Towards Timbre-Invariant Audio Features for Harmony-Based Music

Meinard Müller, *Member, IEEE,* and Sebastian Ewert, *Student Member, IEEE*

***Abstract*—Chroma-based audio features are a well-established tool for analyzing and comparing harmony-based Western music that is based on the equal-tempered scale. By identifying spectral components that differ by a musical octave, chroma features possess a considerable amount of robustness to changes in timbre and instrumentation. In this paper, we describe a novel procedure that further enhances chroma features by significantly boosting the degree of timbre invariance without degrading the features' 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 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. We present a series of experiments to demonstrate that the resulting chroma features outperform various state-of-the art features in the context of music matching and retrieval applications. As a final contribution, we give a detailed analysis of our enhancement procedure revealing the musical meaning of certain pitch-frequency cepstral coefficients.**

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

I. 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 [1], [2], [3]. 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 [1]. In particular, such features are useful in tasks such as music synchronization [4], [5], [6], [7], [8], [3], audio structure analysis and summarization [1], [9], [10], [11], [12], [13], [14], cover song identification [15], [16], [17], or music matching [18], [19], [3], [20], 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 similar to that expressed by certain mel-frequency cepstral coefficients (MFCCs). More precisely, recall that the mel-frequency cepstrum is obtained by taking a discrete cosine transform (DCT) of a log power spectrum on the logarithmic mel scale [21]. A generally accepted observation is that the lower MFCCs are closely related to the aspect of timbre [22], [23]. Therefore, intuitively spoken, one should achieve some degree of timbre-invariance when discarding exactly this information. As one main contribution of this paper, we combine this idea with the concept of chroma features by first replacing the nonlinear mel scale with a nonlinear pitch scale. We then apply a DCT on the logarithmized pitch representation to obtain *pitch-frequency cepstral coefficients* (PFCCs). We then only keep the upper coefficients, apply an inverse DCT, and finally project the resulting pitch vectors onto twelve-dimensional chroma vectors. These vectors are referred to as CRP (c̲hroma DCT-r̲educed log p̲itch) features. The technical details of this procedure are described in Sect. II. The gist of our novel features is illustrated by Fig. 4, which shows two different types of chromagrams for two musically related audio excerpts that differ significantly in instrumentation. Note that the two chromagrams based on conventional chroma features ((a) and (b) of Fig. 4) look rather different, whereas the chromagrams based on our novel CRP features ((c) and (d) of Fig. 4) are quite similar thus indicating the boost towards timbre invariance.

Our novel audio features constitute a valuable tool in all retrieval, matching, and classification applications where one is interested in blending out musical details related to timbre and instrumentation. In particular, we show the potential of our concept by means of the audio matching scenario based on the query-by-example paradigm as introduced in [24]. Given a short query audio clip, the goal is to automatically retrieve all excerpts from all recordings within a given audio collection that musically correspond to the query. Here, one typically has to cope with variations in timbre and instrumentation as they appear in different interpretations, cover songs, and arrangements of a piece of music. As another contribution of this paper, we use the audio matching procedure to systematically evaluate the matching and separation capability of different types of audio features. Among others, we introduce various

M. Müller is with the Saarland University and the Max-Planck Institut für Informatik, 66123 Saarbrücken, Germany (e-mail: meinard@mpi-inf.mpg.de). He is funded by the Cluster of Excellence on Multimodal Computing and Interaction (MMCI).

S. Ewert is with the Multimedia Signal Processing Group, Department of Computer Science III, Bonn University, 53117 Bonn, Germany (e-mail: ewerts@iai.uni-bonn.de). He is funded by the German Research Foundation (DFG CL 64/6-1).

Manuscript received December 15, 2008; revised October 21, 2009.

quality measures that indicate how well semantically correct matches are separated from spurious matches. As it turns out, these quality measures are also good indicators for the degree of timbre invariance exhibited by the respective feature type.

We have conducted a series of experiments to compare our novel CRP features with previously suggested chroma features as well as to reveal the role of the various parameters and measures involved in our procedure. Among others, we investigate the role of the feature rate and the number of coefficients to be pruned as well as the influence of amplitude compression and spectral whitening. Furthermore, we discuss two different cost measures used to compare the resulting features including the binary shift measure introduced in [17]. As one main result, we are able to show that our procedure is conceptually different to previous feature enhancement strategies in the sense that it yields a significant boost towards timbre invariance independent of a particular choice of parameters and measures.

As a final contribution, we analyze in detail the DCT-based reduction step in our enhancement procedure. As it turns out, the most dominant of the upper PFCCs capture interpretable pitch periodicities, whereas PFCCs surrounding the dominant ones account for different phases. We show that a reduction based on only a few relevant DCT basis vectors along with suitable phase-shifted duplicates results in a similar feature enhancement as the strategy of using the entire range of upper DCT basis vectors, see Sect. V. This observation reveals the musical meaning of certain pitch-frequency cepstral coefficients.

The remainder of the paper follows the outline given above. In Sect. II, we start with our main contribution by introducing the novel CRP features and by describing in detail the involved signal processing steps. Then, in Sect. III, we give a short description of the audio matching application, which also lies the foundation for various quality measures used to compare and evaluate the different feature types. In Sect. IV, we report on a series of experiments discussing the influence of various parameters on the quality of the resulting CRP features. Finally, in Sect. V, we research into the underlying principles that achieve the boost towards timbre invariance for harmony-based Western music. Conclusions and prospects on future work are given in Sect. VI. A discussion of related work and further references are given in the respective sections.

II. Feature Design

In this section, we describe our enhancement procedure that allows for increasing the robustness of chroma features to changes in timbre and instrumentation while keeping their discriminative power. To this end, we combine and modify various techniques known from the design of chroma features and mel-frequency cepstral coefficients (MFCCs) in a novel way. In Sect. II-A, we review chroma features and MFCCs and then, in Sect. II-B, go into the technical details of our procedure. Finally, in Sect. II-C, we report on a first baseline experiment conducted on systematically generated audio material.

A. Related Work

Chroma-based audio features are a well-established tool in the music retrieval context [1], [2], [16], [25], [3]. Assuming the equal-tempered scale, the chroma correspond to the set $\{\mathrm{C}, \mathrm{C}^\sharp, \mathrm{D}, \ldots, \mathrm{B}\}$ that consists of the twelve pitch spelling attributes as used in Western music notation. Note that in the equal-tempered scale different pitch spellings such $\mathrm{C}^\sharp$ and $\mathrm{D}^\flat$ refer to the same chroma. A chroma vector can be represented as a 12-dimensional vector $v = (v(1), v(2), \ldots, v(12))$, where $v(1)$ corresponds to chroma C, $v(2)$ to chroma $\mathrm{C}^\sharp$, and so on. A normalized chroma vector $v/||v||_2$ expresses the signal's local energy distribution among the 12 pitch classes, where $||\cdot||_2$ denotes the Euclidean norm. Chroma features account for the well-known phenomenon that human perception of pitch is periodic in the sense that two pitches are perceived as similar in "color" if they differ by an octave [1]. Normalized chroma features can absorb a significant degree of variations in parameters such as dynamics, timbre, as well as articulation and closely correlate to the short-time harmonic content of the underlying audio signal.

There are various ways of computing chroma-based audio features, e. g., by suitably pooling Fourier coefficients obtained from one or several spectrograms [1], [16], [2] or by using constant-Q [26] and multirate filter bank techniques [3], [24]. The properties of the resulting chroma features, sometimes also referred to as *pitch class profiles* (PCPs), depend on many design choices. Integrating a weighting scheme into the coefficient pooling can increase the robustness to noise [16], [2]. Considering harmonics additionally to the fundamental frequency has an influence to the robustness of chroma features to timbre [2]. Furthermore, preprocessing steps in the chroma computation based on spectral whitening [27], the estimation of the instantaneous frequency [16], or peak picking of spectrum's local maxima [2] may have a significant impact on features's quality. Generalized chroma representations with 24 or 36 bins (instead of the usual 12 bins) allow for dealing with differences in tuning [2].

In our experiments, we illustrate the enhancement capabilities of our novel CRP features by comparing them against several state-of-the-art chroma implementations including three chroma implementations (Chroma-IF, Chroma-P, Chroma-E) by Ellis[1], one implementation (Chroma-QM) developed at the Centre for Digital Music, Queen Mary, University of London[2], as well as one implementation (Chroma-MIR) contained in the MIR toolbox[3]. The Chroma-E implementation is based on a Gaussian weighted pooling of magnitude spectrum coefficients. Its extension, Chroma-P, additionally implements a simple spectral peak picking to reduce spectral noise. In the more complex Chroma-IF variant, spectral regions of uniform instantaneous frequency are estimated to separate tonal components from noise. The instantaneous frequency information is also used to account for tuning differences. The fourth implementation, Chroma-MIR, is derived from the magnitude spectrum using a decibel scale. The Chroma-QM

[1] http://www.ee.columbia.edu/~dpwe/resources/matlab/chroma-ansyn/

[2] http://www.vamp-plugins.org

[3] http://www.jyu.fi/hum/laitokset/musiikki/en/research/coe/materials/

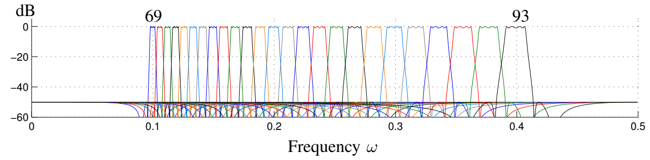

Fig. 1. Magnitude responses in dB for some of the pitch filters (corresponding to MIDI pitches $p \in [69 : 93]$ with respect to a sampling rate of 4410 Hz) of the multirate pitch filter bank used for the chroma computation.

implementation uses the magnitude of the constant-Q transform as described in [26]. For further details and applications of the various chroma variants, we refer to the literature [26], [2], [16], [28], [29]. The exact parameters used with these implementations are given on a separate website[4].

In some sense, MFCC features, which are closely related to the aspect of timbre, can be considered as kind of complementary to chroma features. Originally, MFCCs were developed for speech processing applications [21], [30] and have then found their way into the music domain [31], where they have been used for various music analysis tasks including genre classification [32] and musical instrument recognition [33]. In most implementations, the mel-frequency cepstrum is obtained in the following way. First the power spectrum of the signal is computed using a short-time Fourier transform. Then, to account for properties of the human auditory system, the resulting coefficients are pooled into 20 to 40 nonlinearly spaced frequency bins along the perceptually motivated mel frequency scale [30]. Similarly, a musically motivated frequency scale is used in [34]. Finally, after taking the logarithm on the bin values, a discrete cosine transform (DCT) is applied to yield the MFCCs. A generally accepted observation is that the lower MFCCs are closely related to the aspect of timbre [22], [23]. Therefore, intuitively spoken, one should achieve some degree of timbre-invariance when discarding exactly this information. This is the basic idea of our enhancement procedure to be described next.

