Adaptive Temporal Modeling of Audio Features in the Context of Music Structure Segmentation

Florian Kaiser and Geoffroy Peeters

STMS IRCAM-CNRS-UPMC
1 Place Igor Stravinsky
75004 Paris
surname.name@ircam.fr

Abstract. This paper describes a method for automatically adapting the length of the temporal modeling applied to audio features in the context of music structure segmentation. By detecting regions of homogeneous acoustical content and abrupt changes in the audio feature sequence, we show that we can consequently adapt temporal modeling to capture both fast- and slow- varying structural information in the audio signal. Evaluation of the method shows that temporal modeling is consistently adapted to different musical contexts, allowing for robust music structure segmentation while gaining independence regarding parameter tuning.

Keywords: Music Structure Segmentation, Audio Features Temporal Modeling, Adaptivity

1 Introduction

The problem of music structure segmentation is concerned with the estimation of the largest structural entities that compose a music piece. A verse in Popular music, a bridge in Jazz music or a movement in Classical music constitute such structural entities. As a front-end processing for audio indexing applications such as audio browsing, summarization or annotation, the task knows a growing interest in the Music Information Retrieval research community.

Given the large musical spectrum, research in this field is limited to a very few tangible assumptions on the characteristics of musical structures. Formalization of the task has been however especially active since the introduction in [2] of the audio self-similarity matrix in which the pairwise similarity of an audio feature sequence is computed. Such matrices give a rather understandable visualization of the audio signal's content in terms of self-similarity and repetitions. The similarity matrix largely inspired the two main hypothesis that are made on the sections of a musical structure [9]. For the first hypothesis, i.e. the state hypothesis, a section is characterized by a strong inner-homogeneity within its acoustical content. Feature frames within that section thus activate a single state and methods such as HMM or clustering techniques can be applied to

estimate the states in the feature sequence. The state hypothesis closely relates to the concepts of verse and chorus in popular music. In the second hypothesis, i.e. the sequence hypothesis, sections are solely defined by their repetitions and are composed of sequences of unrelated feature frames. Such sequences can be illustrated by the repetition of a melody. The musical structure is then visualized as stripes on the off-diagonals of the similarity matrix highlighting the repeated patterns in the audio features. A comprehensive overview of music structure segmentation methods that were proposed under both hypothesis can be found in [9].

Musical structures are thus characterized by rather long-term musical patterns that should be captured in the signal description. While the temporal scale of such structural patterns can hardly be extracted by means of low-level audio features solely, part of the research in music structure segmentation has focused on the integration of context in the signal description. In [10] a dynamic feature that models the spectral envelope over a short period of time is proposed to embed contextual timbre information. In [1] Dynamic Texture is applied to model timbral and rhythmical properties of sounds. In [8] a contextual measure of similarity that considers sequences of feature frames instead of single frames in the similarity matrix computation allows to strengthen the visualization of repetitive patterns in the audio features. In [6], the evolution of the tonal context in the audio signal is described by concatenating mid-term chroma sequences in Multi-Probe Histograms [12].

Including that contextual information and bringing the signal description to the temporal level of musical sections thus means that temporal modeling of audio features is applied at some stage. However, the temporal scale of structural sections varies between and within music pieces and the choice of an adequate window length to apply this temporal modeling is crucial. Moreover, though allowing for a better characterization of musical sections, temporal modeling nevertheless significantly damages the detection of boundaries between these. See for example the similarity matrices proposed in [8] and [6]. Modeling of feature sequences of a few seconds is indeed in contradiction with abrupt changes in the audio signal that characterize boundaries between sections. Ideally, the length of temporal modeling should thus be chosen as large for the description of a section, and reduced at the border of sections.

In this paper we propose a method to automatically adapt the window length on which to apply temporal modeling over the audio signal. Detecting both regions of relative stability and strong variance in the audio features, the length of modeling is increased while in the middle of a section and reduced when a boundary might be encountered. Doing so, we show that we can increase the temporal segmentation between sections while keeping the benefit of temporal modeling for the characterization of sections.

We first introduce in Section 2 a system for music structure segmentation that applies temporal modeling with Multi-Probe Histograms and illustrate the need for adaptivity in this context. Our approach for adaptive temporal modeling for the introduced model is then presented in Section 3 and the benefit of the

approach with regard to the task is highlighted. Finally an evaluation of the impact of the adaptive temporal modeling on the music structure segmentation system is proposed in Section 4.

