Unsupervised Music Structure Annotation by Time Series Structure Features and Segment Similarity

Joan Serrà, Meinard Müller, *Member, IEEE,* Peter Grosche, and Josep Ll. Arcos.

***Abstract*—Automatically inferring the structural properties of raw multimedia documents is essential in today's digitized society. Given its hierarchical and multi-faceted organization, musical pieces represent a challenge for current computational systems. In this article, we present a novel approach to music structure annotation based on the combination of structure features with time series similarity. Structure features encapsulate both local and global properties of a time series, and allow us to detect boundaries between homogeneous, novel, or repeated segments. Time series similarity is used to identify equivalent segments, corresponding to musically meaningful parts. Extensive tests with a total of five benchmark music collections and seven different human annotations show that the proposed approach is robust to different ground truth choices and parameter settings. Moreover, we see that it outperforms previous approaches evaluated under the same framework.**

I. Introduction

INFORMATION is very often organized into structures or hierarchies that facilitate its transmission and understanding. In general, humans are very good at identifying such structures, a process that is sometimes unconscious and that allows us to parse and adequately grasp the meaning of a given message. However, considering the vast amount of digital information available nowadays, we progressively need more and more support from machines to select and digest relevant information. Hence, automatically determining the structure of such information becomes a crucial task for current content-processing systems. Among general multimedia contents, music is a paradigmatic and challenging example [1], [2].

Music presents us with a multi-layer and multi-faceted structural organization of its most basic constitutive elements [3], [4]. In Western music, the main high-level structural organization of a piece is the musical form [5], which describes the layout of a composition as divided into sections, segments, or parts (we loosely employ the term *music structure* to refer to this high-level structure). For instance, in popular music, compositions are usually divided into segments that alternate or repeat throughout the piece, commonly called ‘intro’, ‘verse’, ‘chorus’, and ‘bridge’. Structural information of this type can be easily understood by an average music listener [6] and, moreover, allows for several novel applications. Examples include easier intra-piece navigation in music players, automatic generation of representative clips and mash-ups, identification of versions of the same piece, or large-scale musicological research (see [1], [2], [7], [8] for pointers to these and other applications).

According to Paulus et al. [2], there are three basic principles for inferring music structure: *novelty*, *homogeneity*, and *repetition*. This way, existing approaches can be classified into three main categories. Novelty-based approaches aim at detecting transitions between contrasting parts. This is usually done by detecting local differences (or contrasts) based on a moving-window analysis of a suitable feature representation of the music recording. For example, in the seminal work by Foote [9], a short-time checkerboard kernel is moved over the diagonal of a self-similarity matrix to detect corners that correspond to points of novelty. Homogeneity-based approaches, on the other hand, aim at detecting passages that are consistent with respect to some musical property such as rhythm, timbre or harmony [10]. Approaches based on this principle usually employ more refined techniques such as hidden Markov models [11] or dynamic texture mixtures [12]. Finally, repetition-based approaches aim at identifying recurring patterns that often closely correlate to the structure of the piece [7], [13]–[15]. Few studies combine different principles within a single framework such as the approach by Paulus & Klapuri [8], where homogeneity and repetition properties are captured by a single probabilistic fitness measure. Foote's approach [9] can also be seen as combining novelty and homogeneity. Finally, we want to note that Peeters [16], [17] has already classified structure analysis approaches from a more technical point of view, where approaches based on homogeneity are coined as “state approaches” and approaches based on repetition as “sequence approaches”. For in-depth reviews on music structure annotation approaches we refer to [1] and [2].

In this work we are interested in automatically annotating the structure of a musical piece in an unsupervised way, i.e., without employing explicit knowledge of previously annotated pieces. Our goal is to detect the temporal locations of segment boundaries and to assess segment similarities and repetitions within a single piece. Some approaches go one step beyond and try to assign semantic labels to the segments [8], [18] (e.g., trying to guess whether a given segment is a ‘chorus’ or just an ‘intro’). We believe that this difficult final step needs to be grounded on the reliable assessment of

The work of J. Serrà and J. Ll. Arcos has been supported by ICT-2011-8-318770 and 2009-SGR-1434. J. Serrà also acknowledges JAEDOC069/2010. The work of M. Müller and P. Grosche has been supported by the Cluster of Excellence on Multimodal Computing and Interaction at Saarland University and DFG MU 2682/5-1.

J. Serrà and J. Ll. Arcos are with IIIA-CSIC (Campus de la UAB s/n, 08193 Bellaterra, Spain, jserra@iiia.csic.es; arcos@iiia.csic.es). M. Müller is with the International Audio Laboratories Erlangen (Am Wolfsmantel 33, 91058 Erlangen, Germany, meinard.mueller@audiolabs-erlangen.de). P. Grosche is with Saarland University and the Max-Planck Institut für Informatik (Campus E1.4, 66123 Saarbrücken, Germany, pgrosche@mpi-inf.mpg.de).

segment boundaries and similarities, two tasks where current approaches still offer much room for improvement.

Here we exploit a novel class of *structure features* [19] on the basis of which various structure analysis principles can be integrated within a unifying framework. The basic conceptual idea behind structure features is to jointly consider local and global aspects by measuring, for each frame (or window) of a given time series, the relations to all other frames (or windows) of the same time series. This yields a frame-wise, i.e., *local*, feature representation that captures *global* structural characteristics of a time series. The resulting structure features can then be used in combination with standard novelty detection procedures. Note that novelty detection is usually performed on the basis of features that capture local characteristics of the given music signal (e.g., MFCC or chroma features, which capture local characteristics related to timbre or harmony, respectively [20]). Then, applying a local kernel or a derivative function on such feature representations often results in rather noisy novelty curves, making novelty detection a fragile and problematic step. In contrast, our approach goes beyond local musical aspects such as harmony or timbre to incorporate global structural properties of such aspects. This makes the subsequent novelty detection step much more robust and leads, by itself, to structural meaningful segment boundaries. Beyond structure features, we show how to exploit the obtained segment boundaries for music structure labeling. We introduce the use of time series similarity measures for such a task, a dynamic thresholding operation, and a novel and simple approach to force segment transitivity (see below).

Our structure annotation procedure works as follows (Fig. 1). In a first stage, the music recording is converted into a descriptor sequence. As many structure annotation approaches, we employ state-of-the-art descriptors representing tonal/harmonic information [2]. In a second stage, descriptor sequences are transformed into time series of structure features, from which a novelty curve is computed by considering a local derivative. The peak positions of this novelty curve are taken to define segment boundaries. In a third stage, the resulting segments are compared in a pairwise fashion, using a standard time series similarity measure, and divided into groups, with each group containing all segments that are considered repetitions of each other. This procedure resembles, e.g., [8], which also consists of a novelty detection and a classification/grouping stage. However, there is a crucial difference. In [8], the novelty detection stage is used as a mere pre-processing step to cut down the number of boundary candidates, typically including a large number of false positives. Then, in a second phase, an elaborate and computationally-expensive optimization procedure is used to derive segment boundaries and labels. In contrast to this and other approaches, we regard the novelty and boundary detection as a key stage of our approach, with the second classification/grouping stage heavily relying on the obtained boundaries. Indeed, if good segment boundaries are available, one can employ a simple time series alignment/similarity measure to finally annotate the structure of the piece. To obtain reliable segment boundaries in the first stage, our structure features constitute a major ingredient.

The proposed structure annotation approach presents several benefits. First, extensive experiments based on a number of benchmark data sets and evaluation measures show that the proposed approach outperforms previous approaches as reported in the literature. This claim is further supported by the out-of-sample results obtained at the 2012 edition of the Music Information Retrieval Evaluation eXchange[1] (MIREX), an international evaluation campaign for music information retrieval algorithms [21]. Second, the proposed approach is simple and also computationally efficient. No complex machine learning tasks nor expensive optimizations are performed in any of the two stages outlined above (see also Fig. 1). Third, the parameters of the approach are easy to understand and intuitive to set, provided some basic knowledge of the data at hand. A further quality of the proposed approach is that its formulation is generic, in the sense that it does not exploit specific musical knowledge (cf. [19]).

The remainder of the paper is organized as follows. Sec. II explains our method to structure annotation. Sec. III details our evaluation methodology, including the music collections and evaluation measures used. Sec. IV reports and discusses the obtained results, including an extensive assessment of the impact of different parameter choices. Sec. V contains the conclusions of the paper.

II. Proposed Method

A. Music Descriptor Time Series

Before detecting segment boundaries and similarities we need to transform the audio signal into a feature representation that captures musically relevant information (leftmost block, Fig. 1). For that we use pitch class profile (PCP) features, also called chroma features [22], [23]. PCP features are relevant for many music retrieval tasks and, in particular, have been extensively used for music structure annotation [2]. They are usually computed using a moving window, yielding a multi-dimensional time series that captures the harmonic content of the audio signal.

