MULTI-PROBE HISTOGRAMS: A MID-LEVEL HARMONIC FEATURE FOR MUSIC STRUCTURE SEGMENTATION

Florian Kaiser, Thomas Sikora

Communication Systems Group
Technische Universität Berlin
Berlin, Germany
kaiser@nue.tu-berlin.de, sikora@nue.tu-berlin.de

ABSTRACT

The use of mid-level audio features has recently shown promising perspectives for music content description tasks. In this paper we investigate the use of a mid-level harmony-related descriptor for the task of music structure analysis. The idea behind the descriptor is to concatenate the chroma sequence of a musical segment in a Multi Probe Histogram (MPH). Probing local pitch intervals within the chroma sequence, we show that such an abstraction of the audio signal is a good descriptor of the tonal hierarchy that is consequent to a given key or melodic line. It is thus a powerful feature for describing harmonic and melodic progressions within a music piece. After introducing the histograms' computation and enlightening their relation to harmony, we use such a description for the task of music structure segmentation and provide preliminary results that show very encouraging perspectives.

1. INTRODUCTION

Music structure analysis aims at drawing the temporal map of a music piece by extracting its constitutive structural parts. In classical music, such structural forms usually correspond to the first and second movements, development, exposition and so on. In popular music, common structural segments are often referred to as intro, verse, chorus and outro. As a front-end processing for many challenging applications such as content-based information retrieval and browsing, summarization and thumbnailling, the task of music structural analysis has gained an increasing interest in the Music Information Retrieval community.

Most approaches for this task aim at identifying repetitive patterns or segments of homogeneous acoustical information in low-level audio descriptors. While perceptual studies show that sound properties such as timbre and harmony allow to discriminate sections within a music piece in most western music, MFCC and Chroma vectors are often chosen as low-level audio features. Different strategies were proposed to detect structural boundaries and classify sections in the features distributions. A popular approach consists in embedding the audio features in an audio self-similarity matrix [1] [2]. The comparison of the feature vectors in a pairwise manner enlightens repetitive and/or homogeneous segments within the audio data. Structure is then derived using k-means clustering or HMM. Other approaches directly apply clustering techniques to the features distribution. A comparative study of most recent algorithms can be found in [3].

A limitation of using low-level descriptors for the description of a music piece is that acoustical information within musical sections is highly inhomogeneous. Therefore the extraction time-scale of the features, which is rather short, does not always yield a robust description of sections. Mid-level features or dynamic features were thus recently introduced to account for the temporal evolution of the feature vectors. In [4], authors model the temporal evolution of the spectral shape over a fixed time duration window. Varying the window size, authors can derive similarity matrices that either relate to short-term or long-term structures. In [5], Dynamic Texture is applied to the audio to model timbral and rhythmical properties. An alternative solution was proposed in [6] introducing a contextual distance that considers sequences of frames in order to enlighten repetitive patterns in the feature vectors.

We propose in this paper to use a mid-level harmony-related audio feature to describe musical structures, by means of a concatenation of mid-term chroma sequences in Multi-Probe Histograms (MPHs). MPHs were recently introduced in [7] for the task of scalable audio content retrieval and has not been used for musical audio content description yet. Probing local tone intervals within the sequence, such a representation allows to summarize the whole sequence by its dominant tone intervals and is thus an abstraction of its harmonic content.

After introducing the chroma representation of audio signals and the computation of MPHs, we discuss the musical interpretation of the resulting histograms. On the basis of musical knowledge and works on perception of tonal structures, we emphasize the harmony interpretation of MPH's. This is experimentally validated by analyzing the *Well-Tempered Clavier* books using this representation. Once the link between MPH and harmony has been established, MPHs are embedded in similarity matrices and used for the task of music structure segmentation. Evaluation of the system is reported at the end of the paper.