2 Adaptivity and Music Structure Segmentation

Adaptive windowing for temporal modeling is introduced in this paper in the context of the music structure segmentation system introduced in [6]. Temporal modeling is applied in this system in the sense that mid-term chroma features sequences are modeled by means of Multi-Probe Histograms in order to characterize tonal context variations. After a general introduction of the system, we present in this section the components of the system that may benefit from adaptivity in the length of the applied temporal modeling. The presented algorithm will serve in Section 4 to evaluate the impact of automatic temporal modeling length selection for the task.

2.1 System Overview

The general architecture of the system is a rather standard music structure segmentation architecture and is presented in Figure 1. A signal description that relates to the harmonic content is first estimated at the features extraction stage with the calculation of the chroma features.

Fig. 1: Music Structure Segmentation Overview

Tonal context is then derived from this description with the modeling of local chroma frames sequences with Multi-Probe Histograms (MPH). Therefore the chroma frames are split in subsequences of a given length, usually a couple of seconds, and each sequence is concatenated in a MPH. The MPH is a fixed-size representation of a chroma sequence which is determined by the dominant pitch classes transition between all adjacent frames of the sequence. A more detailed description can be found in [12]. Such histograms reflect the tonal structure of local portions of the audio and allow to model the evolution of the tonal context through time and capture slow varying harmonic patterns that relate to musical patterns. This description thus provides a good characterization of the inner-homogeneity of structural sections and also discriminates unrelated sections. Embedding the MPHs in a similarity matrix [2] strengthens the structure visualization. This is well illustrated with the similarity matrix extracted on the song *Things we said today* by the Beatles. Blocks of high similarity indicate

long-term homogeneous tonal sections and a clear structural representation can be visualized.

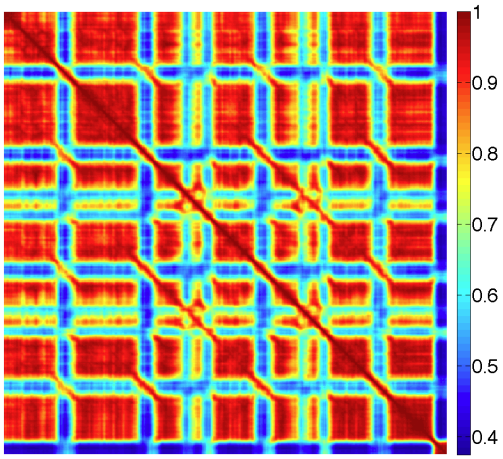


Fig. 2: Similarity Matrix for the song *Things we said today* by the Beatles

Based on this visualization of the structural information, music structure segmentation of a music piece consists in detecting the boundaries between its sections and group these segments together according to the musical structure. These two steps and their potential benefit in adaptive windowing are now presented.

2.2 Temporal Segmentation

A popular method based on similarity matrices for temporal segmentation was introduced in [3]. Arguing that section transitions are characterized by an abrupt change from one homogeneous acoustical content to another homogeneous acoustical content, the assumption behind the method is that boundaries in an audio signal are visualized in similarity matrices as 2-dimensional checkerboards such as the one represented in Figure 3. Such a kernel is actually the ideal template of a transition between two different states of a musical structure. Correlating this gaussian checkerboard kernel along with the main diagonal of the similarity matrix, a novelty curve can be computed and serve for the detection of boundaries in the audio signal.

As illustrated by the boundary detection kernel, temporal segmentation supposes that our signal description allows to visualize both the inner-homogeneity of sections and the precise transition time instants. The use of temporal modeling of chroma frames has however the drawback that it necessarily smoothens the boundaries between sections. In that sense, adapting the length of the modeling in order to increase it while in the middle of a section and reduce it when

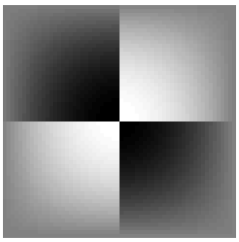


Fig. 3: Novelty detection kernel in [3]

narrowing a boundary may improve the performance of the temporal segmentation.