PCPs are derived from the frequency-dependent energy in a given range of the spectrum, e.g., between 100 and 3000 Hz. This energy is usually mapped into an octave-independent histogram representing the relative intensity of each of the 12 semitones of the Western chromatic scale (12 pitch classes: C, C#, D, D#, etc.). To normalize with respect to loudness, this histogram can be divided by its maximum value, thus leading to values between 0 and 1. In general, PCPs are robust against non-tonal components (e.g., ambient noise or percussive sounds) and independent of timbre and the specific instruments used [22], [23].

In this work we use HPCPs [22], an enhanced version of PCPs that considers the presence of harmonic frequencies (Fig. 2, top). In addition, HPCPs reduce the influence of noisy spectral components and are tuning-independent. We employ the same implementation and parameters as in [24], [25] with 12 pitch classes, a window length of 209 ms, and a hop size of 139 ms. Although we choose an enhanced version

[1] http://www.music-ir.org/mirex/wiki/MIREX_HOME

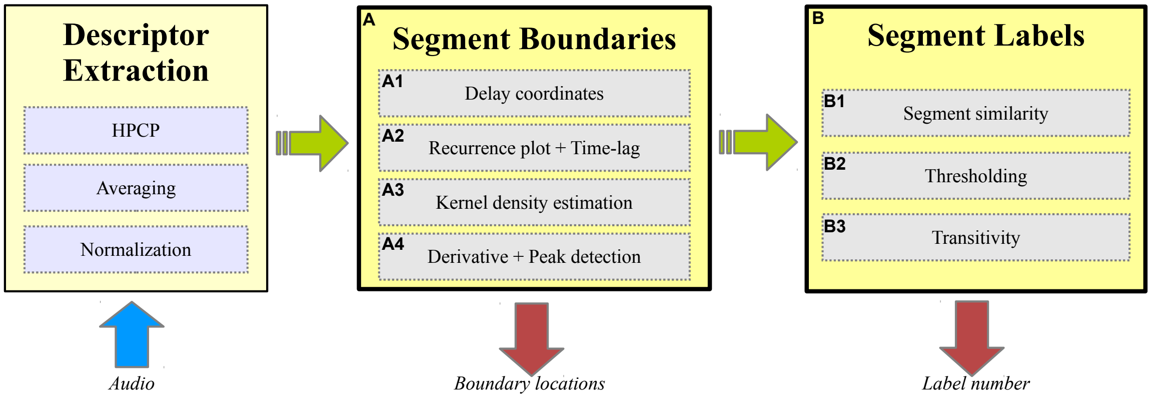

Fig. 1. Block diagram of the proposed method. The input is the audio signal of a musical piece and the output is a set of segment boundaries and labels.

of PCP features, we conjecture that the proposed approach is rather independent of specific feature implementation details. In fact, in preliminary analysis we also reached relatively good accuracies with so-called CENS chroma features [23] or even with Mel-frequency cepstral coefficients [20].

B. Segment Boundaries

Let $X = [\mathbf{x}_1, \dots \mathbf{x}_{N'}]$ be a time series of length N', with potentially multi-dimensional samples $\mathbf{x}_i$ (column vectors; in our case the aforementioned HPCP descriptor values for a given frame). We first improve the information contained in a time series sample $\mathbf{x}_i$ by incorporating information of its most recent past (A1, Fig. 1). This can be easily and elegantly done by using delay coordinates, a technique routinely employed in nonlinear time series analysis [26]. New samples are constructed by vector concatenation as

$$\hat{\mathbf{x}}_i = \left[\mathbf{x}_i^{\mathrm{T}} \quad \mathbf{x}_{i-\tau}^{\mathrm{T}} \quad \cdots \quad \mathbf{x}_{i-(m-1)\tau}^{\mathrm{T}}\right]^{\mathrm{T}}, \tag{1}$$

where $^{\mathrm{T}}$ denotes vector transposition, m is the so-called embedding dimension, and τ is a time delay.

Although there are recipes to estimate the optimal values of m and τ from the information contained in a time series X, we here leave them as parameters (see below). Note that the value of m indicates the amount of past information being considered for the task, which ranges a total time span of $w = (m-1)\tau$. By applying Eq. 1 for $i = w+1, \dots, N'$ we obtain a multi-dimensional time series $\hat{X} = [\hat{\mathbf{x}}_1, \dots, \hat{\mathbf{x}}_N]$ of length $N = N' - w$.

The next step consists in assessing homogeneities and repetitions (A2, Fig. 1). For that we compute a recurrence plot [27], which consists of a square matrix R whose elements $R_{i,j}$ indicate pairwise resemblance between samples at times i and j (Fig. 2, second row). Formally, for $i, j = 1, \dots, N$, we set

$$R_{i,j} = \Theta\left(\varepsilon_{i,j} - \|\hat{\mathbf{x}}_i - \hat{\mathbf{x}}_j\|\right), \tag{2}$$

where $\Theta(z)$ is the Heaviside step function (yielding 1 if $z > 0$ and 0 otherwise), $\|\ \|$ can be any norm (we use the Euclidean norm), and $\varepsilon_{i,j}$ are suitable thresholds.

As done in [25], the threshold $\varepsilon_{i,j}$ for each cell (i, j) is dynamically computed as follows. First, for each sample $\hat{\mathbf{x}}_i$, $i = 1, \dots, N$, we search for its K nearest neighbors in $\hat{\mathbf{x}}_j$, $j = 1, \dots, N$. Then, neighbor mutuality is enforced by setting $R_{i,j} = 1$ only if $\hat{\mathbf{x}}_i$ is a neighbor of $\hat{\mathbf{x}}_j$ and, at the same time, $\hat{\mathbf{x}}_j$ is a neighbor of $\hat{\mathbf{x}}_i$. In our experience with recurrence plots we found this restrictive strategy to be more robust against noise than other variants outlined in [27]. To account for time series of different lengths, we set $K = \kappa N$, i.e., we set the number of nearest neighbors to a fraction $\kappa \in [0, 1]$ of the length of the time series being considered.

The subsequent steps involve the creation of structure features [19] (SF). We first represent the homogeneities and recurrences of R in a circular time-lag matrix L (Fig. 2, third row). Such process is similar to the typical process of constructing a time-lag matrix [7], but incorporates the information of future as well as past time lags. We do it by circularly shifting the rows of R such that

$$L_{i,j} = R_{k+1,j} \tag{3}$$

for $i, j = 1, \dots, N$, where k equals to $i + j - 2$ modulo N.

The circular time-lag matrix L can then be considered as a sample from a bi-variate distribution $\bar{P}$ along the lag and time axes (i and j axes, respectively). This bi-variate distribution would correspond to a probability mass function of time-lag recurrences[2]. The estimate P of the underlying distribution $\bar{P}$ is obtained using bi-variate kernel density estimation [28], a fundamental data smoothing concept where inferences about a population are made based on a finite sample of it (A3, Fig. 1). In our case, P is estimated by convolving L with a bi-variate rectangular Gaussian kernel G:

$$P = L * G. \tag{4}$$

The kernel G is obtained by multiplying two Gaussian windows $\mathbf{g}_{\mathrm{l}}$ and $\mathbf{g}_{\mathrm{t}}$, with sizes s_{l} and s_{t}, corresponding to the lag and time dimensions of L, respectively. This way, G has s_{l} rows and s_{t} columns:

$$G = \mathbf{g}_{\mathrm{l}}\ \mathbf{g}_{\mathrm{t}}^{\mathrm{T}}. \tag{5}$$

The estimated kernel density P can be seen as a time series along the time axis (Fig. 2, fourth row). Structure features

[2] Actually, it is not needed that the values of $\bar{P}$ sum to 1, as normalization only introduces a scale parameter that is eliminated in a subsequent operation.