2. AUDIO SIGNAL REPRESENTATION

2.1. Chroma Features

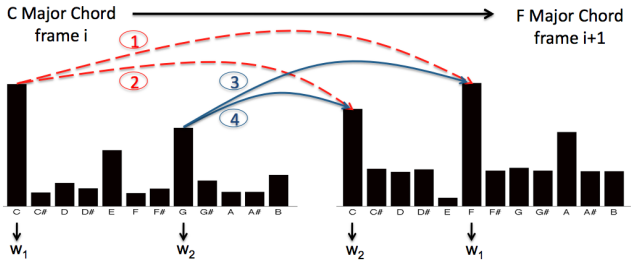
Chroma features are low-level audio descriptors that describe the pitch classes content of an audio data. Each coefficient of a chroma vector sums the signal's spectrum energy in sub-bands corresponding to one of the 12 pitch classes of the well-tempered scale. For the experiments in this paper, chroma features were extracted by means of the chroma toolbox[1]. The analysis of chroma features thus enables to focus on the structure of harmonic-related content within a music piece.

[1] http://www.mpi-inf.mpg.de/∼mmueller/chromatoolbox/

2.2. Multi-Probe Histograms

Chroma features are extracted over neighboring windows of the audio signal. There is thus a strong correlation between adjacent feature frames, unless a strong transition occurs in the audio signal, such as a note change, and in that case, only the value of first dominant chroma bins changes. Multi-Probe Histograms are based on that observation and aim at characterizing these local transitions between dominant bins in chroma sequences by probing the pitch classes intervals between adjacent frames. A deepen description of MPH's computation is beyond the scope of this paper and can be found in [7]. Nevertheless, we will briefly introduce their computation by means of a simple example in order to understand how tonal structures can be reflected in the histograms.

We consider a transition from a C Major chord (frame *i*) to a F Major chord (frame *i+1*) and their corresponding chroma vectors:



From frame *i* to frame *i+1*, there are 12^2 possibilities for the maximum of energy to be transferred from a pitch class to an other. Each of these possible transitions defines a position in the Multi-Probe Histogram. In our example, the energy is logically transferred from the pitch classes of the major triad of a C to the major triad of an F. For the MPH computation, we consider for each adjacent frames of the sequence the transition between the two dominant pitch classes, i.e. C and G for the C Major chord and F and C for the F Major chord. As illustrated above, the transition from frame *i* to frame *i+1* allows 4 possible transfers of energy between those pitch classes. For this iteration, bins of the MPH that will be allocated a new value are defined by these 4 transitions. Considering the transition from the tone C to the tone F, a bin position in the histogram is computed as follows :

$$p = b_C * 12 + b_F \tag{1}$$

with b_C and b_F the positions of the pitch classes C and F in the chroma vector, respectively 1 and 6. Furthermore, in each frame the first dominant bin is allocated the weight w_1 and the second the weight w_2. C and F being the first dominant bins in our example, the added weight for our histogram bin for the transition from C to F is defined as:

$$w = w_1 + w_1 \tag{2}$$

meaning that the histogram is added the value w at its bin p. The operation is then repeated for the remaining 3 pitch classes transitions, and iterated over the remaining frames of the chroma sequence. Note that the actual values of major pitch classes in the chroma vectors do not influence the histogram's value. Only the distance or interval between tones does. Of course, one can increase the number of chroma bins K to be considered from one frame to another, thus probing 2^K intervals at each time frame.

As illustrated in the remainder of this paper, MPH's can either be computed with the whole chroma sequence of an audio data, or just over a portion of the audio signal and used as a mid-level audio feature. Independently of the length of the chroma sequence, computed MPH's size does not vary and is thus composed of 12^2 = 144 bins.

3. MPH AS A HARMONIC FEATURE

Tonality and harmony relate to the combination of pitches in the chord constructions and melodic progressions of a music composition. In this section, we investigate how a MPH abstraction of chroma sequences is related to harmony and tonality. We first shortly review works on tonal structure perception in music cognition research. By means of the *well-tempered clavier books* we then experimentally show how key distances are reflected using the Multi-Probe Histograms as an audio feature.

