IMPROVING MUSIC STRUCTURE SEGMENTATION USING LAG-PRIORS

Geoffroy Peeters
STMS IRCAM-CNRS-UPMC
geoffroy.peeters@ircam.fr

Victor Bisot
STMS IRCAM-CNRS-UPMC
victor.bisot@ircam.fr

ABSTRACT

Methods for music structure discovery usually process a music track by first detecting segments and then labeling them. Depending on the assumptions made on the signal content (repetition, homogeneity or novelty), different methods are used for these two steps. In this paper, we deal with the segmentation in the case of repetitive content. In this field, segments are usually identified by looking for sub-diagonals in a Self-Similarity-Matrix (SSM). In order to make this identification more robust, Goto proposed in 2003 to cumulate the values of the SSM over constant-lag and search only for segments in the SSM when this sum is large. Since the various repetitions of a segment start simultaneously in a self-similarity-matrix, Serra et al. proposed in 2012 to cumulate these simultaneous values (using a so-called structure feature) to enhance the novelty of the starting and ending time of a segment. In this work, we propose to combine both approaches by using Goto method locally as a prior to the lag-dimensions of Serra et al. structure features used to compute the novelty curve. Through a large experiment on RWC and Isophonics testsets and using MIREX segmentation evaluation measure, we show that this simple combination allows a large improvement of the segmentation results.

1. INTRODUCTION

Music structure segmentation aims at estimating the large-scale temporal entities that compose a music track (for example the verse, chorus or bridge in popular music). This segmentation has many applications such as browsing a track by parts, a first step for music structure labeling or audio summary generation [15], music analysis, help for advanced DJ-ing.

The method used to estimate the music structure segments (and/or labels) depends on the assumptions made on the signal content. Two assumptions are commonly used [13] [14]. The first assumption considers that the audio signal can be represented as a succession of segments with homogeneous content inside each segment. This assumption is named "homogeneity assumption" and the estimation approach named "state approach". It is closely related to another assumption, named "novelty", that considers that the transition between two distinct homogeneous segments creates a large "novelty". The second assumption considers that some segments in the audio signal are repetitions of other ones. This assumption is named "repetition assumption". In this case the "repeated" segments can be homogeneous or not. When they are not, the approach is named "sequence approach".

In this paper, we deal with the problem of estimating the segments (starting and ending times) in the case of repeated/ non-homogeneous segments ("sequence" approach).

1.1 Related works

Works related to music structure segmentation are numerous. We refer the reader to [13] or [3] for a complete overview on the topic. We only review here the most important works or the ones closely related our proposal.

Methods relying on the homogeneity or novelty assumption. Because homogeneous segments form "blocks" in a time-time-Self-Similarity-Matrix (SSM) and because transitions from one homogeneous segment to the next looks like a checkerboard kernel, Foote [5] proposes in 2000 to convolve the matrix with a 2D-checkerboard-kernel. The result of the convolution along the main diagonal leads to large value at the transition times. Since, an assumption on the segment duration has to be made for the kernel of Foote, Kaiser and Peeters [9] propose in 2013 to use multiple-temporal-scale kernels. They also introduce two new kernels to represent transitions from homogeneous to non-homogeneous segments (and vice versa). Other approaches rely on information criteria (such as BIC, Akaike or GLR) applied to the sequence of audio features. Finally, labeling methods (such as k-means, hierarchical agglomerative clustering of hidden-Markov-model) also inherently allow performing time-segmentation.

Methods relying on the repetition assumption. Because repeated segments (when non-homogeneous) form sub-diagonals in a Self-Similarity Matrix (SSM), most methods perform the segmentation by detecting these sub-diagonals in the SSM.

If we denote by $S(i,j) = S(t_i, t_j)\ i,j \in [1,N]$ the time-time-SSM between the pairs of times t_i and t_j, the time-lag-SSM [1] is defined as $L(i,l) = L(t_i, l = t_j - t_i)$, since $t_j - t_i \geq 0$ the matrix is upper-diagonal. The lag-matrix can be computed using $L(i,l) = S(i, j = i + l)$ with $i + l \leq N$.