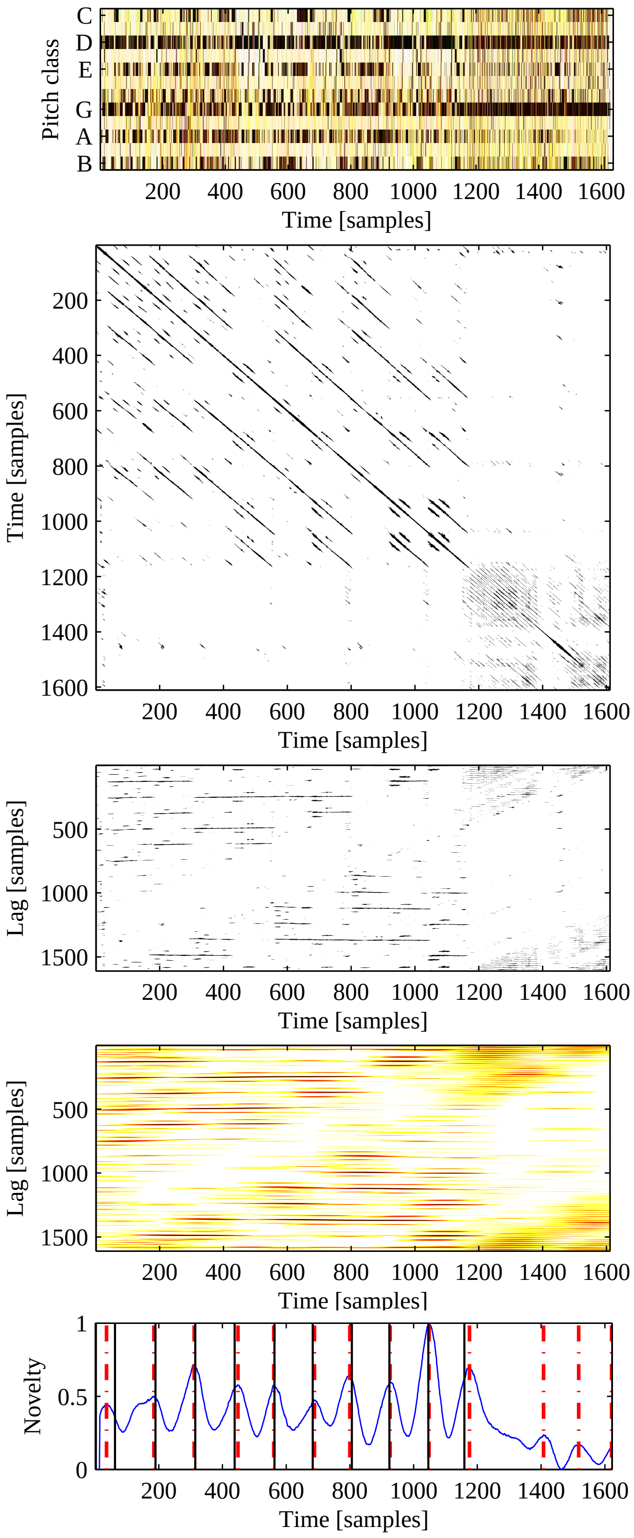

Fig. 2. Illustration of the structure feature computation using "All you need is love" from The Beatles. From top to bottom we show the resulting HPCP time series X, recurrence plot R, time-lag matrix L, structure features P, and novelty curve $\mathbf{c}$. The novelty curve $\mathbf{c}$ also depicts the found boundaries (red dash-dotted lines) and the ground truth boundaries (black solid lines).

$\mathbf{p}_i$ are then defined to be the columns of P, i.e., $P = [\mathbf{p}_1, \ldots, \mathbf{p}_N]$, where $\mathbf{p}_i$ are column vectors. Because of the nature of the recurrence plot, they theoretically encapsulate both homogeneities and repetitions (cf. [19], [27]). Furthermore, by employing a Gaussian kernel, they gain robustness against lag and time deviations, and transitions between them become smooth.

Next, our observation is that structural boundaries of the time series $\hat{X}$ correspond to relative changes in the sequence of structure features P (A4, Fig. 1). To measure these changes we compute the difference between successive structure features $\mathbf{p}_i$. This yields a one-dimensional novelty curve $\mathbf{c} = [c_1, \ldots, c_{N-1}]$, where $c_i = \|\mathbf{p}_{i+1} - \mathbf{p}_i\|$ (we again use the Euclidean norm). The novelty curve $\mathbf{c}$ can be linearly normalized between 0 and 1 by subtracting its minimum value and subsequently dividing it by the resultant maximum (Fig. 2, bottom).

Positions of prominent peaks of $\mathbf{c}$ are finally selected as segment boundaries. Here, we opt for a rather simple peak selection strategy: a sample c_i is considered to be a peak if it is above a certain threshold δ and, at the same time, corresponds to the global maximum of a window of length λ centered at c_i. To compensate for the offset introduced by delay coordinates in Eq. 1, the exact boundary locations of the original time series X are set to the locations of the selected peaks plus $w/2$.

Peak selection yields a set of boundary time stamps which, sorted in increasing order, are denoted by

$$\mathbf{b} = [b_1, \ldots, b_M, b_{M+1}, b_{M+2}] , \tag{6}$$

where M is the total number of peaks found in $\mathbf{c}$ and b_{u+1}, $u = 1, \ldots, M$, represents the location of the u-th peak. For notational convenience we include $b_1 = 1$ and $b_{M+2} = N'$, denoting the start and end of the time series, respectively (M boundaries in $\mathbf{c}$ imply $M+1$ segments, which are represented by $M+2$ boundaries). Notice that, this way, boundaries b_u and b_{u+1} correspond to the beginning and end of the u-th segment.

C. *Segment Repetitions*

Once we obtain estimates for segment boundaries we can employ current time series similarity algorithms for computing segment-segment similarities (B1, Fig. 1). In our case we find it convenient to use the Q_{max} measure presented in [25]. First, it is a generic and configurable time series similarity measure. Second, it has been shown to perform well with PCP time series, outperforming other alignment-based measures in certain tasks (compare e.g., [24] and [25]). Third, it exploits the information contained in the traces of a recurrence plot. Therefore, we can directly reuse R (Eq. 2) as input for the Q_{max} measure.

Given two segments u and v, we express the slice of the recurrence plot R comparing u and v as

$$R^{(uv)} = \begin{bmatrix} R_{b_u,b_v} & \cdots & R_{b_u,b_{v+1}} \\ \vdots & \ddots & \vdots \\ R_{b_{u+1},b_v} & \cdots & R_{b_{u+1},b_{v+1}} \end{bmatrix} , \tag{7}$$

where b_u and b_{u+1} are the limits of segment u, and b_v and b_{v+1} are the limits of segment v (Fig. 3, top). We also define the size of $R^{(uv)}$ as $l_u \times l_v$, with $l_u = b_{u+1} - b_u + 1$ and

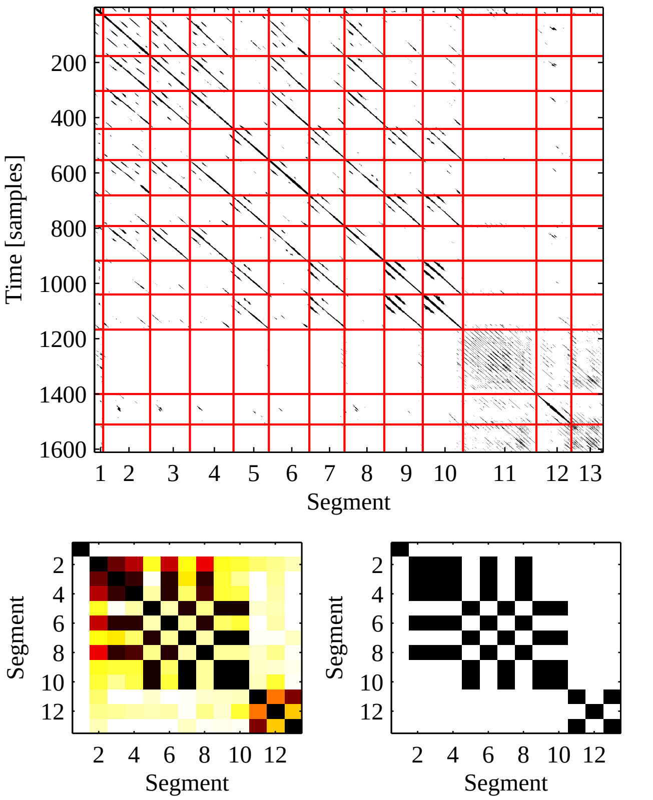

Fig. 3. Example with “All you need is love” from The Beatles. From top to bottom and left to right the plots show the recurrence plot R with the corresponding boundary marks (axes labels and ticks are interchangeable), the segment similarity matrix S, and the transitive binary similarity matrix $\hat{S}$.

