contrary, the corresponding two CRP \((55)\) chromagrams as shown in (c) and (d) coincide to a much larger degree.

3. EXPERIMENTS ON CHORD CHROMA CLASSES

We quantitatively compared our CRP \((n)\) features for various parameters \(n \in [1 : 120]\) with some commonly used chroma types including two freely available chroma implementations by Ellis [3] (Chroma-IF-Ellis, Chroma-P-Ellis) as well as the conventional chroma features described in Sect. 2 (Chroma-Pitch). In all cases, the feature resolution was roughly 10 Hz and all chroma vectors were normalized. The various chroma types will serve as baseline to illustrate the boost of robustness achieved by CRP \((n)\) features.

In our first experiment, we used systematically synthesized audio material. To this end, we created a MIDI file containing all possible single pitches (1-chords), duads (2-chords) and triads (3-chords) within a fixed octave. This resulted in \(12 + \binom{12}{2} + \binom{12}{3} = 220\) chords. The MIDI file was then synthesized in 24 different ways using eight different instruments each playing the file in three different octaves. Here, we used the software Cubase in combination with a high quality sample library. Fixing a specific chroma type, we converted each of the resulting 24 audio files into a chromagram. Next, for each of the 220 chords we formed a class consisting of 48 chroma vectors—one representative chroma vector within the attack and one within the sustain phase of each of the 24 realizations of the respective chord. The classes are referred to as chord chroma classes. The distance between two normalized chroma vectors was computed using the cosine distance \((1 - \langle \cdot, \cdot \rangle)\).

Now, disregarding timbre and dynamics, any two chroma vectors within a chord chroma class are considered as similar, whereas two chroma vectors from different classes are considered as dissimilar. To measure the degree of timbre invariance of a given chroma type, we computed the distances between any two chroma vectors that belong to the same

---

**Fig. 1.** Overview of the computation of the CRP (chroma DCT-reduced log pitch) features.

**Fig. 2.** Various chromagrams of the theme’s beginning of the second Waltz, Jazz Suite No. 2 by Shostakovich. (a)/(b): Conventional chromagram of string/trombone version. (c)/(d): CRP \((55)\) chromagram of string/trombone version. All chroma vectors are normalized.
<table>
<thead>
<tr>
<th>Chroma type</th>
<th>$\mu_O$</th>
<th>$\mu_I$</th>
<th>$\rho$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chroma-II-Ellis</td>
<td>0.66</td>
<td>0.35</td>
<td>1.88</td>
</tr>
<tr>
<td>Chroma-P-Ellis</td>
<td>0.42</td>
<td>0.18</td>
<td>2.38</td>
</tr>
<tr>
<td>Chroma-Pitch</td>
<td>0.75</td>
<td>0.23</td>
<td>3.26</td>
</tr>
<tr>
<td>CRP(35)</td>
<td>0.99</td>
<td>0.10</td>
<td>10.30</td>
</tr>
<tr>
<td>CRP(55)</td>
<td>1.00</td>
<td>0.10</td>
<td>9.83</td>
</tr>
<tr>
<td>CRP(75)</td>
<td>1.00</td>
<td>0.11</td>
<td>8.76</td>
</tr>
</tbody>
</table>

Table 1. Performance of several chroma types in the experiments on chord chroma classes.

4. EXPERIMENTS BASED ON AUDIO MATCHING

Our second experiment was conducted on real audio data and is motivated by an application referred to as audio matching: given a short query audio clip, the goal is to automatically retrieve all musically (harmonically) similar excerpts in different versions and arrangements of the same underlying piece of music [5, 7]. We will compare the CRP($n$) features with conventional chroma features by means of several performance measures that express the matching quality.

As basis for the matching procedure, we use a distance function locally comparing a query sequence with a given database sequence. Let $X = (X(1), X(2), \ldots, X(K))$ and $Y = (Y(1), Y(2), \ldots, Y(L))$ be the feature sequences of the query and the database, respectively. (In our case, the features $X(k), k \in [1 : K]$, and $Y(\ell), \ell \in [1 : L]$, are normalized chroma vectors.) Then, we define a distance function $\Delta : [1 : L] \rightarrow \mathbb{R} \cup \{\infty\}$ between $X$ and $Y$ using a variant of dynamic time warping (DTW):

$$\Delta(\ell) := \frac{1}{K} \min_{a \in [1 : L]} \left( \text{DTW}(X, Y(a : \ell)) \right),$$

(1)

where $Y(a : \ell)$ denotes the subsequence of $Y$ starting at index $a$ and ending at index $\ell \in [1 : L]$. Furthermore, $\text{DTW}(X, Y(a : \ell))$ denotes the DTW distance between $X$ and $Y(a : \ell)$ with respect to a suitable local cost measure (in our case, the cosine distance). For details on DTW and the distance function, we refer to [7]. The interpretation of $\Delta$ is as follows: a small value $\Delta(\ell)$ for some $\ell \in [1 : L]$ indicates that the subsequence of $Y$ starting at frame $a_\ell$ (with $a_\ell \in [1 : \ell]$ denoting the minimizing index in (1)) and ending at frame $\ell$ is similar to $X$.