B. CRP Feature Computation

We now describe in detail all steps needed to compute our novel CRP audio features. For an overview of these steps, we refer to Fig. 2.

Instead of using a mel-frequency scale, our features are based on a pitch-frequency scale. We first decompose the audio signal into 88 frequency bands with center frequencies corresponding to the MIDI pitches $p = 21$ to $p = 108$ (which correspond to the keys of a piano). To properly separate adjacent pitches, we need filters that possess relatively wide passbands around the respective center frequencies, while having a sharp cutoffs in the transition bands and high rejections in the stopbands. In order to design a set of filters satisfying these stringent requirements for all MIDI notes in question, we work with three different sampling rates: 882 Hz for the low subbands $p = 21, \ldots, 59$ (MIDI notes A0-B3), 4410 Hz for the middle subbands $p = 60, \ldots, 95$ (MIDI notes C4-B6), and 22050 Hz for the high subbands $p = 96, \ldots, 108$ (MIDI notes C7-C8). Working with different sampling rates also takes into account that in the analysis of lower frequencies the time resolution naturally decreases. Each of the 88 filters is realized as eighth-order elliptic filter with 1 dB passband ripple and 50 dB rejection in the stopband. To separate the notes, we use a Q factor (ratio of center frequency to bandwidth) of $Q = 25$ and a transition band of half the width of the passband. Fig. 1 shows the magnitude response of some of these filters. For further details, we refer to [3, Section 3.1]. To compensate the large phase distortions inherent to elliptic filters, we use the standard technique known as forward-backward filtering, which can be applied in offline scenarios where the audio signals are entirely known prior to computations. After filtering in the forward direction, the filtered signal is reversed and run back through the filter. The resulting output signal has precisely zero phase distortion and a magnitude modified by the square of the filter's magnitude response. Since the magnitude responses of our filters are close to those of ideal bandpass filters, squaring does not have a large effect on the magnitude response. For further details we refer to standard text books on digital signal processing such as [35]. Furthermore, note that our filters are robust to deviations of up to ± 25 cents[5] from the respective note's center frequency thanks to the relatively wide passbands. This introduces a significant degree of robustness to slight changes in tuning. To cope with tuning deviations of more than 25 cents, one has to revert to automated tuning estimation and compensation techniques, see [2], [17].

In the next step, we compute the short-time mean-square power (local energy) for each of the 88 squared subbands (i. e., the samples of each subband output are squared) using a rectangular window of a fixed length and an overlap of 50%. For example, a window length corresponding to 1 second leads to a feature rate of 2 Hz (2 features per second). In Sect. IV-E, we discuss the role of the feature rate in more detail. As a result, we obtain a sequence of 88 dimensional feature vectors where the entries correspond to MIDI pitches $p = 21$ to $p = 108$. For later usage, we extend each such vector by suitably adding zeros (20 at the beginning and 12 at the end) to obtain a 120 dimensional feature vector where the entries now correspond to MIDI pitches $p = 1$ to $p = 120$. The resulting sequence of pitch vectors is referred to as *pitch representation*, see Fig. 3(a) for an illustration.

To obtain a conventional chroma representation or chromagram, one adds up the corresponding values of the pitch representation that belong to the same chroma yielding a 12-dimensional vector for each analysis window, see Fig. 4 for an illustration. In the following, the resulting features are referred to as *Chroma-Pitch*.

For our novel audio features, we further process the pitch representation before doing the chroma binning. The steps are similar to the ones in the computation of MFCCs. First, the pitch representation is logarithmized, see Fig. 3(b). Here, we replace each entry e by the value $\log(C \cdot e + 1)$, where C is a suitable positive constant. Such a logarithmic compression is conducted to account for the logarithmic sensation of sound

[4] http://www-mmdb.iai.uni-bonn.de/projects/CRP.html

[5] The *cent* is a logarithmic unit to measure musical intervals. The semitone interval of the equally-tempered scale equals 100 cents.

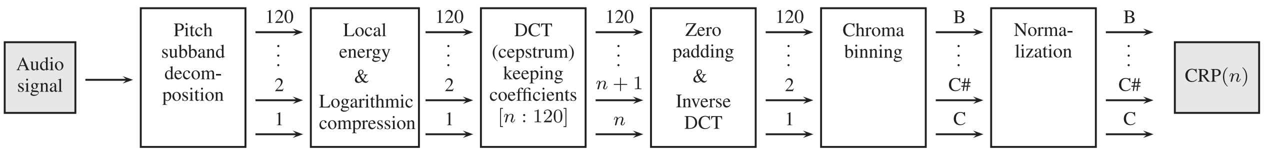
Audio signal
Pitch subband decom-position
120
2
1
Local energy & Logarithmic compression
120
2
1
DCT (cepstrum) keeping coefficients [n : 120]
120
n + 1
n
Zero padding & Inverse DCT
120
2
1
Chroma binning
B
C#
C
Norma-lization
B
C#
C
CRP(n)

Fig. 2. Overview of the steps in the computation of the CRP (chroma DCT-reduced log pitch) features.

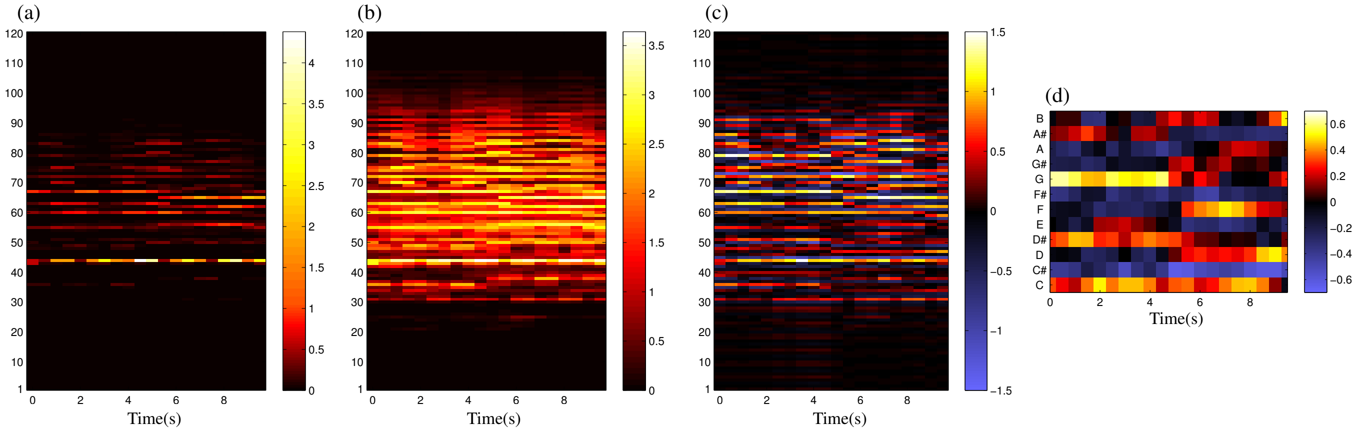

Fig. 3. Various feature representations of the passage E_3 (trombone part in the Yablonsky recording of the Shostakovich Waltz) illustrating the steps in the CRP feature computation. **(a)**: Pitch representation. **(b)**: Pitch representation after the logarithmic compression. **(c)**: Pitch representation after the DCT reduction step keeping coefficients $[55 : 120]$. **(d)**: $\mathrm{CRP}(55)$ features.

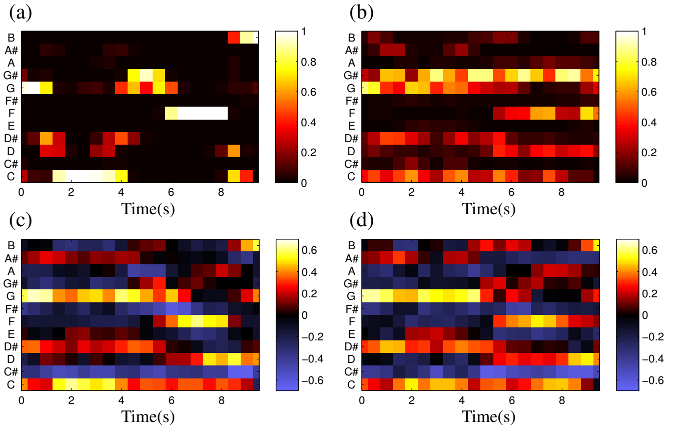

Fig. 4. Various chromagrams of the passages E_1 (clarinet) and E_3 (trombone) in the Yablonsky recording of the Shostakovich Waltz. **(a)**/**(b)**: Conventional chromagram of E_1/E_3. **(c)**/**(d)**: $\mathrm{CRP}(55)$ chromagram of E_1/E_3. All chroma vectors are normalized w.r.t. the Euclidean norm.

intensity [31], [36] and was also used in a similar way in [37]. The role of the parameter C, which is set to $C = 100$ in most of our experiments, is discussed in Sect. IV-D.

Next, we apply a discrete cosine transform (DCT) to each of the 120-dimensional logarithmized pitch vectors resulting in 120 coefficients, which are referred to as *pitch-frequency cepstral coefficients* (PFCCs). The PFCCs have a similar interpretation as the MFCCs. In particular, the lower coefficients are related to timbre as observed by various researchers, see, e. g., [22], [23]. Now, our goal of achieving timbre-invariance is the exact opposite of the goal of capturing timbre. Therefore, we discard the information given by the lower $n-1$ PFCCs for a parameter $n \in [1 : 120]$ by setting them to zero while leaving the upper PFCCs unchanged. Each resulting 120-dimensional vector is then transformed by the inverse DCT to yield an enhanced 120-dimensional pitch vector, see Fig. 3(c). The role of the parameter n is discussed in Sect. IV-C. Furthermore, in Sect. V, we analyze the reduction step in detail and derive a musically meaningful explanation responsible for the final enhancement.

In the last stage, the entries of each enhanced pitch vector are projected onto the twelve chroma bins to yield a 12-dimensional chroma vector. Finally, the chroma vectors are normalized with respect to the Euclidean norm to have unit length. The resulting audio features are referred to as $\mathrm{CRP}(n)$ (chroma DCT-reduced log pitch) features, see Fig. 3(d).

In the experiments to be described, we show that the resulting CRP features have indeed gained a significant amount of robustness to changes in timbre and instrumentation. As a first illustrative example, 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 occurs four times played in four different instrumentations (clarinet, strings, trombone, tutti). Furthermore, there are also significant differences between the four themes with respect to secondary voices. In the considered recording of this piece by Yablonsky, the four occurrences of the theme are referred to as E_1 (5-26), E_2 (39-59), E_3 (129-149), and E_4 (160-180), where the brackets indicate the start and end times in seconds of the respective passage. Fig. 4(a) and (b) show conventional chromagrams of the passages E_1 (theme played by clarinet) and E_3 (theme played by trombone), respectively. Note that the two chromagrams strongly deviate from each other due to large differences in instrumentation and voicing. Contrary, the corresponding two $\mathrm{CRP}(55)$ chromagrams as shown in (c) and (d) of Fig. 4 coincide to a much larger degree.