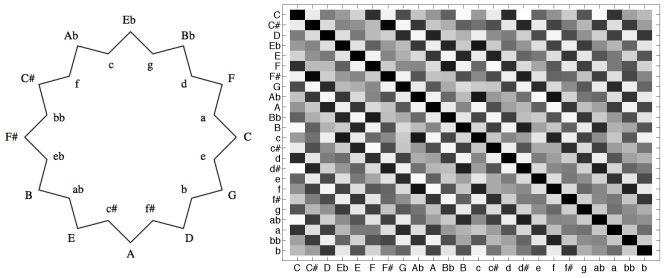
3.1. Tonal Structures

Works and studies on the perception of tone structures in the music cognition literature show a strong interdependency between tones, chords and key. By means of the probe tone experiment, i.e. people are asked to judge of the quality of a note in a given key context, Krumhansl estimated key profiles, or tonal hierarchies, that are produced by a given harmony. Moreover, she states in [8] that such tonal hierarchies generate a map of key distances that is the same of the chord distances, and is verified in the circle-of-fifth. This means that given a tonal context, the transition probabilities between chords and pitch classes of the well-tempered scale are not equally distributed in a music composition. Moreover, a study by Cohen in [9] on how a key becomes established in a music composition showed that the four notes of a music piece (in that case excerpts of the *Well-Tempered Clavier* Books) were sufficient for musically trained listeners to estimate the tonic of the key. This all suggests that a music composition and its tonal structure are characterized by a restricted set of discrete pitches and intervals. The tonal structure is even established in mid-term sections in music pieces. Considering local pitch intervals, MPH's are completely determined by the tonal hierarchy of a key context, and should therefore be a good feature, or abstraction, for describing the particular tonal structure of a music piece.

3.2. Experimental Validation

The Well-Tempered Clavier books consist of preludes and fugues composed in all 24 major and minor keys and are considered as a reference work on harmony. In order to experimentally validate the interpretation of Multi-Probe Histograms as harmonic features, we extracted MPHs on the 24 preludes of the *The Well-Tempered Clavier* books and measured the distance between the pieces comparing their corresponding histogram. The goal is to find wether or not the MPH based comparison of the pieces satisfies the map of key distances that is defined by the circle-of-fifth. Note that Cohen verified in [9] that consonance between Bach Preludes are consistent with the circle-of-fifth.

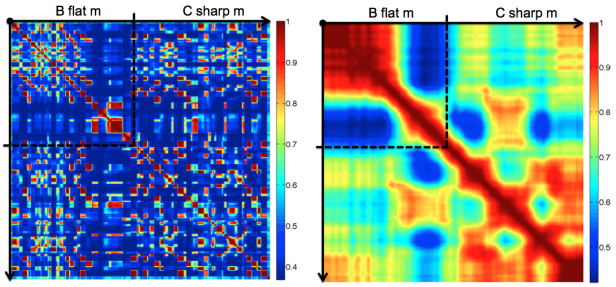
For each piece, the chroma vectors are extracted with a sampling rate of 10Hz and are concatenated in a MPH. A piece is thus modeled by a single 144 bins histogram. Distance between the pieces is measured by calculating the cosine distance between the histograms. The results are shown in Figure 1.



(a) Circle of Fifths

(b) MPH distance matrix between 24 Preludes

Figure 1: Circle of Fifths and MPH similarity between the 24 Preludes of the *Well-Tempered Clavier* Books



(a) Similarity Matrix with Chroma Vectors

(b) MPH Similarity Matrix, L = 50 frames

Figure 2: Similarity matrices computed over a portion of the *Mazurka, Op. 63, No. 3*. Transition between B flat minor and C Sharp minor.

Computing the similarity between the pieces, we verify that keys that are close to each other in the circle-of-fifth, and that are thus consonant, also have a high MPH similarity. On the other hand, pieces composed in keys that are highly distant are highly dissimilar in their MPH representations. Whereas only measuring the similarity between the chroma sequences is not sufficient to yield that result, concatenating the sequences in MPHs thus allows to highlight the relevant tone intervals that are consequent to a given key. And in that sense it can be used as a good and computationally inexpensive harmonic descriptor of music.

4. STRUCTURE DETECTION

A Multi-Probe Histogram can be extracted over chroma sequences of variable sizes. It can thus be used as an abstraction of a whole music piece, but could also define a mid-level audio feature for music content description. In this section we investigate the use of MPH's as a mid-level audio feature for the task of music structural segmentation.