$l_v = b_{v+1} - b_v + 1$ being the lengths of segments u and v, respectively. We then build a cumulative matrix $Q^{(uv)}$ of the same size by initially setting its cells to 0 and recursively applying

$$Q^{(uv)}_{i,j} = \max\left\{Q^{(uv)}_{i-1,j-1}, Q^{(uv)}_{i-2,j-1}, Q^{(uv)}_{i-1,j-2}\right\} + R^{(uv)}_{i,j} \quad (8)$$

for $i = 1, \ldots, l_u$ and $j = 1, \ldots, l_v$, taking $Q^{(uv)}_{q,r} = 0$ for $q, r < 1$. Eq. 8 is the consequence of configuring the original Q_{max} algorithm so that no penalties are applied for local disruptions or mismatches[3] [25]. This way, the Q_{max} measure, defined as $Q^{(uv)}_{\text{max}} = \max\left\{Q^{(uv)}\right\}$, turns into a global similarity measure considering all aligned sample matches from the beginning to the end of the time series being compared (i.e., more in the vein of global measures like dynamic time warping [23] or edit distances).

To obtain a segment similarity measure between 0 and 1 we normalize $Q^{(uv)}_{\text{max}}$ by its maximum possible value (the length of the shortest segment), yielding

$$S_{u,v} = \frac{Q^{(uv)}_{\text{max}}}{\min\{l_u, l_v\}}, \quad (9)$$

the similarity of segments u and v. This operation is performed for all possible pairwise segment comparisons $u, v = 1, \ldots, M+1$, obtaining a segment similarity matrix S (Fig. 3, bottom left). Notice that when $u = v$ the diagonal of $R^{(uv)}$ is completely filled with ones, and hence $S_{u,u} = 1$.

The next step in assessing segment repetitions is to decide which segments are similar enough so that they can be assigned to the same structural label (B2, Fig. 1). To do so we simply threshold the similarity matrix S by $\phi = \mu(S) + \sigma(S)$, where $\mu(S)$ and $\sigma(S)$ are the mean and standard deviation of all the values in S, respectively. The definition of threshold ϕ is rather arbitrary, but with the intention of discarding segment similarities that are below or around the average segment similarity of the time series. Notice that ϕ is dynamic, in the sense that it adapts to the segment-segment similarities of each individual time series. Notice furthermore that ϕ is set automatically for each time series and, therefore, does not need to be tuned in any way. Preliminary experiments with a fixed threshold also worked well but, apart from adding a new parameter to the approach, its correct value turned out to be rather critical and inconsistent across different musical pieces (see also Sec. IV-C for a further possible parameter-based dynamic thresholding approach).

The simple threshold operation above does not enforce transitivity, i.e., if a segment 1 is close (or similar) to a segment 2, and segment 2 close to a segment 3, it does not guarantee that segment 1 is close to segment 3. To enforce transitivity (B3, Fig. 1), we recursively multiply S by itself (matrix multiplication) and threshold the resulting matrix by 1 (i.e., we set its entries to 1 if they are above 1 and 0 otherwise) until no changes are produced in S. We denote this final result as $\hat{S}$ (Fig. 3, bottom right). The final assignment of labels is then straightforward: the label A is given to all segments that correspond to the 1-entries of the first row of $\hat{S}$, the label B to all segments that correspond to the 1-entries of the next row that differs from the previously considered ones, and so forth until all rows of $\hat{S}$ have been processed.

The aforementioned violations in transitivity, which typically result from noisy input data due to musical and acoustic variations, constitute a challenge in music structure analysis, and various strategies have been applied to compute some sort of transitive closure [7], [13], [15], [17]. Our approach reminds of the one by Peeters [17], where higher-order similarity matrices are used to recover missing relations. However, [17] applies this procedure to similarity matrices computed from frame-wise comparisons, and before the segmentation step. At this stage, spurious relations that are due to local deviations in the time series may be boosted by considering higher-order matrices, which may result in a large number of unwanted relations. In contrast, we enforce transitivity after the segmentation stage, considering similarity matrices computed from segment-wise comparisons. At this stage, local deviations in the time series have a smaller impact, which makes the overall transitivity computation much more robust. Furthermore, computing transitivity on the segment level makes our approach much more efficient than frame-based approaches. Finally, we want to mention the audio thumbnailing approach in [29], where separate transitivity enforcement considerations are avoided through a unifying optimization scheme for the concurrent extraction of local relations, segmentation, and grouping.

[3] The reader will easily see that the formulation in Eq. 8 exactly corresponds to the original formulation of Q_{max} in [25]. It directly follows from setting the original parameters $\gamma_o = \gamma_e = 0$.

D. Parameter Setting

From the previous explanations we see that a number of parameters need to be adjusted in our approach. We perform a detailed quantitative analysis of the role of the various parameters in Sec. IV-C. For the moment, we fix the parameters by considering the nature of the time series and the characteristics of the task. For instance, since the time series we will consider fluctuate rapidly, we set $\tau = 1$ samples so that no information from the recent past of $\mathbf{x}_i$ is lost in building $\hat{\mathbf{x}}_i$ (Eq. 1; cf. [26]). Moreover, since we mostly focus on Western popular music, we expect no dramatic speed or tempo changes in our time series. This implies that no strong fluctuations will be present along the lag dimension of L (i.e., diagonal traces in R will not warp dramatically). Therefore, a relatively small value of, say $s_\mathrm{l} = 0.3\,\mathrm{s}$ will suffice for kernel G to track such fluctuations[4] (Eq. 5). For the Gaussian windows $\mathbf{g}_\mathrm{t}$ and $\mathbf{g}_\mathrm{l}$ we follow common practice and use a variance of 0.16, which ensures a value close to 0 at the borders of $\mathbf{g}$ [28]. Finally, to select the peaks of $\mathbf{c}$, we set $\delta = 0.05$, which corresponds to 5% of the maximum amplitude of $\mathbf{c}$ (actually, values of $\delta \in [0, 0.25]$ turn out to have no effect on the results, Sec. IV-C). In addition, we set $\lambda = 6\,\mathrm{s}$, thereby forcing a minimal segment length of 3 s (Sec. II-B). A minimal segment length below 3 s would be conflicting with the standard evaluation setting for boundary detection, as it commonly uses a threshold of 3 s to determine the correct placement of a boundary (Sec. III-B). Moreover, as it has been shown elsewhere [19], [30], placing boundaries at intervals below 3 s induces a very high recall but a low precision, leading to quite high but unrealistic F-measures (see also Sec. III-C). As for a minimal segment length above 3 s, we would directly miss the boundaries of segments shorter than that. Since in our data we find some segments whose lengths are around or slightly above 3 s (cf. Table I), we keep this value as our minimal segment length in all performed experiments.

The previous setting leaves us with three important parameters: m (the amount of recent past we consider for $\mathbf{x}_i$; Eq. 1), κ (which controls the amount of a sample's nearest neighbors in R; Eq. 2), and s_t (the length of the time dimension of the kernel G; Eq. 5). Notice that no parameters need to be set for segment similarity (Sec. II-C). Thus, our method based on structure features is mostly parameterized by SF(m,κ,s_t).

The parameter values we try for m, κ, and s_t can be also justified by the nature of the data and the characteristics of the task. Suitable values for m lie between 0 and 5 s, the latter value accounting for 2% of a time series of 4 min of duration (a typical duration for a song). Going beyond 2% of the time series may introduce irrelevant information and dimensionality problems [26]. Suitable values for κ are found below 0.05, i.e., below 5% of the length of the time series. If we suppose we have K repetitions in a time series of length N, this implies that we will have at least K black dots in the rows of R (in fact, more than K black dots would be desirable due to noise). Therefore, κ should be greater than K/N and, at the same time, not as high as to introduce a lot of noisy dots in the rows of R. In the case of the considered music material we can think of $K \approx 13$ and $N \approx 1600$ (as with our example of Figs. 2 and 3), and hence use $13/1600 \leq \kappa \leq 0.05$. Finally, suitable values for s_t are found around 30 s. If the length of our Gaussian kernel is 30 s, then the Gaussian shape is maximal at 15 s and decreases by 50% at 8 and 23 s [28]. This yields an 'effective' kernel length of approximately 15 s, close to the average time between boundaries in our time series (see Table I).

	Num.	Length	# Bound.	Interval
BEATLES-A	174	158.2 (51.5)	8.2 (2.3)	17.2 (12.3)
BEATLES-B	180	162.9 (56.6)	9.2 (2.3)	16.0 (13.9)
RWC-POP-A	100	242.2 (41.5)	16.1 (4.0)	14.1 (6.8)
RWC-POP-B	100	224.1 (41.4)	16.8 (3.4)	13.7 (7.2)
MAZURKA	2792	160.2 (74.2)	9.7 (4.1)	14.9 (9.0)

TABLE I
DATA SET STATISTICS (STANDARD DEVIATIONS INTO PARENTHESIS): NUMBER OF RECORDINGS, AVERAGE LENGTH, AVERAGE NUMBER OF BOUNDARIES, AND AVERAGE LENGTH OF INTER-BOUNDARY INTERVAL (THE TIME BETWEEN TWO CONSECUTIVE BOUNDARIES). TIME VALUES ARE GIVEN IN SECONDS.