C. Baseline Experiments on Chord Chroma Classes

To illustrate the boost of robustness achieved by our CRP features, we now report on a first baseline experiment conducted on systematically generated audio material. For the moment, we fix certain parameters using a feature rate of 2 Hz, setting $C = 100$ in the logarithmic compression, and considering only the case $n = 55$ in CRP(n). For a detailed analysis of these parameters we refer to Sect. IV, where we report on extensive experiments based on real audio material. We compare the resulting $\mathrm{CRP}(55)$ features with various publicly available implementations of state-of-the-art chroma feature types (Chroma-IF, Chroma-P, Chroma-E, Chroma-QM, Chroma-MIR), which were described in Sect. II-A. Furthermore, we used the conventional chroma features (Chroma-Pitch) obtained from our pitch representation as described in Sect. II-B. For all chroma implementations, we used similar parameters settings and rates [6]. Furthermore, all chroma features were normalized with respect to the Euclidean norm.

To indicate the degree of timbre-invariance of the various chroma implementations, we proceeded as follows. First, 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 with a size of more than 50GB. Fixing a specific feature 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 24 chroma vectors—one representative chroma vector from 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 $1 - \langle \cdot, \cdot \rangle$ (also referred to as cosine distance). Note that the entries of CRP features may be negative, so that the distance between two normalized CRP vectors lies in the range $[0, 2]$.

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 feature type, we computed the distances between any two chroma vectors that belong to the same chord chroma class. Let μ_I be the mean and σ_I the standard deviation over the resulting $220 \cdot \binom{24}{2}$ distances. Note that μ_I should be small in the case that the feature type has a high degree of timbre invariance. Similarly, let μ_O be the mean and σ_O the standard deviation over the distances of any two chroma vectors from different chord chroma classes. Note that μ_O should be large to indicate a high discriminative power of a feature type. Finally, we formed the quotient $\delta := \mu_\mathrm{I}/\mu_\mathrm{O}$ which expresses the within-class distance μ_I relative to the across-class distance μ_O. Note that a small value of δ is desirable in view of our evaluation.

TABLE I
QUALITY OF SEVERAL FEATURE TYPES IN THE EXPERIMENTS ON CHORD CHROMA CLASSES.

Feature type	μ_I	σ_I	μ_O	σ_O	δ
Chroma-IF	0.299	0.193	0.654	0.188	0.457
Chroma-P	0.174	0.133	0.464	0.160	0.374
Chroma-E	0.168	0.129	0.452	0.159	0.373
Chroma-MIR	0.107	0.078	0.268	0.137	0.398
Chroma-QM	0.124	0.098	0.396	0.146	0.313
Chroma-Pitch	0.232	0.194	0.749	0.161	0.309
$\mathbf{CRP(55)}$	**0.078**	**0.069**	**1.002**	**0.131**	**0.077**

Table I shows the values μ_I, μ_O, and δ for various feature types. Note that for our $\mathrm{CRP}(55)$ features the within-class distance ($\mu_\mathrm{I} = 0.078$) is much smaller while the across-class distance ($\mu_\mathrm{O} = 1.002$) is much larger than for all other conventional chroma types. This clearly demonstrates that our CRP features differ fundamentally from previous chroma types. As shown in Sect. IV, the boost of timbre invariance can also be observed when using real audio material.

III. Application: Audio Matching

The identification and retrieval of semantically related music data is of major concern in the field of music information retrieval. Loosely speaking, one can distinguish between two different scenarios. In the *global* matching scenario one compares and relates entire instances (on the document level) of a piece of music such as entire audio recordings or MIDI files. For example, in *cover song identification* the goal is to identify all performances of the same piece by different artists with varying interpretations, styles, instrumentation, and tempos [16], [17]. In the *local* matching scenario one compares and relates different subsegments contained in the same or in different instances of a piece. For example, in *audio matching* the goal is to automatically retrieve all passages (subsegments) from all audio documents that musically correspond to a given query excerpt [24]. Of course, the two scenarios seamlessly merge into each other. For example, Serrà et al. [17] use a local matching strategy for global document retrieval. The quality of the respective matching procedure depends on various factors including the underlying feature representation, the cost measure used to compare two feature vectors, as well as the distance function used to relate the various feature sequences.

In this paper, we study the behavior of our novel feature enhancement strategy within the audio matching scenario. In Sect. III-A, we define the distance function that underlies the matching procedure and provides a powerful tool for compactly assessing the matching capability of the used feature type. Then, in Sect. III-B, we derive various quality measures from the distance function, which turn out to be good indicators for the degree of timbre invariance exhibited by the respective feature type.

A. Distance Function

Let Q be a query a clip (typically a short audio excerpt) and let $(D_1, D_2, \ldots, D_N)$ be a collection of database documents (typically a large number of audio recordings). To simplify things, we assume that we have only one large database document D by concatenating $D_1, \ldots, D_N$, where we keep track of document boundaries in a supplemental data structure.

[6] http://www-mmdb.iai.uni-bonn.de/projects/CRP.html

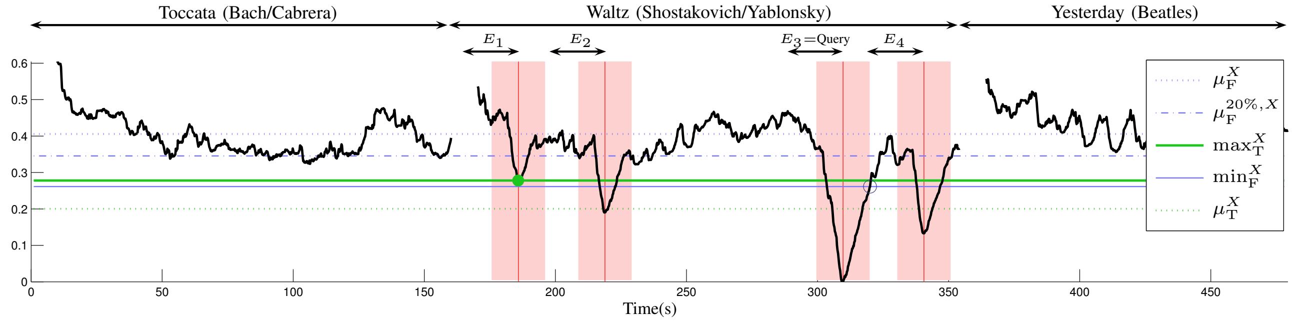

Fig. 5. Distance function with respect to the query E_3 for a database sequence corresponding to audio recordings of three different pieces (Bach Toccata played by Cabrera, Shostakovich Waltz conducted by Yablonsky, Yesterday by the Beatles). Indices corresponding to the four true matches are indicated by the four vertical red lines. The false alarm region consists of all indices outside the neighborhoods that are indicated by light red. The various quality measures are indicated by the horizontal lines. The green dot and the blue circle indicate the positions in the distance function that correspond to $\max_T^X$ and $\min_F^X$, respectively.

The goal of audio matching is to find all subsegments or passages within D that are similar to Q.

The first step of the audio matching procedure is to transform the query and the database document into suitable feature sequences $X = (X(1), X(2), \ldots, X(K))$ with $X(k) \in \mathcal{F}$ for $k \in [1 : K] := \{1, 2, \ldots, K\}$ and $Y = (Y(1), Y(2), \ldots, Y(L))$ with $Y(\ell) \in \mathcal{F}$ for $\ell \in [1 : L]$, respectively. Here, $\mathcal{F}$ denotes the underlying feature space. For example, in the case of normalized chroma features one has $\mathcal{F} = [0, 1]^{12}$. Furthermore, let $c : \mathcal{F} \times \mathcal{F} \to \mathbb{R}$ denote a *cost measure* on $\mathcal{F}$. If not stated otherwise, we revert to the cost measure $1 - \langle\cdot,\cdot\rangle$ (which is the cosine measure for normalized vectors). In Sect. IV, we also consider a binary shift measure similar to the one as introduced in [17]. As basis for the matching procedure, we use a distance function that locally compares the query sequence X with subsequences of the database sequence Y. More precisely, we define a distance function $\Delta : [1 : L] \to \mathbb{R} \cup \{\infty\}$ between X and Y using dynamic time warping (DTW):

$$\Delta(\ell) := \frac{1}{K} \min_{a \in [1:\ell]} \Big(\mathrm{DTW}\big(X\,,\, Y(a : \ell)\big)\Big), \tag{1}$$

where $Y(a : \ell)$ denotes the subsequence of Y starting at index a and ending at index $\ell \in [1 : L]$. Furthermore, $\mathrm{DTW}(X, Y(a : \ell))$ denotes the DTW distance between X and $Y(a : \ell)$ with respect to the cost measure c. To avoid degenerations in the DTW alignment, we use the modified step size condition with step sizes $(2, 1)$, $(1, 2)$, and $(1, 1)$ (instead of the classical step sizes $(1, 0)$, $(0, 1)$, and $(1, 1)$). Note that the distance function Δ can be computed efficiently using dynamic programming. For details on DTW and the distance function, we refer to [3, Section 4.4].

The interpretation of Δ is as follows: a small value $\Delta(\ell)$ for some $\ell \in [1 : L]$ indicates that the subsequence of Y starting at frame a_ℓ (with $a_\ell \in [1 : \ell]$ denoting the minimizing index in (1)) and ending at frame ℓ is similar to X. To determine the best match between Q and D, one simply has to look for the index $\ell_0 \in [1 : L]$ minimizing Δ. Then the best match is the audio clip corresponding to the feature subsequence $(Y(a_{\ell_0}), \ldots, Y(\ell_0))$. The value $\Delta(\ell_0)$ is also referred to as the *cost* of the match. To look for the second best match, we exclude a neighborhood around the index ℓ_0 from further consideration to avoid large overlaps with the best match. In our implementation, we exclude half the query length to the left and right by setting the corresponding Δ-values to ∞. To find subsequent matches, the above procedure is repeated until a certain number of matches is obtained or a specified distance threshold is exceeded. Note that the extracted matches can be naturally ranked according to their cost.

We illustrate the definition of Δ by means of our Shostakovich example introduced in Sect. II-B. We consider three different database documents that refer to audio recordings of three different pieces (Bach Toccata played by Cabrera, Shostakovich Waltz conducted by Yablonsky, Yesterday by the Beatles). First, we transform the three audio recordings into suitable feature sequences, which are concatenated to form a single database feature sequence Y. Furthermore, using the passage E_3 (trombone) from the Yablonsky recording as query, we derive a query feature sequence X for the query E_3. Fig. 5 shows the resulting distance function Δ. Within the three documents, there are four semantically correct matches, namely the passages E_1, E_2, E_3, and E_4 within the Waltz. Indeed, these four passages are revealed by four local minima of Δ. However, note that due to the above mentioned differences in timbre, some of these local minima are not well developed and have relatively large Δ-values such as the one corresponding to E_1. This is problematic as will be detailed in the next section. For example, iteratively extracting matches as described above, E_3, E_4, and E_2 appear as the top three matches. However, the next match is a false positive match (corresponding to the index 320 next to the right neighborhood boundary of E_3), before E_1 is identified as the fifth match.