4.1. MPH as a mid-level audio feature

For the description of a music piece, we embed the MPH's as mid-level audio features in an audio self-similarity matrix. This means that instead of calculating the feature frame pairwise distance, as proposed in [1], each time instant is now modeled by the MPH calculated on a sequence of L frames that surrounds it. Thus including more contextual information in the measure of similarity, we intend to provide a more homogeneous description of structural parts. We can illustrate that with a simple example: let's consider a same single note, for example an A, that is played in two different sections of a music piece, and thus in two different melodic, and eventually tonal, contexts. The pairwise similarity between the feature frames extracted over these two notes will be maximal and rise confusion, whereas if one introduces contextual information, awareness of the past- and forthcoming tonal structure is considered and the similarity between the two time instants is reduced.

We show a concrete example in Figure 2. The audio excerpt is a 30 seconds excerpt of Chopin's *Mazurka, Op. 63, No. 3* in which a tonality change from a B flat minor to a C Sharp minor occurs. In Figure 2.a, the standard similarity matrix as proposed in [1] is computed on the chroma features. In Figure 2.b, we compute our proposed similarity matrix with MPH's computed over chroma sequences of length $L = 50$ frames, which corresponds to 5 seconds.

The scales of B flat minor and C Sharp minor contain similar pitch classes. There is thus a high similarity in the chroma vectors and it is not clear from the chroma similarity matrix when in the music piece the tonality change occurs. But introduction of contextual information by means of the MPHs considerably reduces the similarity between the two sections. Moreover, each section tend to be more represented as a block of high similarity, satisfying the definition of a state representation of structure as introduced in [10]. For comparison, similarity matrices were also generated using the mid-term mean-value of Chroma vectors as features. While self-similarity within sections is also strengthened, high confusion between the two sections remains. We thus hope that the MPH representation will robustly enhanced the description of music pieces.

4.2. Segmentation and Structure Clustering

The Segmentation step aims at estimating the potential boundaries between the structural parts of a music piece. We use for that purpose the audio novelty approach as described in [11]. This method has already shown good performances for the task of structure segmentation.

We use the MPH enhanced similarity matrices as input for the structure clustering algorithm described in [12]. The algorithm is based on a nonnegative matrix factorization (NMF) of similarity matrices that separates musical sections in the matrix. The approach tends to work better when sections are displayed as blocks of high similarity in the matrix. This is referred to in the literature as the state representation of structure. Computing the standard similarity matrix on the low-level descriptors, structure is however rarely displayed in that manner. As shown above, introducing contextual information with the MPHs strengthens such a state representation and we therefore hope to improve the performance of the structure segmentation using our MPH based similarity matrix.

5. EVALUATION

5.1. Evaluation set-up and evaluation metrics

In order to compare our approach to the state of the art, we run the evaluation on the *TUT Beatles*[2] data set that consists of 174

[2]http://www.cs.tut.fi/sgn/arg/paulus/structure.html

songs from The Beatles. There is no ideal performance measure of music structure analysis algorithms. In fact musical structures being highly hierarchical, it is hard to find a match between the hierarchy-level of the annotation and the estimated structure. Nevertheless a compromise has been found in using the pairwise precision (P), recall (R) and F-measure (F), and the over- and under-segmentation scores (So and Su) introduced in [13].

The results are compared with the system in [14] that won the MIREX[3] music structure segmentation evaluation task in 2009 for the same dataset, and with the same clustering algorithm we use [12], but ran on standard chroma similarity matrices.

5.2. Results

The evaluation is reported in Table 1.

