# Estimating the structural segmentation of popular music pieces under regularity constraints

Gabriel Sargent, Frédéric Bimbot, Emmanuel Vincent

# Estimating the structural segmentation of popular music pieces under regularity constraints

Gabriel Sargent, Frédéric Bimbot and Emmanuel Vincent

***Abstract*—Music structure estimation has recently emerged as a central topic within the field of Music Information Retrieval. Indeed, as music is a highly structured information stream, knowledge of how a music piece is organized represents a key challenge to enhance the management and exploitation of large music collections.**

**This article focuses on the benefits that can be expected from a *regularity constraint* on the structural segmentation of popular music pieces. Specifically here, we study how a constraint which favors structural segments of comparable size provides a better conditioning of the boundary estimation process.**

**Firstly, we propose a formulation of the structural segmentation task as an optimization process which separates the contribution from the audio features and the one from the constraint. We illustrate how the corresponding cost function can be minimized using a Viterbi algorithm.**

**We present briefly its implementation and results in three systems designed for and submitted to the MIREX 2010, 2011 and 2012 evaluation campaigns. Then, we explore the benefits of the regularity constraint as an efficient mean for combining the outputs of a selection of systems presented at MIREX between 2010 and 2015, yielding a level of performance competitive to that of the state-of-the-art on the "MIREX10" dataset (100 J-Pop songs from the RWC database).**

***Index Terms*—Music structure estimation, structural segmentation, optimization, Viterbi algorithm, regularity constraint, multi criteria approach, fusion, MIREX evaluation campaign**

## I. Introduction

THE recent advance of information and communication technologies has increased the production, the storage and the accessibility of musical contents. As it is hard for a human to handle very large amounts of data, new solutions are needed to browse and exploit huge catalogs of music pieces efficiently. These solutions require informative descriptions of music, which can be collected from textual references or automatically extracted from the audio.

This article focuses on the macroscopic structural segmentation of music pieces in audio form. This task consists in producing a description of the overall organization of a music piece from an audio recording as a sequence of segments which are labeled according to their similarity, using a limited set of arbitrary symbols - each symbol representing a class of similar segments. This task can be viewed as an inverse problem, i.e. recovering the latent structure of the music piece [1].

Knowing the structure of a music piece provides a better understanding of its content, from which it is possible to produce concise yet reliable representations for catalog management, music analysis and music generation. Indeed, it can enhance the management of collections of music pieces by providing relevant features for measuring the similarity between its elements [2], offering indexes to access specific parts of the audio stream or enabling the generation of meaningful summaries [3], which provide a quick and representative glimpse of the full pieces. Music structure can also guide the generation of remixes or the improvement of other MIR tasks such as source separation [4], chord analysis [5] or audio/MIDI transcription into scores.

However, a music piece is a complex object. In particular, it provides multiple scales of analysis, from elementary notes and silences to larger patterns which share correspondences as well as a particular logic as regards their occurrences over time. Many criteria can be chosen to define these patterns across various musical genres. This creates ambiguity in the notion of structure in music, leading human annotators to describe the macroscopical structure of a single piece according to various criteria [6].

In this article, we analyze the benefits of incorporating a *regularity constraint* within the macroscopic structural segmentation of popular music pieces. In its simplest version, this constraint favors structures with segment sizes distributed around a typical value. Such a strategy was initially formulated (and termed as "structural pulsation period" [7]) to deal with the scale ambiguity arising from the manual annotation of the *semiotic* structure, defined at a macroscopic level [1].

In this article, Sections II and III describe the state-of-the-art of music structure annotation and estimation, respectively. Section IV defines the regularity constraint and describes its implementation as a Viterbi algorithm. After a general overview of recent MIREX campaigns in Section V, we briefly focus in Section VI on the systems designed at IRISA and implemented for MIREX 2010 to 2012 campaigns, for a series of diagnostic analyses of the regularity constraint. These five sections sum up work developed in [8].

In the last section (Section VII), we explore and highlight the benefits of the regularity constraint as an efficient means of fusing the outputs of several systems which participated in MIREX between 2010 and 2015. The results yield a level of performance comparable to that of the state-of-the-art on the "MIREX10" dataset.

Gabriel Sargent and Frédéric Bimbot are with the Institut de Recherche en Informatique et Systèmes Aléatoires (IRISA), Rennes, France (emails : gabriel.sargent@irisa.fr, frederic.bimbot@irisa.fr).

Emmanuel Vincent is with the Institut National de Recherche en Informatique et en Automatique (INRIA) Grand-Est, Nancy, France. (email : emmanuel.vincent@inria.fr)

Manuscript submitted in December 2015.

**Three Levels of Musical Experience**

| | Events per second | Seconds per event |
|---|---|---|
| | 16,384 | 1/16,384 |
| **EVENT FUSION** | 8,192 | 1/8,192 |
| (early processing) | 4,096 | 1/4,096 |
| | 2,048 | 1/2,048 |
| Functional units = | 1,024 | 1/1,024 |
| individual *events* and | 512 | 1/512 |
| *boundaries*; pitches, | 256 | 1/256 |
| simultaneous intervals, | 128 | 1/128 |
| loudness changes, etc. | 64 | 1/64 |
| | 32 | 1/32 |
| | 16 | 1/16 |
| **MELODIC and RHYTHMIC** | 8 | 1/8 |
| **GROUPING** | 4 | 1/4 |
| (short-term memory) | 2 | 1/2 |
| Functional units = | 1 | 1 |
| *patterns*; rhythmic and | 1/2 | 2 |
| melodic groupings, | 1/4 | 4 |
| phrases. | 1/8 | 8 |
| | 1/16 | 16 |
| **FORM** | 1/32 | 32 |
| (long-term memory) | 1/64 | 1 min 4 sec |
| Functional units = | 1/128 | 2 min 8 sec |
| large scale *constancies*; | 1/256 | 4 min 16 sec |
| sections, movements, | 1/512 | 8 min 32 sec |
| entire pieces. | 1/1,024 | 17 min 4 sec |
| | 1/2,048 | 34 min 8 sec |
| | 1/4,096 | 1 hr 8 min 16 sec |

Fig. 1. Snyder's "levels of musical experience" [9, p. 12].

## II. Music structure

A music piece can be considered as a sequence of short sound events played by one or several sound sources — instruments — so as to create a perceptual experience towards listeners. During listening, these events tend to be grouped cognitively according to their temporal and/or spectral relationships. The resulting patterns range over variable durations leading to numerous temporal scales of perception required to analyze a music piece. As depicted in Fig. 1, Snyder defines "three levels of musical experience" related to the immediate perception and the memorization processes of humans [9]. These levels are :

- the "event fusion" level, relating to the temporal scale where the sound events can not be perceptually distinguished from each other. Its elements correspond to durations up to $1/32 \approx 30$ ms.
- the "melodic and rhythmic grouping" level, referring to patterns that result from the cognitive binding of events. The complexity of these patterns is limited by the capacities of the short-term memory, which limits in practice their duration below 16 seconds.
- the "form" level, where only a few elements and relations are coded within the long-term memory. It corresponds to durations of 16 seconds and above. This level coincides with the macroscopic scale we focus on in the present work.

The notion of macroscopic structure within a music piece is ill-defined. Indeed, music is a complex and ambiguous stream of information that can be described on multiple musical *dimensions* such as timbre, harmony, melody, rhythm, tonality, nuance or loudness. Moreover, one can describe a structural segment according to :

- some high-level acoustic properties such as the instruments playing or the singers singing,
- its function within the music piece, e.g. introduction, chorus, verse, bridge, theme, variation, *coda* ...
- its context, as it can be bounded by timbral breakdowns or the location of harmonic patterns over the piece,
- its internal consistency through the statistical homogeneity or the systemic behavior of its musical dimensions.

Such possibilities to define a structural segment explain why the intuitive annotation of music structure by several humans rarely converges to a single description straightaway [6].

Several approaches for the structural analysis of music pieces exist in the domain of Musicology, as presented by Bent and Drabkin in [10]. However they are principally designed for classical music, which makes them hard to use for other genres. In the same context, Peiszer notices the difficulty for a non-musicologist to establish the suitability of a particular analysis method compared to another [11] but emphasizes the relevance of the *paradigmatic analysis*[1] proposed by Ruwet [12], which consists in segmenting a music piece according to the repetitions of its musical content. One can also mention the view of Middleton who considers that "while repetition is a feature of all music, of any sort, a high level of repetition may be a specific mark of 'the popular' [...]" [13, p. 139].

Research work from the Music Information Retrieval (MIR) community attempts to respond to the lack of consensus on the specification of macroscopic structure. Peeters and Deruty proposed a multidimensional characterization of music structure [14], considering in parallel several structure types reflecting different "view-points" on music structure. They are respectively based on the repetitions of chord sequences (*acoustical similarity*), the roles of instruments over time, and the function (*musical role*) of the segments. By convention, structural borders are synchronous to downbeats. To deal with ambiguities on the scale of analysis, additional boundaries can be used to mark possible subdivisions of the annotated segments from the acoustical similarity structure. Smith *et al.* proposed an annotation format that follows a similar philosophy [15], where a structural annotation is composed of multiple tracks, named "musical similarity", "function" and "lead instrumentation". In particular, the musical similarity structure is annotated on two coarse temporal scales as follows. First, the music piece is divided into structural segments labeled using an infinite vocabulary. Then, a coarser structure is obtained by re-labeling the segments using a set of symbols whose size is limited to five, in order to "guide the

[1]Paradigmatic : pertaining to the relationship among elements that can be substituted with each other in a given context. In general, elements in a paradigmatic relationship form a substitution class. In music, this can be understood as one aspect of structural analysis which consists in segmenting a piece on the basis of the repetitions or similarities of its musical content. See for instance [11], [12], [13].

annotator towards a certain level of abstraction" [15]. From a different viewpoint, Bruderer *et al.* focused on the perceptual characterization of structural boundaries within music pieces, through the analysis of annotations produced by 18 listeners for 6 songs from various genres [16]. They relate the number of times a boundary has been marked to its perceived saliency and note that "there is a wide range of salience across different boundaries", suggesting that all the structural boundaries may not be perceptually salient. They also remarked that the terms justifying the annotation of boundaries mainly relate to "a change in timbre", "repetition", "change in dynamics" and "rest", i.e. cues of various nature.

To deal with the ambiguity of music structure, Bimbot et al. [1] assume the existence of a latent discrete structure within music pieces, called *semiotic*, for which the structural cues mentioned above are surface expressions. The analysis of the semiotic structure is performed under a structuralist point of view : the music piece is considered as the output of a system whose elements, the structural segments, have comparable sizes and share similarity and temporal relationships as the result of an underlying piece-specific code.

A semiotic sequence can look like :

**A B C D E F B C D E G D E D E H**

where the letters encode the relative location of the segments as well as their degree of similarity - each symbol represents a class of similarity.

The multiplicity of the temporal scales of perception implies the need to choose a particular granularity for the structural analysis. A finer granularity can lead to redundant decompositions, such as

**AA'BB'CC'DD'EE'FF'BB'CC'DD'EE'GG'DD'EE'DD'EE'HH'**

whereas a coarser granularity may converge towards irregular decompositions composed of more heterogeneous segments like

**AB CDE FB CDE G DE DE H**

which can be rewritten as :

**A B C B D E E F**

Hence, the granularity associated to the semiotic structure is driven by the tradeoff between the following properties :

- accuracy : a high informativeness of the structural segmentation
- concision : the use of few structural symbols to characterize the musical content
- regularity : segments of comparable size and/or, more generally speaking, conforming to a specific segment model

Around these concepts, the semiotic approach for the structural analysis of music pieces is based on general principles so as to cover a large panel of music genres. It incorporates assumptions which limit scale ambiguity and increase the reproducibility of the annotation process [1] :

- The manual annotation of the music piece is performed by locating structural segments whose musical content is repeated over time and which show an internal consistency w.r.t. the *System and Contrast* model [17], whose properties are then used to label the various music segments [18].
- A property of regularity is enforced by an additional assumption favoring segments whose sizes tend to be distributed around a small number of values[2]. This assumption is applied as a smooth constraint in order to deal with the variety of musical structures : it can be partly relaxed over a music piece to favor occasionally the global consistency and the concision of the overall structure.