B. Quality Measures

In view of the audio matching application, the following two properties of Δ are of crucial importance. First, the semantically correct matches (in the following referred to as *true matches*) should correspond to local minima of Δ close to zero thus avoiding false negatives. We capture this property by defining μ_T^X and $\max_T^X$ to be the average and maximum of Δ, respectively, over all indices that correspond to the local minima of the true matches for a given query X. Second, Δ should be well above zero outside a neighborhood of the desired local minima thus avoiding false positives. Recall from Sect. III-A that we use half the query length to the left and right to define such a neighborhood. The region outside these neighborhoods is referred to as *false alarm region*. We then define μ_{F}^X and $\min_{\mathrm{F}}^X$ to be the average respective minimum of Δ over all indices within the false alarm region. For our Shostakovich example shown in Fig. 5, these values are indicated by suitable horizontal lines. In order to separate the true matches from spurious matches, it is clear that μ_{T}^X and $\max_{\mathrm{T}}^X$ should be small whereas μ_{F}^X and $\min_{\mathrm{F}}^X$ should be large. We express these two properties within a single number, respectively, by defining the quotients $\alpha^X := \mu_{\mathrm{T}}^X / \mu_{\mathrm{F}}^X$ and $\gamma^X := \max_{\mathrm{T}}^X / \min_{\mathrm{F}}^X$.

In view of a good separability, α^X and γ^X should be close to zero. In the case $\gamma^X < 1$, all true matches appear as the top most matches. Contrary, $\gamma^X > 1$ indicates that at least one false positive match appears before all true matches are retrieved. Note that the quality measure γ^X is rather strict in the sense that one single outlier (either a true match of high cost or a spurious match of low cost in the false alarm region) may completely corrupt the value of γ^X. Contrary, the quality measure α^X is rather soft in the sense that despite of having a low value α^X one may obtain a large number of false positive matches. As a trade-off between the two quality measures α^X and γ^X, we introduce a third quality measure β^X. To this end, we sort the indices within the false alarm region by increasing cost and define $\mu_{\mathrm{F}}^{p\%,X}$ to be the average Δ only over the lower $p\%$ of the indices, $p \in [0, 100]$, see also Fig. 5. Note that for $p = 100$, one simply obtains $\mu_{\mathrm{F}}^{p\%,X} = \mu_{\mathrm{F}}^X$. Finally, we define $\beta^X := \mu_{\mathrm{T}}^X / \mu_{\mathrm{F}}^{p\%,X}$. In our experiments, we used $p = 1$ considering only 1% of the indices within the false alarm region. Note that β^X is a much better measure for indicating possible false positive matches than α^X while being more robust to outliers than γ^X.

In Sect. IV, we apply these quality measures on the basis of a carefully selected set of queries and a manually annotated collection of audio recordings in order to determine the degree of timbre invariance of various features types.

IV. Experiments

In Sect. II-C, we have reported on a first baseline experiment using systematically generated audio material. In this section, we report on a series of experiments based on real audio recordings to indicate how our novel CRP features behave in comparison to previously introduced chroma features as well as to explore the role of various parameters. First, in Sect. IV-B, we show that our CRP features outperform various publicly available state-of-the-art chroma features [16], [29], [28] with regard to timbre invariance. Then, we discuss the dependency of the CRP features' quality on the number of coefficients to be pruned in the reduction step (Sect. IV-C), on the value of the constant used in the logarithmic compression (Sect. IV-D), and on the feature rate (Sect. IV-E). Only recently, Serrà et al. [17] have introduced a novel binary shift measure for comparing chroma features. In Sect. IV-F, we show that our CRP features also yield significant quality improvements in combination with this novel cost measure. Finally, we investigate the effect of the CRP features on precision and recall values in the context of the audio matching application (Sect. IV-G). Altogether, the experiments show that our enhancement strategy yields a significant boost towards timbre invariance independent of a particular choice of parameters and measures.

A. Experimental Setup

For evaluating and comparing various types of chroma features, we compiled a collection of audio recordings that comprises harmony-based music of various genres. Here, the objective was to include music material that, on the one hand, contains a large number of harmonically related excerpts, which, on the other hand, reveal significant differences in timbre and instrumentation. For one thing, the collection contains pieces such as the Waltz by Shostakovich or the Bolero by Ravel, where a theme is repeated in different instrumentations. For another thing, for each piece there are at least two different versions such as different arrangements or cover songs. For example, on the classical music side, the collections contains an orchestra version as well as a piano version of the first movement of Beethoven's Fifth Symphony, Brahms' Hungarian dance No. 5, or Wagner's Prelude of the Meistersinger. On the popular music side, there are the original version and at least one cover song of pieces by the Beatles, Queen, Genesis, Indigo Girls, and Gloria Gaynor. Altogether, the collection consists of 32 recordings amounting to 166 minutes of music[7].

We carefully selected 101 audio excerpts with an average length of 30 seconds, which were used as queries in our matching experiments. The data collection was then manually annotated by specifying all relevant matches (referred to as true matches, see Sect. III-A) for each of the queries. At this point, we emphasize that the main object of our experiments is to assess the degree of timbre invariance and the discriminative power of the various chroma features. In other words, we are interested in evaluating the underlying features and use the matching procedure only for the purpose of comparing features. Therefore, we employ a controlled and manageable database with a clear notion of true matches, where the true matches represent various kinds of variations with regard to timbre and instrumentation.

For each query X, we compute the values μ_{T}^X, μ_{F}^X, $\mu_{\mathrm{F}}^{1\%,X}$, $\min_{\mathrm{F}}^X$, and $\max_{\mathrm{T}}^X$ as well as the quality measures α^X, β^X, and γ^X using the entire collection as the database documents, see Sect. III-B. Averaging over all 101 queries, we obtain the corresponding numbers denoted by μ_{T}, μ_{F}, $\mu_{\mathrm{F}}^{1\%}$, $\min_{\mathrm{F}}$, and $\max_{\mathrm{T}}$, as well as α, β, and γ. Note that α is not the quotient of μ_{T} and μ_{F}, but the average of the α^X. Analogously, this also holds for β and γ.

B. Comparison between Feature Types

We compared our $\mathrm{CRP}(n)$ features for various parameters $n \in [1 : 120]$ with various state-of-the-art chroma types using

[7] http://www-mmdb.iai.uni-bonn.de/projects/CRP.html

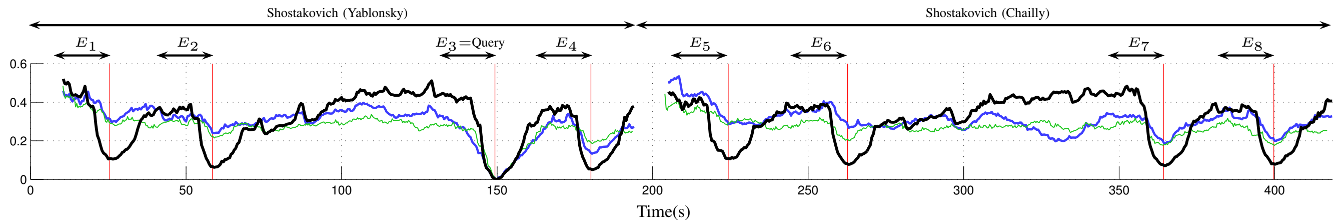

Fig. 6. Several distance functions shown for two recordings (Yablonsky, Chailly) of the Shostakovich example using the excerpt E_3 as query. The following feature types were used: Chroma-IF (thin green), Chroma-Pitch (blue) and $\mathrm{CRP}(55)$ (bold black). For the query, there are 8 annotated excerpts (true matches).

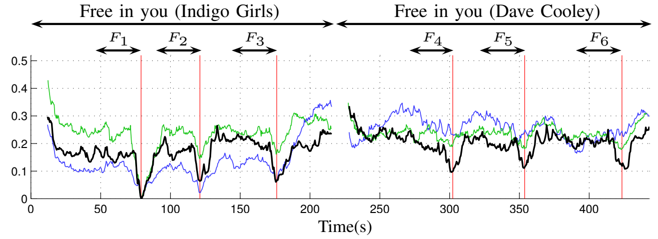

Fig. 7. Different distance functions using the excerpt F_3 as query. Only the part of the database is shown that consists of two versions (original by Indigo Girls, cover by Dave Cooley) of the piece "Free in you". Altogether, there are 6 true matches denoted by F_1 to F_6. The following feature types were used: Chroma-IF (thin green), Chroma-Pitch (blue) and $\mathrm{CRP}(55)$ (bold black).

TABLE II
OVERVIEW OVER THE VARIOUS QUALITY MEASURES FOR DIFFERENT TYPES OF CHROMA FEATURES (FEATURE RATE ≈ 2 HZ, $C = 100$).

	μ_T	μ_F	α	μ_T	$\mu_\mathrm{F}^{1\%}$	β	$\max_\mathrm{T}$	$\min_\mathrm{F}$	γ
Chroma-IF	0.150	0.313	0.487	0.150	0.198	0.767	0.177	0.162	1.098
Chroma-P	0.042	0.132	0.320	0.042	0.072	0.610	0.055	0.060	0.954
Chroma-E	0.045	0.133	0.346	0.045	0.075	0.642	0.059	0.062	0.979
Chroma-MIR	0.032	0.078	0.415	0.032	0.042	0.834	0.040	0.035	1.234
Chroma-QM	0.065	0.147	0.457	0.065	0.090	0.750	0.080	0.075	1.087
Chroma-Pitch	0.120	0.433	0.282	0.120	0.219	0.568	0.164	0.171	0.985
$\mathrm{CRP}(35)$	0.110	0.671	0.167	0.110	0.287	0.397	0.147	0.223	0.691
$\mathrm{CRP}(55)$	**0.092**	**0.626**	**0.150**	**0.092**	**0.244**	**0.388**	**0.124**	**0.187**	**0.693**
$\mathrm{CRP}(75)$	0.076	0.550	0.143	0.076	0.172	0.459	0.106	0.136	0.840
$\mathrm{CRP}(95)$	0.044	0.472	0.099	0.044	0.081	0.546	0.066	0.058	1.180

the same parameter settings as described in Sect. II-C (feature rate of 2 Hz, $C = 100$, feature vectors of Euclidean norm 1).

Before giving a systematic evaluation, we illustrate the matching capability of different feature types by means of our Shostakovich example, see Sect. II-B. Our database contains two different recordings (Yablonsky, Chailly) of the Waltz with 8 annotated excerpts corresponding to the theme, where E_1 to E_4 denote the corresponding excerpts in the Yablonsky and E_5 to E_8 in the Chailly recording. Now, using E_3 (trombone) as query, one has eight true matches. Using conventional chroma features such as Chroma-IF or Chroma-Pitch most of the expected local minima are not significant or not even existing (e. g., E_5), see Fig. 6. Now, using our $\mathrm{CRP}(n)$ features, one obtains for all eight true matches (even for E_5) much more concise local minima, see the black curve of Fig. 6. This demonstrates that the particular choice of a feature type has a significant impact on the final matching quality. A similar effect can be noticed in Fig. 7, which shows the distance function for the song "Free in you" by the Indigo Girls and a cover version of the same piece by Dave Cooley. In the original version, the voice is accompanied by an acoustic guitar and some moderate percussion, whereas in the cover song there are additional voices, percussion is much more dominant, and the guitar is replaced by distorted electronic synthesizer effects. Also for this popular music example, using conventional chroma features (Chroma-IF, Chroma-Pitch) results in a distance function without well defined local minima for the true matches (especially for the cover version). On the contrary, using our $\mathrm{CRP}(n)$ features leads to local minima at the positions of the true matches that are clearly separated from the false alarm region.

Table II shows different quality measures for six types of conventional chroma features and for our novel $\mathrm{CRP}(n)$ features for selected parameters $n \in [1 : 120]$. For example, using the conventional chroma features Chroma-P, the average cost of the true matches is $\mu_\mathrm{T} = 0.042$, whereas the average distance in the false alarm region is $\mu_\mathrm{F} = 0.132$. The average quotient amounts to $\alpha = 0.320$[8]. Other conventional chroma features exhibit a larger average cost for the true matches such as $\mu_\mathrm{T} = 0.120$ for Chroma-Pitch. However, in this case the average distance within the false alarm region also increases remarkably amounting to $\mu_\mathrm{F} = 0.433$. As a result, the average quotient of $\alpha = 0.282$ for Chroma-Pitch is lower thus expressing a higher discrimination capability than the one for Chroma-P. Now, looking at the quality measures for our novel $\mathrm{CRP}(n)$ features, one can recognize a significant improvement. For example, in the case $n = 55$ one obtains $\alpha = 0.150$, which is nearly half of $\alpha = 0.282$ obtained from Chroma-Pitch. In other words, the discrimination capability of $\mathrm{CRP}(55)$ features is nearly twice as good as in the case of Chroma-Pitch.

According to the measure α, the $\mathrm{CRP}(n)$ features seem to perform best for the parameter $n = 95$ among all selected parameters listed in Table II. However, looking at the γ-measure, one obtains $\gamma = 1.180$ for $n = 95$, which is much worse than $\gamma = 0.693$ for $n = 55$. As already noted in Sect. III-B, the α-measure does not warrant a clear separation between true matches and spurious matches. In contrast, the γ-measure yields an explicit separation distance, but may be corrupted by a single outlier. In the following, our main focus is on the β-measure using only the lower $p = 1\%$ of the indices in the false alarm region, which constitutes a suitable compromise between the α- and γ-measure, see Sect. III-B.

With respect to the β-measure, the $\mathrm{CRP}(55)$ features with $\beta = 0.388$ perform best among all listed feature types. For

[8] Recall that the values are obtained by averaging over all queries. The average quotient $\alpha = 0.320$ does not coincide with the quotient of the averages $\mu_\mathrm{T}/\mu_\mathrm{F} = 0.318$.

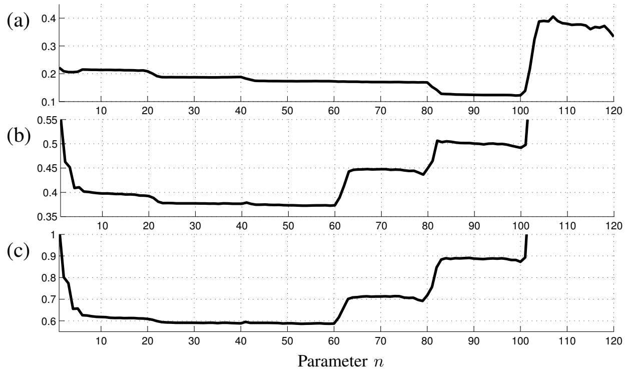

Fig. 8. Influence of the parameter $n \in [1 : 120]$ (horizontal axis) on the performance measures **(a):** α, **(b):** β, and **(c):** γ. The range of each vertical axis has been limited to show more details of the relevant parts of the respective curve.

TABLE III
INFLUENCE OF THE PARAMETER C USED IN THE LOGARITHMIC COMPRESSION ON THE QUALITY OF $\mathrm{CRP}(55)$ FEATURES.

C	μ_T	μ_F	α	μ_T	$\mu_\mathrm{F}^{1\%}$	β	max_T	min_F	γ
1	0.114	0.654	0.179	0.114	0.279	0.414	0.155	0.219	0.718
10	0.098	0.642	0.156	0.098	0.258	0.389	0.133	0.199	0.686
100	**0.092**	**0.626**	**0.150**	**0.092**	**0.244**	**0.388**	**0.124**	**0.187**	**0.693**
1000	0.089	0.606	0.151	0.089	0.234	0.397	0.120	0.180	0.705
10000	0.089	0.583	0.157	0.089	0.226	0.412	0.119	0.173	0.733
100000	0.087	0.557	0.162	0.087	0.216	0.425	0.116	0.166	0.758

TABLE IV
INFLUENCE OF THE FEATURE RATE ON THE QUALITY OF $\mathrm{CRP}(55)$ FEATURES.

	μ_T	μ_F	α	μ_T	$\mu_\mathrm{F}^{1\%}$	β	max_T	min_F	γ
10 Hz	0.121	0.642	0.191	0.121	0.315	0.384	0.159	0.253	0.633
5 Hz	0.107	0.636	0.171	0.107	0.291	0.374	0.143	0.229	0.637
2 Hz	**0.092**	**0.626**	**0.150**	**0.092**	**0.244**	**0.388**	**0.124**	**0.187**	**0.693**
1 Hz	0.083	0.619	0.138	0.083	0.197	0.443	0.112	0.147	0.827
0.5 Hz	0.074	0.625	0.123	0.074	0.154	0.513	0.106	0.110	1.061

the best conventional feature (Chroma-Pitch), one already has $\beta = 0.568$. The difference between $\mathrm{CRP}(35)$ and $\mathrm{CRP}(55)$ is not significant. Here, the parameter $n = 55$ may be preferable since less coefficients are needed to yield the same discriminative power. In the next section, we investigate the role of the reduction parameter $n \in [1 : 120]$ in more detail.

C. Dependency on DCT Reduction

In the last section, we have compared and discussed the discrimination capability of various conventional chroma features and of $\mathrm{CRP}(n)$ features for selected parameters $n \in [1 : 120]$. We now look closer at the role of this parameter, which determines the number of PFCCs to be pruned in the reduction step, see Sect. II-B. To this end, we computed the quality measures α, β, and γ in dependence of $n \in [1 : 120]$. The resulting curves are shown in Fig. 8. The curve for α may indicate that the discriminative power, in average, is optimal for parameters $n \in [83 : 99]$. However, as already discussed in Sect. IV-B, a low α-measure does not warrant a clear separation between true matches and spurious matches. More meaningful indicators are the β- and γ-measures. Here, the corresponding curves show that one obtains the best separation between true and spurious matches for parameters $n \in [23 : 59]$. In the following experiments, we use the parameter $n = 55$, which exhibits low values with respect to all three quality measures. Actually, in Sect. V, we will discuss the musical meaning of a small number of dominant PFCCs, which also explains the “jumps” in the curves of Fig. 8. These findings can then be used to further reduce the number of coefficients without a degradation of the discriminative power.

D. Dependency on Logarithmic Compression

It is a well known fact that loudness is perceived in a logarithmic fashion [36]. Therefore, after a suitable decomposition of the audio signal, one often applies a logarithmic energy or amplitude compression. For example, such a step is involved in the computation of MFCCs [31] or in deriving onset signals as used for beat tracking and meter analysis [37]. In Sect. II-B, we employed such a compression step after the subband decomposition replacing each entry e in the resulting pitch representation by the value $\log(C \cdot e + 1)$. To investigate the role of the positive constant C, we computed $\mathrm{CRP}(55)$ features using different constants C and derived the corresponding quality measures α, β, and γ. From these measures, which are listed in Table III, we conclude that for any choice of C between 10 and 1000 one obtains features of a similar quality. In our experiments, we therefore use the value $C = 100$. Similar findings are reported by Klapuri et al. [37].

Another approach used for dynamics compression is referred to as *spectral whitening*. We implemented a version of the whitening procedure similar to [27] locally normalizing the pitch subbands obtained from our filterbank decomposition according to short-time variances of the subbands. Actually, this procedure is related to the logarithmic amplitude compression as both flatten (or whiten) the spectral energy distribution. Indeed, using spectral whitening instead of logarithmic compression did not have a significant impact on the various quality measures. Therefore, in the following, we only consider the algorithmically simpler logarithmic compression.

E. Dependency on Feature Rate

Next, we investigate the influence of the features rate on the final quality of $\mathrm{CRP}(n)$ features. In Table IV, we only report on the results for the parameter $n = 55$; other feature types were found to show a similar behavior. Recall from Sect. II-B that the final feature rate can be adjusted by modifying the size of the rectangular window used to compute the local energies in the pitch subbands. With respect to the β- and γ-measure, the resulting CRP features perform almost equally well for feature rates ranging from 10 Hz down to 2 Hz. However, further decreasing the feature rate results in noticeable degradations. For example, one has $\beta = 0.374$ for 5 Hz and a slightly higher $\beta = 0.388$ for 2 Hz, whereas the value significantly drops to $\beta = 0.443$ for 1 Hz. In our experiments, we therefore revert to the feature rate of 2 Hz. In comparison to higher features rates, 2 Hz features not only possess a comparable quality, but also keep the data at

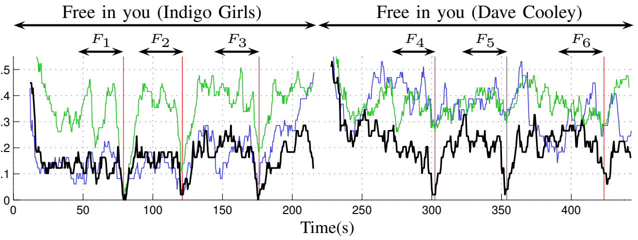

Fig. 9. Different distance functions using the excerpt F_3 as query with respect to the binary shift measure c_{bs}, continuing the example from Fig. 7. The following feature types were used: Chroma-IF (thin green), Chroma-Pitch (blue) and $\mathrm{CRP}(55)$ (bold black).

TABLE V
OVERVIEW OVER THE QUALITY OF VARIOUS FEATURE TYPES EMPLOYING THE BINARY SHIFT MEASURE IN THE MATCHING PROCESS.

	μ_{T}	μ_{F}	α	μ_{T}	$\mu_{\mathrm{F}}^{1\%}$	β	$\max_{\mathrm{T}}$	$\min_{\mathrm{F}}$	γ
Chroma-IF	0.171	0.528	0.327	0.171	0.283	0.598	0.233	0.215	1.081
Chroma-P	0.069	0.558	0.124	0.069	0.264	0.256	0.121	0.189	0.633
Chroma-E	0.087	0.552	0.159	0.087	0.275	0.312	0.142	0.203	0.689
Chroma-MIR	0.089	0.525	0.171	0.089	0.222	0.408	0.142	0.148	1.135
Chroma-QM	0.151	0.531	0.289	0.151	0.286	0.527	0.215	0.211	1.071
Chroma-Pitch	0.091	0.583	0.160	0.091	0.231	0.431	0.164	0.152	1.220
CRP(35)	0.011	0.572	0.019	0.011	0.126	0.115	0.029	0.060	1.710
CRP(55)	**0.011**	**0.570**	**0.019**	**0.011**	**0.123**	**0.097**	**0.031**	**0.057**	**1.246**
CRP(75)	0.029	0.574	0.052	0.029	0.140	0.227	0.076	0.061	3.371
CRP(95)	0.058	0.580	0.101	0.058	0.139	0.416	0.108	0.052	6.694

a manageable size, thus making the subsequent steps in the matching procedure more efficient.

F. Dependency on Cost Measure

So far, we have used the cosine measure as cost measure to compare two chroma vectors. Only recently, Serrà et al. [17] have introduced a novel binary cost measure that only assumes two values. Basically, the idea is to consider all cyclically shifted versions of the two vectors to be compared [38]. Then, the two original chroma vectors are regarded as similar (binary cost measure assumes the value 0) if they best correlate without any shift relative to each other, otherwise they are regarded as dissimilar (binary cost measure assumes the value 1). This cost measure has turned out to be suitable in global matching tasks such as cover song identification [17]. A similar concept considering minimizing shift indices has been introduced in the context of music structure analysis, see [39].

In this section, we first define a binary cost measure similar to [17], which we refer to as *binary shift measure*. Then, we show that our CRP features also yield significant quality improvements in combination with this novel cost measure. In the following, all chroma vectors are assumed to be normalized with respect to the Euclidean norm. As in Sect. III-A, let $\mathcal{F} = [0,1]^{12}$ denote the feature space and $c : \mathcal{F} \times \mathcal{F} \to \mathbb{R}$ the cosine measure. We define the *cyclic shift* $\sigma : \mathcal{F} \to \mathcal{F}$ by

$$\sigma((v(1), v(2), \ldots, v(12))) := (v(2), \ldots, v(12), v(1))$$

for a chroma vector $v = (v(1), \ldots, v(12)) \in \mathcal{F}$. By iteratively applying σ, one obtains σ^i, $i \in \mathbb{N}_0$, where i is referred to as the *shift index*. Obviously, $\sigma^{12} = \sigma^0$ is the identity on $\mathcal{F}$. Now, when comparing two chroma vectors $v, w \in \mathcal{F}$, one first computes the minimizing shift index:

$$\mathrm{msi}(v, w) := \mathrm{argmin}_{i \in [0:11]} \Big(c(v, \sigma^i(w)) \Big).$$

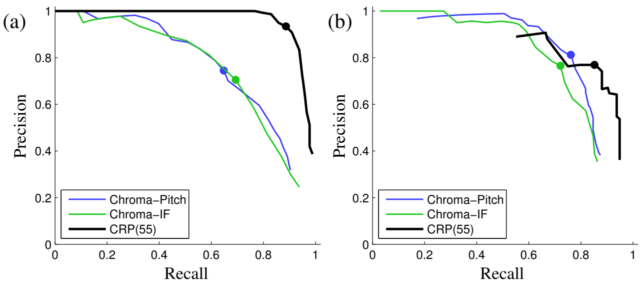

Fig. 10. Quality of several feature types in terms of precision (vertical axis) and recall (horizontal axis) values. **(a):** PR-diagrams when using the cosine measure c. **(b):** PR-diagrams when using the binary shift measure c_{bs}. The dot within a PR-diagram indicates the respective maximal F-value $\mathrm{F}_{\max}$.

Then, the binary shift measure $c_{\mathrm{bs}} : \mathcal{F} \times \mathcal{F} \to \{0, 1\}$ is defined by

$$c_{\mathrm{bs}}(v, w) := \begin{cases} 0 & \text{for } \mathrm{msi}(v, w) = 0, \\ 1 & \text{for } \mathrm{msi}(v, w) \neq 0. \end{cases}$$

We now repeat the computation of the quality measures α, β, and γ, where we use the binary shift measure c_{bs} instead of c. The result is shown in Table V. There are several interesting observations. First, it is striking that for all features types the α-measures with respect to c_{bs} are much lower than the ones with respect to c (compare Table V and Table II). Second, also using c_{bs} as cost measure, the $\mathrm{CRP}(n)$ features by far outperform conventional chroma features with respect to the α- and β-measure. Again, the parameter $n = 55$ leads to very good overall results. For example, one has $\beta = 0.097$ for $\mathrm{CRP}(55)$ using c_{bs}, which yields the lowest β-value among the listed feature types.

At first sight surprisingly, the behavior of the γ-measure is quite different. Here, when using c_{bs} instead of c, conventional chroma features seem to be superior to CRP features. This can be explained as follows. Recall from Sect. III-B that the γ-measure suffers in the sense that a single outlier may completely corrupt the value of γ. Now, the binary shift measure c_{bs} assuming only the two values zero and one is a rather coarse measure compared to the cosine measure c. As a consequence, the distance function Δ typically decreases in regions that are harmonically related to the query (but it may even increase in regions that are harmonically unrelated to the query). On the positive side, this generally lowers the cost of true matches. On the negative side, this often produces a few (not necessarily many) false positive matches of quite low cost. These false positive matches corrupt the γ-measure, but do not have such a large effect on the β-measure. This phenomenon is also illustrated by Fig. 9 (continuing the Indigo Girls example shown in Fig. 7), where we employ the binary shift measure c_{bs} instead of c. Note that the c_{bs}-based distance functions yield a much better average separation (almost all true matches have a cost very close to zero) than the c-based counterparts. However, in particular in the original version, the c_{bs}-based distance function dangerously approaches zero at some positions within the false alarm regions.

G. Effect on Precision and Recall

To indicate the potential of the CRP features for music retrieval applications, we investigate the effect of our enhancement strategy in terms of precision and recall values. In the following experiment, we use the queries and the manually annotated database from Sect. IV-A. The annotations constitute the ground truth on the exact positions of the true (relevant) matches for each query. Now, for a fixed feature type, we compute the distance function Δ for each of the queries . Then, for a given positive distance threshold τ, we subsequently derive all matches having a cost below τ as described in Sect. III-A. Using the ground truth information, we then compute the precision value P_τ and the recall value R_τ for the set of retrieved matches. From these values one obtains the F-measure $\mathrm{F}_\tau := \frac{2\cdot\mathrm{P}_\tau\mathrm{R}_\tau}{\mathrm{P}_\tau+\mathrm{R}_\tau}$. Starting with a threshold τ close to zero and increasing it little by little, one obtains a family of precision (P) and recall (R) values, which can be graphically visualized by a PR-diagram.

We have computed such PR-diagrams for various types of chroma features. Fig. 10(a) shows three representative diagrams for two conventional chroma features (Chroma-IF, Chroma-Pitch) and for the $\mathrm{CRP}(55)$ features. As the diagrams indicate, one obtains much better PR-values using the enhanced CRP features than in the case of conventional chroma features. A good indicator for this is the maximal F-value $\mathrm{F}_{\max} := \max_\tau(\mathrm{F}_\tau)$, which is indicated by a dot within the respective PR-diagram in Fig. 10. In our experiments, one obtains $\mathrm{F}_{\max} = 0.70$ and $\mathrm{F}_{\max} = 0.69$ for the conventional chroma features Chroma-IF and Chroma-Pitch, respectively. On the other hand, one obtains $\mathrm{F}_{\max} = 0.91$ for our $\mathrm{CRP}(55)$ features, which is an improvement of more than 30% over the conventional features.

Finally, Fig. 10(b) shows the corresponding PR-diagrams using the binary shift measure c_{bs} instead of the cosine measure c. Also in this case, the $\mathrm{CRP}(55)$ features still outperform the conventional features, in particular with regard to recall. However, as explained in Sect. IV-F, there tend to be a notable number of false positives when using c_{bs}, which is also reflected in the PR-diagrams. For example, when using $\mathrm{CRP}(55)$ features in combination with c_{bs}, there are already quite a number of false positive matches having cost zero. This experiment also indicates that the binary shift measure, even though being a very powerful tool in global matching scenarios, tends to be too coarse for local matching scenarios, see Sect. III.

V. Detailed Analysis

In this section, we give a detailed analysis of the DCT-based reduction step, which plays a central role in our enhancement procedure. In Sect. V-A, we show that for harmony-based music the upper PFCCs are dominated by a few dominating coefficients. As it turns out, these coefficients correspond to pitch periodicities that allow for a musically meaningful interpretation (Sect. V-B). Furthermore, we show that the PFCCs surrounding the dominating ones account for different phases or pitch transpositions (Sect. V-C).

TABLE VI
Frequency and period for selected DCT basis vectors.

DCT basis vector	$\mathbf{c}_{21}$	$\mathbf{c}_{41}$	$\mathbf{c}_{61}$	$\mathbf{c}_{81}$	$\mathbf{c}_{101}$	$\mathbf{c}_{120}$
frequency	0.083	0.167	0.250	0.333	0.417	0.496
period	12	6	4	3	2.4	≈ 2

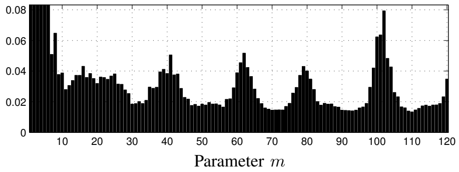

Fig. 11. Entries $\bar{y}(m)$ of the vector $\bar{y}$ for $m \in [1:120]$ (horizontal axis). The value $\bar{y}(m)$ indicates the average absolute correlation of the normalized pitch vectors of the database recordings with the DCT basis vector $\mathbf{c}_m$.

A. Relation of DCT Basis Vectors to Pitch Vectors

In our enhancement procedure, the logarithmized pitch vectors are transformed by means of a discrete cosine transform (DCT). This transform is represented by an orthogonal 120×120 matrix denoted by DCT_{120}, where the m^{th} row of DCT_{120} can be thought of as a 1-sampled cosine function of frequency $freq(m) = \frac{m-1}{2\cdot 120}$, $m \in [1:120]$. In the following, this vector is denoted by $\mathbf{c}_m$ and referred to as the m^{th} DCT basis vector. The period of $\mathbf{c}_m$ is given by $period(m) = \frac{1}{freq(m)}$. Now, computing the matrix-vector product $y = \mathrm{DCT}_{120} \cdot x$ for a pitch vector $x \in \mathbb{R}^{120}$, the m^{th} coefficient $y(m)$ of y expresses to which degree x and $\mathbf{c}_m$ correlate.

To get some hints on a possible semantic meaning of the DCT basis vectors, we conducted the following experiment. First, we computed the logarithmized pitch vectors as described in Sect. IV for each audio recording of our database (Sect. IV-A). Then, we normalized each of these vectors with respect to the Euclidean norm, applied a DCT to obtain a PFCC vector, and replaced each entry of the coefficient vector by its absolute values. Finally, we averaged all resulting vectors over the entire database to obtain a single 120-dimensional vector, say $\bar{y}$. This vector is shown (in a horizontal form) in Fig. 11. The entry $\bar{y}(m)$ can be interpreted to represent the average absolute correlation of the normalized pitch vectors with the DCT basis vector $\mathbf{c}_m$.

The lower PFCCs, which are related to loudness and timbre, are left unconsidered in the CRP features, and we also disregard them in the following analysis. As revealed by Fig. 11, some of the upper PFCCs indicate that certain DCT basis vectors show a strikingly high average correlation with the pitch vectors. This particularly holds for all DCT basis vectors $\mathbf{c}_m$ with $m \in \mathcal{S} := \{41, 61, 81, 101, 120\}$. Actually, as seen from Table VI, all these basis vectors are 12-periodic (or nearly 12-periodic in the case $m = 120$).

B. Musical Meaning of Dominating DCT Basis Vectors

The dominance of the 12-periodic DCT basis vectors does not come all of a sudden, but originates from certain musical properties of the underlying audio material. We now give some explanations for this dominance. Recall that our

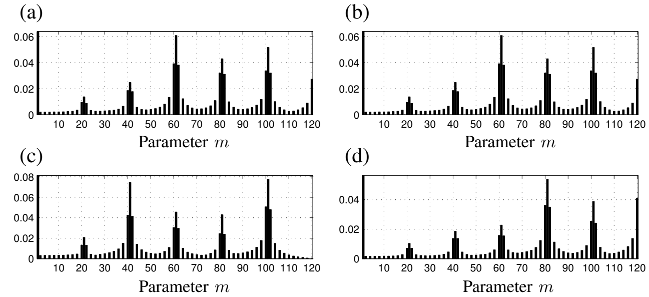

Fig. 12. Average absolute correlation of the pitch vectors of various sets with the DCT basis vectors $\mathbf{c}_m$ for $m \in [1:120]$. **(a)** Major 3-chords. **(b)** Minor 3-chords. **(c)** Tonic/dominant 2-chords. **(d)** Major seventh 4-chords.

enhancement strategy is based on the pitch-frequency scale, which has a much closer relation to harmony-based music than the mel-frequency scale. In our setting, the DCT basis vectors capture certain periodicities of a pitch vector along the 120-dimensional pitch scale. The 12-periodicity is strongly connected to the octave interval that plays a crucial role in musical sounds and harmony-based music [1]. First, playing a musical note on an instrument typically produces a sound involving several frequencies known as harmonics, where the harmonics are integer multiples of the fundamental frequency. Since many of the harmonics are in an octave relationship, a pitch vector computed from a musical sound typically contains some quasi-periodic patterns of period 12. Second, Western music is often based on the use of specific *chords*, i. e., pattern of notes that are played simultaneously. Typical examples are major and minor 3-chords or major seventh 4-chords. Also the tonic-dominant relationship plays a fundamental role in Western harmony-based music. This implies the importance of certain pitch intervals including the fifth (pitch distance 7), the fourth (pitch distance 5), the major third (pitch distance 4), the minor third (pitch distance 3), and the octave (pitch distance 12).

Because of these two reasons—the nature of harmonics and the nature of harmony-based music—the pitch vectors derived from such music recording often exhibit quasi-periodic patterns, which are captured by the dominant DCT coefficients. To illustrate this fact, we conducted some experiments similar to the one described in Sect. V-A. Instead of using pitch vectors from real audio recordings, we constructed sets of pitch vectors that correspond to specific harmonic chords. For example, the C-major chord is represented by a pitch vector were all entries are set to one that correspond either to pitch class C, E, or G; all other entries were set to zero. Other chords are represented in the same fashion. Now, using the set of pitch vectors covering all major chords, we computed the average absolute correlation vector $\bar{y}$. The vector $\bar{y}$, which is shown Fig. 12(a), clearly exhibits the dominance of $\mathbf{c}_m$ for $m = 61$, $m = 81$, and $m = 101$. Here, the basis vector $\mathbf{c}_{61}$ accounts for the major third (distance 4) and $\mathbf{c}_{81}$ for the minor third (distance 3). Interestingly, the basis vector $\mathbf{c}_{101}$ of period 2.4 accounts for the tonic-dominant relationship, which is based on a fifth (distance 7) and a fourth (distance 5) to the next octave. Here, note that twice the period 2.4 picks up the fourth (distance $4.8 \approx 5$) and three times the period 2.4 picks up the fifth (distance $7.2 \approx 7$). Similar experiments were conducted with a set of minor 3-chords, a set of tonic-dominant 2-chords, and a set of major-seventh 4-chords, see (b)-(d) of Fig. 12. For example, Fig. 12(d) reveals a striking dominance of $\mathbf{c}_{81}$ of period 3, which indeed reflects the importance of the minor third in seventh chords. Finally, the basis vector $\mathbf{c}_{120}$ of approximate period 2 accounts for the dominance of whole steps (distance 2) in harmonic chords. Finally, we emphasize that the 12-periodic basis vectors $\mathbf{c}_m$ for $m \in \mathcal{S} = \{41, 61, 81, 101, 120\}$ additionally account for the octave relationship. This not only explains the musical importance of these basis vectors but also the "jumps" in the curves of Fig. 8 at the corresponding index positions.

TABLE VII
QUALITY MEASURES FOR VARIOUS CRP VARIANTS.

	μ_{T}	μ_{F}	α	μ_{T}	$\mu_{\mathrm{F}}^{1\%}$	β	$\max_{\mathrm{T}}$	$\min_{\mathrm{F}}$	γ
$\mathrm{CRP}(55)$	0.092	0.626	0.150	0.092	0.244	0.388	0.124	0.187	0.693
$\mathrm{CRP}(\mathcal{S})$	0.105	0.667	0.158	0.105	0.249	0.459	0.145	0.199	0.799
$\mathrm{CRP}^{\sin}(\mathcal{S})$	0.112	0.673	0.169	0.112	0.290	0.397	0.150	0.225	0.696
$\mathrm{CRP}(\overline{\mathcal{S}})$	0.110	0.677	0.165	0.110	0.292	0.387	0.148	0.227	0.682

C. Phase Shift Simulation by DCT Basis Vectors

So far, we have seen that the DCT basis vectors $\mathbf{c}_m$ for $m \in \mathcal{S}$ capture the musically important pitch periodicities. At this point, one may assume that a reduction using only the few dominating DCT basis vectors may yield a similar enhancement than using the entire range of upper PFCCs. To investigate this assumption, we have conducted the following experiment. In the construction of the CRP features, we only kept the five PFCCs corresponding to the set $\mathcal{S}$ and discarded the other 115 PFCCs by setting them to zero. We then applied the inverse DCT, the chroma binning, and the normalization as before, see Fig. 2. The resulting features are referred to as $\mathrm{CRP}(\mathcal{S})$ features. Finally, we computed the various quality measures for the $\mathrm{CRP}(\mathcal{S})$ features, see Table VII. Even though the $\mathrm{CRP}(\mathcal{S})$ features achieve some improvement over conventional chroma features (cf. Table II), there is a significant degradation in the β- and γ-measures compared to $\mathrm{CRP}(55)$ features. For example, one has $\beta = 0.388$ for $\mathrm{CRP}(55)$ features, whereas $\beta = 0.459$ for $\mathrm{CRP}(\mathcal{S})$ features (Table VII).

The main reason for this degradation can be explained as follows. First, recall that a DCT basis vector is obtained by sampling a suitable cosine functions of certain frequency and phase. Using only one DCT basis vector of a fixed phase for a specific pitch periodicity, one is not able to deal with phase shifts, which can be interpreted as pitch transpositions in our scenario. For example, the DCT basis vector $\mathbf{c}_{101}$ is able to capture periodicities stemming from a C-major chord, but has difficulties in capturing the same periodicities in the case of a D-major chord. One can deal with phase shifts as is done in Fourier analysis [3]: one simply complements each DCT basis vectors by an additional phase-shifted (shifted by $\pi/2$) duplicate. In our scenario, we introduce an additional phase-shifted basis vector $\mathbf{s}_m$ for each DCT basis vector $\mathbf{c}_m$, $m \in \mathcal{S}$. Here, $\mathbf{s}_m$ is obtained by sampling a sine function

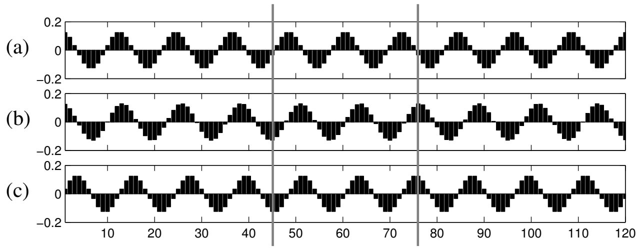

Fig. 13. **(a)**: DCT basis vector $\mathbf{c}_{21}$. **(b)**: DCT basis vector $\mathbf{c}_{20}$. **(c)**: Phase-shifted version of $\mathbf{c}_{21}$ shifted by $\pi/2$. Note that the vectors shown in (b) and (c) nearly coincide in the part between the two gray vertical lines.

that corresponds to the cosine function used to obtain $\mathbf{c}_m$. Now, we project the pitch vectors onto the space spanned by the ten basis vectors $\mathbf{c}_m$ and $\mathbf{s}_m$, $m \in \mathcal{S}$. (Before, we only used the five vectors $\mathbf{c}_m$.) Then, we continue with the usual chroma binning and normalization to obtain features denoted by $\mathrm{CRP}^{\sin}(\mathcal{S})$. As shown by Table VII, the $\mathrm{CRP}^{\sin}(\mathcal{S})$ features exhibit much better β- and γ-measures than the $\mathrm{CRP}(\mathcal{S})$ features and qualitatively come up to the original $\mathrm{CRP}(55)$ features.

Finally, we investigate how this phase shift information is recovered in the case of the purely cosine-based $\mathrm{CRP}(n)$ features. Looking at Fig. 11 and Fig. 12, one can notice that the dominating PFCCs corresponding to the DCT basis vectors $\mathbf{c}_m$, $m \in \mathcal{S}$, are flanked at both sides by further relevant PFCCs. Exemplarily, we look at $\mathbf{c}_{21}$ and its adjacent basis function $\mathbf{c}_{20}$, see (a) and (b) of Fig. 13. The two underlying cosine functions differ only slightly in their frequency. As a consequence, $\mathbf{c}_{20}$ behaves like a phase-shifted version of $\mathbf{c}_{21}$ in the middle part of the pitch scale. In this part, the vector $\mathbf{c}_{20}$ nearly coincides with $\mathbf{s}_{21}$, cf. (b) and (c) of Fig. 13. In other words, phase-shifts in the middle part of the pitch scale are simulated by DCT basis vectors with a slightly changed frequency. This property is particularly important in view of real-world music recordings, where most of the energy is concentrated in the middle part of the pitch scale. We close our discussion by a final experiment, which reinforces our explanations. Here, we used in the reduction step the set $\overline{\mathcal{S}} := \{40:42, 60:62, 80:82, 100:102, 119:120\}$ instead of the set $\mathcal{S}$. The resulting features, which are denoted as $\mathrm{CRP}(\overline{\mathcal{S}})$ features, indeed exhibit a similar β- and γ-measure as the $\mathrm{CRP}^{\sin}(\mathcal{S})$ and the $\mathrm{CRP}(55)$ features, see Table VII.

VI. Conclusions and Future Work

In this paper, we introduced a novel enhancement procedure for significantly increasing the robustness of conventional chroma features to changes in timbre and instrumentation. Here, our main ideas were first to compute cepstral coefficients based on a pitch-frequency scale, second to discard the lower PFCCs, and third to deduce from the remaining upper PFCCs the chroma-based CRP features. As it turned out, the upper PFCCs are dominated by only a few coefficients that reflect harmonically relevant pitch-periodicities as prominent in Western music. Revealing the musical meaning of certain PFCCs not only puts our procedure in a nutshell, but also allows for further reducing the number of PFCCs without a degradation of the discriminative power of the resulting CRP features.

Extensive experiments showed that our enhancement strategy yields a significant boost towards timbre invariance independent of a particular choice of parameters and measures. Using our novel CRP features, one can significantly improve the performance in all those matching and classification applications, where one wants to be invariant with regard to instrumentation and tone color. Exemplarily, this was shown for an audio retrieval application, where precision and recall values substantially increased when using our CRP features instead of conventional chroma features. For the future, we plan to apply CRP features also for other MIR tasks such as cover song identification [16], [17], structure analysis [11], [12], [13], [14], and cross-domain music matching [18], [20].

Generally, the direct comparison of audio features as well as the assessment of the features' properties is a difficult and time-consuming problem. Here, as a further conceptual contribution of this paper, our evaluation framework constitutes a powerful tool for comparing and studying the behavior of audio features in a compact form and systematic way. Using a DTW-based distance function, we derived various quality measures that express separation and matching capabilities on the basis of real-world music material. Note that the musical meaning of the measures very much depend on the particular choice of the underlying audio material. In this paper, we carefully selected and annotated music recordings to obtain quality measures that indicate the degree of timbre invariance exhibited by the respective feature type. By suitably changing the audio material and the annotations, our framework can easily be adjusted to also facilitate the evaluation of audio features with regard to other musical aspects such as timbre (not timbre-invariance as in this paper), rhythm, or melodic similarity.

References

[1] M. A. Bartsch and G. H. Wakefield, "Audio thumbnailing of popular music using chroma-based representations," *IEEE Transactions on Multimedia*, vol. 7, no. 1, pp. 96–104, 2005.

[2] E. Gómez, "Tonal description of music audio signals," Ph.D. dissertation, Universitat Pompeu Fabra (UPF), 2006.

[3] M. Müller, *Information Retrieval for Music and Motion.* Springer, 2007.

[4] V. Arifi, M. Clausen, F. Kurth, and M. Müller, "Synchronization of music data in score-, MIDI- and PCM-format," *Computing in Musicology*, vol. 13, pp. 9–33, 2004.

[5] R. Dannenberg and C. Raphael, "Music score alignment and computer accompaniment," *Communications of the ACM, Special Issue*, vol. 49, no. 8, pp. 39–43, 2006.

[6] S. Dixon and G. Widmer, "MATCH: A music alignment tool chest," in *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, London, GB, 2005, pp. 492–497.

[7] N. Hu, R. Dannenberg, and G. Tzanetakis, "Polyphonic audio matching and alignment for music retrieval," in *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, 2003, pp. 185–188.

[8] F. Kurth, M. Müller, C. Fremerey, Y. Chang, and M. Clausen, "Automated synchronization of scanned sheet music with audio recordings," in *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, Vienna, Austria, 2007, pp. 261–266.

[9] W. Chai, "Semantic segmentation and summarization of music: methods based on tonality and recurrent structure," *IEEE Signal Processing Magazine*, vol. 23, no. 2, pp. 124–132, 2006.

[10] R. Dannenberg and N. Hu, “Pattern discovery techniques for music audio,” in *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, Paris, France, 2002, pp. 63–70.

[11] M. Goto, “A chorus section detection method for musical audio signals and its application to a music listening station,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 5, pp. 1783–1794, 2006.

[12] M. Müller and F. Kurth, “Towards structural analysis of audio recordings in the presence of musical variations,” *EURASIP Journal on Advances in Signal Processing*, vol. 1 (Article ID 89686), 2007.

[13] G. Peeters, “Sequence representation of music structure using higher-order similarity matrix and maximum-likelihood approach,” in *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, Vienna, Austria, 2007, pp. 35–40.

[14] C. Rhodes and M. Casey, “Algorithms for determining and labelling approximate hierarchical self-similarity,” in *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, Vienna, Austria, 2007, pp. 41–46.

[15] M. Casey and M. Slaney, “Song intersection by approximate nearest neighbor search,” in *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, Victoria, Canada, 2006, pp. 144–149.

[16] D. Ellis and G. Poliner, “Identifying cover songs with chroma features and dynamic programming beat tracking,” in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Hawai’i, USA, 2007, pp. 1429–1432.

[17] P. H. J. Serrà, E. Gómez and X. Serra, “Chroma binary similarity and local alignment applied to cover song identification,” *IEEE Transactions on Audio, Speech and Language Processing*, vol. 16, pp. 1138–1151, August 2008.

[18] C. Fremerey, M. Müller, F. Kurth, and M. Clausen, “Automatic mapping of scanned sheet music to audio recordings,” in *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, Philadelphia, USA, 2008, pp. 413–418.

[19] F. Kurth and M. Müller, “Efficient index-based audio matching,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 2, pp. 382–395, 2008.

[20] J. Pickens, J. P. Bello, G. Monti, T. Crawford, M. Dovey, M. Sandler, and D. Byrd, “Polyphonic score retrieval using polyphonic audio,” in *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, Paris, France, 2002, pp. 140–149.

[21] S. B. Davis and P. Mermelstein, “Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 4, pp. 357–366, 1980.

[22] J.-J. Aucouturier and F. Pachet, “Improving timbre similarity: How high’s the sky,” *Journal of Negative Results in Speech and Audio Sciences*, vol. 1, 2004.

[23] H. Terasawa, M. Slaney, and J. Berger, “The thirteen colors of timbre,” in *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, USA, 2005, pp. 323–326.

[24] M. Müller, F. Kurth, and M. Clausen, “Audio matching via chroma-based statistical features,” in *Proceedings of the International Conference on Music Information Retrieval (ISMIR), London, GB*, 2005, pp. 288–295.

[25] K. Lee and M. Slaney, “Acoustic chord transcription and key extraction from audio using key-dependent HMMs trained on synthesized audio,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 2, pp. 291–301, 2008.

[26] J. C. Brown, “Calculation of a constant Q spectral transform,” *Journal of the Acoustical Society of America*, vol. 89, no. 1, 1991.

[27] A. Klapuri, “Multipitch analysis of polyphonic music and speech signals using an auditory model,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 2, pp. 255–266, 2008.

[28] C. Cannam, C. Landone, M. Sandler, and J. P. Bello, “The Sonic Visualiser: A visualisation platform for semantic descriptors from musical signals,” in *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, Victoria, Canada, 2006.

[29] O. Lartillot and P. Toiviainen, “MIR in Matlab (II): A toolbox for musical feature extraction from audio,” in *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, Vienna, Austria, 2007, pp. 127–130.

[30] L. R. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*. Prentice Hall Signal Processing Series, 1993.

[31] B. Logan, “Mel frequency cepstral coefficients for music modeling,” in *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, Plymouth, USA, 2000, pp. 11–23.

[32] G. Tzanetakis and P. Cook, “Musical genre classification of audio signals,” *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 5, pp. 293–302, July 2002.

[33] A. Eronen and A. Klapuri, “Musical instrument recognition using cepstral coefficients and temporal features,” in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Istanbul, Turkey, 2000.

[34] N. C. Maddage, C. Xu, M. S. Kankanhalli, and X. Shao, “Content-based music structure analysis with applications to music semantics understanding,” in *Proceedings of the ACM international conference on Multimedia*, New York, USA, 2004, pp. 112–119.

[35] J. G. Proakis and D. G. Manolakis, *Digital Signal Processsing*. Prentice Hall, 1996.

[36] E. Zwicker and H. Fastl, *Psychoacoustics, facts and models*. Springer Verlag, 1990.

[37] A. P. Klapuri, A. J. Eronen, and J. Astola, “Analysis of the meter of acoustic musical signals.” *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 1, pp. 342–355, 2006.

[38] M. Goto, “A chorus-section detecting method for musical audio signals,” in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Hong Kong, China*, 2003, pp. 437–440.

[39] M. Müller and M. Clausen, “Transposition-invariant self-similarity matrices,” in *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, Vienna, Austria, 2007, pp. 47–50.

Meinard Müller studied mathematics (Diplom) and computer science (Ph.D.) at Bonn University, Germany. In 2002/2003, he conducted postdoctoral research in combinatorics at the Mathematical Department of Keio University, Japan. In 2007, he finished his Habilitation at Bonn University in the field of multimedia retrieval writing a book titled *Information Retrieval for Music and Motion*, which appeared as Springer monograph. Currently, Meinard Müller is a member of the Saarland University and the Max-Planck Institut für Informatik, where he leads the research group *Multimedia Information Retrieval & Music Processing* within the Cluster of Excellence on *Multimodal Computing and Interaction*. His recent research interests include content-based multimedia retrieval, audio signal processing, music processing, music information retrieval, and motion processing.

Sebastian Ewert received the M.Sc. degree (Diplom) in computer science from Bonn University, Germany, in 2007. He is currently pursuing his doctoral degree in the Multimedia Signal Processing Group headed by Prof. Michael Clausen, Bonn University, under the supervision of Meinard Müller. Sebastian Ewert has been a researcher in the field of music information retrieval since 2008. His research interests include audio signal processing and machine learning with applications to automated music processing. His particular interests concern the design of musically relevant audio features as well as music synchronization and source separation techniques.