MULTIPLE HYPOTHESES AT MULTIPLE SCALES FOR AUDIO NOVELTY COMPUTATION WITHIN MUSIC

Florian Kaiser and Geoffroy Peeters

STMS IRCAM-CNRS-UPMC
1 Place Igor Stravinsky
75004 Paris
florian.kaiser@ircam.fr

ABSTRACT

Novelty-based segmentation of audio signals has proven good performances for the estimation of boundaries of structural sections within music pieces. However, boundaries are detected only if structural sections satisfy the condition of sufficient acoustic inner-homogeneity. While this constraint is very restrictive and not representative of all musical contents, we propose in this paper to extend the detection of acoustic novelty to transitions between homogeneous and non-homogeneous sections and vice versa. Moreover, the length of the considered sections for the boundary detection is crucial, we also introduce a multi-scale novelty approach that allows to capture boundaries between sections of different temporal scales in a same segmentation. Evaluation of the combination of these two methods proves convincing results for temporal segmentation of music pieces. Embedding the algorithm in a music structure segmentation system, we show that performances can be consistently improved for this task.

Index Terms— Audio Novelty, Audio Segmentation, Music Structure Segmentation

1. INTRODUCTION

Music structure segmentation is the task of estimating the largest temporal musical entities that can be segmented in a music piece, e.g. verse or refrain in popular music. Of great interest for many applications such as audio summarization, similarity or synchronization, the task has raised a growing interest in the Music Information Retrieval community in recent years. Beyond applications, the challenge in this particular task resides in the music itself that implies major acoustic changes at all temporal levels. With the beat-, note-, or melody- level, the musical content is highly hierarchical. Structure segmentation aims at estimating the highest degree of this hierarchy.

In order to match this level of hierarchy, a technique that is commonly applied in music structure segmentation consists in first estimating the boundaries of structural sections and limit the search space for the structure labeling. Such a temporal segmentation has proven major impact on music structure segmentation and its performance is thus crucial. Music structure segmentation methods were developed on three main different hypotheses on musical structures [1]: Novelty-, Homogeneity-, and Repetition- based methods. Consequently, methods for the estimation of structural boundaries make use of the same hypotheses.

The novelty principle is often applied in audio temporal segmentation and consists in estimating time points of significant acoustic changes. In applications such as speaker segmentation, an efficient method consists in comparing adjacent segments with the Bayesian Information Criterion to decide whether or not they belong together [2]. In music structure segmentation however, the novelty principle was extended by Foote to the notion of ”acoustic contrast” [3] and is implicitly related to the homogeneity hypothesis of structural sections. Time points of major acoustic contrast are indeed defined as junctions between two homogeneous audio segments that present sufficient dissimilarity when compared. Foote proposed to detect such boundaries by embedding the audio signal in an audio Self-Similarity Matrix (SSM) [4] in which acoustic contrast is visualized as the transition between blocks of high similarity. Note that such a visualization of structural sections as blocks is formalized by Peeters in [5] as the ”state” representation. Novelty over time is calculated with the convolution of the SSM with a novelty kernel inspired from this representation of boundaries. Segmentation is then estimated by means of the novelty curve. This method was successfully applied in many music structure segmentation systems for boundary retrieval [6] [7] [8] [9]. Sargent et al combine in [10] a novelty segmentation with a prior regularity constraint on the temporal scale of the segmentation. After each detected boundary, the temporal segmentation is optimized given a regularity constraint at the music piece level.

The other dominant approach to music structure segmentation is based on the repetition paradigm for which a structural section defines itself by its repetitions. This principle of repetition is very closely related to the notion of ”sequence” representation developed by Peeters in [5]. Indeed, what characterizes repetitions is often a particular sequence of notes or chords that is non-homogeneous. Visualization of the audio signal in a SSM computed on chroma sequences for example reveals the sequences and their repetitions as stripes on the main diagonal and off-diagonals of the SSM. The detection of music boundaries in the repetition paradigm is thus treated as the problem of estimating the start and end of these stripes in the SSM [11] [12] [13]. More recently, Serra et al. proposed in [14] a structure feature for the detection of sequences in time series. Audio features are temporally filtered and embedded in a circular time-lag matrix that considers both future and past time-lags. The matrix is convolved with a bivariate gaussian kernel to highlight sequences and the structure feature is then constructed as the observation of each time sample in this representation. Boundaries are finally detected by computing the difference between consecutive frames of the structure feature. Advantage of the method is that boundaries of repetitions can be detected and to some extent boundaries of homogeneous sections within the sequences as well.