	MPH based Similarity Matrix	Standard Matrix Ref [12]	Ref [14]
F	63.3%	60.8%	60.0 %
P	59.3%	61.5%	56.1%
R	72.4%	64.6%	71.0%
So	68.3%	61%	73,9%
Su	58.8%	59.9%	61.7%

Table 1: Evaluation of the proposed approach and comparison with the state-of-the-art

The general increase in the F-measure is of 3% in comparison with the reference systems. While the algorithm seems to behave in a similar manner as in [14] (comparable Precison and Recall rates), the nature of the segmentation changes using MPHs instead of raw chroma vectors for the similarity matrix computation. Indeed, introduction of the MPH increases the Recall rate of 8% in comparison with [12] with a reasonable loss in Precision (2%). This means that our approach deals better with over-segmentation issues. Over-segmentation is indeed often a problem in structure segmentation because of the inner structure of musical sections. While this structure is reflected in the estimated segmentation, it doesn't match the hierarchy level of the annotated structure.

It is also to be noted that our approach reflects structure in the harmonic progression of the songs. However, in this database, many structural information is also contained in the instrumentation changes. It would therefore be appropriate in further work to study the impact of MPH with mono-instrumental recordings.

6. CONCLUSION

In this paper we showed that concatenating chroma sequences in Multi-Probe Histograms is efficient for describing tonal and harmonic properties of sounds. Varying the length of the studied chroma sequences, MPHs can either be utilized as global descriptors of music pieces, or as a mid-level feature for music content description. The first evaluation of its application to the task of structure segmentation shows very promising results. It is however important to keep in mind that evaluation methods for the task of structure segmentation are still under active discussions, and the performance measures do not reflect all aspects of the relevancy of an estimated segmentation. Indeed, there is a lack of accuracy in the definition of musical structures and annotation procedures. Further work will include the evaluation of the approach on a large mono-instrumental classical music database. Thus focusing on the harmonic aspects of structure, with no eventual confusion introduced by instrumentation changes, one could then run a deeper investigation on the benefits of using MPHs for the task of music structure analysis.

[3] http://www.music-ir.org/mirex/wiki/2009:Music_Structure_Segmentation_Results

7. ACKNOWLEDGMENT

This work was supported by the European Commission under contract FP7-21644 PetaMedia.

8. REFERENCES

[1] Jonathan Foote, "Visualizing music and audio using self-similarity," in *ACM Multimedia (1)*, 1999, pp. 77–80.

[2] Geoffroy Peeters, "Sequence representation of music structure using higher-order similarity matrix and maximum-likelihood approach," in *ISMIR*, 2007.

[3] Jouni Paulus, Meinard Müller, and Anssi Klapuri, "Audio-based music structure analysis," in *ISMIR*, 2010.

[4] Geoffroy Peeters, "Toward automatic music audio summary generation from signal analysis," in *ISMIR*, 2002, pp. 94–100.

[5] Luke Barrington, Antoni B. Chan, and Gert Lanckriet, "Modeling music as a dynamic texture," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 3, 2010.

[6] Meinard Müller and Frank Kurth, "Enhancing similarity matrices for music audio analysis," in *in Proc. IEEE ICASSP*, 2006.

[7] Y. Yu, M. Crucianu, V. Oria, and E. Damiani, "Combining multi-probe histogram and order-statistics based lsh for scalable audio content retrieval," in *ACM Multimedia*, 2010.

[8] Carol L. Krumhansl and Roger N. Shepard, "Quantification of the hierarchy of tonal functions within a diatonic context," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 5, no. 4, pp. 579–594, 1979.

[9] Annabel J. Cohen, "Tonality and perception: Musical scales primed by excerpts from the well-tempered clavier of j.s.bach," *Psychological Research*, vol. 53, no. 4, pp. 305–314, 1991.

[10] Geoffroy Peeters, "Deriving musical structures from signal analysis for music audio summary generation: "sequence" and "state" approach," in *CMMR*, 2003.

[11] Jonathan Foote, "Automatic audio segmentation using a measure of audio novelty," in *ICME*, 2000, p. 452.

[12] Florian Kaiser and Thomas Sikora, "Music structure discovery in popular music using non-negative matrix factorization," in *ISMIR*, 2010.

[13] Hanna M. Lukashevich, "Towards quantitative measures of evaluating song segmentation," in *ISMIRSMIR*, 2008.

[14] Matthias Mauch, Katy Noland, and Simon Dixon, "Using musical structure to enhance automatic chord transcription," in *ISMIR*, 2009.