2.3 Segment Grouping

Once a temporal segmentation of the audio signal is estimated, grouping of segments according to the musical structure is obtained with the algorithm proposed in [5]. The method is based on the state hypothesis and uses the fact that information in similarity matrices is highly redundant over time in this hypothesis. Indeed, in the case of ideal states, structure is represented in the similarity matrix as uniform blocks. A single row or column of the similarity matrix thus contains the information of the whole state. Intuitively, the similarity matrix is thus ideally spanned by much lower dimensional basis, with a dimensionality that relates to the number of states in the musical structure. The method thus proposes to perform the dimension reduction of the similarity matrix by means of its Non-negative Matrix Factorization (NMF), and allows to group segments according to the structure by applying hierarchical clustering on the segments projected on this new basis.

The introduction of temporal modeling in the signal description is fully coherent with this algorithm in the sense that it strengthens the redundancy in the description of sections. While different sections of a same music piece may have various temporal scales, adaptive windowing can potentially enhance the description of the homogeneity in each section and consequently improve the segment grouping.

3 Adaptive Windowing

We present in this section our method for automatically adapting the length of temporal modeling with Multi-Probe Histograms over audio signals. After briefly introducing the approach, we describe the novelty score computation that allows to discriminate between homogeneous regions and abrupt changes in the audio features sequence. A method to consequently select the length of temporal modeling is then introduced.

3.1 Approach

Temporal modeling is generally applied by first splitting the audio feature sequence into a set of overlapping sub-sequences of a given length of N feature frames. Computing the Multi-Probe Histogram on the sub-sequences, one then captures information of slower variation than the one contained in the original audio feature sequence and thus embeds what we call contextual information. The core of the problem that is addressed in this paper now resides in the fact that the length of the subsequences should not be chosen as constant over the whole signal in order to capture both slow- and fast- varying structural information when it is needed.

One can easily illustrate the problem with a feature sub-sequence that would overlap two sections of the music piece. In that case, the goal of window adaptation consists in detecting in the sub-sequence the change point induced by the boundary in order to determine a stop criterion for the modeling. This way, frames of the sub-sequence that belong to the new section can be excluded from the current model. The consistency of each section modeling is consequently increased and the discrimination between the sections is preserved and not smoothened in the histogram modeling.

The first step of our method thus consists in the detection of potential break points within the feature subsequence that is currently modeled in a novelty score manner. This is discussed in the next subsection. Doing so over the whole signal, one can then consequently adapt the length of the sub-sequences over the audio signal. This part will be discussed in the second subsection.

3.2 Novelty Score Computation

Local Novelty Detection We consider the whole feature sequence extracted on a music piece and divide it into K overlapping sub-sequences of length N, N being the maximum chroma subsequence length that may be considered. Regardless of the musical structure, the dynamics of music usually implies that the original feature sequence is composed of an alternation of relatively stable regions and abrupt changes. Therefore, any frame in any sub-sequences potentially constitutes a border between these regions. To estimate these borders, our approach first consists in considering each feature sub-sequence independently, and probe significant changes within its elements.

Considering the k^{th} feature sub-sequence, we consider each of its frames as the potential division point of the sub-sequence. A novelty measure between all divisions in 2 of the subsequence then allows to estimate for all division point whether or not the subsequence is better modeled by a single histogram or by two histograms. The novelty score nov_k is computed for the k^{th} feature sub-sequence as follows:

$$nov_k(i) = distance(MPH_1^i, MPH_{i+1}^N) \tag{1}$$

with MPH_a^b the Multi-Probe Histogram computed on the sequence from frame a to frame b, and with the frame $i \in [\tau, N - \tau]$, τ being the minimum sequence length for the modeling. It is to be noted that the subsets are not necessarily of equal lengths and one should thus choose a normalized model or use a dynamics-independent distance. In our case, the novelty is thus computed by means of the cosine distance.