As we will see in Secs. IV-B and IV-C, the specific setting of the parameters is often not crucial, as long as they are fixed within suitable ranges. In particular, different parameter combinations yield comparable results, and stable accuracies are obtained for rather wide parameter ranges. In the above discussion, suitable parameter values and ranges have been motivated in a musically-informed way.

III. EXPERIMENTAL SETUP

A. Music Collections

To evaluate the proposed approach we employ three benchmark music collections with boundary and structure annotations: Beatles, RWC-Pop, and Mazurka. The Beatles data set corresponds to all the recordings in the 12 original albums of the band. There are two versions for ground truth annotations of this data set, which are denoted as BEATLES-A[5] and BEATLES-B[6] (Table I). Since many works in the literature have been evaluated using these annotations, the performance of our approach can be compared with the current state-of-the-art.

The second data set consists of all recordings of the Real World Computing Popular Music Database [31]. These recordings represent Japanese mainstream music and, to a less extent, American chart hits. We also use two versions of annotations as ground truth, which are denoted by RWC-POP-A[7] and RWC-POP-B[8] (Table I). At the moment RWC-POP-B only contains boundary annotations. The RWC recordings and the two annotations are publicly available.

The third data set, which we denote by MAZURKA, comprises many recorded performances for 49 mazurkas by Frédéric Chopin. This collection was assembled by the

[4] For ease of interpretation we express all time-related parameter values in seconds (τ, which is better understood in samples, is the only exception here). Parameter values in samples can be easily obtained dividing by the time series sampling rate or, in other words, multiplying by the descriptor hop size.

[5] http://www.cs.tut.fi/sgn/arg/paulus/beatles_sections_TUT.zip

[6] http://isophonics.net/content/reference-annotations

[7] http://staff.aist.go.jp/m.goto/RWC-MDB/AIST-Annotation

[8] http://musicdata.gforge.inria.fr

Mazurka Project[9] and contains a total of 2792 audio recordings (Table I). For each of the 49 Mazurkas, the musical form was first manually annotated by a human expert on the basis of the musical score of the piece and, later, these score-based annotations were transferred to all performances available for this piece using an automated procedure. To this end, the score was first exported to a MIDI file, the MIDI file was temporally aligned to a given audio recording using music synchronization techniques [32], and the resulting alignment information was used to temporally warp the score-based annotations to match the respective audio recording.

Apart from the music collections and annotations used here, we also submitted the proposed approach to the annual Music Information Retrieval Evaluation eXchange [21] (MIREX). MIREX is an international evaluation campaign for music information retrieval algorithms, coupled with the International Society for Music Information Retrieval conference (ISMIR), and hosted by the University of Illinois at Urbana Champaign. In 2012, the so-called structural segmentation task was run with three music collections [10] : the MIREX09 collection, the aforementioned RWC-POP collection, and the recent MIREX12 (SALAMI) data set. The MIREX09 collection is a mixture of part of the collections used by Paulus and by Peiszer, including Beatles' and other pop songs up to a total of 220 audio recordings. The MIREX12 collection contains over 1,000 annotated pieces collected within the SALAMI project[11], covering a range of musical styles.

B. Evaluation Measures

For comparing to existing approaches we use common evaluation measures (see [30] for a summary). For boundary annotation we use hit rates and median deviations. With hit rates, segment boundaries are accepted to be correct if they are within a certain threshold from a boundary in the ground truth annotation. Common thresholds are 0.5 and 3 s [33], [34]. Based on the matched hits, standard precision, recall, and F-measure are computed for each music recording and averaged across the whole data set[12]. Since we observed that some annotations were not accurate at a resolution of 0.5 s, we only report on precision, recall, and F-measure using a 3 s threshold (P_{B}, R_{B}, and F_{B}, respectively). Additionally, two median deviation values are computed, counting the median times from true-to-guess and from guess-to-true [34]. However, due to space constraints we do not show these measures here. A full account of the average performance of our method using all the evaluation measures, as well as the performance for each individual piece, can be downloaded from the web[13].

For label annotation we use pairwise frame clustering as well as over- and under-segmentation measures. Handling the result and the ground truth in 0.3 s frames we compute standard pairwise precision, recall, and F-measure as defined in [11] (P_{L}, R_{L}, and F_{L}, respectively). Additionally, we compute over- and under-segmentation measures as proposed in [35]. However, due to space constraints, we here only report the former and refer to the results web document for the latter.

C. Baseline evaluations

One should be cautious with some of the aforementioned evaluation measures, specially with boundary evaluation measures. For instance, placing a boundary every second can already yield a boundary recall $R_{\mathrm{B}} = 1$ and a boundary F-measure $F_{\mathrm{B}} \approx 0.5$ [30]. For the sake of comparison with existing approaches we revert to these measures, but provide two additional baseline evaluations: placing a certain number of random boundaries using the average number of boundaries reported in Table I (Baseline 1) and placing a boundary every 3 s (Baseline 2). In all these baselines the same label is assigned to each segment. Additional information on evaluation baselines can be found in [30] or [19].

One should also note that two different human annotators could disagree in the annotation of the structure of a musical piece (see [36] and also Table I, where we see that the numbers for two different ground truths do not totally agree). To obtain a reference of human performance when evaluating against our ground truth, we propose to evaluate across different annotations. That is, given two annotations A and B for the same music data set, we use annotation B as a result when evaluating with the ground truth provided by annotation A and vice versa (e.g., using BEATLES-B as a candidate result and BEATLES-A as a ground truth).

IV. RESULTS AND DISCUSSION

A. Illustrating Example

Before presenting the quantitative results, we first give an illustrating example based on the HPCP time series computed from "All you need is love" by The Beatles (Figs. 2 and 3). We see that homogeneities (gray areas) and repetitions (straight diagonal lines) are already visible in the recurrence matrix R, and before computing the novelty curve $\mathbf{c}$. In this case, the majority of the detected boundaries correspond to real boundaries (Fig. 2, bottom). Interestingly, the two false positive boundaries towards the end (samples 1408 and 1518) correspond to a short musical quotation of the main chorus of the song "She loves you", which The Beatles included in the final fade-out of this recording. These actually musically meaningful boundaries had not been annotated by humans, which again highlights the problems of collecting ground truth annotations for this task.

Continuing with the same example, but focusing on segment labeling (Fig. 3, bottom), we see that the proposed approach makes a good distinction between verse (segments 3, 4, 6, and 8) and chorus sections (segments 5, 7, 9, and 10). Our approach annotates segment 2 being the same as 3, 4, 6, and 8, despite it being annotated differently by humans. However, it turns out that segment 2 shares the same chords and instrumentation as those verse segments, but only incorporates the

[9] http://mazurka.org.uk

[10] http://www.music-ir.org/mirex/wiki/2012:MIREX2012_Results

[11] http://ddmal.music.mcgill.ca/salami

[12] This specially includes the F-measure, which is also computed for each song independently and averaged afterwards. Noticeably, if one computes it from the averaged precision and recall measures, one obtains abnormally high Fs.

[13] http://www.iiia.csic.es/~jserra/downloads/2012_TMM_StructureAnnotation_Results.zip

Approach	Boundaries			Labels		
	P_{B}	R_{B}	F_{B}	P_{L}	R_{L}	F_{L}
Baseline 1 (8 bound.)	0.418	0.458	0.427	0.344	0.989	0.499
Baseline 2 (3 s)	0.343	0.998	0.505	0.348	0.982	0.502
Levy & Sandler [11]^a	0.586	0.832	0.581	0.600	0.627	0.597
Paulus & Klapuri [8]	0.521	0.612	0.550	0.729	0.546	0.599
Peiszer [37]^a	0.515	0.824	0.617	0.611	0.623	0.597
SF(2,0.01,32)	0.681	0.729	0.696	0.709	0.659	0.658
SF(2,0.02,29)	0.691	0.808	0.737	0.707	0.741	0.699
SF(2.5,0.03,32)	0.714	0.797	0.745	0.693	0.775	**0.707**
SF(3,0.04,35)	0.734	0.791	**0.753**	0.651	0.800	0.690
Human	0.889	0.937	0.911	0.902	0.870	0.876

TABLE II

RESULTS WITH BEATLES-A. BEST F-MEASURES ARE HIGHLIGHTED IN BOLD. THE SUPERSCRIPT ^{a} DENOTES RESULTS REPORTED BY SMITH [30].