2. RELATED WORK

Methods for the temporal segmentation of music pieces are thus developed under either the constraint of homogeneity or repetition of the structural sections. However, music is diverse and we observe that none of these hypotheses can be uniquely applied to all musical structures. In contrast, structural sections of a music piece may often alternate between homogeneous and non-homogeneous sections.

As a matter of fact, recent work by Paulus [8] highlighted the benefit brought by a non-unique hypothesis approach for the comparison of structural segments in music structure segmentation. We therefore propose in this paper to extend the acoustic contrast measured in novelty-based methods to the particular transitions between homogeneous and non-homogeneous sections and vice-versa, independently of repetitions.

In the next section we therefore propose two new novelty kernels for measuring the acoustic contrast under the constraint of homogeneous/non-homogeneous transition and vice versa. A multi-temporal scale approach for computing novelty by means of these kernels and a method for fusing the information brought by all kernels is then presented in Section 4. After illustrating the approach with an example in Section 5, evaluation procedure and results are presented in Section 6. Section 7 finally concludes this paper and draws perspectives for future research.

3. MULTIPLE HYPOTHESES FOR NOVELTY

In this section we first recall the novelty-based segmentation as used in music structure segmentation systems. We then introduce two new novelty kernels that test further hypotheses on acoustic contrasts between structural sections with the consideration of Homogeneous/Non-Homogeneous transitions.

3.1. Novelty-based Segmentation

The novelty approach to audio temporal segmentation was originally proposed by Foote in [3] and allows to detect transitions between homogenous segments of an audio signal. The idea is that such transitions visualized in a SSM locally resembles a 2×2 checkerboard. Indeed, the boundary contrasts two sections of high self-similarity that form two distinct blocks, and are at the same time highly dissimilar forming blocks of low similarity when compared. Foote proposes to detect these boundaries by correlating along the diagonal of the SSM a kernel matrix inspired from this visual interpretation. In the following of this paper we will denote SSMs as $\mathbf{S}$ and novelty kernels as $\mathbf{K}$.

The canonical kernel $\mathbf{K}$ that was proposed is a $2{\times}2$ checkerboard that is easily enlarged to any size L by means of the kronecker product. $\mathbf{K}$ is also usually weighted with a symmetric gaussian radial function that gives less importance to the kernel's edges. Such a novelty kernel example is shown in Figure 1.a with a length L = 60 samples and radial gaussian weighting with σ = 0.5. Multiplication of $\mathbf{K}$ with the SSM measures the acoustic contrast at the centre point of the kernel in the sense of the acoustic homogeneity on either sides of the kernel's center point and the dissimilarity between the two segments. Correlation of the kernel along the diagonal of the SSM thus yields a novelty curve in which boundary candidates can be selected by peak-picking high values.

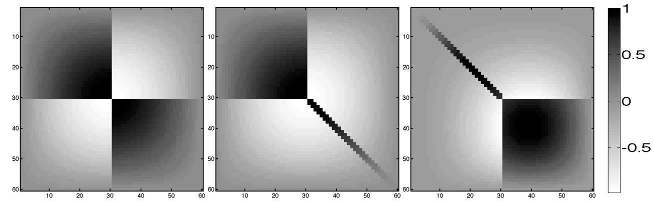

Fig. 1: Novelty detection kernels

3.2. Multiple Hypotheses Kernels $\mathbf{K_i}$

The novelty kernel introduced by Foote allows for the detection of transitions between homogeneous sections. We now propose to extend the notion of acoustic contrasts to the transitions between homogeneous and non-homogeneous sections. We follow the same approach as Foote and introduce two new novelty kernels for the detection of such transitions (see Figure 1).