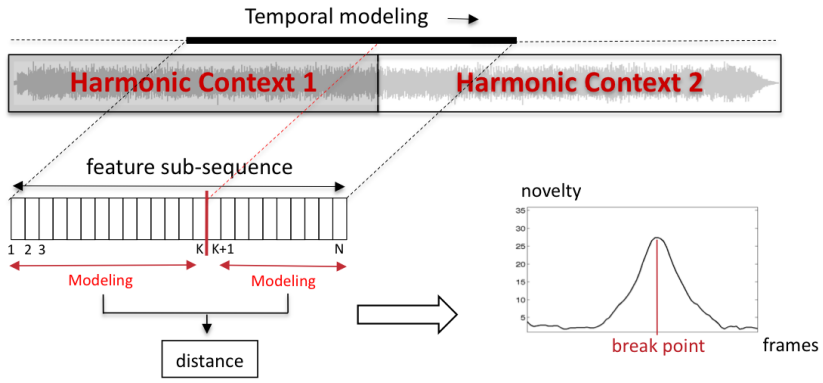


Fig. 4: Local novelty score computation within an audio signal

The approach is illustrated in Figure 4 with the example of two regions of different harmonic contexts in a feature sequence. Peaks that arise in the local novelty curve indicate likely changes in the audio signal with respect to the chosen audio features and temporal modeling. In contrast, regions of low novelty indicate temporal sequences of homogeneous acoustical content.

Sustainable Novelty Score Calculating a novelty score within each subsequence of N frames, we obtain K local novelty scores within the feature sequence. Say feature sub-sequences overlap with a hop size of one frame, each feature frame is thus probed $N - 2\tau$ times as a potential border between two contexts, each time as part of a different sub-sequence. In order to ensure that the borders that will be selected for the window adaptation are sustainable over all local novelty scores, we fuse all novelty scores in a single one that covers the whole audio signal. The novelty for each frame is therefore defined as the sum of the $N - 2\tau$ novelty values computed for it in each sub-sequence it is part of.

Considering feature frames as break points between different context ensures the reliability of the novelty score. The benefit of estimating boundaries as part of different contexts was already highlighted in [11].

3.3 Window Adaptation

Peaks and Troughs Detection Computing the audio features novelty score described above, one extracts break points in the feature sequence. In practice, these breakpoints are used to adapt the length of the temporal modeling applied to the audio features over time.

The novelty score is therefore used to discriminate between regions of rather homogeneous acoustical content, i.e. troughs in the novelty score, and regions of potential changes, i.e. peaks in the novelty score. Peaks and troughs are detected in the novelty score by means of the adaptive threshold estimation described in [4]. Therefore, a peak P in the novelty score is defined as the local maxima between two consecutive local minima, and a trough T as a local minima between two local maxima. Peaks-troughs are then selected as local maxima-minima values that exceed neighboring local minima-maxima values of at least a threshold δ as described in equation 2 and illustrated in Figure 5.

$$P_i \quad \equiv \quad T_i + \delta \leq P_i \cap T_{i+1} + \delta \leq P_i \tag{2}$$

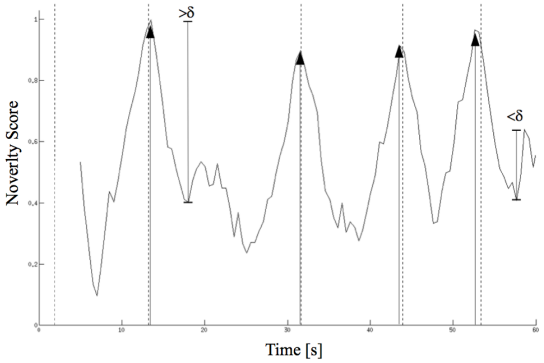


Fig. 5: Peaks and trough detection in the novelty curve

The threshold δ for the peak detection is automatically selected by clustering the novelty score into the classes of lowest and highest values, and setting δ to the mean value of the largest cluster. Supposing that there are more troughs than peaks in the novelty score, i.e. changes should not constantly occur in the music piece, we use a lower threshold value for the troughs detection in order to be less discriminative in their selection.

Window Length Selection Peaks and troughs in the novelty score constitute our reference points for adapting the window length N of the temporal modeling. Indeed both fast-varying and slow-varying structural information can be captured by setting N to a minimal value at peaks and to a maximal value at

troughs. To extend the window adaptation to the remaining time instants and thus the whole signal, only the time information of peaks and troughs is kept and their values set to N_{min} for peaks and N_{max} for troughs, with N_{min} and N_{max} respectively the minimum and maximum modeling lengths. Cubic interpolation of these points over time then allows to estimate an adapted window length for all time instant of the signal. Such an interpolation ensures a smooth variation of the temporal modeling length over time.

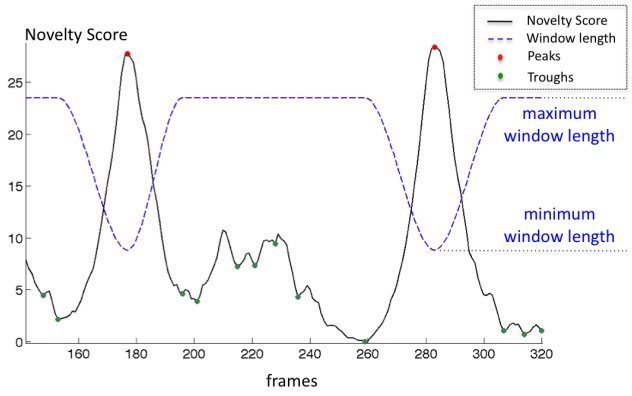


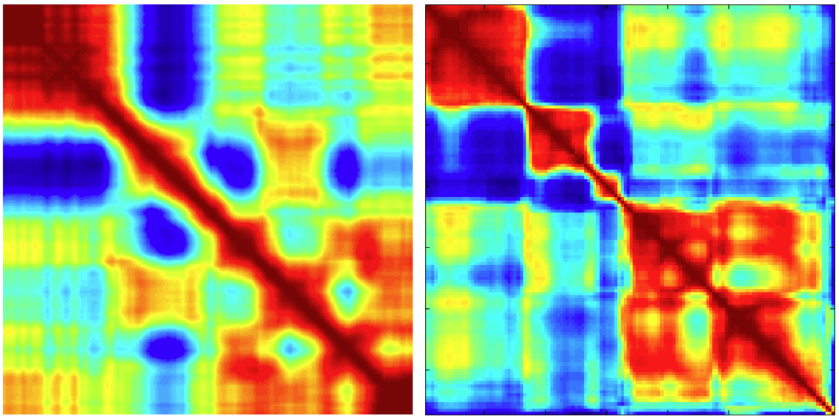
Fig. 6: Novelty score based window adaptation

The window selection method is illustrated in Figure 6. A novelty score and the peaks and troughs detected are represented. The interpolation of time points of maximal and minimal window lengths shows that the method allows to ideally select the window length according to the acoustical content. Examples of similarity matrices based on this adaptive temporal modeling concept shown in the next section of the paper illustrate that the method is indeed efficient to prevent smoothing effects between sections.

3.4 Application to Music Structure Visualization

A good example of the smoothing effect implied by temporal modeling is shown in Figure 7.a. The music piece used for this similarity matrix example is a 30 seconds excerpt of the *Mazurka op.63 n.3* composed by Chopin that is centered on a transition between two sections of different tonal contexts. The chroma vectors sampled at 10 Hz are extracted on the audio file and Multi-Probe Histograms computed on a sliding window of 5s, i.e. 50 frames, are embedded in a similarity matrix. While the discrimination between sections is globally clearly visible, one can hardly determine the exact boundary between the two sections.

We present in Figure 7.b a similarity matrix computed on the same audio example but this time using the adaptive temporal modeling for the computation



7.a MPHs over sequences of 50 chroma frames

7.b: MPHs over adapted chroma sequences

Fig. 7: MPH Similarity Matrices, Excerpt of Chopin's *Mazurka op.63 n.3*

of Multi-Probe Histograms. The maximum window length is set to 80 frames and the minimum window length to 10 frames. Doing so, the inner-homogeneity of the two sections is still well captured by our modeling and the boundary visualization between the two sections is considerably increased. Indeed, the visualization highlights two clear boundaries in the transition between the two sections, suggesting a progressive evolution between the two tonal contexts.

This thus suggests that our method may increase the precision of the temporal segmentation of similarity matrices that use temporal modeling of audio features while keeping the benefit for the characterization of structural sections.

4 Evaluation

We now present an evaluation of the impact of the adaptive temporal modeling introduced in this paper on the music structure segmentation system presented in Section 2. Therefore the system is evaluated using both a constant modeling length and adaptive modeling for the Multi-Probe Histograms computation. After introducing the evaluation test set and the evaluation metrics, results will be presented and discussed.