Approach	Boundaries			Labels		
	P_{B}	R_{B}	F_{B}	P_{L}	R_{L}	F_{L}
Baseline 1 (9 bound.)	0.446	0.526	0.471	0.332	1.000	0.491
Baseline 2 (3 s)	0.355	1.000	0.516	0.338	0.995	0.495
Chen & Li [38]	-	-	-	0.610	0.690	0.630
Mauch et al. [39]^a	-	-	-	0.610	0.770	0.660
Weiss & Bello [40]	-	-	-	0.570	0.690	0.600
SF(2,0.01,32)	0.723	0.755	0.728	0.700	0.663	0.656
SF(2,0.02,29)	0.733	0.825	0.766	0.695	0.751	0.700
SF(2.5,0.03,32)	0.753	0.816	**0.774**	0.681	0.787	**0.711**
SF(3,0.04,35)	0.764	0.802	0.773	0.641	0.809	0.691
Human	0.937	0.889	0.911	0.870	0.902	0.876

TABLE III

RESULTS WITH BEATLES-B. BEST F-MEASURES ARE HIGHLIGHTED IN BOLD. THE SUPERSCRIPT ^{a} DENOTES RESULTS REPORTED BY WEISS & BELLO [40].

Approach	Boundaries			Labels		
	P_{B}	R_{B}	F_{B}	P_{L}	R_{L}	F_{L}
Baseline 1 (16 bound.)	0.463	0.438	0.443	0.289	1.000	0.447
Baseline 2 (3 s)	0.411	1.000	0.576	0.293	1.000	0.451
Barrington et al. [12]	-	-	-	-	-	0.620
Paulus & Klapuri [8]	0.717	0.578	0.630	0.603	0.721	0.637
Peiszer [37]^a	0.613	0.807	0.680	-	-	-
SF(2,0.01,32)	0.827	0.730	0.766	0.755	0.659	**0.691**
SF(2,0.02,29)	0.817	0.773	**0.785**	0.728	0.665	0.680
SF(2.5,0.03,32)	0.829	0.745	0.776	0.707	0.676	0.678
SF(3,0.04,35)	0.810	0.691	0.737	0.667	0.701	0.669
Human	0.921	0.891	0.899	-	-	-

TABLE IV

RESULTS WITH RWC-POP-A. BEST F-MEASURES ARE HIGHLIGHTED IN BOLD. THE SUPERSCRIPT ^{a} DENOTES RESULTS REPORTED BY SMITH [30].

Approach	Boundaries			Labels		
	P_{B}	R_{B}	F_{B}	P_{L}	R_{L}	F_{L}
Baseline 1 (17 bound.)	0.466	0.500	0.476	-	-	-
Baseline 2 (3 s)	0.402	1.000	0.569	-	-	-
Kaiser et al. [41]	0.687	0.658	0.661	-	-	-
Sargent et al. [42]^a	0.622	0.623	0.612	-	-	-
Sargent et al. [42], [43]^b	0.697	0.584	0.628	-	-	-
SF(2,0.01,32)	0.817	0.763	0.782	-	-	-
SF(2,0.02,29)	0.799	0.790	0.787	-	-	-
SF(2.5,0.03,32)	0.827	0.782	**0.797**	-	-	-
SF(3,0.04,35)	0.807	0.732	0.762	-	-	-
Human	0.891	0.921	0.899	-	-	-

TABLE V

RESULTS WITH RWC-POP-B. BEST F-MEASURE IS HIGHLIGHTED IN BOLD. THE SUPERSCRIPTS ^{a,b} DENOTE RESULTS REPORTED IN THE MIREX 2011 AND 2012 CAMPAIGNS [21], RESPECTIVELY.

choirs (skipping the main voice). We also see that the proposed approach clearly differentiates the long ostinato ending and fade-out (segments 11 to 13) from the rest. Segment 1 is different from the rest because it is a distinct brass band introduction and segment 12 is different from the rest because of the aforementioned "She loves you" quotation.

B. Accuracy Assessment

Let us now turn to some quantitative results. We start considering the results obtained with the BEATLES collection (Tables II and III) and then comment on the other data sets. First, we note that our approach (SF) clearly outperforms the random evaluation baselines. For boundaries, the highest F_{B} for a baseline is 0.516 (Baseline 2, BEATLES-B), whereas the lowest F_{B} for the proposed approach is 0.696 (SF(2,0.01,32), BEATLES-A). For labels, the highest F_{L} for a baseline is 0.502 (Baseline 2, BEATLES-A), whereas the lowest F_{L} for the proposed approach is 0.658 (SF(2,0.01,32), BEATLES-A). Human performance is still higher than our approach's, but the difference gets tighter. For instance, with BEATLES-A we get $F_{\mathrm{L}} = 0.707$ and human agreement is at $F_{\mathrm{L}} = 0.876$.

Second, we observe that, independently of the ground truth we use (BEATLES-A or BEATLES-B), different parameter combinations for SF yield similar results. Actually, many of these results turned out to be not statistically significantly different[14] between them. This highlights the robustness of the proposed approach against different parameter settings and, also, against different ground truth annotations.

Third, along with the results of SF we also report the best results found in the literature with that collection, using the same evaluation framework. We observe that SF clearly outperforms the other approaches, with a statistically significant difference between average performances. For instance, Paulus & Klapuri [8] achieved $F_{\mathrm{L}} = 0.599$ with BEATLES-A and Mauch et al. [39] were reported [40] to have $F_{\mathrm{L}} = 0.66$ with BEATLES-B. Our approach yields an F_{L} close to 0.7 or higher with both annotations.

Replicating our evaluation with RWC-POP and MAZURKA collections confirms all the aforementioned statements (Tables IV, V, and VI): SF is far above the baseline, different parameter combinations yield comparable accuracies, and SF clearly outperforms the best results published so far under the same evaluation framework. Achieving similar high accuracies in such a cross-collection evaluation further highlights the strength of the proposed approach and its robustness across different parameter combinations. Moreover, the accuracies achieved with BEATLES and RWC-POP seem to go beyond popular music genres, as suggested by the comparable results achieved with the MAZURKA collection.

The proposed approach also achieved very good results in MIREX 2012 (Table VII). The results for RWC-POP matched the ones reported here, with only minor differences due to some parameter modifications (at the time of submitting our algorithm we used $\tau = 1$, $s_{\mathrm{l}} = 0.33$, $\delta = 0.1$, and $\lambda = 6$). The results for the MIREX09 collection were statistically significantly better than the rest of participants. Specifically,

[14] Unless stated otherwise, statistical significance is assessed with a T-test at $p < 0.05$ and assuming a Gaussian distribution of the evaluation measures. When standard deviations are not reported in the literature, equal variances as with our approach are assumed.

Approach	Boundaries			Labels		
	P_{B}	R_{B}	F_{B}	P_{L}	R_{L}	F_{L}
Baseline 1 (10 bound.)	0.479	0.472	0.461	0.330	0.999	0.482
Baseline 2 (3 s)	0.431	1.000	0.582	0.333	0.992	0.484
SF(2,0.01,32)	0.703	0.653	0.659	0.752	0.652	0.681
SF(2,0.02,29)	0.715	0.719	**0.699**	0.772	0.693	0.713
SF(2.5,0.03,32)	0.724	0.695	0.692	0.758	0.716	**0.719**
SF(3,0.04,35)	0.725	0.661	0.675	0.733	0.733	0.717

TABLE VI
RESULTS WITH MAZURKA. BEST F-MEASURES ARE HIGHLIGHTED IN BOLD. SINCE WE ONLY HAVE ONE ANNOTATION, HUMAN ACCURACY CANNOT BE ESTIMATED FOR THIS COLLECTION (SEC. III-C).

Approach	Music collection			
	MIREX09	RWC-B	RWC-A	MIREX12
SF(2.5,0.03,32)	**0.653**	**0.766**	0.675	**0.581**
SF(2,0.01,32)	0.633	0.759	**0.688**	0.528
Kaiser et al. [41]-3	0.572	0.661	0.605	0.531
Martin et al. [44]	0.556	0.545	0.583	0.572
Kaiser et al. [41]-1	0.554	0.661	0.603	0.502
Kaiser et al. [41]-4	0.551	0.661	0.562	0.554
Kaiser et al. [41]-2	0.544	0.661	0.583	0.528
Sargent et al. [43]	0.515	0.628	0.535	0.460
Ono et al. ^{u}	0.464	0.531	0.507	0.501
Mauch et al. [39]	0.612	0.605	-	-
Sargent et al. [42]	0.501	0.612	-	-
Martin et al. [44]	0.555	0.486	-	-

TABLE VII
RESULTS FOR MIREX COLLECTIONS AS REPORTED AT ISMIR 2012 (TOP ROWS). FOR COMPREHENSIVENESS WE ALSO INCLUDE THE THREE BEST ACCURACIES AMONG PREVIOUS MIREX EDITIONS (BOTTOM ROWS). BEST F_{L} VALUES ARE HIGHLIGHTED IN BOLD. THE SUPERSCRIPT ^{u} DENOTES "UNPUBLISHED".