As an example, one can imagine a music piece composed of regular structural segments, some of which are incompletely realized from time to time, as in :

$$\mathbf{A\ B\ \frac{C}{2}\ A\ B\ C\ D\ \frac{C}{2}\ D}$$

In this article, we focus on the automatic estimation of the semiotic segment boundaries. As is developed in Section IV, we incorporate in the formulation of the segmentation algorithm a mechanism which favors segmentations with comparable segment sizes, i.e., segment durations concentrated around a target value named the *structural period*.

## III. Outline of audio-based structural segmentation methods

This section provides a brief outline of the main components involved in automatic systems for the structural segmentation of music pieces[3].

Structure estimation systems are mainly composed of two steps. First, a feature extraction step transforms the audio signal of a music piece into a sequence of feature vectors $X = \{x_t\}_{1 \leq t \leq T}$ modeling some acoustical properties over time. Second, a structural analysis step produces a description of the structure of the piece by means of a sequence of consecutive segments $S = \{s_k\}_{1 \leq k \leq K}$ covering the whole piece.

The type of structure implies the choice of *musical dimensions* to analyze, the way a structural segment is characterized (*segment detection criteria*) and the additional constraints used to converge towards a single macroscopic structure (*structural constraints*).

### A. Musical dimensions

Most music structure estimation approaches rely on timbral, tonal and/or rhythmical descriptions of the music pieces over time. These descriptions are obtained by decomposing the audio signal into short temporal frames and extracting a vector of features for each of them. The frame duration and the hop size are either fixed beforehand with absolute values (typically a few hundred milliseconds [22]) or depend on the beat scale previously estimated by specific algorithms [23].

[2]The "size" of a segment is expressed using a temporal unit comparable to the scale of onbeats, i.e. the first and the third beats of 4-beat bars, to be synchronous to the pace of the music. It is conceptually different, yet correlated to, the notion of segment duration (absolute time).

[3]For more detailed overviews, see for instance [19], [20], [21].

One can describe the musical content of an audio frame in relation to its "overall timbre". This dimension is the result of the timbres of the instruments playing as well as the way they are played, mixing spectral and temporal information. Multiple features are used to describe the overall timbre of music excerpts [24]. In the context of structure estimation, they model the spectral shape of an audio signal using statistical and shape metrics such as spectral centroid, rolloff, contrast, valley [25], or perceptually-motivated features such as the popular Mel-Frequency Cepstral Coefficients (MFCC) [26], originally developed for speech analysis.

The tonal content of an audio excerpt is generally modeled by a Chroma vector, also referred to as Pitch Class Profile. This 12-dimensional vector quantifies the energy of the audio signal for each of the 12 semitones of the chromatic scale [27]. Several variants have been proposed for its computation [28], [29].

Some music estimation approaches also consider the rhythmic content and its evolution by means of loudness curves [30], rhythmograms obtained from the autocorrelation of the perceptual spectral flux [31], musical accent detection criteria [32] or tempogram-based features [33]. Besides, some dynamic features are designed to include information on the local temporal evolution of features as in [25].

From a transversal point of view, recent work has introduced a model of short-term memory effect within the low-level description of music pieces, which can be viewed as a way to include some kind of remanence in modeling musical dimensions : [34] associates to every temporal frame its feature vector concatenated to the ones of its recent past, and evokes "substantial increases in accuracy" w.r.t. the estimated structural boundaries.

### B. Segment detection criteria

Structural segmentation systems consider various criteria that can be split into two families, depending on how a structural segment is defined : either as *consistency* criteria, or as *repetition* criteria.

Consistency criteria assume that the targeted structural segments are musical entities showing an internal coherence. The literature generally assimilates consistency to statistical homogeneity of the features within segments. As a consequence, two successive structural segments are commonly assumed to differ by a significant change in statistics. This assumption leaves out any fine temporal modeling of the segments, and is also referred as "state representation" [25]. There are principally two trends for the detection of homogeneous segments. The first one consists in locating zones of specific texture within similarity matrices using image filtering or matrix decomposition methods [35], [22], [36]. The second approach aims at detecting homogeneous portions of the sequence of feature vectors using multi-level classification processes, relying on variations of clustering, or on Hidden Markov Models (HMM) [3], [37], [38].

In the case of repetition criteria, the structural segments are assumed to be repeated elsewhere within the music piece. The analysis of repetitions, also named "sequence approach" [25], accounts for the temporal evolution of the feature vectors within the segments. The repeated patterns can be inferred directly from the sequence of feature vectors, using methods based on HMM [39] and sparse decompositions [40], or by locating diagonal bands with high scores in similarity matrices [41], [23], [42], [5].

Analyzing music structure in terms of either homogeneity or repetition shows variable efficiency across the diversity of music pieces. As a consequence, recent "hybrid" approaches have been proposed that rely on multiple families of criteria. Several approaches merge segment distance matrices obtained from homogeneity and repetition-based segmentations in a single matrix representation [43], [37]. Others perform the linear combination of several segment detection criteria based on a similar formulation [44], where criteria based on the generalized likelihood ratio are used to account for homogeneity breakdowns, repetition boundaries and short events. Different criteria can be used at various steps of the structural segmentation process. For example, repetitions of consecutive small homogeneous entities are searched across the music piece [45], or homogeneous segments are located along a sequence of feature vectors encoding the repetitiveness of the musical content over time [46].

Beyond this inventory, new approaches aim at learning segment detection criteria directly from examples provided by manual structural annotations of a catalog of music pieces. Namely, [47] relies on a convolutional neural network trained on multiple descriptions of music pieces (such as mel-log spectrograms and similarity matrices computed at several levels of granularity) as input, and manually annotated structures at two scales of analysis as output.

### C. Structural constraints

As discussed in Section II, the structure of music can be approached at different time scales. The specification of the structural segments is not sufficient to converge towards a single description of music structure and additional assumptions are considered — explicitly or implicitly — to solve this problem.

As a consequence, the analysis of music structure is typically driven by constraints on the number of structural segments, their duration, the number of segment labels (i.e. classes) and their size. These constraints are not independent from each other : looking for larger segments tends to decrease the total number of segments found within a music piece, and decreasing the number of labels mechanically increases the number of segments per class.

In the literature, the duration of the macroscopic structural segments can be constrained through the specification of minimal or maximal duration values [41], [40], through reference sizes like multiples of 4 beats[4] [5], or by the temporal resolution of the analysis performed on the features [48], [34], [45]. Some approaches model the structural segmentation as an optimization process incorporating penalties on the use of a large number of segments [31], numerous segment classes

[4]A group of 4 beats often corresponds to a bar in popular music.

[40], or on structures showing low intra-class similarities and inter-class dissimilarities [26], [3].

Alternatively, the scale ambiguity of music structure can be dealt with by using a regularity assumption, favoring segments whose size are concentrated around a particular value, referred to as *structural period* in this article. Several methods incorporate structural constraints implementing more or less strictly this assumption. [40] performs a sparse decomposition of the chromagrams using components whose size is fixed to a value between 10 and 120 beats — a size of 70 beats is found to yield optimal segmentation performance. [38] influences the labeling of each time frame w.r.t. its neighbors in order to favor large segments. The size of the neighborhood is set to 8 and 16 beats, and the authors note that "[its] exact value is unlikely to be significant". The structural segmentation in [49] is performed by a hidden semi-Markov model incorporating a distribution on the duration of the states. A state represents segments of a same class, and the distribution of segment durations is initialized in the same way for all states, giving only high probability for multiples of a reference value, which is, "over a test set of popular music [...] reliably found to be a four bar phrase length [...]". Finally, [50] performs a post-processing step on its boundary estimations, shifting them by one downbeat forward or backward, and keeping the changes favoring a segmentation whose elements have a size of 2, 4, 8 and 16 downbeats.

To sum up this brief state-of-the-art of segmentation methods, most approaches achieve a compromise between, on one hand, a data-fitness score and, on the other hand, the compliance to structural constraints. We therefore propose in the next section to formulate explicitly the structural segmentation problem as a cost minimization process where these two terms are identified separately and can be dealt with independently from each other.

## IV. Formulation and implementation

### A. Problem formulation

Denoting as $X = \{x_t\}_{1\leq t\leq T}$ the sequence of feature vectors for a given music piece, a segmentation consists in a set $S = \{s_k\}_{1\leq k\leq K}$ of consecutive and non-overlapping segments covering the whole piece, i.e. $s_k = [t_k, t_{k+1}[$ with

$$1 = t_1 < t_2 < ... < t_k < t_{k+1} < ... < t_K < t_{K+1} = T + 1$$

The search for the optimal segmentation $S^*$ can be approached as a cost minimization problem : any segmentation $S$ is associated to a particular cost $C(S)$, which we assume to be the sum of the costs $\{\Gamma(s_k)\}_k$ of its $K$ segments $\{s_k\}_k$ :

$$C(S) = \sum_{k=1}^{K} \Gamma(s_k) \tag{1}$$

If we further assume that $\Gamma(s_k)$ decomposes into a first term related to data-fitness and a second one to structural compliance, it is natural to formulate $\Gamma(s_k)$ as :

$$\Gamma(s_k) = (1-\lambda)\Phi(s_k) + \lambda\Psi(s_k) \tag{2}$$

where

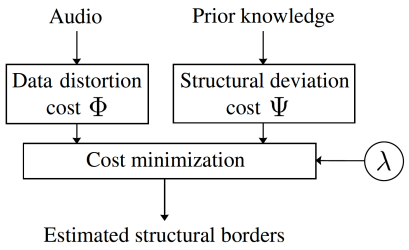



Fig. 2. Block diagram of the structural boundary estimation approach

- $\Phi$ is a data distortion cost which returns low values for sequences $s_k$ that match well the expected acoustic properties of feature vectors within structural segments.
- $\Psi$ is a structural deviation cost which penalizes segments that do not comply with the properties of the targeted structure, yielding for instance higher values for segments whose size is not in accordance with the scale at which the structure is intended to be described.
- $\lambda$ is a tuning parameter ranging between 0 and 1, which balances the relative importance of $\Phi$ and $\Psi$.

This generic formulation, schematized as a block diagram in Fig. 2, encompasses a large variety of structure types depending on the choice made in terms of criteria and constraints. An example of use of this formulation in a specific case can be found in [31] where the cost $\Phi(s_k)$ is the average of the self-similarity matrix obtained on the sequence of features associated to candidate segment $s_k$, and the constraint $\Psi(s_k)$ is a fixed constant, the "cost of segment split", which penalizes segmentations with many segments, i.e., fine-grained segmentations.

### B. Cost minimization using a Viterbi algorithm

The search for the segmentation with minimal cost can be implemented by means of a Viterbi algorithm [44].

We denote $S_t$ the partial segmentation of minimal cost $C_t$ related to $X_1^t = [x_1, x_t[$, i.e. the portion of $X$ from the beginning of the piece up to time instant $t$, and we let $s_{t,h}$ be the segment associated to the sequence of features $X_{t-h}^t = [x_{t-h}, x_t[$ preceding the time index $t$ within a window of length $h$. We also denote as $H$ the maximal possible window length[5]. The algorithm progresses through the following steps :

- *Initialization*
  By convention, we set $S_1 = \emptyset$ and $C_1 = 0$.
- *For* $t = 2 : T + 1$
  Evaluate the relative position $h$ of the predecessor of $t$ minimizing the cumulative cost $C_t$ of the segmentation $S_t$. As depicted in Fig. 3, let $\{s_{t,h}\}_{1\leq h\leq H'}$ be the set of admissible predecessors for time index $t$, where $H' = \min(t-1, H)$. We assume to know the optimal segmentations $\{S_{t-h}\}_{1\leq h\leq H'}$ as well as their associated cumulative costs $\{C_{t-h}\}_{1\leq h\leq H'}$. The best partial segmentation $S_t$ is built by choosing the extension $s_{t,h}$ of

[5]Typically, $H = T$, but smaller values can be considered like the structural period $\tau$ or a small multiple of it.