Having this interpretation in mind, the two following properties of $\Delta$ are of crucial importance in view of the audio matching application. First, the semantically correct matches (in the following referred to as true matches) should correspond to local minima of $\Delta$ close to zero (thus avoiding false negatives). We capture this property by defining $\mu_X^\ell$ and $\max_X^\ell$ to be the average respective maximum of $\Delta$ over all indices that correspond to the local minima of the true matches for a given query $X$. Second, $\Delta$ should be well above zero outside a neighborhood of the desired local minima (thus avoiding false positives). Here, we define $\mu_O^\ell$ and $\min_O^\ell$ to be the average respective minimum of $\Delta$ over all indices outside these neighborhoods. From the above, it is clear that $\mu_X^\ell$ and $\max_X^\ell$ should be small whereas $\mu_O^\ell$ and $\min_O^\ell$ should be large. Similar to Sect. 3, we express these two properties within a single number, respectively, by defining the quotients $\mu_X^\ell = \mu_X^\ell / \mu_X^\ell$ and $\min_X^\ell = \max_X^\ell / \min_X^\ell$.

We illustrate the definition of $\Delta$ by means of our Shostakovich example from Sect. 2. Suppose our database consists of two interpretations (Yablonsky, Chailly) of the Waltz. Recall that the theme appears four times in the piece. Let $E_1$ to $E_5$ denote the corresponding excerpts in the first and $E_6$ to $E_8$ in the second recording. Now, using $E_3$ (trombone) as query, one has eight true matches. Using conventional chroma features, seven of the eight matches (except of $E_3$) are indeed indicated by local minima of the resulting distance function $\Delta$, see the gray curve of Fig. 3. However, due to the above mentioned differences in timbre, most of these local minima are not well developed and have relatively large $\Delta$-values. Now, using our CRP($n$) features, one obtains for all eight true matches (even for $E_3$) much more concise local minima, see the black curve of Fig. 3. This demonstrates that the choice of a chroma type has a significant impact on the final matching quality.

In order to quantitatively evaluate the CRP($n$) features for various parameters $n \in [1 : 120]$ and the other chroma types, we generated a database consisting of 31 audio recordings of 12 different pieces comprising classical and popular music by Bach, Shostakovich, Wagner, Queen, Genesis, Beatles, and others. For each piece there are at least two different recordings typically comprising the original version and an arrangement (e.g., piano version of an orchestral piece) or cover song. For each piece, we picked an excerpt and manually annotated all musically similar excerpts within all the recordings. These excerpts are the true matches when using any of these excerpts as query. For example, the database contains two versions (Yablonsky, Chailly) of the Shostakovich example with the above mentioned 8 annotated excerpts. Altogether, we annotated 90 excerpts with an average length of 30 seconds. We used any of these excerpts as query, computed the distance functions over all database audio files, and derived the values $\mu_X^\ell$, $\max_X^\ell$, $\mu_O^\ell$, $\min_O^\ell$, $\rho_X^\ell$, and $\rho_O^\ell$. Averaging over all 90 queries, we obtain the corresponding numbers $\mu_I$, $\max_I$, $\mu_O$, $\min_O$, $\rho_I$, and $\rho_O$. Note that $\rho_I$ is not the quotient of...
In a final experiment, we computed for each $n \in [1 : 120]$ the performance measure $\rho_{\text{min}}$ using the corresponding CRP($n$) features. The resulting curve, which is shown in Fig. 4, indicates that one obtains the best separation between true and spurious matches for parameters $n \in [22 : 60]$.

### 5. CONCLUSIONS AND FUTURE WORK

In this paper, we introduced a new type of chroma feature, which shows a higher degree of robustness to changes in timbre than conventional chroma features. Using our novel CRP features, one can significantly improve the performance in matching and classification applications, where one wants to be invariant to instrumentation and tone color. Actually, the essence of this improvement is best explained by Fig. 3. For the future, we plan to apply CRP features for various tasks in music information retrieval. We will also further explore and improve CRP features. Here, first experiments indicate that one may further reduce the number of coefficients without a degradation of the discriminative power.

### 6. REFERENCES


![Fig. 3. Different distance functions shown for two recordings (Yablonsky, Chailly) of the Shostakovich example using the excerpt $E_3$ as query. The following chroma types were used: Chroma-IF-Ellis (thin gray), Chroma-Pitch (bold gray) and CRP(55) (black). For the query, there are 8 annotated excerpts.](image)

![Fig. 4. Dependence of the performance measure $\rho_{\text{min}}$ on the parameter $n \in [1 : 120]$ using CRP($n$) features.](image)

<table>
<thead>
<tr>
<th>Chroma type</th>
<th>$\mu_{\text{O}}$</th>
<th>$\mu_{\text{I}}$</th>
<th>$\mu$</th>
<th>$\rho_{\text{I}}$</th>
<th>$\rho_{\text{min}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chroma-IF-Ellis</td>
<td>0.41</td>
<td>0.25</td>
<td>0.16</td>
<td>0.20</td>
<td>2.74</td>
</tr>
<tr>
<td>Chroma-P-Ellis</td>
<td>0.19</td>
<td>0.11</td>
<td>0.06</td>
<td>0.08</td>
<td>3.15</td>
</tr>
<tr>
<td>Chroma-Pitch</td>
<td>0.48</td>
<td>0.26</td>
<td>0.15</td>
<td>0.20</td>
<td>3.84</td>
</tr>
<tr>
<td>CRP(35)</td>
<td>0.68</td>
<td>0.31</td>
<td>0.14</td>
<td>0.18</td>
<td>5.81</td>
</tr>
<tr>
<td>CRP(55)</td>
<td>0.64</td>
<td>0.27</td>
<td>0.12</td>
<td>0.16</td>
<td>6.24</td>
</tr>
<tr>
<td>CRP(75)</td>
<td>0.57</td>
<td>0.21</td>
<td>0.10</td>
<td>0.14</td>
<td>5.97</td>
</tr>
</tbody>
</table>

Table 2. Performance of several chroma types in the experiments based on audio matching.