Parameter	Unit	Meaning
m	s	Embedding dimension (Eq. 1)
τ	s	Time delay (Eq. 1)
κ	%	Percentage of nearest neighbors (Eq. 2)
δ	-	Peak picking threshold
λ	s	Peak picking minimal window length
s_{l}	s	Smoothing length for lag dimension (Eq. 5)
s_{t}	s	Smoothing length for time dimension (Eq. 5)

Setting	Configuration: SF(m,κ,s_{t})
A	SF(2,0.01,32)
B	SF(2,0.02,29)
C	SF(2.5,0.03,32)
D	SF(3,0.04,35)

TABLE VIII
SUMMARY OF PARAMETERS AND PARAMETER SETTINGS.

Bypassed blocks	Boundaries			Labels		
	P_{B}	R_{B}	F_{B}	P_{L}	R_{L}	F_{L}
Baseline 1 (8 bound.)	0.418	0.458	0.427	0.344	0.989	0.499
Baseline 2 (3 s)	0.343	0.998	0.505	0.348	0.982	0.502
A1 off	0.658	0.752	0.693	0.760	0.449	0.540
A2 off	0.645	0.554	0.575	0.552	0.705	0.591
A3 off	0.371	0.917	0.522	0.515	0.740	0.561
A1+A2 off	0.386	0.873	0.528	0.491	0.724	0.542
A2+A3 off	0.375	0.801	0.503	0.500	0.717	0.546
A1+A2+A3 off	0.372	0.866	0.514	0.464	0.792	0.542
SF(2.5,0.03,32)	0.714	0.797	0.745	0.693	0.775	0.707
Human	0.889	0.937	0.911	0.902	0.870	0.876

TABLE IX
ACCURACIES WITH BEATLES-A WHEN SWITCHING OFF STRUCTURE FEATURE COMPUTATION BLOCKS (SEE FIG. 1). FOR COMPARISON PURPOSES, WE ALSO SHOW BASELINE AND HUMAN ACCURACIES, AS WELL AS A REFERENCE PERFORMANCE FROM TABLE II.

SF obtained an $F_{\text{L}} = 0.653$, whereas the best result reported for this collection was $F_{\text{L}} = 0.612$, obtained by Mauch et al. [39] in 2010. The results for the MIREX12 collection were more tight, with SF scoring $F_{\text{L}} = 0.581$ and Martin et al. [44] scoring $F_{\text{L}} = 0.572$. In general, for many of the statistical tests we could try, this difference is not found to be statistically significant. However, noticeably, [44] does not perform so well with the other MIREX test collections. Furthermore, the accuracies for all algorithms on the MIREX12 collection are much lower than than the ones obtained with the collections used in this paper and the other MIREX collections, what brings up the question of the criteria used for annotating such collection and whether it conforms to the rest of the annotations used elsewhere.

C. Blocks and Parameters Impact

To assess the accuracy of the proposed method (Tables II–VI) we have made use of four fixed parameter configurations (Table VIII). In such configurations, parameter values were intuitively determined from the musically-informed considerations of Sec. II-D. This has been useful for showing that the proposed method can reach high accuracies with no extensive parameter tuning (Sec. IV-B). We now go one step further and study the impact of each parameter in each of the four studied configurations (Table VIII). This will allow us to identify critical steps in our method and to gain empirical knowledge on the useful parameter ranges. In addition, for improving our understanding of the method, we can switch off several of its blocks (Fig. 1). This will additionally allow us to compare the proposed approach against simpler variants of it. In the following, we use BEATLES-A for quantitative evaluation. We found similar parameter behaviors with the other music collections.

We start by looking at the accuracies obtained with parameter configuration SF(2.5,0.03,32) when switching off some of the segment boundary computation blocks (A1 to A4, Fig. 1). In general, we observe that accuracies seriously drop when leaving out any of the proposed steps (Table IX). Switching off A1, we maintain a relatively high boundary detection accuracy, $F_{\text{B}} = 0.693$. However, this is not coupled with a correct segment labeling, where we get $F_{\text{L}} = 0.540$, approaching to the random baseline. This indicates that block A1 (delay coordinates, Eq. 3) does not strongly affect boundary placement, but is crucial for segment similarity, the key ingredient of the proposed structure labeling step. As for the remaining switch off options, we see that they lead to results not far from the random baseline accuracies in both boundary detection and segment labeling (Table IX). Noticeably, these switch off options include using the raw music descriptor time series as features for novelty detection (A1+A2+A3 off), thus replacing the proposed structure features by standard HPCPs, and a smoothed descriptor time series version (A1+A2 off). Overall, these results show that every single block, and hence all steps of our approach, are necessary for achieving the good accuracies reported in Sec. IV-B.

We next study the impact every single parameter has on the overall accuracy. For that we employ the same four parameter configurations we used in Sec. IV-B and systematically explore

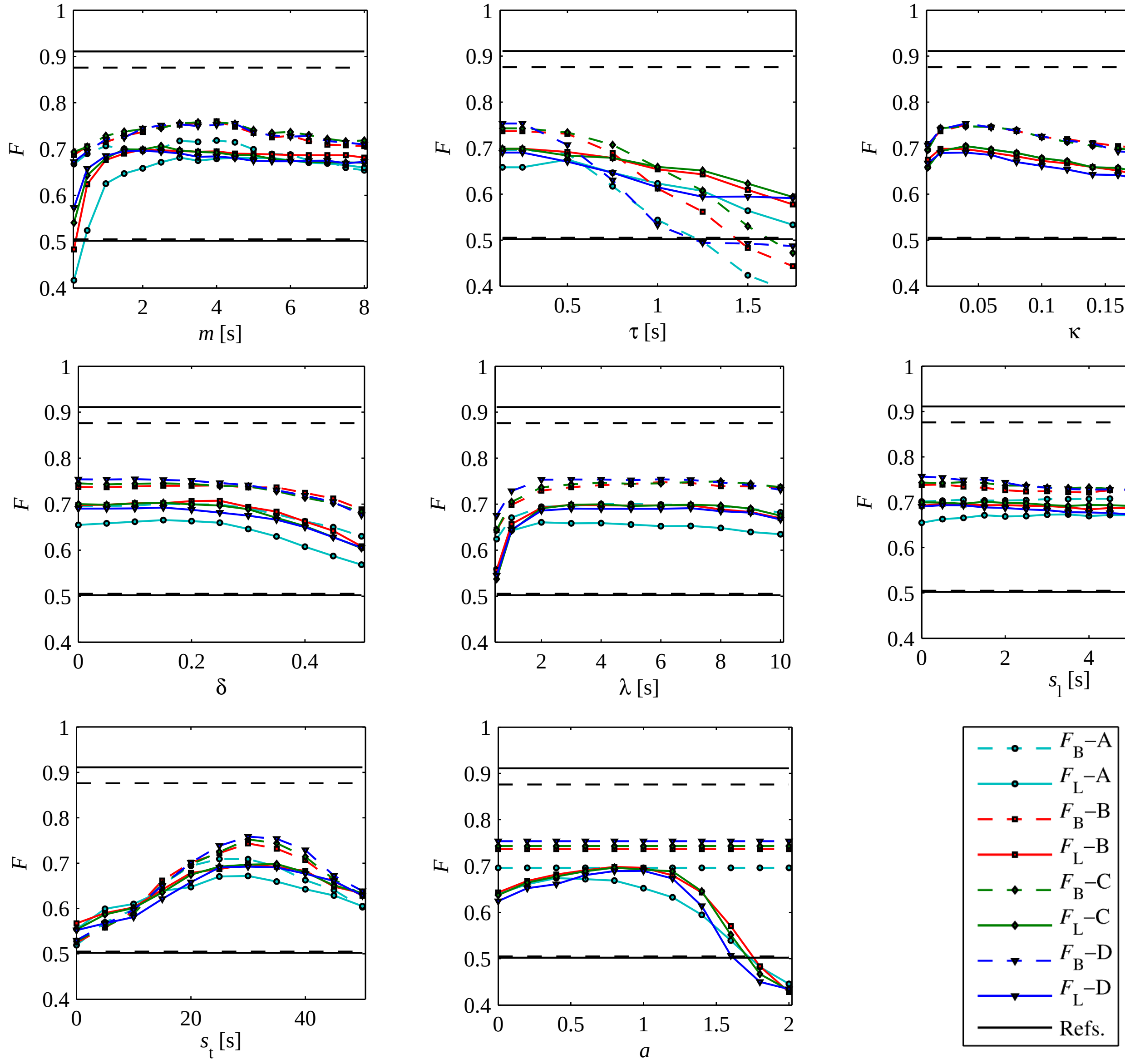

Fig. 4. Impact of different parameter values. From left to right and top to bottom these correspond to m, τ, κ, δ, λ, s_l, s_t, and a (see Table VIII and the text). Dashed lines correspond to F_B and solid lines to F_L. The four different colors/symbols correspond to the four different fixed configurations used in Sec. IV-B (Tables II–VI), and are denoted by A, B, C, and D (see Table VIII and the legend shared by all plots). The black horizontal lines on top correspond to human performance reference. The black horizontal lines at the bottom correspond to the random baseline reference.