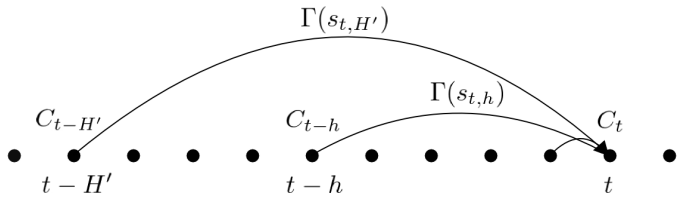



Fig. 3. Admissible predecessors for $t$ and their costs

the former partial segmentation $S_{t-h}$ totalling the lowest cost. We evaluate successively :

1) $\Gamma(s_{t,h})$ for $1 \leq h \leq H'$
2) $h^*(t) = \text{argmin}_{1 \leq h \leq H'}\{C_{t-h} + \Gamma(s_{t,h})\}$
3) $C_t = C_{t-h^*(t)} + \Gamma(s_{t,h^*(t)})$

and we set $S_t = S_{t-h^*(t)} \cup \{s_{t,h^*(t)}\}$.

The optimal segmentation $S^*$ of $X$ with minimal cost $C^*$ is obtained by backtracking the optimal predecessors using $h^*(t)$. Denoting (by anticipation) $K^*$ as the number of segments in $S^*$, the associated time indexes $\{t_k\}$ are found thanks to the following recursion :

1) $t_{K^*+1} = T + 1$ (i.e. $s_{K^*}$ ends in $T$)
2) $t_k = t_{k+1} - h^*(t_{k+1})$, for $1 \leq k \leq K^*$

In practice, $K^*$ is only obtained at the end of the backtracking process as the actual number of segments in $S^*$.

The total cost $C^*$ can be normalized (for instance by $T$, if it were to be compared across songs). As we are only interested in the segmentation, no normalization is implemented in the present work.

### C. Designing the regularity function

The data distortion cost $\Phi$ can detect segments belonging to potentially any temporal scale. In order to constrain the scale of the structure analysis, we incorporate a structural deviation cost $\Psi$ favoring the regularity of segments, i.e., segment sizes distributed around a target value[6] noted $\tau$. Such a constraint implements the structural period assumption related to the semiotic structure as presented in Section II. With $s$ denoting a structural segment of size $m$, the cost $\Psi$ can rely on functions that respect the following properties :

1) $\Psi(s) = 0$ if $m = \tau$
2) $\Psi(s) > 0$, with increasing values as $m$ deviates from $\tau$

### D. Relationship to existing methods

A number of approaches previously used in MIR for structural segmentation of music also rely on a Viterbi algorithm.

Early works such as [51] or [3] have been followed by a number of methods using ergodic HMMs where the probability of staying in a particular state depends on a single value (self-transition probability).

However, a few approaches rely on explicit constraints within the Viterbi algorithm [52] or on a duration model of the hidden states [49] to favor typical segment sizes.

[6]The choice of a single target value is motivated by the observation of the annotations of the semiotic structure for the RWC popular dataset, presented in Section V-A, where 73 songs over 100 show a strongly dominant structural period.

- Shiu et. al represent a music piece according to its similarity matrix [52]. They interpret the $x$ and $y$ axis of the matrix in terms of time and states respectively, with similarity values being state probabilities over time. A Viterbi algorithm is then used to find the path with maximal cumulative probability within the bottom left corner of the matrix. Paths parallel to the main diagonal are favored through a constraint on state transition probabilities to cross diagonal sub-bands of high value, hence detecting repetitions.
- Levy and Sandler use an extension of the Viterbi algorithm to decode the sequence of macroscopic states describing the current music piece from their Hidden Semi-Markov Model (HSMM) [49]. Each state is modeled by a probability distribution and a distribution of durations. The observations related to a particular state are considered to be conditionally independent.

The HSMM approach is probably the closest one to that presented in this paper : via a negative logarithm, the state probability distributions can be turned into data-distortion costs while the duration distributions provide the contributions to the structural deviation cost. Our formalism generalizes this conception to any kind of cost functions $\Phi$ and $\Psi$, whether they are defined in a probabilistic framework or not.

## V. Overview of MIREX evaluations

MIREX (Music Information Retrieval Evaluation eXchange) is an initiative stemming from the MIR community which was launched in 2005 with the goal of comparing state-of-the-art algorithms and systems on a variety of tasks, by inviting research and development teams to submit implementations of their approaches on common benchmarks [53], [54].

The work presented in this article was originally developed in the context of the MIREX "structural segmentation" task[7] over the period 2010–2012 and has been further deployed on more recent MIREX results produced over the period 2013–2015.

This section provides a rapid summary of the MIREX campaigns in "structural segmentation" over 2010–2015. We briefly present a selection of datasets and metrics used as benchmarks in this context, as well as a shortlist of 17 MIREX systems, which we focus on later in this article.

### A. Evaluation datasets

Four datasets of structural annotations were used for MIREX from 2010 to 2015.

The MIREX09 dataset corresponds to the grouping of several existing datasets produced by Tampere University of Technology, Vienna University of Technology and Queen Mary University of London. The dataset gathers 297 popular music pieces. The track list is currently unavailable, however "music of The Beatles makes up a significant proportion of the MIREX 2009 dataset" [55]. If MIREX09 was the first substantial dataset used by MIREX, the variety of the sources implies some inconsistency in the structural annotation

[7]http://music-ir.org/mirex/wiki/2015:Structural_Segmentation

methodology, that can vary significantly across the music pieces considered. To our knowledge, the methodologies employed for the production of the corresponding annotations are not publicly available today.

The MIREX10 dataset is composed of the 100 J-pop songs from the RWC Popular dataset proposed by the AIST [56]. The annotations used are the ones produced by an early version of the semiotic structure annotation methodology in the scope of the Quaero project [7]. Some minor revisions were applied to this dataset since 2010, in conformity with recent evolutions of the methodology [18], but the MIREX10 dataset itself remained unchanged.

Two other datasets have also been used in MIREX since 2012. The first one is made of the original structural segmentations of the RWC Popular dataset produced by the AIST, which focuses on the annotation of choruses and verses, synchronizing their borders to manually annotated beats [57]. The second consists in the annotations of over 1000 music pieces from the SALAMI dataset, which considers various genres and recording conditions (live or not), and partly covers the MIREX09 and MIREX10 datasets [15]. It was annotated by CIRMMT/McGill University, in terms of "musical similarity" (according to two temporal scales), "musical function" and "lead instrumentation", as mentioned in Section II. This dataset is referred to as MIREX12 in this paper.

Reference annotations from MIREX09, MIREX10 and MIREX12 have been made available by Johan Pauwels and can be downloaded from a Github repository[8].

### B. Performance metrics for boundary estimation

As considered by MIREX, several metrics can be used to evaluate the quality of a structural segmentation[9]. This requires some measure of the differences between estimated boundaries provided by an automatic system and reference boundaries as annotated in the benchmarking data.

In this article, we focus on the *boundary hit rates* which refer to the well-known Precision ($P$), Recall ($R$) and F-measure ($F$) metrics adapted to the matching of structural boundaries. Denoting as $b_E$ and $b_R$ the estimated and reference boundaries respectively, $P$, $R$ and $F$ are defined as follows :

$$P = \frac{b_E \cap b_R}{|b_E|}; R = \frac{b_E \cap b_R}{|b_R|}; F = \frac{2PR}{P+R} \qquad (3)$$

where $|X|$ corresponds to the number of elements of $X$. Two boundaries match if they are contained within a tolerance window of fixed duration (either 0.5 s [58] or 3 s [38]). During the calculation of $P$ and $R$, each estimated boundary can be matched to only one reference boundary and vice versa[10].

### C. Participants

Since their beginning, MIREX evaluations on structural segmentation have gathered typically half a dozen submissions every year, sometimes many more.

| Name | Ref. | Name | Ref. | Name | Ref. |
|---|---|---|---|---|---|
| CC1 | [60] | IRISA11 | [61] | NB2 | [62] |
| CL1 | [37] | IRISA12 | [63] | OYZS1 | - |
| GP6 | [64] | KSP1 | [65] | RBH2 | [66] |
| GP7 | [64] | MHRAF1 | [67] | SMGA1 | [68] |
| GS1 | [69] | MND1 | [5] | WB1 | [40] |
| IRISA10 | [70] | MP2 | [46] | - | - |

TABLE I
ACRONYMS AND REFERENCE OF THE MIREX STRUCTURAL SEGMENTATION SYSTEMS CONSIDERED IN THIS ARTICLE.

We have considered a selection of 17 of these systems (including ours), among the 40 submissions evaluated during the MIREX campaigns between 2010 and 2015. This subset was defined according to the individual system performances and/or to their specificities. In particular, when some systems were submitted several times with some variations in the tuning of their parameters, we ignored duplicates that obtained similar or lower performance. These systems are listed in Table I and named according to their acronym in MIREX, except for IRISA10, IRISA11 and IRISA12 which were originally labeled as BV1, SBVRS1 and SBV1 and are renamed in this article for more clarity. While Table I also provides a reference to the system descriptions and/or approaches[11], a short summary of each of them is also proposed as supplementary material to this article [59].

Following the broad categories defined in Section III-B, systems CC1, CL1, GP6-7, KSP1 and RBH2 rely mainly on homogeneity criteria, whereas systems IRISA11, MHRAF1, MND1, NB2 and WB1 are essentially based on repetition criteria. Systems GS1, IRISA10, IRISA12, MP2 and SMGA1 can be viewed as hybrid combinations of several segmentation criteria. It is beyond the scope of this article to carry out an extensive comparison of these methods, but we show in Section VII how this diversity can be exploited in a "cooperative" way, through fusion.

### D. Comments on MIREX performance

Tables II, III and IV summarize the performance levels obtained on the MIREX09, MIREX10 and MIREX12 datasets by the systems considered in the previous section, in terms of boundary hit rates with two levels of tolerance, 0.5 s and 3 s. The values we report here are the ones from the MIREX website[12] with an accuracy of 0.1.

In these two tables, some general trends can be observed :

- Performance levels with a tolerance of 3 s are naturally better than those with 0.5 s. It is worth noting that 0.5 s is approximately the duration of a beat in a popular song, whereas 3 s can be expected to be above the length of a typical bar.
- One can also observe that the overall performance of all systems is higher on MIREX10 than on MIREX09. Such a behavior may be related to the structural profile

[8] https://github.com/jpauwels/mirex-tools

[9] http://www.music-ir.org/mirex/wiki/2015:Structural_Segmentation#Boundary_retrieval

[10] The evaluation scripts used for this article are based on a version developed by Jouni Paulus provided to us directly by MIREX.

[11] Except one which has not been released, to our knowledge.

[12] The original values are published in the Structural Segmentation task item of http://www.music-ir.org/mirex/wiki/2010:MIREX2010_Results, http://www.music-ir.org/mirex/wiki/2011:MIREX2011_Results and http://www.music-ir.org/mirex/wiki/2012:MIREX2012_Results.

| MIREX09 dataset | | | | | | | |
|---|---|---|---|---|---|---|---|
| | | tol = 0.5 s | | | tol = 3 s | | |
| Year | Participants | $F$ | $P$ | $R$ | $F$ | $P$ | $R$ |
| 2010 | GP7 | 18.1 | 14.4 | 25.7 | 50.1 | 40.0 | 70.3 |
| | IRISA10 | 21.7 | 18.1 | 29.2 | 56.7 | 47.0 | 75.7 |
| | MND1 | 32.5 | 33.5 | 33.4 | 60.7 | 62.6 | 62.6 |
| | WB1 | 20.0 | 19.6 | 21.8 | 47.5 | 46.3 | 51.6 |
| 2011 | CL1 | 15.1 | 15.6 | 15.7 | 41.0 | 41.8 | 42.7 |
| | GP6 | 17.4 | 13.7 | 25.4 | 48.7 | 38.2 | 70.4 |
| | IRISA11 | 23.1 | 20.2 | 28.6 | 53.3 | 46.6 | 66.4 |
| 2012 | IRISA12 | 22.7 | 21.6 | 25.0 | 55.4 | 52.7 | 61.2 |
| | KSP1 | 28.2 | 24.1 | 35.8 | 59.1 | 50.6 | 75.0 |
| | MHRAF1 | 22.0 | 22.5 | 22.8 | 52.5 | 53.4 | 54.7 |
| | OYZS1 | 19.1 | 25.0 | 17.5 | 44.1 | 54.9 | 41.1 |
| | SMGA1 | 22.8 | 20.5 | 26.7 | 64.5 | 58.0 | 75.1 |
| 2013 | MP2 | 28.1 | 26.3 | 32.0 | 55.4 | 52.1 | 62.5 |
| | RBH2 | 25.4 | 23.1 | 30.4 | 56.9 | 50.6 | 69.3 |
| 2014 | NB2 | 23.7 | 22.1 | 26.7 | 63.6 | 59.5 | 71.2 |
| 2015 | CC1 | 19.6 | 16.4 | 26.4 | 59.3 | 48.9 | 79.8 |
| | GS1 | 52.3 | 49.9 | 57.2 | 64.9 | 62.0 | 70.9 |

TABLE II
PERFORMANCE ON THE MIREX09 DATASET OF THE MAIN MIREX SUBMISSIONS (STRUCTURAL SEGMENTATION TASK) FROM 2010 TO 2015.

| MIREX10 dataset | | | | | | | |
|---|---|---|---|---|---|---|---|
| | | tol = 0.5s | | | tol = 3s | | |
| Year | Participants | $F$ | $P$ | $R$ | $F$ | $P$ | $R$ |
| 2010 | GP7 | 22.8 | 23.3 | 23.3 | 57.1 | 57.5 | 59.1 |
| | IRISA10 | 23.4 | 24.3 | 23.7 | 61.0 | 61.9 | 62.2 |
| | MND1 | 35.9 | 44.1 | 32.3 | 60.5 | 73.6 | 54.4 |
| | WB1 | 29.1 | 36.2 | 24.9 | 58.2 | 72.0 | 50.0 |
| 2011 | CL1 | 23.1 | 30.8 | 19.1 | 43.4 | 57.1 | 36.3 |
| | GP6 | 18.8 | 17.5 | 21.0 | 53.4 | 49.8 | 59.5 |
| | IRISA11 | 32.4 | 32.6 | 33.2 | 61.2 | 62.2 | 62.3 |
| 2012 | IRISA12 | 25.3 | 28.4 | 23.3 | 62.8 | 69.8 | 58.4 |
| | KSP1 | 36.3 | 37.6 | 36.2 | 66.1 | 68.7 | 65.8 |
| | MHRAF1 | 28.9 | 40.4 | 23.2 | 54.5 | 75.8 | 43.8 |
| | OYZS1 | 29.9 | 41.6 | 24.3 | 53.1 | 73.9 | 43.3 |
| | SMGA1 | 26.8 | 28.7 | 25.6 | 76.6 | 81.6 | 73.5 |
| 2013 | MP2 | 35.5 | 45.8 | 29.6 | 59.0 | 76.1 | 49.3 |
| | RBH2 | 37.5 | 39.2 | 36.8 | 67.3 | 70.0 | 66.4 |
| 2014 | NB2 | 31.4 | 35.0 | 29.1 | 70.0 | 78.1 | 64.6 |
| 2015 | CC1 | 22.4 | 22.1 | 23.9 | 60.0 | 58.2 | 65.0 |
| | GS1 | 69.7 | 80.4 | 62.7 | 79.3 | 91.9 | 71.1 |

TABLE III
PERFORMANCE ON THE MIREX10 DATASET OF THE MAIN MIREX SUBMISSIONS (STRUCTURAL SEGMENTATION TASK) FROM 2010 TO 2015.

of MIREX10 songs which are more regular (see Section V-A), and possibly to the higher consistency of the annotation methodology used for this dataset.

- The typical F-measure gain across six campaigns is about 15% with $tol = 3$ s, and about 20% for $tol = 0.5$ s, up to 35% for the GS1 system in 2015. This illustrates the significant progress made by structural segmentation methods, as measured by the MIREX campaigns over the recent years.
- At the time of the evaluations, IRISA10 and IRISA11 were ranked amongst the 3 best MIREX systems on MIREX09 and MIREX10 datasets and in particular, IRISA11 appeared as the best performing system on MIREX10 in 2011.
- The performance levels reported on the SALAMI-based dataset (MIREX12) are significantly lower than those of the two other datasets. Also, the system rankings are unalike : see Kendall's rank correlation coefficient shown in Table V[13].

## VI. DIAGNOSTIC ANALYSIS

In this section, we develop some diagnostic elements of the IRISA11 system on the MIREX10 dataset, so as to investigate further the impact of the regularity constraint. We focus specifically on the optimal tuning of the system, we compare the basic regularity approach to a slightly more sophisticated histogram-based segment duration model, and we investigate experimentally the difference of behavior of the three IRISA systems by comparing their segmentation output. While the three IRISA approaches are presented next, the full system descriptions are detailed in [59].

[13]Systems from 2010 were later evaluated on the entire SALAMI dataset (over 1300 songs). However, if the results are available in [55], they are not reported in Table IV as the dataset is not exactly the same as the one used by MIREX.

| MIREX12 (SALAMI) dataset | | | | | | | |
|---|---|---|---|---|---|---|---|
| | | tol = 0.5 s | | | tol = 3 s | | |
| Year | Participants | $F$ | $P$ | $R$ | $F$ | $P$ | $R$ |
| 2010 2011 | SALAMI dataset not yet released | | | | | | |
| 2012 | IRISA12 | 15.7 | 13.6 | 21.0 | 43.4 | 37.8 | 57.4 |
| | KSP1 | 27.9 | 22.3 | 43.7 | 49.0 | 39.2 | 76.7 |
| | MHRAF1 | 18.8 | 19.4 | 19.9 | 42.3 | 44.5 | 44.0 |
| | OYZS1 | 28.7 | 45.8 | 25.3 | 43.7 | 64.1 | 39.7 |
| | SMGA1 | 19.2 | 15.6 | 28.2 | 49.2 | 40.4 | 70.3 |
| 2013 | MP2 | 31.7 | 29.5 | 39.0 | 51.9 | 48.5 | 63.4 |
| | RBH2 | 26.0 | 21.1 | 38.5 | 49.7 | 40.4 | 73.6 |
| 2014 | NB2 | 26.1 | 22.7 | 34.1 | 52.7 | 46.4 | 67.7 |
| 2015 | CC1 | 21.3 | 16.9 | 33.0 | 50.7 | 40.6 | 77.1 |
| | GS1 | 54.1 | 49.6 | 64.5 | 62.3 | 57.3 | 73.9 |

TABLE IV
PERFORMANCE ON THE MIREX12 DATASET OF THE MAIN MIREX SUBMISSIONS (STRUCTURAL SEGMENTATION TASK) FROM 2012 TO 2015. SYSTEMS SUBMITTED BEFORE 2012 WERE NOT RUN ON THIS DATASET.

### A. A brief description of the IRISA systems

*IRISA10:* timbral and tonal information are respectively described by beat-synchronous sequences of MFCCs and Chroma vectors[14]. The data distortion cost $\Phi$ combines three segment detection criteria formulated as Generalized Likelihood Ratios : a timbral homogeneity breakdown criterion, a tonal repetition breakdown criterion and a short event detection

[14]We used the MA-Toolbox by Pampalk to compute the MFCCs [71] and Matlab scripts by Ellis for the extraction of Chroma vectors and beats [72].

| | $tol = 0.5$ s | $tol = 3$ s |
|---|---|---|
| MIREX09 vs MIREX10 | 0.11 | 0.38 |
| MIREX09 vs MIREX12 | -0.07 | 0.56 |
| MIREX10 vs MIREX12 | 0.11 | 0.20 |

TABLE V
KENDALL'S RANK CORRELATION COEFFICIENTS FOR MIREX SYSTEMS CONSIDERED IN SECTION V-C WITH RESPECT TO THEIR AVERAGE $F$-MEASURE ON MIREX09, MIREX10 AND MIREX 12, FOR TOLERANCE VALUES OF 0.5 S AND 3 S.

criterion relying on the timbre. The structural deviation cost $\Psi$ is a parabolic function, shifted to reach its minimum at $\tau$ which is estimated using a spectral analysis of the homogeneity breakdown criterion.

*IRISA11:* the audio stream is first described as a sequence of estimated chords expressed at the scale of onbeats[15]. Then, the optimal segmentation is searched using a data distortion cost $\Phi$ measuring the repetitiveness of the chord sequence of each segment and a structural deviation cost $\Psi$ favoring segments whose sizes are close to the structural period $\tau$. The value of $\tau$ is set to 16 temporal units (onbeats), and $\Psi$ is a non-convex function.

*IRISA12:* this system mainly differs from IRISA11 by its features and its data distortion cost $\Phi$. Here, the audio stream is described by a variant of Chroma vectors[16] expressed at the scale of onbeats. $\Phi$ allows to search for segments whose inner organization can be modeled through a short sequence of mid-term entities following a "System and Contrast" pattern [17] (i.e., patterns like $aabb$, $aabc$, $abab$ and $abac$). The structural deviation cost $\Psi$ is the same as IRISA11, with $\tau$ empirically set to 16 onbeats.

### B. Optimizing the regularity constraint

To further analyze the benefits brought about by a simple regularity constraint $\Psi$, we study the behaviour of the IRISA11 system for various weights and shapes of $\Psi$.

Let $m$ be the size of a segment. In the IRISA11 system, we consider regularity cost functions which are symmetric as regards $\tau$, and are defined as :

$$\Psi_\alpha(m) = \left|\frac{m}{\tau} - 1\right|^\alpha \quad (4)$$

Parameter $\alpha$ controls the convexity of the function : $\Psi_\alpha$ is non-convex if $0 < \alpha < 1$, and it is convex if $\alpha > 1$. $\Psi_\alpha$ is represented in Fig. 4 for $\alpha = \{0.5, 1, 2\}$. On the basis of our annotation experience, we set $\tau$ to 16 time units regarding the temporal scale (i.e. 16 onbeats). Non-convex versions of $\Psi$ penalize more severely segments which deviate from $\tau$, whereas convex ones are more "lenient" to small irregularities. More specifically, values of $\alpha$ below 1 tend to favor structures for which irregularities can be stronger but less frequent, whereas the convex case ($\alpha$ above 1) tends to favor structures where the irregularities are less radical but can be more widely distributed over the whole segmentation. As formulated in Eq. 2, $\Psi_\alpha$ is combined with the data distortion term $\Phi$ with relative weights $\lambda$ and $(1-\lambda)$ respectively.

To explore the impact of $\alpha$ and $\lambda$ on the performance

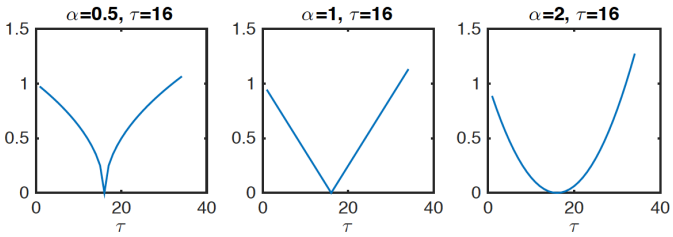



Fig. 4. Shape of the regularity cost function $\Psi_\alpha$ for $\alpha \in \{0.5, 1, 2\}$ and $\tau = 16$.

of IRISA11 system[17], we ran segmentation experiments for $\alpha \in [0, 3]$ with a hop size of 0.1, and $\lambda \in [0, 1]$ with a hop size of 0.05. Fig. 5 shows the evolution of the average F-measure obtained on MIREX10 for an optimal tuning of $\alpha$ (top), and of $\lambda$ (bottom). It can be observed that the maximal average F-measures decrease for extreme values of $\lambda$, for the two tolerances $tol$.

Optimal performance levels[18] are reached for $\lambda = 0.80$ ($F_{\text{av}} = 32.1$ % with $\alpha_{\text{opt}} = 0.6$) for a tolerance value of 0.5 s and $\lambda = 0.70$ ($F_{\text{av}} = 61.5$ % with $\alpha_{\text{opt}} = 0.1$) for a tolerance value of 3 s.

Fig. 6, top, shows the optimal values of $\alpha$ maximizing the average F-measure over all values of $\lambda$. Although the curves are a bit erratic, most values of $\alpha_{\text{opt}}$ happen to fall below 1 for $tol = 0.5$ s, indicating an advantage for non-convex cost functions (this is less marked for $tol = 3$ s). Moreover, it seems that very low values of $\alpha$ are particularly relevant when $\lambda \simeq 0.5$

Fig. 5, bottom, shows the evolution of the maximal average F-measure when $\lambda$ is tuned optimally for each $\alpha$. If this evolution is quite flat for $tol = 3$ s, a more significant increase is noticeable for low values of $\alpha$ for $tol = 0.5$ s. In any case, the highest performance is obtained for non-convex functions ($\alpha < 1$) when $\lambda$ is optimal. The optimal values of $\lambda$ given $\alpha$ are depicted in Fig. 6, bottom. These values fall in the interval [0.65,1], hence giving a significant importance to the regularity constraint in the segmentation criteria.

### C. Histogram-based segment duration model

In this section, we explore a more complex regularity cost, derived from the actual distribution of the structural segment sizes from the reference annotations. We wish to investigate whether a finer *a priori* knowledge of structural information can lead to better estimates of the structural boundaries. To test this assumption, we compare the performance of IRISA11,

[15]The chord estimation is performed by the algorithm by Ueda *et al.* [73], beats/downbeats used for onbeat estimation are extracted using Matlab scripts by Davies [74], [75].

[16]*Chroma Pitch* features extracted with the Chroma Toolbox by Müller *et al.* [29].

[17]The evaluation of IRISA11 on MIREX10 using our own equipment (Linux 64 bits for the chord estimation, MacOS 64 bits for the remaining) leads to some minor variations versus those published by MIREX and reproduced in Table IV. We obtained the following average results on the dataset using the same evaluation metrics : $F = 31.2\%$, $P = 31.4\%$, $R = 32.0\%$. Internal tests have shown that small variations occurring in the estimation of chords, beats and downbeats impact the Viterbi decoding described in Section IV-B. However, we did not have access to the features extracted by our algorithm on MIREX servers for further investigation.

[18]The difference between the maximal values obtained in this section and the ones from MIREX are due to the hop size used to browse the values of $\lambda$ as well as the variations in the features we extracted locally, as stated in the previous footnote.

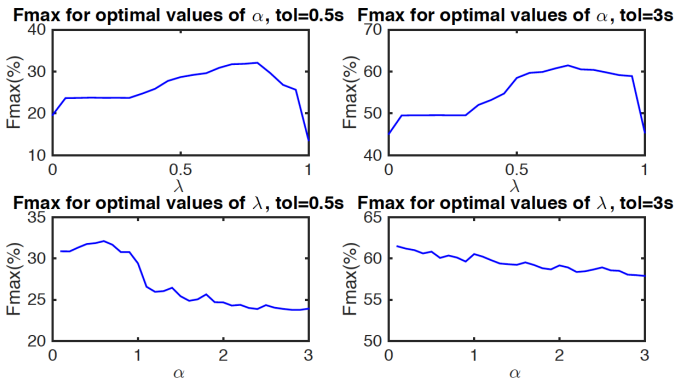



Fig. 5. Maximal average F-measure on the MIREX10 dataset, as a function of $\lambda$ for an optimal tuning of $\alpha$ (top) and as a function of $\alpha$ for optimal values of $\lambda$ (bottom)

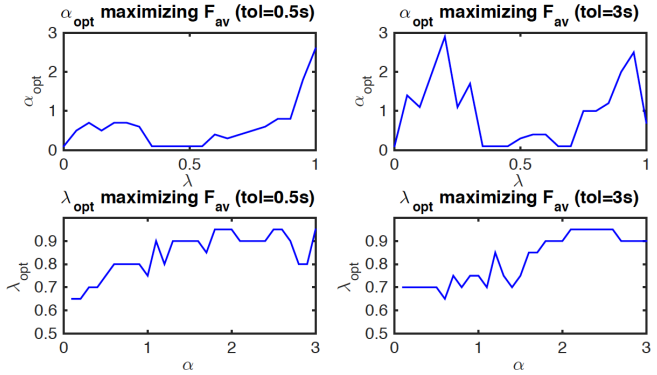



Fig. 6. Optimal values of $\alpha$ (top) and $\lambda$ (bottom) as a function of one another on the MIREX10 dataset.

as described in the previous subsection, to the one obtained by the same system in which $\Psi_\alpha$ is replaced by a cost function learned from the reference annotations of MIREX10. We consider two versions of this cost : both can be viewed as deriving from "oracle priors", i.e. the prior knowledge of the distribution of segment durations for each reference song (ORACLE1) or over the whole MIREX10 corpus (ORACLE2).

More specifically, the first regularity cost function $\Psi_{\text{ORACLE1}}$ is obtained by computing, for each music piece, the actual histogram $D$ of the segment sizes[19] from the reference annotations. $D$ is then converted into a regularity cost function by the following equation :

$$\Psi_{\text{ORACLE1}}(m) = 1 - \frac{D(m)}{D(\tau)} \tag{5}$$

where $m$ is the size of a segment in onbeats and $\tau = \arg\max_m\{D(m)\}$.

The second regularity cost function, $\Psi_{\text{ORACLE2}}$, is obtained by computing the histogram $\bar{D}$ of segment sizes over the annotations of the entire MIREX10 dataset. It is also an "oracle" setting, but a single distribution is considered for the

[19]The segment sizes are expressed according to the scale of onbeats previously estimated, each temporal boundary being approximated to its closest onbeat.

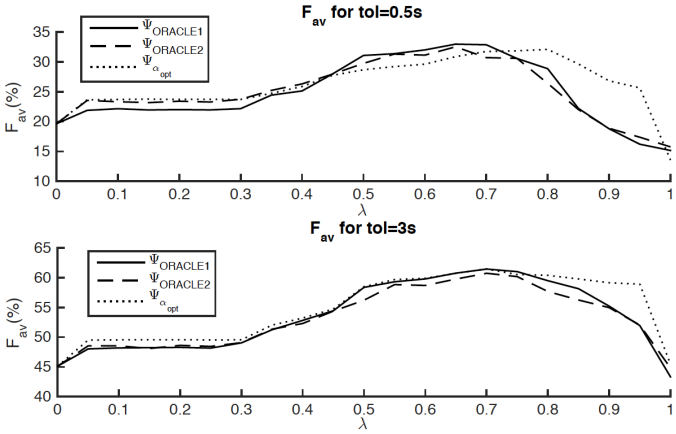



Fig. 7. Evolution of the average F-measure over MIREX10 as a function of $\lambda$, using song-based oracle regularity costs $\Psi_{\text{ORACLE1}}$ and dataset-based oracle regularity costs$\Psi_{\text{ORACLE2}}$, for tol=0.5s (top) and tol=3s (bottom). F-measures obtained with $\Psi_{\alpha_{\text{opt}}}$ are also depicted.

| Constraint | $F$max(tol = 0.5s) | $F$max(tol = 3s) |
|---|---|---|
| $\Psi_\alpha$ | 32.1 | **61.5** |
| $\Psi_{\text{ORACLE1}}$ | **33.0** | 61.5 |
| $\Psi_{\text{ORACLE2}}$ | 32.5 | 60.7 |

TABLE VI
AVERAGE F-MEASURES ON MIREX10 FOR THREE REGULARITY CONSTRAINTS: $\Psi_\alpha$ (WITH OPTIMAL PARAMETERS), SONG-BASED ORACLE HISTOGRAM $\Psi_{\text{ORACLE1}}$ AND DATASET-BASED ORACLE HISTOGRAM $\Psi_{\text{ORACLE2}}$.

whole dataset. In the same way as for $\Psi_{\text{ORACLE1}}$, we define :

$$\Psi_{\text{ORACLE2}}(m) = 1 - \frac{\bar{D}(m)}{\bar{D}(\tau)} \tag{6}$$

The evaluation of IRISA11 with these two regularity costs is summarized in Fig. 7. This figure shows the evolution of the average F-measure according to $\lambda$ for $\Psi_{\text{ORACLE1}}$ and $\Psi_{\text{ORACLE2}}$ given the tolerance of 0.5 s (top) and 3 s (bottom). In both cases, the maximal values are reached with $\lambda = 0.65$ for $tol = 0.5$ s and $\lambda = 0.70$ for $tol = 3$ s. The corresponding performance values are given in Table VI. The two oracle priors yield a level of performance comparable to the one obtained with the basic regularity constraint $\Psi_\alpha$, i.e. with no clear advantage for the histogram-based model[20].

### D. Comparing IRISA systems

As a way to compare the similarities/differences in behavior of the three IRISA systems, we have computed F-measures for each pair of systems, assuming that one of them corresponds to the reference and the other one to the estimated structure. Table VII gathers the resulting F-measures on MIREX10. Significant divergences can be noted across the three systems : in particular, all F-measures fall below 60% for a segmentation tolerance of 3 s.

Such differences may be explained by the variety of data-distortion costs implemented within the three systems as well as variants in the structural deviation costs. However, it

[20]This result may be connected to work published while the present article was under review [76].

| Systems | $tol = 0.5$ s | $tol = 3$ s |
|---|---|---|
| IRISA10 vs IRISA11 | 16.0 | 55.2 |
| IRISA10 vs IRISA12 | 13.9 | 54.7 |
| IRISA11 vs IRISA12 | 26.1 | 59.2 |

TABLE VII
COMPARISON OF THE SEGMENTATION OUTPUT OF THE THREE IRISA SYSTEMS: RELATIVE AVERAGE F-MEASURES ON THE MIREX10 DATASET

must be noted that IRISA11 and IRISA12 provide the most correlated outputs, which may come from the use of similar (non-convex) regularity constraints and comparable temporal resolutions in the two systems.

Another implication of these results is to encourage the combination of the three systems in a fusion framework, so as to exploit their potential complementarity. In this context, the regularity constraint appears as a natural way to guide the fusion process.

## VII. SYSTEM FUSION UNDER A REGULARITY CONSTRAINT

As distinct systems deliver potentially complementary information on the structural segmentation of songs, the regularity constraint may indeed provide an efficient way of combining multiple systems towards a robust segmentation output.

In this section, we consider combinations of $M$ segmentation systems to which we apply a simple late-fusion approach using the regularity constraint $\Psi$ in combination with a measure of "boundary hypothesis density" derived from the multiple outputs of the fused systems (acting as $\Phi$). Though extremely simple in its principle, this type of blind late fusion method seems to have been surprisingly under-explored for the structural segmentation of music[21].

We first describe the fusion algorithm in more details. We then apply it to the three IRISA systems and, in a last step, we use the fusion to combine a large selection of recent MIREX systems and compare the result of the combination to the state-of-the-art performance level.

### A. A simple fusion algorithm

In the forthcoming fusion experiments, we use a late fusion strategy, i.e., we consider solely the set of outputs of the $M$ segmentation systems as the input of the fusion system.

Considering a temporal integration window of duration $w$ centered on $t$, we compute the segmentation cost $\Phi_{\text{FUS}}(t)$ :

$$\Phi_{\text{FUS}}(t) = 1 - \frac{q(t)}{Q} \tag{7}$$

where $q(t)$ is the total number of segment boundaries estimated by all the fused systems within the window $w$, and $Q$ is the maximum of $q(t)$ over the entire song. Note that $q$ and $Q$ are integers and that $Q \leq M$.

$\Phi_{\text{FUS}}(t)$ takes only $Q$ distinct values and turns out to be a piecewise constant function as it keeps steady when the number of boundary hypotheses remains identical for successive time shifts of the integration window $w$. $\Phi_{\text{FUS}}$ shows local minima which range over time intervals and do not provide precise estimation of fused boundaries. In order to obtain an optimal segmentation in a certain sense, the $\Phi_{\text{FUS}}$ cost function can be combined with a regularity constraint $\Psi$ in the way described earlier in Section VI-B and handled by means of the Viterbi algorithm described in Section IV-B. This fusion approach is particularly simple and straightforward and it enables the fusion of any set of systems, treated as purely independent black-boxes.

Following Eq. 2, the combination of $\Phi$ and $\Psi$ allows a compromise between two trends : if many systems estimate a local segmentation boundary around a given instant $t$, $\Phi$ provides a low value which reinforces the segmentation hypothesis for the fusion system. However, this can be mitigated by a high value of $\Psi$ indicating an unlikely location of the hypothesized segment boundary. Conversely, a low level of boundary detection from the collection of fused systems (i.e., a high value of $\Phi_{\text{FUS}}$) can be compensated by a high *a priori* on the presence of a segment boundary, as enforced by the regularity constraint (low value of $\Psi$).

In the present section, we use a regularity constraint of the $\Psi_\alpha$ type, such as formulated in Eq. 4. Parameter $\lambda$ controls the relative importance of the two costs. The size of the temporal integration window $w$ ranges typically from 0.5 s to 8 s. The temporal resolution $\delta$, i.e. the shift between two successive estimations of $\Phi_{\text{FUS}}(t)$, is set to 0.1 s so as to finely track the evolution of the segmentation hypotheses density. The structural pulsation $\tau$ is allowed to vary between 7 s and 20 s. These various possibilities result in a relatively large number of hyper-parameters, whose combinations have been "reasonably" explored in the reported experiments hereafter.

Re-focusing once again on the MIREX10 dataset, we carry out tests on a number of system combinations under two possible modes : an "oracle" mode, where the hyper-parameters are tuned on the test-data themselves, and a cross-validation mode, where the MIREX10 dataset is split into odd-numbered and even-numbered songs so that the hyper-parameters can be tuned on one subset and used to test the other one (and vice versa).

### B. Fusion of IRISA systems

Let us denote as IRISA_F the fusion system combining the outputs of the three IRISA systems. Table VIII shows the results obtained with the oracle and the cross-validation set-ups respectively, and table XI gathers the corresponding hyper-parameters.

The top of table VIII shows that when IRISA_F is tuned optimally to maximize the average F-measure with $tol = 0.5$ s, it outperforms the best IRISA system by over 6.5 % absolute F-measure. An even higher improvement (over 7.5 %) is measured in the case of $tol = 3$ s. The optimal values related to these performances correspond to slightly convex regularity costs, with values of $\alpha$ of 1.8 for $tol = 0.5$ s and 1.2 for $tol = 3$ s.

Fig. 8 shows the effect of each hyper-parameter $\alpha$ (top-left), $\lambda$ (top-right), $w$ (bottom-left) and $\tau$ (bottom-right) on the average F-measure when all the other hyper-parameters

[21] In a neighbouring context, [77] uses Mean Mutual Agreement to select the best beat segmentation among the outputs from multiple beat trackers.

| System | tol = 0.5s | | | tol = 3s | | |
|---|---|---|---|---|---|---|
| | $F(\%)$ | $P(\%)$ | $R(\%)$ | $F(\%)$ | $P(\%)$ | $R(\%)$ |
| Oracle set-up | | | | | | |
| IRISA_F* | **38.8** | 42.1 | 36.6 | **70.3** | 71.0 | 70.9 |
| Cross-validation set-up | | | | | | |
| IRISA_F | **38.0** | 41.0 | 35.9 | **69.4** | 71.6 | 68.6 |
| IRISA systems | | | | | | |
| IRISA10 | 23.4 | 24.3 | 23.7 | 61.0 | 61.9 | 62.2 |
| IRISA11 | **32.3** | 32.6 | 33.2 | 61.2 | 62.2 | 62.3 |
| IRISA12 | 25.3 | 28.4 | 23.3 | **62.8** | 69.8 | 58.4 |

TABLE VIII
PERFORMANCE OF IRISA SYSTEM FUSION ON THE MIREX10 DATASET : FUSION WITH ORACLE SET-UP (TOP), FUSION WITH CROSS-VALIDATION (MIDDLE) AND INDIVIDUAL SYSTEMS (BOTTOM). THE CORRESPONDING PARAMETERS ARE GIVEN IN TABLE XI.

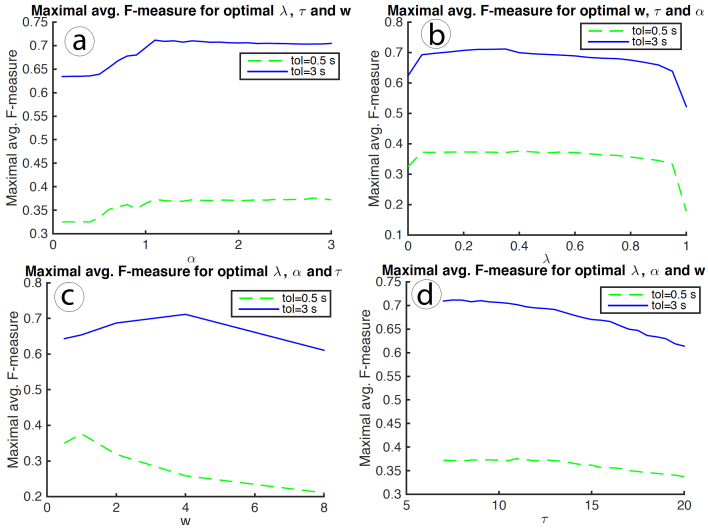



Fig. 8. IRISA_F : Evolution of the maximal average F-measure on MIREX10 : a) according to parameters $\alpha$, b) according to $\lambda$, c) according to $w$ and d) according to $\tau$.

are optimally tuned. The curves are globally smooth and show a rather flat maximum, except for $w$ (Fig. 8.c) showing a slightly distinctive peak. The curve of optimal performance as a function of $\alpha$ (Fig. 8.a) is flat for values above 1.1 around the maximum, implying that convex regularity cost functions are more efficient than non-convex ones. A flat behavior can be observed for $\lambda$ (Fig. 8.b) varying between 0.05 and 0.4, indicating that the fine tuning of $\lambda$ is not essential. The curve from Fig. 8.c shows that the analysis window length $w$ plays a significant role on performance, as curves associated to $tol = 0.5$ s and $tol = 3$ s show clear global maxima for $w = 1$ and 4 respectively. Finally, the last curve (Fig. 8.d) shows that the optimal performances are reached for values of $\tau$ between 7 and 12 s for $tol = 0.5$ s, and between 7 and 10 s for $tol = 3$ s.

In terms of performance, the same trends are observed in the cross-validation set-up, as shown in the center part of Table VIII), with a F-measure improvement of 5.8 % and 6.6 % for $tol = 0.5$ s and $tol = 3$ s respectively. Thus, the fusion of IRISA systems clearly outperforms the best IRISA systems taken individually.

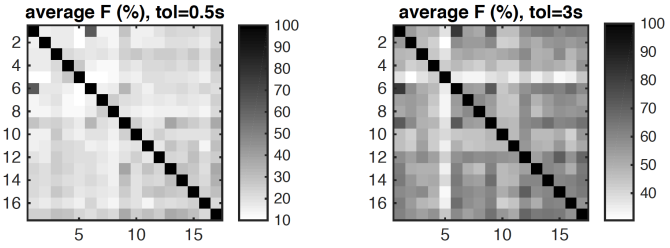



Fig. 9. Cross-comparison of the segmentation outputs from all the MIREX systems considered on the MIREX10 dataset. Left : average F-measures for $tol = 0.5$ s; right : average F-measures for $tol = 3$ s. The systems are presented from left to right and from top to bottom in the same order as in Table III.

### C. Challenging the leader : a massive fusion of MIREX systems

As a further investigation, we applied our naive fusion approach to the whole collection of MIREX systems presented in Section V-C. Fig. 9 depicts the comparison of the 17 systems, in terms of their relative average F-measures (similarly to Table VII). These matrices illustrate visually the degree of (de-)correlation of the system outputs with one another.

Two configurations are considered : FUSION_17, which combines the outputs of all 17 MIREX systems, and FUSION_16 which combines all of them except the best performing one (GS1), so as to evaluate if the current state-of-the-art method could be challenged by joining the efforts of a pool of other MIREX systems[22]. Fusion results provided in Tables IX and X indicate the following trends :

- FUSION_16 shows a noticeable improvement in fusion performance compared to the best individual system for the two tolerances, and provides, for $tol = 3$ s, a level of performance of 80.3 % (in cross-validation mode) which is competitive to that of the state-of-the-art GS1 system (not included in FUSION_16), yielding 79.3%.
- FUSION_17 (which includes GS1) improves significantly[23] the performance of GS1, from 79.3 % to 81.5 %, i.e. +2.2%, for a tolerance of 3 s, but it remains clearly below that of GS1 for $tol = 0.5$ s.

These experiments indicate that the regularity constraint is an efficient method to guide the fusion of multiple segmentation systems on the MIREX10 dataset. While fusion provides "strength in unity", the regularity constraint helps focusing the effort around time instants where structural boundaries are considered to be more likely. Considering the success of these experiments despite the very "naive" design of the tested approach, we believe these results should encourage the MIR community towards a further exploration of late fusion schemes to exploit and combine segmentation data from multiple sources.

[22]Individual system outputs were obtained on Pauwel's GitHub https://github.com/jpauwels/mirex-tools

[23]For a performance level $p = 80\%$ and for $n_{\text{test}} = 1800$ tests, the 95% prediction interval calculated as $\pm 1.96 \frac{\sqrt{p(1-p)}}{\sqrt{n_{\text{test}}}}$ is equal to $\pm 1.9\%$.

| | tol = 0.5 s | | | tol = 3 s | | |
|---|---|---|---|---|---|---|
| System | $F$ | $P$ | $R$ | $F$ | $P$ | $R$ |
| Oracle set-up | | | | | | |
| FUSION_16* | **48.5** | 54.5 | 44.2 | **80.7** | 83.8 | 79.0 |
| Cross-validation set-up | | | | | | |
| FUSION_16 | **48.2** | 55.0 | 43.4 | **80.3** | 84.0 | 78.1 |
| Best individual systems | | | | | | |
| SMGA1 | 26.8 | 28.7 | 25.8 | **76.6** | 81.6 | 73.5 |
| RBH2 | **37.5** | 39.2 | 36.8 | 67.8 | 70.0 | 66.4 |

TABLE IX
PERFORMANCE OF FUSION_16 ON MIREX10 : FUSION WITH ORACLE SET-UP (TOP), FUSION WITH CROSS-VALIDATION (MIDDLE) AND BEST INDIVIDUAL SYSTEMS IN FUSION_16 (BOTTOM). THE CORRESPONDING PARAMETERS ARE GIVEN IN TABLE XI.

| | tol = 0.5 s | | | tol = 3 s | | |
|---|---|---|---|---|---|---|
| System | $F$ | $P$ | $R$ | $F$ | $P$ | $R$ |
| Oracle set-up | | | | | | |
| FUSION_17* | **54.3** | 62.6 | 48.4 | **81.7** | 84.6 | 80.1 |
| Cross-validation set-up | | | | | | |
| FUSION_17 | **52.8** | 58.8 | 48.6 | **81.5** | 84.9 | 79.6 |
| Best individual system | | | | | | |
| GS1 | **69.7** | 80.4 | 62.7 | **79.3** | 91.9 | 71.1 |

TABLE X
PERFORMANCE OF FUSION_17 ON MIREX10 : FUSION WITH ORACLE SET-UP (TOP), FUSION WITH CROSS-VALIDATION (MIDDLE) AND STATE-OF-THE-ART GS1 SYSTEM (BOTTOM). THE CORRESPONDING PARAMETERS ARE GIVEN IN TABLE XI.

| System | $tol$ | $\tau^*$ | $w^*$ | $\alpha^*$ | $\lambda^*$ |
|---|---|---|---|---|---|
| Oracle set-ups | | | | | |
| IRISA_F* | 0.5 | 11.0 | 1.0 | 1.8 | 0.360 |
| FUSION_16* | 0.5 | 9.5 | 1.0 | 1.1 | 0.415 |
| FUSION_17* | 0.5 | 10.0 | 1.0 | 1.2 | 0.385 |
| IRISA_F* | 3 | 7.5 | 4.0 | 1.2 | 0.310 |
| FUSION_16* | 3 | 7.0 | 4.0 | 2.7 | 0.085 |
| FUSION_17* | 3 | 7.0 | 4.0 | 2.6 | 0.095 |
| Cross-validation set-ups | | | | | |
| IRISA_F (A) | 0.5 | 12.0 | 1.0 | 1.7 | 0.415 |
| IRISA_F (B) | 0.5 | 9.0 | 1.0 | 1.5 | 0.290 |
| FUSION_16 (A) | 0.5 | 8.5 | 1.0 | 1.1 | 0.400 |
| FUSION_16 (B) | 0.5 | 9.5 | 1.0 | 1.1 | 0.415 |
| FUSION_17 (A) | 0.5 | 12.0 | 1.0 | 1.1 | 0.465 |
| FUSION_17 (B) | 0.5 | 9.0 | 1.0 | 1.1 | 0.415 |
| IRISA_F (A) | 3 | 8.0 | 4.0 | 1.2 | 0.235 |
| IRISA_F (B) | 3 | 8.0 | 4.0 | 1.2 | 0.330 |
| FUSION_16 (A) | 3 | 7.0 | 4.0 | 3.0 | 0.065 |
| FUSION_16 (B) | 3 | 7.0 | 4.0 | 3.0 | 0.065 |
| FUSION_17 (A) | 3 | 7.0 | 4.0 | 2.6 | 0.085 |
| FUSION_17 (B) | 3 | 7.0 | 4.0 | 2.6 | 0.095 |

TABLE XI
OPTIMAL PARAMETERS FOR THE VARIOUS SYSTEMS EVALUATED IN THIS ARTICLE ON MIREX10. (*) : ORACLE SET-UP. (A) : PARAMETERS TUNED ON ODD NUMBERED SONGS (MIREX NUMBERING). (B) : PARAMETERS TUNED ON EVEN NUMBERED SONGS (MIREX NUMBERING).

## VIII. CONCLUSION

While the precise definition of musical structure remains a challenging research topic, this article has aimed at illustrating a number of benefits that can be gained from the explicit introduction of a regularity constraint in the specification and the implementation of the structural segmentation of musical contents. On our experimental corpus (composed of pop songs), the *a priori* concentration of segment size around a typical value appears to be a valuable criterion for constraining automatic segmentation from audio. The regularity constraint can be implemented by means of a Viterbi algorithm, and smoothly combined with a variety of other segment detection criteria (similarity, consistency, etc...) in an independent way, offering a rather generic scheme, compatible with many existing methods.

In this article, the regularity criterion remains very basic and could be refined for music genres and pieces where regularity is not as systematic as in pop music. Moreover, it must be explicitly stressed that experimental focus has been put on MIREX10 data, which have precisely been annotated with the concern of defining structural segments around a predetermined scale, and this certainly enhances the relevance of regularity in the estimation process. Future work could focus on more sophisticated models of segmental organization (in particular, hierarchical models) and investigate how they can be used efficiently to constrain and elicit adequate segmentations of music in more complex situations and at several scales simultaneously.

An additional contribution of this work is to highlight the potential benefits that structural segmentation can gain from system fusion, orchestrated here again by a regularity constraint, where different kinds of segment models and probabilistic schemes could be imagined to re-weight and combine the various hypotheses. Fusion schemes under structural constraints may also be considered for the exploitation of collaborative annotations.

## ACKNOWLEDGMENTS

This work was partly achieved within the scope of the Quaero project funded by OSEO, French State agency for innovation. The authors thank Matthew E. P. Davies and Yoshi Ueda for sharing with them their beat/downbeat and chord estimation systems which have been used in some of the IRISA systems described in this paper.

## REFERENCES

[1] F. Bimbot, E. Deruty, G. Sargent, and E. Vincent, "Semiotic structure labeling of music pieces: Concepts, methods and annotation conventions," in *Proceedings of the 13th International Society for Music Information Retrieval Conference (ISMIR)*, Oct. 2012, pp. 235–240.

[2] J. P. Bello, "Measuring structural similarity in music," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, no. 7, pp. 2013–2025, 2011.

[3] G. Peeters, A. La Burthe, and X. Rodet, "Toward automatic music audio summary generation from signal analysis," in *Proceedings of the 3rd International Conference on Music Information Retrieval (ISMIR)*, Oct. 2002, pp. 94–100.

[4] A. Liutkus, Z. Rafii, R. Badeau, B. Pardo, and G. Richard, "Adaptive filtering for music/voice separation exploiting the repeating musical structure," in *Proceedings IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Mar. 2012, pp. 53–56.

[5] M. Mauch, K. Noland, and S. Dixon, "Using musical structure to enhance automatic chord transcription," in *Proceedings of the 10th International Society for Music Information Retrieval Conference (ISMIR)*, 2009, pp. 231–236.

[6] J. B. L. Smith, "Explaining listener differences in the perception of musical structure," Ph.D. dissertation, Department of Electronic Engineering, Queen Mary, University of London, 2014.

[7] F. Bimbot, O. Le Blouch, G. Sargent, and E. Vincent, "Decomposition into autonomous and comparable blocks: A structural description of music pieces," in *Proceedings of the 11th International Society for Music Information Retrieval Conference (ISMIR)*, Aug. 2010, pp. 189–194.

[8] G. Sargent, "Estimation de la structure de morceaux de musique par analyse multi-critères et contrainte de régularité," Ph.D. dissertation, Université de Rennes 1, 2013.

[9] B. Snyder, *Music and Memory : an Introduction*. MIT Press, 2000.

[10] I. Bent and W. Drabkin, *Analysis*. Macmillian Reference Limited, 1988 (reprinted 1998).

[11] E. Peiszer, "Automatic audio segmentation : segment boundary and structure detection in popular music," Master's thesis, Vienna University of Technology, Austria, August 2007.

[12] N. Ruwet, "Methods of analysis in musicology," *Music Analysis*, vol. 6, no. 1/2, pp. 3–9+11–36, 1987.

[13] R. Middleton, *Studying Popular Music*. Open University Press, 1990.

[14] G. Peeters and E. Deruty, "Is music structure annotation multi-dimensional? A proposal for robust local music annotation." in *Proceedings of the 3rd International Workshop on Learning Semantics of Audio Signals (LSAS)*, December 2009, pp. 75–90.

[15] J. B. L. Smith, J. A. Burgoyne, I. Fujinaga, D. De Roure, and J. S. Downie, "Design and creation of a large-scale database of structural annotations," in *Proceedings of the 12th International Society for Music Information Retrieval Conference (ISMIR)*, October 2011, pp. 555–560.

[16] M. J. Bruderer, M. McKinney, and A. Kohlrausch, "Structural boundary perception in popular music," in *Proceedings of the 7th International Conference on Music Information Retrieval (ISMIR)*, October 2006, pp. 198–201.

[17] F. Bimbot, E. Deruty, G. Sargent, and E. Vincent, "System & contrast : A polymorphous model of the inner organization of structural segments within music pieces," *Music Perception*, vol. 32, no. 5, pp. 631–661, 2016, this paper extends former work initially reported in IRISA PI-1999 [Research Report], December 2012, 40 pages, hal-00868398.

[18] F. Bimbot, G. Sargent, E. Deruty, C. Guichaoua, and E. Vincent, "Semiotic Description of Music Structure: An Introduction to the Quaero/METISS Structural Annotations," in *Proceedings of the 53rd AES International Conference on Semantic Audio, article number P1-1*, 2014.

[19] R. Dannenberg and M. Goto, "Music structure analysis from acoustic signals," in *Handbook of Signal Processing in Acoustics*, D. Havelock, S. Kuwano, and M. Vrlander, Eds. Springer, 2008, vol. 1, pp. 305–331.

[20] 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)*, August 2010, pp. 625–636.

[21] M. Müller, *Fundamentals of Music Processing: Audio, Analysis, Algorithms, Applications*. Springer, 2015.

[22] F. Kaiser and T. Sikora, "Music structure discovery in popular music using non-negative matrix factorization," *Proceedings of the 11th International Society on Music Information Retrieval (ISMIR)*, pp. 429–434, October 2010.

[23] M. A. Bartsch and G. H. Wakefield, "Audio thumbnailing of popular music using chroma-based representations," in *IEEE Transactions on multimedia*, vol. 7, no. 1, February 2005, pp. 96–104.

[24] K. Patil, D. Pressnitzer, S. Shamma, and M. Elhilali, "Music in our ears: The biological bases of musical timbre perception," *PLoS Computational Biology*, vol. 8, no. 11, pp. 1–16, november 2012.

[25] G. Peeters, "Deriving musical structures from signal analysis for music audio summary generation: Sequence and state approach," *Lectures Notes in Computer Science*, vol. 2771/2004, pp. 1–25, 2004.

[26] J. Paulus and A. Klapuri, "Music structure analysis using a probabilistic fitness measure and an integrated musicological model," in *Proceedings of the 9th International Conference on Music Information Retrieval (ISMIR)*, September 2008, pp. 369–374.

[27] M. A. Bartsch and G. H. Wakefield, "To catch a chorus : Chroma-based representations for audio thumbnailing," in *Proceedings of the 2001 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, October 2001, pp. 15–18.

[28] E. Gómez, "Tonal description of music audio signals," Ph.D. dissertation, Universitat Pompeu Fabra, Barcelona, 2006.

[29] M. Müller and S. Ewert, "Chroma Toolbox: MATLAB implementations for extracting variants of chroma-based audio features," in *Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR)*, October 2011, pp. 215–220.

[30] T. Jehan, "Creating music by listening," Ph.D. dissertation, Massachusetts Institute of Technology, September 2005.

[31] K. Jensen, "Multiple scale music segmentation using rhythm, timbre and harmony," *EURASIP Journal on Advances in Signal Processing*, vol. 2007, pp. 1–11, 2007.

[32] J. Paulus and A. Klapuri, "Acoustic features for music piece structure analysis," in *In proceedings of the 11th International Conference on Digital Audio Effects (DAFx-08)*, September 2008, pp. 309–312.

[33] M. Tian, G. Fazekas, D. A. A. Black, and M. Sandler, "On the use of the tempogram to describe audio content and its application to music structural segmentation," in *Proceedings of the 2015 International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2015, pp. 419–423.

[34] J. Serrà, M. Müller, P. Grosche, and J. L. Arcos, "Unsupervised music structure annotation by time series structure features and segment similarity," *IEEE Transactions on Multimedia, Special Issue on Music Data Mining*, vol. 16, no. 5, pp. 1229–1240, August 2014.

[35] J. T. Foote and M. L. Cooper, "Media segmentation using self-similarity decomposition," in *Proceedings of the SPIE Storage and Retrieval for Multimedia Databases*, Jan. 2003, pp. 167–175.

[36] O. Nieto and T. Jehan, "Convex non-negative matrix factorization for automatic music structure identification," in *Proceedings of the 2013 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2013, pp. 236–240.

[37] R. Chen and M. Li, "Music structural segmentation by combining harmonic and timbral information," in *Proceedings of the 12th International Society for Music Information Retrieval Conference (ISMIR)*, Oct 2011, pp. 477–482.

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

[39] J.-J. Aucouturier and M. Sandler, "Finding repeating patterns in acoustic audio signals: applications for audio thumbnailing," in *Proceedings of the Audio Engineering Society (AES) 22nd International Conference of Virtual, Synthetic and Entertainment Audio*, 2002, pp. 412–421.

[40] R. Weiss and J. Bello, "Identifying repeated patterns in music using sparse convolutive non-negative matrix factorization," in *Proceedings of the 11th International Society on Music Information Retrieval (ISMIR)*, October 2010, pp. 123–128.

[41] M. Goto, "A chorus-section detecting method for musical audio signals," in *Proceedings of the 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, April 2003, pp. 437–440.

[42] L. Lu, M. Wang, and H.-J. Zhang, "Repeating pattern discovery and structure analysis from acoustic music data," in *Proceedings of the 2004 Multimedia Information Retrieval Workshop*, Oct. 2004, pp. 275–282.

[43] F. Kaiser and G. Peeters, "A simple fusion method of state and sequence segmentation for music structure discovery," *Proceedings of the 14th International Society on Music Information Retrieval (ISMIR)*, pp. 257–262, October 2013.

[44] G. Sargent, F. Bimbot, and E. Vincent, "A regularity-constrained Viterbi algorithm and its application to the structural segmentation of songs," in *Proceedings of the 12th International Society for Music Information Retrieval Conference (ISMIR)*, Oct. 2011, pp. 483–488.

[45] B. McFee and D. P. Ellis, "Analyzing song structure with spectral clustering," in *Proceedings of the 15th International Society for Music Information Retrieval Conference (ISMIR)*, 2014, pp. 405–410.

[46] ——, "DP1, MP1, MP2 entries for MIREX 2013 structural segmentation and beat tracking," in *Proceedings of the Music Information Retrieval Evaluation eXchange (MIREX)*, 2013.

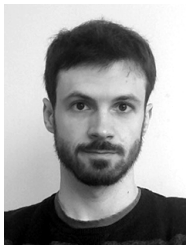
[47] T. Grill and J. Schlüter, "Music boundary detection using neural networks on combined features and two-level annotations," in *Proceedings of the 16th International Society on Music Information Retrieval (ISMIR)*, October 2015, pp. 531–537.

[48] F. Kaiser and G. Peeters, "Multiple hypothesis at multiple scales for audio novelty computation within music," *Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 23–235, May 2013.

[49] M. Levy and M. Sandler, "New methods in structural segmentation of musical audio," in *Proceedings of European Signal Processing Conference (EUSIPCO)*, September 2006, pp. 318–326.

[50] M. E. P. Davies, P. Hamel, Y. Kazuyoshi, and M. Goto, "AutoMashUpper: Automatic creation of multi-song music mashups," *IEEE/ACM*

*Transactions on Audio, Speech, and Language Processing (TASLP)*, vol. 22, no. 12, pp. 1726–1737, December 2014.
[51] J.-J. Aucouturier and M. Sandler, "Segmentation of musical signals using hidden Markov models," in *Proceedings of the Audio Engineering Society (AES) 110th Convention*, May 2001, 8 pages.
[52] Y. Shiu, H. Jeong, and C.-C. Jay Kuo, "Similarity matrix processing for music structure analysis," in *Proceedings of the 14th ACM International Conference on Multimedia*, Oct. 2006, pp. 69–76.
[53] J. S. Downie, K. West, A. Ehmann, and E. Vincent, "The 2005 Music Information Retrieval Evaluation eXchange (MIREX 2005): Preliminary overview," in *Proceedings of the 6th International Conference on Music Information Retrieval (ISMIR)*, September 2005, pp. 320–323.
[54] J. S. Downie, "The music information retrieval evaluation exchange (2005-2007): A window into music information retrieval," *Acoustical Science and Technology*, vol. 29, no. 4, pp. 247–255, 2008.
[55] A. Ehmann, M. Bay, J. S. Downie, I. Fujinaga, and D. De Roure, "Music structure segmentation algorithm evaluation : expanding on MIREX 2010 analyses and datasets," in *Proceedings of the 12th International Society for Music Information Retrieval Conference (ISMIR)*, Miami, United States, Oct. 2011, pp. 561–566.
[56] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka, "RWC music database: Popular, classical, and jazz music databases," in *Proceedings of the 3rd International Conference on Music Information Retrieval (ISMIR)*, October 2002, pp. 287–288.
[57] M. Goto, "AIST Annotation for the RWC Music Database," in *Proceedings of the 7th International Conference on Music Information Retrieval (ISMIR)*, October 2006.
[58] D. Turnbull, G. Lanckriet, E. Pampalk, and M. Goto, "A supervised approach for detecting boundaries in music using difference features and boosting," *Proceedings of the 8th International Conference on Music Information Retrieval (ISMIR)*, pp. 51–54, 2007.
[59] G. Sargent, F. Bimbot, and E. Vincent, "Supplementary material to the article: Estimating the structural segmentation of popular music pieces under regularity constraints," 2016. [Online]. Available: https://hal.inria.fr/hal-01368683
[60] C. Cannam, E. Benetos, M. Mauch, M. E. P. Davies, S. Dixon, C. Landone, K. Noland, and D. Stowell, "MIREX 2015: VAMP Plugins from the Centre for Digital Music," in *Proceedings of the Music Information Retrieval Evaluation eXchange (MIREX)*, October 2015.
[61] G. Sargent, F. Bimbot, and E. Vincent, "A music structure inference algorithm based on symbolic data analysis," in *Proceedings of the Music Information Retrieval Evaluation eXchange (MIREX)*, october 2010.
[62] O. Nieto and J. P. Bello, "MIREX 2014 entry: 2D Fourier magnitude coefficients," in *Proceedings of the Music Information Retrieval Evaluation eXchange (MIREX)*, 2014.
[63] G. Sargent, F. Bimbot, and E. Vincent, "A music structure inference algorithm based on morphological analysis," in *Proceedings of the Music Information Retrieval Evaluation eXchange (MIREX)*, october 2010.
[64] G. Peeters, "MIREX 2010 music structure segmentation task : IRCAMsummary submission," in *Proceedings of the Music Information Retrieval Evaluation eXchange (MIREX)*, Oct. 2010.
[65] F. Kaiser, T. Sikora, and G. Peeters, "MIREX 2012 - Music structural segmentation task : IRCAMstructure submission," in *Proceedings of the Music Information Retrieval Evaluation eXchange (MIREX)*, 2012.
[66] B. Rocha, N. Bogaards, and A. Honingh, "Segmentation and timbre similarity in electronic dance music," in *Proceedings of the Sound and Music Computing Conference (SMC 2013)*, 2013, pp. 754–761.
[67] B. Martin, P. Hanna, M. Robine, and P. Ferraro, "Structural analysis of harmonic features using string matching techniques," in *Proceedings of the Music Information Retrieval Evaluation eXchange (MIREX)*, 2011.
[68] J. Serrà, M. Müller, P. Grosche, and J. L. Arcos, "The importance of detecting boundaries in music structure annotation," in *Proceedings of the Music Information Retrieval Evaluation eXchange (MIREX)*, oct 2012.
[69] T. Grill and J. Schlüter, "Structural segmentation with convolutional neural networks MIREX submission," in *Proceedings of the Music Information Retrieval Evaluation eXchange (MIREX)*, October 2015.
[70] G. Sargent, F. Bimbot, and E. Vincent, "A structural segmentation of songs using generalized likelihood ratio under regularity assumptions," in *Proceedings of the Music Information Retrieval Evaluation eXchange (MIREX)*, october 2010.
[71] E. Pampalk, "Ma toolbox (last accessed: Sept. 2016)," 2007. [Online]. Available: http://www.pampalk.at/ma/
[72] D. P. Ellis, "Music beat tracking and cover song identification (last accessed: Sept. 2016)," 2007. [Online]. Available: http://labrosa.ee.columbia.edu/projects/coversongs/
[73] Y. Ueda, Y. Uchiyama, T. Nishimoto, N. Ono, and S. Sagayama, "HMM-based approach for automatic chord detection using refined acoustic features," in *Proceedings of the 2010 International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2010, pp. 5506–5509.
[74] M. E. P. Davies, "Towards automatic rhythmic accompaniment," Ph.D. dissertation, Department of Electronic Engineering, Queen Mary, University of London, 2007.
[75] A. M. Stark, M. E. P. Davies, and M. D. Plumbley, "Real-time beat-synchronous analysis of musical audio," in *Proceedings of the 12th International Conference on Digital Audio Effects (DAFx)*, September 2009, pp. 299–304.
[76] J. B. L. Smith and M. Goto, "Using priors to improve estimates of music structure," in *Proceedings of the 17th International Society for Music Information Retrieval Conference (ISMIR)*, October 2016, pp. 554–560.
[77] J. R. Zapata, A. Holzapfel, M. E. P. Davies, J. L. Oliveira, and F. Gouyon, "Assigning a confidence threshold on automatic beat annotation in large datasets," in *Proceedings of the 6th International Society for Music Information Retrieval Conference (ISMIR)*, October 2012, pp. 157–162.



**Gabriel Sargent** received the state engineering degree from ENSEA Cergy-Pontoise, France, in 2009, and a Ph.D. degree in Signal Processing and Telecommunications from the University of Rennes 1, France, in 2013. From 2013 to 2015, he collaborated with LaBRI, Talence, France and CNAM-CEDRIC, Paris, France, in the context of the international project MEX-CULTURE. He currently works on the analysis and linking of multimedia documents with the Linkmedia team in IRISA, Rennes, France, in the context the NexGenTV project and the A||GO web-platform. His research interests include music and multimedia documents indexing, segmentation and classification.

**Frédéric Bimbot** After graduating as a telecommunication engineer in 1985 (ENST, Paris, France), Frédéric BIMBOT received in 1988 a PhD in Signal Processing, on the topic of speech synthesis. He also obtained in 1987 a B.A. in Linguistics (Sorbonne Nouvelle University, Paris III).
In 1990, he joined CNRS (French National Center for Scientific Research) as a permanent researcher and worked with Telecom Paris Tech for 7 years. He then moved to IRISA, a joint laboratory between CNRS, INRIA and the University of Rennes 1. He also repeatedly visited AT&T – Bell Laboratories between 1990 and 1999. He is now a Senior Researcher with CNRS. His research is dedicated to speech and audio analysis, speaker recognition, audio source separation and music content modeling, with a particular focus on music structure. He is also the Head of the "Digital Signals and Images, Robotics" Department at IRISA.

**Emmanuel Vincent** is a Research Scientist with Inria (Nancy, France). He received the Ph.D. degree in music signal processing from the Institut de Recherche et Coordination Acoustique/Musique (Paris, France) in 2004 and worked as a Research Assistant with the Centre for Digital Music at Queen Mary, University of London (United Kingdom), from 2004 to 2006. His research focuses on probabilistic machine learning for speech and audio signal processing, with application to real-world audio source localization and separation, noise-robust speech recognition, and music information retrieval. He is a founder of the series of Signal Separation Evaluation Campaigns and CHiME Speech Separation and Recognition Challenges. He was an associate editor for IEEE Transactions on Audio, Speech, and Language Processing.