4.1 Test Set

The music structure segmentation system is evaluated on two different test sets:

- 18 pieces of the RWC Classic[1] annotated within the SALAMI Project[2]

[1] http://staff.aist.go.jp/m.goto/RWC-MDB/AIST-Annotation/
[2] http://ddmal.music.mcgill.ca/salami

- 180 songs of the Beatles annotated within the Isophonics corpus[3]

Besides the annotation of temporal boundaries, the structural information is annotated in the state hypothesis. All feature frames of a same section are thus labeled under a same state. Note that no semantic or functional labeling of the structure will be evaluated here. Only the retrieval of temporal boundaries and the grouping of segments under the correct structural states are evaluated.

4.2 Evaluation Metrics

Temporal Segmentation Evaluation The temporal segmentation step produces a set of boundaries between sections. The following terms are defined for its evaluation: the True Positives (TP) as the number of correctly retrieved boundaries within a given tolerance range, the False Positives (FP) as the number of unexpected retrieved boundaries, and the False Negatives (FN) as the number of missing boundaries. The precision P and recall R are then defined as in equations 3 and 4.

$$P = \frac{TP}{TP + FP} \tag{3}$$

$$R = \frac{TP}{TP + FN} \tag{4}$$

Both precision and recall can then be combined in the F-Measure defined in equation 5.

$$F = \frac{2PR}{P + R} \tag{5}$$

Segment Grouping Evaluation A largely consensual method to evaluate the frame labeling consists in using the pairwise precision, recall and F-measure introduced in [7]. Considering F_a the set of identically labelled frames in the reference annotation, and F_e the set of identically labelled frames in the estimated structure, the pairwise precision, recall and F-measure, respectively noted pw_P, pw_R and pw_F are then defined as:

$$pw\_P = \frac{|F_e \bigcap F_a|}{|F_e|} \tag{6}$$

$$pw\_R = \frac{|F_e \bigcap F_a|}{|F_a|} \tag{7}$$

$$pw\_F = \frac{2pw\_Ppw\_R}{pw\_P + pw\_R} \tag{8}$$

[3] http://isophonics.net/

4.3 Set Up

The "Constant Modeling" system is run with similarity matrices computed on chroma sequences Multi-Probe Histograms of three different lengths, i.e. 4, 6 and 8 seconds. The "Adaptive Modeling" system applies our adaptive temporal modeling length method with a maximal window of 8 seconds and a minimum window of 2s. The segment grouping is set to form 4 structural sections for each piece. The temporal segmentation evaluation is considered with a tolerance range of 2 seconds.

4.4 Results and Discussion

Structure segmentation performance evaluation for the RWC Classic and Isophonics Beatles datasets are reported in Tables 1 and 2 respectively.

	F	P	R
Constant Modeling - 2s	35,54	34,87	39,47
Constant Modeling - 4s	31,27	31,04	35,35
Constant Modeling - 6s	29,55	30,86	31,99
Constant Modeling - 8s	26,27	28,15	26,27
Adaptive Modeling	32,26	35,86	32,91

1.a: Temporal Segmentation [%]

	pw_F	pw_P	pw_R
Constant Modeling - 2s	47,33	56,74	45,04
Constant Modeling - 4s	48,54	50,90	52,34
Constant Modeling - 6s	49,36	52,32	53.09
Constant Modeling - 8s	51,48	55,30	53,69
Adaptive Modeling	52.45	54,81	55.14

1.b: Segment Grouping [%]

Table 1: Structure Segmentation Evaluation [%] - **RWC Classic Set**

The results that holds for both datasets is that the performance of temporal segmentation and segment grouping varies with the length of the modeling for the "Constant Modeling" method. Moreover, the lengths of modeling that provide the best temporal modeling results are not the same as the one that provide the best segment grouping performance. Indeed, for the RWC Classic dataset, reducing the length of modeling strongly increases the temporal segmentation performance. However, the homogeneity of structural sections is better captured with large temporal modeling, and hence, segment grouping is better handled

with large windows as illustrated by the results. In that context, our adaptive modeling acts coherently with the musical context and allows for both a good temporal segmentation and the best segment grouping performance.

