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 out 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 author is funded by the Cluster of Excellence on Multimodal Computing and Interaction (MMCI)

[†] The author is funded by the German Research Foundation (DFG CL 64/6-1)

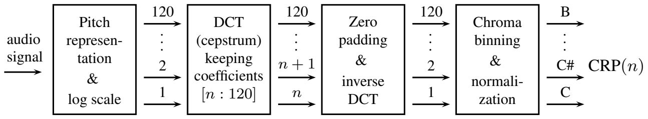
audio signal
Pitch represen-tation & log scale
120
⋮
2
1
DCT (cepstrum) keeping coefficients [$n:120$]
120
⋮
$n+1$
n
Zero padding & inverse DCT
120
⋮
2
1
Chroma binning & normali-zation
B
⋮
C#
C
CRP(n)

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

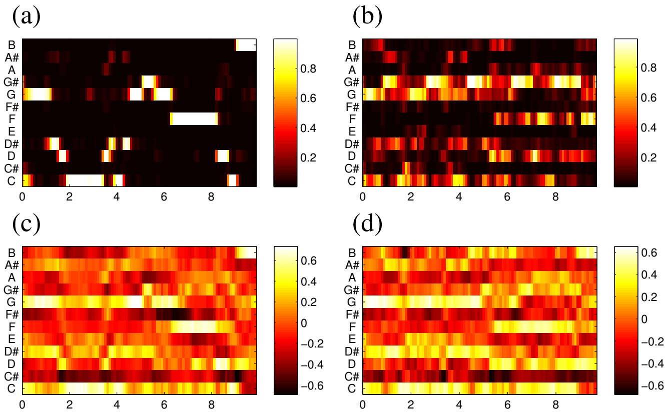

Fig. 2. Various chromagrams of the theme's beginning of the second Waltz, Jazz Suite No. 2 by Shostakovich. (a)/(b): Conventional chromagram of string/trombone version. (c)/(d): CRP(55) chromagram of string/trombone version. All chroma vectors are normalized.

rectangular window corresponding to 200 ms with a 50% overlap. The resulting feature representation has a resolution of 10 Hz (10 features per second) and is referred to as *pitch representation*. To obtain a conventional chroma representation (chromagram), one adds up the corresponding values of the pitch representation that belong the same chroma yielding a 12-dimensional vector for each analysis window. We refer to [7] for details and to Fig. 2 for an illustration.

For our novel audio features, we process the pitch representation before doing the chroma binning. The steps are similar to the ones in the computation of MFCCs [6], where one uses a mel scale instead of a pitch scale. First the pitch representation is logarithmized. Here, we replace each value v by $\log(C \cdot v + 1)$ with a positive constant C. In our experiments, $C = 1000$ turned out to be a suitable value, even though any value between 100 and 10000 produced a similar result. Then, we apply a discrete cosine transform (DCT) of size 120 to each of the 120-dimensional logarithmized pitch vectors. The resulting 120 coefficients have a similar interpretation as the MFCCs. In particular, the lower coefficients are related to timbre as observed by various researchers, see [1, 9] and the references therein. 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 $n-1$ coefficients for a parameter $n \in [1:120]$ by setting them to zero while leaving the upper coefficients unchanged. Each resulting 120-dimensional vector is then transformed by 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

Chroma type	μ_{O}	μ_{I}	ρ
Chroma-IF-Ellis	0.66	0.35	1.88
Chroma-P-Ellis	0.42	0.18	2.38
Chroma-Pitch	0.75	0.23	3.26
CRP(35)	0.99	0.10	10.30
CRP(55)	1.00	0.10	9.83
CRP(75)	1.00	0.11	8.76

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

chord chroma class. Let μ_{I} be the average over the resulting $220 \cdot \binom{48}{2}$ distances. Note that μ_{I} should be small in the case that the chroma type has a high degree of timbre invariance. Similarly, we computed the average distance μ_{O} over any two chroma vectors from different chord chroma class. Note that μ_{O} should be large to guarantee discriminate power of a chroma types. Finally, we form the quotient $\rho := \mu_{\mathrm{O}}/\mu_{\mathrm{I}}$ which expresses the across-class distance μ_{O} relative to the within-class distance μ_{I}. Table 1 shows the values μ_{I}, μ_{O}, and ρ for various chroma types. Note that the within-class distance drastically decreases for our CRP(n), while retaining the discriminative power even for large n.

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 $\Delta : [1 : L] \to \mathbb{R} \cup \{\infty\}$ between X and Y using a variant of dynamic time warping (DTW):

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

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 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 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.

Having this interpretation in mind, the two following properties of Δ 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 Δ close to zero (thus avoiding false negatives). We capture this property by defining μ_{I}^X and $\max_{\mathrm{I}}^X$ to be the average respective maximum of Δ 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). Here, we define μ_{O}^X and $\min_{\mathrm{O}}^X$ to be the average respective minimum of Δ over all indices outside these neighborhoods. From the above, it is clear that μ_{I}^X and $\max_{\mathrm{I}}^X$ should be small whereas μ_{O}^X and $\min_{\mathrm{O}}^X$ should be large. Similar to Sect. 3, we express these two properties within a single number, respectively, by defining the quotients $\rho_\mu^X = \mu_{\mathrm{O}}^X / \mu_{\mathrm{I}}^X$ and $\rho_{\min}^X = \max_{\mathrm{I}}^X / \min_{\mathrm{O}}^X$.

We illustrate the definition of Δ 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_4 denote the corresponding excerpts in the first and E_5 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_5) are indeed indicated by local minima of the resulting distance function Δ, 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 Δ-values. Now, using our 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. 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 μ_{I}^X, $\max_{\mathrm{I}}^X$, μ_{O}^X, $\min_{\mathrm{O}}^X$, ρ_μ^X, and $\rho_{\min}^X$. Averaging over all 90 queries, we obtain the corresponding numbers μ_{I}, $\max_{\mathrm{I}}$, μ_{O}, $\min_{\mathrm{O}}$, ρ_μ, and $\rho_{\min}$. Note that ρ_μ is not the quotient of

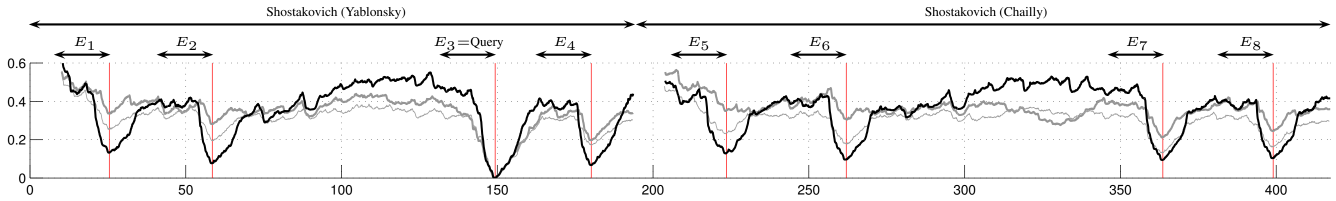

Fig. 3. Different distance functions shown for two recordings (Yablonsky, Chailly) of the Shostakovich example using the excerpt E_3 as query. The following chroma types were used: Chroma-IF-Ellis (thin gray), Chroma-Pitch (bold gray) and CRP(55) (black). For the query, there are 8 annotated excerpts.

Chroma type	μ_O	$\min_\mathrm{O}$	μ_I	$\max_\mathrm{I}$	ρ_μ	ρ_min
Chroma-IF-Ellis	0.41	0.25	0.16	0.20	2.74	1.33
Chroma-P-Ellis	0.19	0.11	0.06	0.08	3.15	1.35
Chroma-Pitch	0.48	0.26	0.15	0.20	3.84	1.53
CRP(35)	0.68	0.31	0.14	0.18	5.81	1.99
CRP(55)	0.64	0.27	0.12	0.16	6.24	2.00
CRP(75)	0.57	0.21	0.10	0.14	5.97	1.62

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

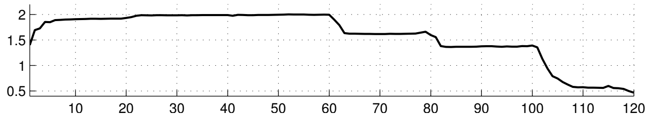

Fig. 4. Dependence of the performance measure ρ_min on the parameter $n \in [1 : 120]$ using CRP(n) features.

μ_O and μ_I, but the average of the ρ_μ^X. Analogously, this also holds for ρ_min.

Table 2 shows these numbers using various chroma type. For example, using the conventional chroma features (Chroma-Pitch), the average distance of the true matches is $\mu_\mathrm{I} = 0.15$, whereas the average distance outside the matches is $\mu_\mathrm{O} = 0.48$, resulting in a quotient $\rho_\mu = 3.84$. In view of the audio matching application, the values $\max_\mathrm{I}$ and $\min_\mathrm{O}$ are even more expressive: in case the maximal distance over the true matches is below the minimal distance outside the true matches (in this case one has $\rho_\mathrm{min}^X > 1$), all true matches will appear as the top matches. In this case, the true matches are separated from spurious matches. With respect to this measure, our novel CRP(n) features achieve a significant improvement. For example, one obtains $\rho_\mathrm{min} = 2.00$ when using CRP(55) features, whereas one has $\rho_\mathrm{min} = 1.53$ when using the conventional chroma features (Chroma-Pitch).

In a final experiment, we computed for each $n \in [1 : 120]$ the performance measure ρ_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

[1] 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.

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

[3] D. Ellis, "Chroma features analysis and synthesis," `http://www.ee.columbia.edu/~dpwe/`, 2007.

[4] D. Ellis and G. Poliner, "Identifying Cover Songs With Chroma Features and Dynamic Programming Beat Tracking," in *Proc. IEEE ICASSP*, 2007.

[5] 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, Feb. 2008.

[6] B. Logan, "Mel frequency cepstral coefficients for music modeling," in *Proc. ISMIR*, Plymouth, USA, 2000.

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

[8] J. Serrà, E. Gómez, P. Herrera, and X. Serra, "Chroma binary similarity and local alignment applied to cover song identification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 6, pp. 1138–1151, Aug. 2008.

[9] H. Terasawa, M. Slaney, and J. Berger, "The thirteen colors of timbre," in *Proc. IEEE WASPAA, New Paltz, NY, USA*, 2005, pp. 323–326.