To estimate the canonical form of a novelty kernel that would allow for the detection of such contrasts, we consider the theoretical form of such transitions in similarity matrices. Ideally, a transition from an homogeneous to a non-homogeneous section is indeed visualized as a block of high similarity, i.e. state, followed by a stripe on the main diagonal. This observation can be formalized in the $4{\times}4$ SSM $\mathbf{S}$ described in Equation (1) that is centered on an ideal transition between a state and a non-homogeneous segment. We denote $\mathbf{s}$ the expected value of samples that contain structural information and α the expected value of noise samples. Note that the similarity measure for the computation of $\mathbf{S}$ is usually contained in [0 1].

$$\mathbf{S} = \begin{bmatrix} \mathbf{s} & \mathbf{s} & \alpha & \alpha \\ \mathbf{s} & \mathbf{s} & \alpha & \alpha \\ \alpha & \alpha & \mathbf{s} & \alpha \\ \alpha & \alpha & \alpha & \mathbf{s} \end{bmatrix}, \quad E\big(S(t_i,t_j)\big) = \begin{cases} \mathbf{s}, \text{ if } (t_i,t_j) \in \text{struct} \\ \alpha, \text{ if } (t_i,t_j) \in \text{noise} \end{cases} \tag{1}$$

From this observation SSM, we deduce a canonical novelty kernel for the detection of transitions between homogeneous and non-homogeneous segments by simply replacing $\mathbf{s}$ by 1 and α by -1 for now. Note that homogeneous/non-homogeneous and non-homogeneous/homogeneous transitions detection are reversed but equivalent problems and we only detail detection of the latter.

3.3. Kernel Weighting

An implicit property of the kernel proposed by Foote is that it perfectly balances the probability of sections being in the state or non-state hypothesis in the SSM, i.e. mean value of the kernel is zero. In contrast, the canonical kernel introduced in the last subsection does not give equal weights to state and sequence representations. Indeed, the mean value of the kernel distribution is largely negative, and the kernel will have more energy while in the middle of a stripe on the main diagonal than while in the middle of a state section when convolved with a SSM. However, the energy of the kernel should be maximum at the transition between these two representations and rapidly decreasing while sliding towards the diagonal stripe for the detection.

We thus apply a weighting of the kernel that forces its mean value to zero. Samples of the SSM whose values are expected to form states are weighted in the kernel with a coefficient κ, and remaining samples of the kernel with the coefficient ν. Calculating the mean value of the weighted kernel $\mathbf{K_i}$ introduced in Equation (2) in a 4×4 version, we deduce in Equations in (3) the condition on κ and ν for the mean value of the kernel μ_K to equal zero, with L the length of the kernel and $l = \frac{L}{2}$.

$$\mathbf{K_i} = \begin{bmatrix} \kappa & \kappa & \nu & \nu \\ \kappa & \kappa & \nu & \nu \\ \nu & \nu & \kappa & \nu \\ \nu & \nu & \nu & \kappa \end{bmatrix} \tag{2}$$

$$\mu_K = \frac{1}{4l^2}\big(l(l+1)\kappa + l(3l-1)\nu\big) = 0 \;\Rightarrow\; \nu = -\frac{l+1}{3l-1}\kappa \tag{3}$$

Since the main diagonal of the SSM usually equals one, we set κ to one and deduce ν with the length of the kernel. Examples of Homogeneous / Non-Homogeneous and Non-Homogeneous / Homogeneous transition kernels smoothed by a radial gaussian function (σ=0.5) are respectively shown in Figures 1.b and 1.c. We now denote the kernels $\mathbf{K_i}$ as: ($\mathbf{K_a}$) Foote's kernel, ($\mathbf{K_b}$) Homogeneous/Non-Homogeneous kernel, and ($\mathbf{K_c}$) Non-Homogeneous/Homogeneous kernel.

4. TEMPORAL SEGMENTATION

Correlation with a SSM of kernels introduced in the last section measures the acoustic novelty over time according to different hypotheses. In this section we propose a method to measure these novelties at various temporal scales and combine the information of all kernels in a single temporal segmentation.

4.1. Multi-scale kernels $\mathbf{K_{iL}}$

Novelty estimation by means of boundary detection kernels is of course influenced by the shape of the kernel but also by its size. Indeed, music content being highly hierarchical, enlarging the size of the kernel allows to smooth low temporal level events such as notes and focus on the novelty between events of a couple of seconds. However, the temporal scale of structural sections is not known in advance and may vary within music pieces.

Instead of setting the length L of kernels to a fixed value, we thus compute multi-scale novelty curves for kernels $\mathbf{K_{iL}}$ of varying lengths. The lengths vary between L = 7.5s (30 samples @4Hz) and L = 30s (120 samples @4Hz). We call $f_{iL}(t)$ the novelty computed with the convolution of a kernel $\mathbf{K_{iL}}$ with a SSM. Novelty distributions are then reduced to a unidimensional novelty curve that still reflects the different temporal scales by summing over all kernel lengths the novelty at each time sample. For each kernel i, the novelty function $n_i(t)$ is thus defined as: $n_i(t) = \sum_L f_{iL}(t)$.

4.2. Boundaries selection and fusion

Summing the novelty functions of all kernels for the detection of boundaries would smooth the information brought by each kernel. We therefore apply a late fusion strategy for the combination of the temporal segmentations. Such a strategy was for example applied in [15] for merging structural segments. We therefore estimate boundary candidates for each detection kernel independently and then merge the boundaries in a single temporal segmentation. To do so, we apply the adaptive peak tracking technique described in [16] to extract boundary candidates on the novelty functions $n_i(t)$ of each kernel.

For the fusion of boundaries we denote the set of boundary candidates extracted on all three novelty functions as T_a, T_b and T_c respectively obtained with the kernels $\mathbf{K_a}$, $\mathbf{K_b}$ and $\mathbf{K_c}$. Since identical boundaries may be detected by all kernels but with a slight temporal deviation, the simple union of boundary sets is not precise enough for the fusion. We therefore define in Equation (4) the operators $\cap_\Delta$ and $\setminus_\Delta$ that respectively extract the intersection and relative complement of boundary sets within a tolerance range Δ. with T_1 and T_2 two sets of boundaries. When two boundaries t_1 and t_2 are found at the intersection of two boundary sets, we only retain t_1. The final segmentation T is obtained with equation (5)

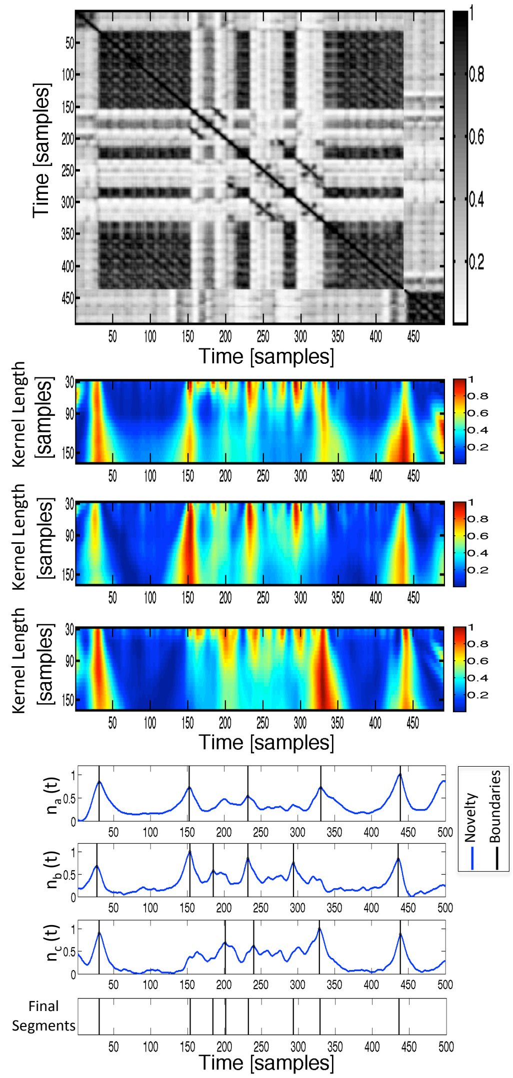

Fig. 2: From top to bottom: (1) Self-similarity matrix, Multi-scale Novelty scores computed with: (2) Foote kernel [3], (3) Homogeneous/Non-Homogeneous kernel and (4) Non-Homogeneous/Homogeneous kernel, (5) to (7) Summed novelties and detected boundaries, (8) Final segmentation

$$\begin{aligned} T_1 \cap_\Delta T_2 &= \{(t_1,t_2) : (t_1 - t_2) < \Delta, t_1 \in T_1, t_2 \in T_2\} \\ T_1 \setminus_\Delta T_2 &= \{t : t \in T_1, [t - \frac{\Delta}{2}, t + \frac{\Delta}{2}] \ni T_2\} \end{aligned} \tag{4}$$

$$T = \Big((T_a \cap_\Delta T_b) \cap_\Delta T_c\Big) \bigcup \Big((T_a \setminus_\Delta T_b) \setminus_\Delta T_c\Big) \tag{5}$$

We set the tolerance range at Δ = 2 seconds (1 bar @120bpm) to limit over-segmentation.

5. CASE STUDY

In this section we briefly illustrate the multiple hypotheses kernels for novelty computation at different time scales with a music piece example. The audio signal used is an excerpt of the song 32 of the RWC Popular dataset[1] on which chroma features are extracted. A SSM sampled at 4Hz is computed by means of the cosine distance. The SSM and the novelty distribution $f_{iL}(t)$ for all detection kernels with varying lengths from 30 samples (7.5s) to 160 samples (40s) are shown together with the novelty functions and detected boundaries in Figure 2. We observe with this example that structural sections indeed may alternate state and sequence representation. Moreover, main boundaries (samples 150, 330 and 440) are detected by both Foote's kernel and either one of the kernels introduced in this paper. Boundaries detected by the new kernels however allow for a much finer segmentation of the audio signal. For example at sample 290, a boundary is clearly highlighted with the Homogeneous/Non-Homogeneous kernel (second novelty curve of Figure 2). The use of Foote's kernel (first novelty curve of Figure 2) suggests this boundary for very small kernel sizes but does not detect it. Secondly, the multi-scale novelty curves illustrate the importance of the kernel lengths and that different lengths produce different segmentations.

6. EVALUATION

We now present an evaluation of the novelty approach introduced in this paper for the task of temporal segmentation of music pieces. Post segment grouping is also applied in order to evaluate the impact of our approach on structure segmentation performances.

6.1. Protocol

Test Set: In order to compare our method with current results obtained at the 2012 MIREX[2] evaluation for structural segmentation, we use the structural annotations provided in the Isophonics[3] dataset that consists of 294 popular music songs (the Beatles, Queen, Michael Jackson...)

Evaluation Metrics: The temporal segmentation is evaluated by means of the precision P, recall R and F-Measure F. We distinguish two tolerance ranges of 0.5 and 3 seconds in these measures for the True Positives, False Positives and False Negatives calculation. The segment grouping is evaluated by means of the pairwise Precision, Recall and F-Measure introduced in [17].

Algorithms: SSMs sampled at 4Hz are extracted on timbre-related features as described in [18]. Temporal segmentation is then estimated with the method presented in this paper, denoted as "ICASSP13" in the results tables. The same matrices are segmented with the standard novelty approach in the method denoted as "MIREX12-Ircam". For both segmentations, the structural grouping of segments is performed as described in [18]. We compare the performance of these two systems with the algorithm of Serra et al. [14] evaluated on the same dataset at the 2012 MIREX.

6.2. Results and Discussion

Temporal segmentation and segment grouping evaluation are reported in Tables 1 and 2 respectively. The methods "ICASSP13" and "MIREX12-Ircam" only differ with the introduction of our novelty kernels and multi-scale novelty at the temporal segmentation step. Evaluation in Table 1 suggests that this extension of the novelty approach strongly improves the quality of the segmentation gaining respectively 4% and 5% in the F-Measures at 0.5s and 3s. Moreover, precision@3s gains 11% with a relative loss of recall of 4%. Segmentation with a single kernel in "MIREX12-Ircam" is set with a kernel length of L=60 samples (=15s) and tends to over-segmentation with high recall compared to precision. The multi-kernels and multi-scales segmentation in contrast thus seems more balanced between over- and under- segmentation. Performances of our method compares with the method of Serral et al. [14] rather well and proves that novelty is a consistent approach for segmentation of music. We also note that our method performs comparatively very well @0.5s. Concerning the segment grouping, the segmentation proposed in this paper increased the F-Measure of about 2%. This is a very encouraging result. Results are however still below the performance of [14] that applies string matching techniques. While our method uses the homogeneity principle for grouping, it would be interesting to extract with our kernels information on the homogeneity or non-homogeneity of segments. Segment grouping techniques adapted to this information could then strongly improve the structure segmentation performance.

Method	F@0,5s	P@0,5s	R@0,5s
Serra et al. [14]	22,82	20,51	26,65
MIREX12 - Ircam [18]	28,16	24,14	35,77
ICASSP13	32,41	29,29	38,87

Method	F@3s	P@3s	R@3s
Serra et al. [14]	64,49	57,95	75,05
MIREX12 - Ircam [18]	59,13	50,64	75,00
ICASSP13	64,41	61,54	71,44

Table 1: Temporal Segmentation Evaluation [%]

Method	F	P	R
Serra et al. [14]	65,28	61,80	74,64
MIREX12 - Ircam [18]	57,18	56,61	62,26
ICASSP13	59,86	59,80	64,57

Table 2: Segment Grouping Evaluation [%]

7. CONCLUSION

We proposed in this paper an extension for novelty-based audio segmentation approaches to the detection of transitions between homogeneous and non-homogeneous sections. Moreover, we proposed a multi-scale novelty computation to account for structural changes at different temporal scales. While music structure segmentation systems are often developed under the constraint of a unique hypothesis on the structural sections, we show in our evaluation the benefit of a multiple hypotheses approach for the temporal segmentation of music. Moreover, we believe that this approach allows to detect more than temporal boundaries and contains information on the acoustic nature of segments that could trigger efficient segment grouping techniques adapted to diverse musical contexts.

8. ACKNOWLEDGMENTS

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

[1] http://staff.aist.go.jp/m.goto/RWC-MDB/AIST-Annotation/
[2] http://www.music-ir.org/mirex/wiki/2012:MIREX2012_Results
[3] http://isophonics.net/

9. REFERENCES

[1] Jouni Paulus, Meinard Müller, and Anssi Klapuri, "Audio-based music structure analysis," in *Proceedings of the 11th International Society for Music Information Retrieval Conference (ISMIR)*, 2010.

[2] Scott Shaobing Chen and P.S. Gopalakrishnan, "Clustering via the bayesian information criterion with applications in speech recognition," in *Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on*, May 1998, vol. 2, pp. 645–648 vol.2.

[3] Jonathan Foote, "Automatic audio segmentation using a measure of audio novelty," in *Proceedings of the IEEE International Conference on Multimedia and Expo*, 2000.

[4] Jonathan Foote, "Visualizing music and audio using self-similarity," in *Proceedings of the ACM Multimedia*, 1999, pp. 77–80.

[5] Geoffroy Peeters, *Deriving Musical Structures from Signal Analysis for Music Audio Summary Generation: "Sequence" and "State" Approach*, vol. 2771 of *Lecture notes in Computer Science*, pp. 143–166, Springer, 2004.

[6] M. Cooper and J. Foote, "Summarizing popular music via structural similarity analysis," in *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2003.

[7] Geoffroy Peeters, "Toward automatic music audio summary generation from signal analysis," in *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, 2002, pp. 94–100.

[8] Jouni Paulus and Anssi Klapuri, "Music structure analysis using a probabilistic fitness measure and a greedy search algorithm," *IEEE Transactions on Audio, Speech & Language Processing*, vol. 17, no. 6, pp. 1159–1170, 2009.

[9] Florian Kaiser and Thomas 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.

[10] Gabriel Sargent, Frederic Bimbot, and Emmanuel 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.

[11] Masataka Goto, "Chorus-section detecting method for musical audio signals," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2003.

[12] Meinard Müller and Michael Clausen, "Transposition-invariant self-similarity matrices," in *Proceedings of the 8th International Conference on Music Information Retrieval (ISMIR)*, Vienna, Austria, Sept. 2007, pp. 47–50.

[13] B. Ong, *Structural Analysis and Segmentation of Music Signals*, Ph.D. thesis, Music Technology Group - Universitat Pompeu Fabra, 2007.

[14] Joan Serra, Meinard Müller, Peter Grosche, and Josep Ll. Arcos, "Unsupervised detection of music boundaries by time series structure features," in *AAAI International Conference on Artificial Intelligence*, 2012.

[15] Ruofeng Chen and Ming Li, "Music structural segmentation by combining harmonic and timbral information," in *Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR)*, 2011.

[16] A.L. 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*, 2001, vol. 3, pp. 2194–2195 vol.3.

[17] M. Levy and M. Sandler, "Structural segmentation of musical audio by constrained clustering," *IEEE Transactions on Audio, Speech & Language Processing*, vol. 16, no. 2, pp. 318–326, 2008.

[18] Florian Kaiser, Thomas Sikora, and Geoffroy Peeters, "Mirex 2012 - music structural segmentation task: Ircamstructure submission," in *MIREX at ISMIR 2012*, 2012.