	F	P	R
Constant Modeling - 2s	43,17	40,39	50,38
Constant Modeling - 4s	43,64	41,54	49,73
Constant Modeling - 6s	44,05	43,06	48,95
Constant Modeling - 8s	41,48	41,55	44,91
Adaptive Modeling	43,36	42,77	47,47

2.a: Temporal Segmentation [%]

	pw_F	pw_P	pw_R
Constant Modeling - 2s	59.37	52.68	73.94
Constant Modeling - 4s	60.34	54.28	73.53
Constant Modeling - 6s	59.29	52.47	74.16
Constant Modeling - 8s	57.54	51.60	70.36
Adaptive Modeling	61.17	55.59	73.51

2.b: Segment Grouping [%]

Table 2: Structure Segmentation Evaluation - **Isophonics Beatles Set**

Concerning the Isophonics Beatles set, the good temporal segmentation and segment grouping obtained with the 4 seconds constant modeling suggests that harmonic patterns in this dataset are of a lower temporal scale than in the RWC Classic dataset. There again, the adaptive method consistently adapted modeling windows to this musical context and shows the best segment grouping performance and a good temporal segmentation performance. This all suggests that the approach captures both long term musical patterns and abrupt changes in the audio content without any strict assumption on the temporal scale of the sections.

It is to be noted that the impact of adaptive modeling is sensitively higher on the segment grouping performance. Indeed, temporal segmentation performances did match but not overcome the performance of the best constant modeling algorithm. In contrast, adaptive modeling did outperform for both datasets the segment grouping performance given by fixed-length temporal modeling. While we ideally could consider the two tasks separately, i.e. segment the music piece with a small temporal modeling length and increase it for the segment grouping, the method introduced in this paper gives a single performant and computationally effective solution for both problems.

5 Conclusion

We proposed in this paper a method for the automatic adaptation of the temporal modeling of audio features for music segmentation purposes. Modeling chroma features by means of Multi-Probe Histograms we detect regions of homogeneous acoustical content where the length of the modeling can be increased, and changing contexts where the length should be reduced. We may in that manner describe both fast- and slow- varying musical patterns in music signals. The evaluation in the context of music structure segmentation indeed showed that our method adapts the temporal modeling length to the acoustical context and allows to capture both long-term harmonic patterns and precise breaks between these patterns. While parameter tuning for the modeling length is hardly generalizable for large datasets, our method allows us to be independent on any strict assumption on the temporal scale of structural sections. We believe that the approach could be efficiently extended to other temporal modeling and for example allow to enhance the description of timbral contexts. More generally, adaptivity in music signals description is a challenging research for music information retrieval. As an example, music interaction systems that face different kinds of musical events through time could benefit from this research.

6 Acknowledgments

This work was partly supported by the Quaero Program funded by Oseo French agency.

References

1. L. Barrington, A. B. Chan, and G. Lanckriet. Modeling music as a dynamic texture. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(3), 2010.
2. J. Foote. Visualizing music and audio using self-similarity. In *Proceedings of the ACM Multimedia*, pages 77–80, 1999.
3. J. Foote. Automatic audio segmentation using a measure of audio novelty. In *Proceedings of the IEEE International Conference on Multimedia and Expo*, 2000.
4. A. Jacobson. Auto-threshold peak detection in physiological signals. In *Engineering in Medicine and Biology Society, 2001. Proceedings of the 23rd Annual International Conference of the IEEE*, volume 3, pages 2194–2195 vol.3, 2001.
5. F. Kaiser and T. Sikora. Music structure discovery in popular music using non-negative matrix factorization. In *Proceedings of the 11th International Society for Music Information Retrieval Conference (ISMIR)*, aug 2010.
6. F. Kaiser and T. Sikora. Multi-probe histograms: A mid-level harmonic feature for music structure segmentation. In *Proceedings of the 14th International Conference on Digital Audio Effects (DAFx)*, Paris, France, Sept. 2011.
7. M. Levy and M. Sandler. Structural segmentation of musical audio by constrained clustering. *IEEE Transactions on Audio, Speech & Language Processing*, 16(2):318–326, 2008.

8. M. Mueller and F. Kurth. Enhancing similarity matrices for music audio analysis. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2006.
9. J. Paulus, M. Müller, and A. Klapuri. Audio-based music structure analysis. In *Proceedings of the 11th International Society for Music Information Retrieval Conference (ISMIR)*, 2010.
10. G. Peeters. Toward automatic music audio summary generation from signal analysis. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, pages 94–100, 2002.
11. G. Sargent, F. Bimbot, and E. Vincent. A regularity-constrained viterbi agorithm and its application to the structural segmentation of songs. *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, 2011.
12. 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.