In [6], Goto proposes to detect the repetitions in a time-lag-SSM using a two-step approach. He first detects the various lags l_k at which potential repetitions may occur. This is done by observing that when a repetition (at the same-speed) occurs, a vertical line (at constant lag) exists in the time-lag-SSM (see Figure 1). Therefore, the sum over the times of the time-lag-SSM for this specific lag will be large. He proposes to compute the function $f(l) = \sum_{t_i \in [0,N-l]} \frac{1}{N-l} L(t_i, l)$. A peak in $f(l)$ indicates that repetitions exist at this specific lag. Then, for each detected peaks l_k, the corresponding column of $L(t_i, l_k)$ is analyzed in order to find the starting and ending times of the segments.

Serra et al. method [16] for music structure segmentation also relies in the time-lag-SSM but works in the opposite way. In order to compute the lower-diagonal part of the matrix ($t_j - t_i < 0$), They propose to apply circular permutation. The resulting matrix is named circular-time-lag-matrix (CTLM) and is computed using $L^*(i,l) = S(i, k+1)$, for $i, l \in [1, N]$ and $k = i + l - 2$ modulo N. They then use the fact that the various repetitions of a same segment start and end at the same times in the CTLM. They therefore define a N-dimensional feature, named "structure feature" $\underline{g}(i)$, defined as the row of the CTLM at t_i. Start and end of the repetitions create large frame-to-frame variations of the structure feature. They therefore compute a novelty curve defined as the distance between successive structure features $\underline{g}(i)$: $c(i) = ||\underline{g}(i+1) - \underline{g}(i)||^2$ (see Figure 1). Large values in this curve indicate starts or ends times of repetitions.

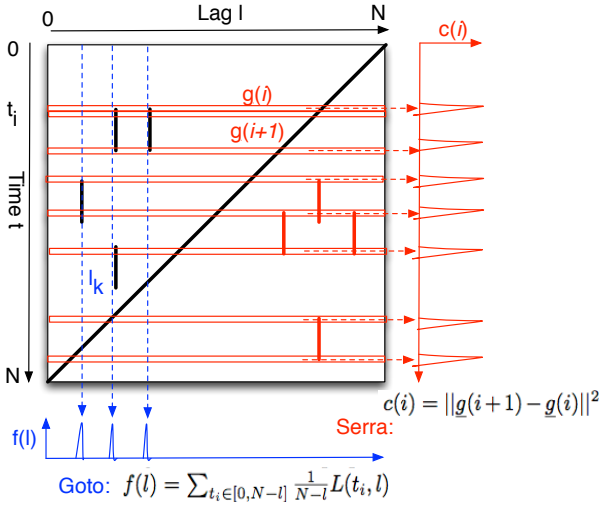

Figure 1. Illustration of Goto method [6] on a time-lag Self-Similarity-Matrix (SSM) and Serra et al. method [16] on a circular-time-lag-matrix (CTLM).

1.2 Paper objective and organization

In this paper, we deal with the problem of estimating the segments (starting and ending times) in the case of repeated/ non-homogeneous segments ("sequence" approach). We propose a simple, but very efficient, method that allows using Goto method as a prior lag-probability of segments in Serra et al. method. Indeed, Serra et al. method works efficiently when the "structure" feature $\underline{g}(i)$ is clean, i.e. contains large values when a segment crosses $\underline{g}(i)$ and is null otherwise. Since, this is rarely the case, we propose to create a prior assumption $f(l)$ on the dimensions of $\underline{g}(i)$ that may contain segments. To create this prior assumption, we use a modified version of Goto method applied locally in time to the CTLM (instead of to the time-lag-SSM).

Our proposed method for music structure segmentation is presented in part 2. We then evaluate it and compare its performance to state-of-the-art algorithms in part 3 using the RWC-Popular-Music and Isophonics/Beatles test-sets. Discussions of the results and potential extensions are discussed in part 4.

2. PROPOSED METHOD

2.1 Feature extraction

In order to represent the content of an audio signal, we use the CENS (Chroma Energy distribution Normalized Statistics) features [12] extracted using the Chroma Toolbox [11]. The CENS feature is a sort of quantized version of the chroma feature smoothed over time by convolution with a long duration Hann window. The CENS features $x_a(t_i)$ $i \in [1, N]$ are 12-dimensional vector with a sampling rate of 2 Hz. $x_a(t_i)$ is in the range $[0, 1]$. It should be noted that these features are l^2-normed [1].

2.2 Self-Similarity-Matrix

From the sequence of CENS features we compute a time-time Self-Similarity-Matrix (SSM) [4] $S(i,j)$ using as similarity measure the scalar-product [2] between the feature vector at time t_i and t_j: $S(i,j) =< x_a(t_i), x_a(t_j) >$.