the role of individual parameters within the suitable ranges outlined in Sec. II-D (Table VIII). First, as mentioned, we see that the embedding dimension m can affect the structural segment labeling, but has a minor effect on boundary detection (Fig. 4, top left). Indeed, F_L values approach the random baseline as $m \rightarrow 0\,\mathrm{s}$, whereas F_B is more or less maintained across the entire considered range. Notice that an $m = 0\,\mathrm{s}$ would correspond to no embedding enhancement, i.e., using the raw time series, and therefore switching off A1 as above. Besides, we see that suitable values for m lie between 2 and 5 s, confirming our intuition of Sec. II-D. Noticeably, accuracy values seem to be in a stable plateau for $m > 2\,\mathrm{s}$.

The next parameter we study is τ, the amount of delay between embedding coordinates (Eq. 1). In particular, we see that the longer the delay, the less the accuracy we get (Fig. 4, top center). However, values of $\tau < 0.5\,\mathrm{s}$, corresponding to less than 4 samples, yield comparable and stable accuracies.

Next, we turn our attention to κ, the number of nearest neighbors used in our recurrence plot (Eq. 2). We can detect a performance peak between $0.02 \leq \kappa \leq 0.06$ (Fig. 4, top right), coinciding with our hypotheses of Sec. II-D. Considering a larger number of neighbors, $\kappa > 0.06$, gradually decreases accuracy, but not drastically. Regarding the peak threshold δ (Eq. 6), we confirm that values of $\delta \in [0, 0.25]$ do not affect the results (Fig. 4, middle left). The accuracies also behave very smoothly with the peak picking window length λ (Fig. 4, middle center), except for very small windows (we discussed this aspect in Sec. II-D). Similarly, the kernel density estimation parameter s_l has a marginal effect, provided it is not set to an unreasonably high value $s_\mathrm{l} \gg 6\,\mathrm{s}$ (Fig. 4, middle right). The other density estimation parameter, s_t, yields stable accuracies in the range of $20 \leq s_\mathrm{t} \leq 40\,\mathrm{s}$ (Fig. 4, bottom left).

Noticeably, values of $s_{\mathrm{t}} \rightarrow 0$ (corresponding to switching off A4) lead to very poor accuracies (see also Table IX).

All the previous parameters relate to the boundary detection stage (A blocks, Fig. 1). As mentioned, the segment labeling part is parameter-free, thus no configurable options exist (B blocks, Fig. 1). However, one may introduce some alternatives in order to get an impression on how different components may affect final results. As an example, we replaced the segment similarity threshold $\phi = \mu(S) + \sigma(S)$ by $\phi = \mu(S) + a\sigma(S)$, introducing a new parameter a controlling the threshold magnitude. Hence, we can now assess the impact of different thresholds and see how critical the overall thresholding operation is (Fig. 4, bottom center). Values of $a \approx 1$ yield the best F_{L} accuracies, with less pronounced drops for $a < 1$ than for $a > 1$ (of course, as this newly introduced parameter does not affect boundary placement, F_{B} remains the same). In general, we see that a wide range of values $0.5 \leq a \leq 1.2$ provide good accuracies, thus showing that the specific dynamic choice of ϕ is not crucial.

V. Conclusion

In this paper we introduced a novel approach to structure annotation based on structure features and segment similarity. First, we showed that structure features measuring global characteristics of a time series becomes a powerful tool in combination with local measurements as usually done for novelty detection. This local/global combination leads to a robust estimation of segment boundaries, as confirmed by our experimental results. Second, we showed how to combine boundary estimations a labeling procedure based on structural segment similarities. To this extent, we adapted an existing time series similarity measure and developed suitable thresholding and transitivity strategies. Third, we studied the impact of each parameter under alternative settings and assessed the importance of the different steps of our method. Finally, we conducted an exhaustive empirical evaluation of our approach with three different music collections and five distinct structure annotations. The overall results outperformed any results published in the literature using the same evaluation framework (both in boundary detection and segment labeling tasks). The MIREX 2012 evaluation results further confirmed this aspect in an out-of-sample and non-optimized scenario, using different collections.

References

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

[2] J. Paulus, M. Müller, and A. Klapuri, "Audio-based music structure analysis," in *Proc. of the Int. Soc. for Music Information Retrieval Conf. (ISMIR)*, 2010, pp. 625–636.

[3] A. D. Patel, *Music, language, and the brain*. Oxford, UK: Oxford University Press, 2007.

[4] P. Ball, *The music instinct: how music works and why we can't do without it*. London, UK: Bodley Head, 2010.

[5] D. Arnold, A. Latham, and J. Dunsby, "Form," in *The Oxford Companion to Music*, A. Latham, Ed. Oxford Music Online, 2012. [Online]. Available: http://www.oxfordmusiconline.com/subscriber/article/opr/t114/e2624

[6] G. Boutard, S. Goldszmidt, and G. Peeters, "Browsing inside a music track, the experimentation case study," in *Proc. of the Workshop on Learning the Semantics of Audio Signals (LSAS)*, 2006, pp. 87–94.

[7] M. Goto, "A chorus section detection method for musical audio signals and its application to a music listening station," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 14, no. 5, pp. 1783–1794, 2006.

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

[9] J. Foote, "Automatic audio segmentation using a measure of audio novelty," in *Proc. of the IEEE Int. Conf. on Multimedia and Expo (ICME)*, 2000, pp. 452–455.

[10] K. Jensen, "Multiple scale music segmentation using rhythm, timbre, and harmony," *EURASIP Journal on Advances in Signal Processing*, vol. 2007, p. 73205, 2007.

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

[12] L. Barrington, A. B. Chan, and G. Lanckriet, "Modeling music as a dynamic texture," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 18, no. 3, pp. 602–612, 2010.

[13] R. B. Dannenberg and N. Hu, "Pattern discovery techniques for music audio," in *Proc. of the Int. Conf. on Music Information Retrieval (ISMIR)*, 2002, pp. 63–70.

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

[15] M. Müller and F. Kurth, "Towards structural analysis of audio recordings in the presence of musical variations," *EURASIP Journal on Advances in Signal Processing*, vol. 2007, p. 89686, 2007.

[16] G. Peeters, "Deriving musical structure from signal analysis for music audio summary generation: "sequence" and "state" approach," in *Computer Music Modeling and Retrieval*, ser. Lecture Notes in Computer Science, 2004, vol. 2771, pp. 143–166.

[17] ——, "Sequence representation of music structure using higher-order similarity matrix and maximum-likelihood approach," in *Proc. of the Int. Conf. on Music Information Retrieval (ISMIR)*, 2007, pp. 35–40.

[18] N. C. Maddage, "Automatic structure detection for popular music," *IEEE Multimedia*, vol. 13, no. 1, pp. 65–77, 2006.

[19] J. Serrà, M. Müller, P. Grosche, and J. L. Arcos, "Unsupervised detection of music boundaries by time series structure features," in *Proc. of the AAAI Int. Conf. on Artificial Intelligence*, 2012, pp. 1613–1619.

[20] M. Müller, D. P. W. Ellis, A. Klapuri, and G. Richard, "Signal processing for music analysis," *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 6, pp. 1088–1110, 2011.

[21] J. S. Downie, A. F. Ehmann, M. Bay, and M. C. Jones, "The music information retrieval evaluation exchange: some observations and insights," in *Advances in Music Information Retrieval*, ser. Studies in Computational Intelligence, Z. W. Ras and A. A. Wieczorkowska, Eds. Berlin, Germany: Springer, 2010, vol. 274, pp. 93–115.

[22] E. Gómez, "Tonal description of polyphonic audio for music content processing," *INFORMS Journal on Computing*, vol. 18, no. 3, pp. 294–304, 2004.

[23] M. Müller, *Information retrieval for music and motion*. Berlin, Germany: Springer, 2007.

[24] J. Serrà, E. