Towards Timbre-Invariant Audio Features for Harmony-Based Music

Sebastian Ewert[1], Meinard Müller[2], Michael Clausen[1]
[1] *Bonn University, Germany, Email: {ewerts,clausen}@iai.uni-bonn.de*
[2] *Saarland University and MPI Informatik, Germany, Email: meinard@mpi-inf.mpg.de*

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, 3, 4, 6, 7]. An important step of the chroma feature calculation is the grouping of spectral energy components that belong to the same pitch class or chroma of the equal tempered scale. Here, the octave identification introduces a high degree of invariance to changes in timbre and instrumentation [2]. In particular, such features are useful in tasks such as cover song identification [3, 7] or audio matching [4, 6], 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 introduce a strategy to further increase this invariance by combining the concept of chroma features with the well-known concept of mel-frequency cepstral 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 [5]. The lower MFCCs are known to capture information on timbre [1, 8]. 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 the next section. After that, 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. Conclusions and prospects on future work are given in the last section.

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

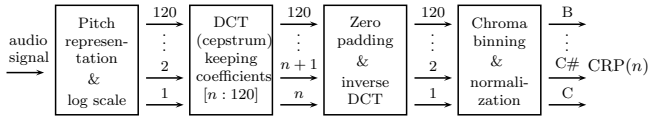

Figure 1: Overview of the computation of the CRP (chroma DCT-reduced log pitch) features.

1 to 120 using a suitable multirate filter bank. We then calculate the local energy for each of the subbands resulting in a sequence of 120-dimensional pitch feature vectors. We refer to this sequence as *pitch representation*. To obtain a conventional chroma representation, one adds up the components of a pitch feature vector that belong to the same chroma yielding a 12-dimensional chroma feature vector. We refer to [6] for details and to Fig. 2 (a) and (b) 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 [5], where one uses a mel scale instead of a pitch scale. First, the pitch representation is logarithmized by replacing each value v by $\log(C \cdot v + 1)$ with a positive constant C. In our experiments we used $C = 100$. Then, we apply a discrete cosine transform (DCT) 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, 8] 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. The theme of this piece appears four times played in four different instrumentations (clarinet, strings, trombone, tutti). Due to these differences, the resulting conventional chromagrams deviate strongly from each other. This is illustrated by Fig. 2 (a) and (b) showing the conventional chromagrams of the theme's beginning of the first (clarinet) 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.

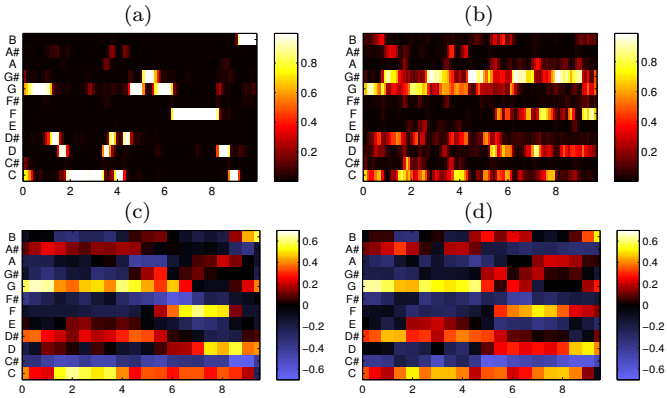

Figure 2: Various chromagrams of the theme's beginning of the second Waltz, Jazz Suite No. 2 by Shostakovich. (a)/(b): Conventional chromagram of clarinet/trombone version. (c)/(d): CRP(55) chromagram of clarinet/trombone version. All chroma vectors are normalized.

Experiments

We compared our CRP features to conventional chroma features using 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 [4, 6]. Here, the idea is to transform the query and an audio database into suitable feature sequences. Next, the query feature sequence is locally compared to the database feature sequence using a variant of *Dynamic Time Warping (DTW)* resulting in a distance value for each local comparison. Now, if a distance value is below a given threshold $\tau > 0$ then the corresponding excerpt from the database is returned as a match. For details we refer to [6].

For evaluation purposes 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. Altogether, the collection consists of 32 recordings amounting to 166 minutes of music. 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 for each of the queries. Using this database we conducted an experiment, to indicate the potential of the CRP features for music retrieval applications in terms of precision and recall. To this end, we computed for a fixed feature type the local distance values for each of the queries. Then, for a given positive distance threshold τ, we derive all matches having a distance below τ as described above. 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. As the diagram

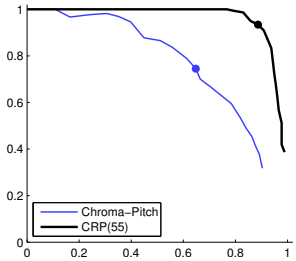

Figure 3: Retrieval performance in terms of precision (vertical axis) and recall (horizontal axis) values for conventional chroma and CRP(55) features.

indicates, one obtains much better PR-values using CRP features than in the case of conventional chroma features. A good indicator for this is the maximal F-value, which is indicated by a dot within the respective PR-diagram in Fig. 3. In our experiments, we obtained $\mathrm{F}_{\max} = 0.70$ for the conventional chroma features and $\mathrm{F}_{\max} = 0.91$ for the CRP(55) features, which is an improvement of more than 30% over the conventional features.

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. 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.

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, 2005.

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

[4] 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.

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

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

[7] 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, 2008.

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