In order to highlight the diagonal-repetitions in the SSM while reducing the influence of noise values, we then apply the following process.

1. We apply a low-pass filter in the direction of the diagonals and high-pass filter in the orthogonal direction. For this, we use the kernel $[-0.3, 1, -0.3]$ replicated 12 times to lead to a low-pass filter of duration 6 s.

2. We apply a threshold $\tau \in [0, 1]$ to the resulting SSM. τ is chosen such as to keep only β % of the values of the SSM. Values below τ are set to a negative penalty-value α. The interval $[\tau, 1]$ is then mapped to the interval $[0, 1]$.

3. Finally, we apply a median filter over the diagonals of the matrix. For each value $S(i,j)$, we look in the backward and forward diagonals of δ-points duration each $[(i-\delta, j-\delta) \dots (i,j) \dots (i+\delta, j+\delta)]$. If more than 50% of these points have a value of α, $S(i,j)$ is also set to α.

By experiment, we found $\beta = 6\%$ (percentage of values kept), $\alpha = -2$ (lower values) and $\delta = 10$ frames (interval duration [3]) to be good values.

[1] $\sum_{a=[1,12]} x_a^2(t_i) = 1$

[2] Since the vectors are l^2-normed, this is equivalent to the use of a cosine-distance.

[3] Since the sampling rate of $x_a(t_i)$ is 2 Hz, this corresponds to a duration of 5 s. The median filter is then applied on a window of 10 s total duration.

2.3 Proposed method: introducing lag-prior

As mentioned before, Serra et al. method works efficiently when the "structure" feature $\underline{g}(i)$ is clean, i.e. contains large values when a segment crosses $\underline{g}(i)$ and is null otherwise. Unfortunately, this is rarely the case in practice.

If we model the structure feature $\underline{g}(i)$ as the true contribution of the segments $\underline{\hat{g}}(i)$ and a background noise (modeled as a centered Gaussian noise) $\mathcal{N}_{\mu=0,\sigma}$: $\underline{g}(i) = \underline{\hat{g}}(i) + \mathcal{N}_{\mu=0,\sigma}$, one can easily shows that the expectation of $c(i) = ||\underline{g}(i+1) - \underline{g}(i)||^2$ is equal to

- $K + 2\sigma^2$ for the starting/ending of K segments at t_i
- $2\sigma^2$ otherwise.

If σ (the amount of background noise in the CTLM) is large, then it may be difficult to discriminate between both case for small K. In the opposite, the expectation of the values of Goto function $f(l) = \sum_{t_i} L^*(t_i, l)$ remains independent of σ hence on the presence of background noise (in the case of a centered Gaussian noise).

We therefore propose to use $f(l)$ as a prior on the lags, i.e. the dimensions of $\underline{g}(i)$. This will favor the discrimination provided by $c(i)$ (in Serra et al. approach, all the lags/dimensions of $g(i)$ are considered equally).

For this, we consider, the circular time-lag (CMLT) $L^*(t, l)$ as a joint probability distribution $p(t, l)$.

Serra et al. novelty curve $c(i)$ can be expressed as

$$c_1(t) = \int_l \left| \frac{\partial}{\partial t} p(t,l) \right|^2 dl \qquad (1)$$

In our approach, we favor the lags at which segments are more likely. This is done using a prior $p(l)$:

$$c_2(t) = \int_l p(l) \left| \frac{\partial}{\partial t} p(t,l) \right|^2 dl \qquad (2)$$

In order to compute the prior $p(l)$ we compute $f(l)$ as proposed by Goto but applied to the CMLT. In other words, we compute, the marginal of $p(t, l)$ over t: $p(l) = \int_{t=0}^{t=N} p(t, l)dt$.

As a variation of this method, we also propose to compute the prior $p(l)$ locally on t: $p_t(l) = \int_{\tau=t-\Delta}^{\tau=t+\Delta} p(\tau, l)dt$. This leads to the novelty curve

$$c_3(t) = \int_l p_t(l) \left| \frac{\partial}{\partial t} p(t,l) \right|^2 dl \qquad (3)$$

By experiment, we found $\Delta = 20$ (corresponding to 10 s), to be a good value.

2.4 Illustrations

In Figure 2, we illustrate the computation of $c_1(t)$, $c_2(t)$ and $c_3(t)$ on a real signal (the track 19 from RWC Popular Music).

In Figure 2 (A) we represent Serra et al. [16] method. On the right of the time-lag-circular-matrix (CTLM), we represent the novelty curve $c_1(t)$ (red-curve) and superimposed to it, the ground-truth segments (black dashed lines).

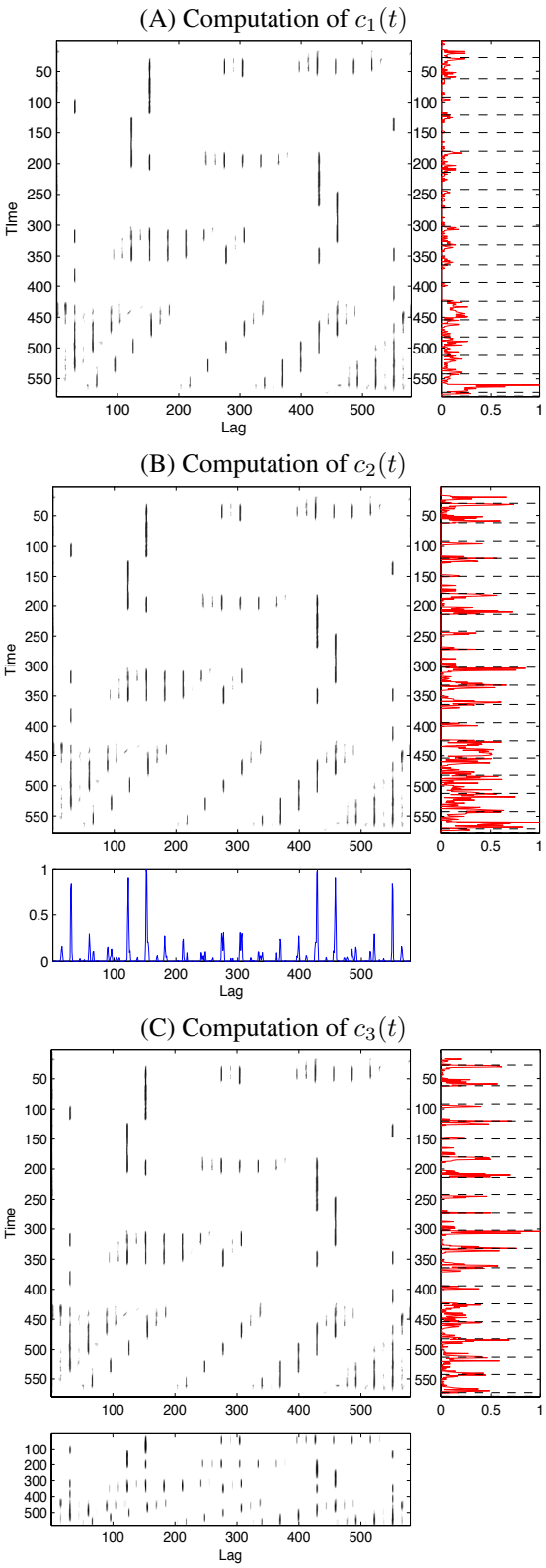

Figure 2. Illustration of the computation of $c_1(t)$, $c_2(t)$ and $c_3(t)$ on Track 19 from RWC Popular Music. See text of Section 2.4 for explanation.

In Figure 2 (B) we represent the computation of $c_2(t)$ (using a global lag-prior). Below the CTLM we represent the global prior $p(l)$ (blue curve) obtained using Goto method applied to the CMLT. On the right of the CTLM

we represent $c_2(t)$ using this global lag-prior. Compared to the above $c_1(t)$, we see that $c_2(t)$ allows a larger discrimination between times that correspond to ground-truth starts and ends of segments and that do not.

In Figure 2 (C) we represent the computation of $c_3(t)$ (using a local lag-prior). Below the CTLM we represent the local prior $p_t(l)$ in matrix form obtained using Goto method applied locally in time to the CMLT. On the right of the CTLM we represent $c_3(t)$ using this local lag-prior. Compared to the above $c_1(t)$ and $c_2(t)$, we see that $c_3(t)$ allows an even larger discrimination.

2.5 Estimation of segments start and end times

Finally, we estimate the starting and ending time of the repetitions from the novelty curves $c_1(t)$, $c_2(t)$ or $c_3(t)$. This is done using a peak picking process. $c_*(t)$ is first normalized by min-max to the interval $[0, 1]$. Only the values above 0.1 are considered. t_i is considered as a peak if $i = \arg\max_j c_*(t_j)$ with $j \in [i-10, i+10]$, i.e. if t_i is the maximum peak within a $\pm$ 5 s duration interval.

The flowchart of our Music Structure Segmentation method is represented in the left part of Figure 3.

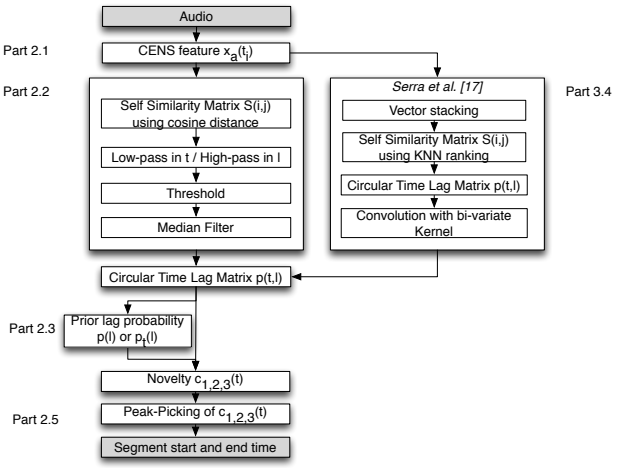

Figure 3. Flowchart of the proposed Music Structure Segmentation method.

3. EVALUATION

In this part, we evaluate the performances of our proposed method for estimating the start and end times of music structure segments. We evaluate our algorithm using the three methods described in part 2.3: – without lag-prior $c_1(t)$ (this is equivalent to the original Serra et al. algorithm although our features and the pre-processing of the CTLM differ from the ones of Serra et al.), – with global lag-prior $c_2(t)$, – with local lag-prior $c_3(t)$.

3.1 Test-Sets

In order to allow comparison with previously published results, we evaluate our algorithm on the following test-sets:

RWC-Pop-A: is the RWC-Popular-Music test-set [8], which is a collection of 100 music tracks. The annotations into structures are provided by the AIST [7].

RWC-Pop-B is the same test-set but with annotations provided by IRISA [2] [4] .

Beatles-B Is the Beatles test-set as part of the Isophonics test-set, which is a collection of 180 music tracks from the Beatles. The annotations into structure are provided by Isophonics [10].

3.2 Evaluation measures

To evaluate the quality of our segmentation we use, as it is the case in the MIREX (Music Information Retrieval Evaluation eXchange) Structure Segmentation evaluation task, the Recall (R), Precision (P) and F-Measure (F). We compute those with a tolerance window of 3 and 0.5 s.

3.3 Results obtained applying our lag-prior method to the SSM as computed in part 2.2.

In Table 1 we indicate the results obtained for the various configurations and test-sets. We compare our results with the ones published in Serra et al. [16] and to the best score obtained during the two last MIREX evaluation campaign: MIREX-2012 and MIREX-2013 on the same test-sets [5] [6] .

For the three test-sets, and a 3 s tolerance window, the use of our lag-prior allows a large increase of the F-measure:
RWC-Pop-A: $c_1(t) : 66.0\%$, $c_2(t) : 72.9\%$, $c_3(t) : 76.9\%$.
RWC-Pop-B: $c_1(t) : 67.3\%$, $c_2(t) : 72.6\%$, $c_3(t) : 78.2\%$.
Beatles-B: $c_1(t) : 65.7\%$, $c_2(t) : 69.8\%$, $c_3(t) : 76.1\%$.

For the 0.5 s tolerance window, the F-measure also increase but in smaller proportion.

The F-measure obtained by our algorithm is just below the one of [16], but our features and pre-processing of the SSM much simpler. This means that applying our lag-priors to compute $c_{2,3}(t)$ on Serra et al. pre-processed matrix could even lead to larger results. We discuss this in the next part 3.4. We see that for the two RWC test-sets and a 3 s tolerance window, our algorithm achieves better results than the best results obtained in MIREX (even the ones obtained by Serra et al. – SMGA1). It should be noted that the comparison for the Beatles-B test-set cannot be made since MIREX use the whole Isophonics test-set and not only the Beatles sub-part.

Statistical tests: For a @3s tolerance window, the differences of results obtained with $c_3(t)$ and $c_2(t)$ are statistically significant (at 5%) for all three test-sets. They are not for a @0.5s tolerance window.

Discussion: For the RWC-Pop-B test-set, using $c_3(t)$ instead of $c_1(t)$ allows increasing the F@3s for 88/100 tracks, for the Beatles-B for 144/180 tracks. In Figure 4,

[4] These annotations are available at `http://musicdata.gforge.inria.fr/structureAnnotation.html`.

[5] The MIREX test-set named "M-2010 test-set Original" corresponds to RWC-Pop-A, "M-2010 test-set Quaero" to RWC-Pop-B.

[6] SMGA1 stands for [Joan Serra, Meinard Mueller, Peter Grosche, Josep Lluis Arcos]. FK2 stands for [Florian Kaiser and Geoffroy Peeters]. RBH1 stands [Bruno Rocha, Niels Bogaards, Aline Honingh].

Table 1. Results of music structure segmentation using our lag-prior method applied to the SSM as computed in part 2.2.

RWC-Pop-A						
Method	F @3s	P @3s	R @3s	F @0.5s	P @0.5s	R @0.5s
Serra et al. [16]	0.791	0.817	0.783			
MIREX-2012 (SMGA1 on M-2010 test-set Original)	0.7101	0.7411	0.7007	0.2359	0.2469	0.2319
MIREX-2013 (FK2 on M-2010 test-set Original)	0.6574	0.8160	0.5599	0.3009	0.3745	0.2562
$c_1(t)$ (without lag-prior)	0.660	0.700	0.648	0.315	0.338	0.308
$c_2(t)$ (with global lag-prior)	0.729	0.739	0.737	0.349	0.354	0.353
$c_3(t)$ (with local lag-prior)	**0.769**	0.770	0.78	**0.386**	0.392	0.390
RWC-Pop-B						
Method	F @3s	P @3s	R @3s	F @0.5s	P @0.5s	R @0.5s
Serra et al. [16]	0.8	0.81	0.805			
MIREX-2012 (SMGA1 on M-2010 test-set Quaero)	0.7657	0.8158	0.7352	0.2678	0.2867	0.2558
MIREX-2013 (RBH1 on M-2010 test-set Quaero)	0.6727	0.7003	0.6642	0.3749	0.3922	0.3682
$c_1(t)$ (without lag-prior)	0.673	0.6745	0.689	0.238	0.223	0.263
$c_2(t)$ (with global lag-prior)	0.726	0.704	0.766	0.250	0.231	0.281
$c_3(t)$ (with local lag-prior)	**0.782**	0.782	0.816	**0.281**	0.264	0.31
Beatles-B						
Method	F @3s	P @3s	R @3s	F @0.5s	P @0.5s	R @0.5s
Serra et al. [16]	0.774	0.76	0.807			
$c_1(t)$ (without lag-prior)	0.657	0.674	0.658	0.232	0.240	0.238
$c_2(t)$ (with global lag-prior)	0.698	0.696	0.718	0.254	0.258	0.265
$c_3(t)$ (with local lag-prior)	**0.761**	0.745	0.795	**0.262**	0.259	0.278

we illustrate one of the examples for which the use of $c_3(t)$ decreases the results over $c_1(t)$. As before the discrimination obtained using $c_3(t)$ (right sub-figure) is higher than the ones obtained using $c_1(t)$ (left sub-figure). However, because of the use of the prior $p_t(l)$ which is computed on a long duration window ($[t-\Delta, t+\Delta]$ represents 20 s), $c_3(t)$ favors the detection of long-duration segments. In the example of Figure 4, parts of the annotated segments (black dashed lines) are very short segments which therefore cannot be detected with the chosen duration Δ for $p_t(l)$.

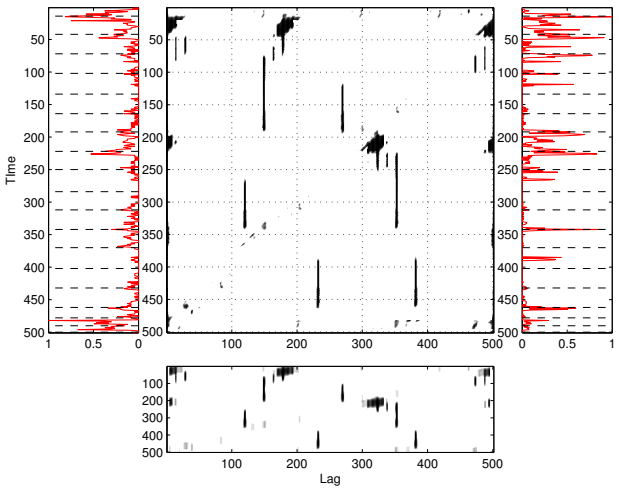

Figure 4. Illustration of a case for which $c_3(t)$ (right sub-figure) decrease the results over $c_1(t)$ (left sub-figure). $F@3s(c_1(t)) = 0.93$ and $F@3s(c_3(t)) = 0.67$ [Track 20 form RWC-Pop-B].

3.4 Results obtained applying our lag-prior method to the SSM as computed by Serra et al. [16]

In order to assess the use of $c_{2,3}(t)$ as a generic process to improve the estimation of the segments on a SSM; we applied $c_*(t)$ to the SSM computed as proposed in [16] instead of the SSM proposed in part 2.2. The SSM will be computed using the CENS features instead of the HPCP used in [16]. For recall, in [16] the recent past of the features is taken into account by stacking the feature vectors of past frames (we used a value m corresponding to 3 s). The SSM is then computed using a K nearest neighbor algorithm (we used a value of $\kappa = 0.04$). Finally the SSM matrix is convolved with a long bivariate rectangular Gaussian kernel $G = \mathbf{g}_t \mathbf{g}_l^T$ (we used s_l =0.5 s s_t =30 s and $\sigma^2 = 0.16$). $c_*(t)$ is then computed from the resulting SSM. The flowchart of this method is represented in the right part of Figure 3.

Results are given in Table 2 for the various configurations and test-sets. $c_1(t)$ represents Serra et al. method [16]. As one can see, the use of a global prior ($c_2(t)$) allows to increase the results over $c_1(t)$ for the three test-sets and the two tolerance window (@3s and @0.5s). Surprisingly, this time, results obtained with a local prior ($c_3(t)$) are lower than the ones obtained with a global prior ($c_2(t)$). This can be explained by the fact that Serra et al. method applies a long duration low-pass filter (s_t =30s) to the SSM. It significantly delays in time the maximum value of a segment in the SSM, hence delays $p_t(l)$, hence delays $c_3(t)$. In the opposite, because $c_2(t)$ is global, it is not sensitive to Serra et al. delay.

Statistical tests: For a @3s tolerance window, the difference of results obtained with $c_2(t)$ (0.805) and $c_1(t)$ (0.772) is only statistically significant (at 5%) for the Beatles-B test-set. For a @0.5s tolerance window, the differences are statistically significant (at 5%) for all three test-sets.

Table 2. Results of music structure segmentation using our lag-prior method applied to the SSM as computed by [16] .

RWC-Pop-A						
Method	F @3s	P @3s	R @3s	F @0.5s	P @0.5s	R @0.5s
$c_1(t)$ (without lag-prior) with Serra front-end	0.780	0.846	0.742	0.254	0.271	0.246
$c_2(t)$ (with global lag-prior) with Serra front-end	**0.784**	0.843	0.750	**0.289**	0.316	0.275
$c_3(t)$ (with local lag-prior) with Serra front-end	0.735	0.827	0.682	0.245	0.300	0.215
RWC-Pop-B						
Method	F @3s	P @3s	R @3s	F @0.5s	P @0.5s	R @0.5s
$c_1(t)$ (without lag-prior) with Serra front-end	0.799	0.795	0.818	0.338	0.326	0.359
$c_2(t)$ (with global lag-prior) with Serra front-end	**0.823**	0.846	0.820	**0.389**	0.408	0.381
$c_3(t)$ (with local lag-prior) with Serra front-end	0.797	0.856	0.765	0.336	0.369	0.318
Beatles-B						
Method	F @3s	P @3s	R @3s	F @0.5s	P @0.5s	R @0.5s
$c_1(t)$ (without lag-prior) with Serra front-end	0.772	0.792	0.773	0.371	0.365	0.394
$c_2(t)$ (with global lag-prior) with Serra front-end	**0.805**	0.813	0.817	**0.439**	0.430	0.450
$c_3(t)$ (with local lag-prior) with Serra front-end	0.799	0.790	0.827	0.422	0.416	0.442

4. CONCLUSION AND FUTURE WORKS

In this paper, we have proposed a simple, but very efficient, method that allows using Goto 2003 method as a prior lag-probability on Serra et al. structure feature method. We provided the rational for such a proposal, and proposed two versions of the method: one using a global lag prior, one using a local lag prior. We performed a large-scale experiment of our proposal in comparison to state-of-the-art algorithms using three test-sets: RWC-Popular-Music with two sets of annotations and Isophonics/Beatles. We showed that the introduction of the lag-prior allows a large improvement of the F-Measure results (with a tolerance window of 3 s) over the three sets. Also, our method improves over the best results obtained by Serra et al. or during MIREX-2012 and MIREX-2013.

Future works will concentrate on integrating this prior lag probability on an EM (Expectation-Maximization) algorithm to estimate the true $p(t, l)$. Also, we would like to use this segmentation as a first step to a segment labeling algorithm.

Acknowledgements This work was partly founded by the French government Programme Investissements d'Avenir (PIA) through the Bee Music Project.

5. REFERENCES

[1] Mark A. Bartsch and Gregory H. Wakefield. To catch a chorus: Using chroma-based representations for audio thumbnailing. In *Proc. of IEEE WASPAA (Workshop on Applications of Signal Processing to Audio and Acoustics)*, pages 15–18, New Paltz, NY, USA, 2001.

[2] Frédéric Bimbot, Emmanuel Deruty, Sargent Gabriel, and Emmanuel Vincent. Methodology and conventions for the latent semiotic annotation of music structure. Technical report, IRISA, 2012.

[3] Roger Dannenberg and Masataka Goto. Music structure analysis from acoustic signal. In *Handbook of Signal Processing in Acoustics Vol. 1*, pages 305–331. Springer Verlag, 2009.

[4] Jonathan Foote. Visualizing music and audio using self-similarity. In *Proc. of ACM Multimedia*, pages 77–80, Orlando, Florida, USA, 1999.

[5] Jonathan Foote. Automatic audio segmentation using a measure of audio novelty. In *Proc. of IEEE ICME (International Conference on Multimedia and Expo)*, pages 452–455, New York City, NY, USA, 2000.

[6] Masataka Goto. A chorus-section detecting method for musical audio signals. In *Proc. of IEEE ICASSP (International Conference on Acoustics, Speech, and Signal Processing)*, pages 437–440, Hong Kong, China, 2003.

[7] Masataka Goto. Aist annotation for the rwc music database. In *Proc. of ISMIR (International Society for Music Information Retrieval)*, pages pp.359–360, Victoria, BC, Canada, 2006.

[8] Masataka Goto, Hiroki Hashiguchi, Takuichi Nishimura, and Ryuichi Oka. Rwc music database: Popular, classical, and jazz music databases. In *Proc. of ISMIR (International Society for Music Information Retrieval)*, pages pp. 287–288, Paris, France, 2002.

[9] Florian Kaiser and Geoffroy Peeters. Multiple hypotheses at multiple scales for audio novelty computation within music. In *Proc. of IEEE ICASSP (International Conference on Acoustics, Speech, and Signal Processing)*, Vancouver, British Columbia, Canada, May 2013.

[10] Matthias Mauch, Chris Cannam, Matthew Davies, Simon Dixon, Christopher Harte, Sefki Klozali, Dan Tidhar, and Mark Sandler. Omras2 metadata project 2009. In *Proc. of ISMIR (Late-Breaking News)*, Kobe, Japan, 2009.

[11] Meinard Müller and Sebastian Ewert. Chroma toolbox: Matlab implementations for extracting variants of chroma-based audio features. In *Proc. of ISMIR (International Society for Music Information Retrieval)*, Miami, Florida, USA, 2011.

[12] Meinard Müller, Franz Kurth, and Michael Clausen. Audio matching via chroma-based statistical features. In *Proc. of ISMIR (International Society for Music Information Retrieval)*, London, UK, 2005.

[13] Jouni Paulus, Meinard Müller, and Anssi Klapuri. Audio-based music structure analysis. In *Proc. of ISMIR (International Society for Music Information Retrieval)*, Utrecht, The Netherlands, 2010.

[14] Geoffroy Peeters. *Deriving Musical Structures from Signal Analysis for Music Audio Summary Generation: Sequence and State Approach*, pages 142–165. Lecture Notes in Computer Science. Springer-Verlag Berlin Heidelberg 2004, 2004.

[15] Geoffroy Peeters, Amaury Laburthe, and Xavier Rodet. Toward automatic music audio summary generation from signal analysis. In *Proc. of ISMIR (International Society for Music Information Retrieval)*, pages 94–100, Paris, France, 2002.

[16] Joan Serra, Meinard Muller, Peter Grosche, and Josep Ll. Arcos. Unsupervised detection of music boundaries by time series structure features. In *Proc. of AAAI Conference on Artificial Intelligence*, 2012.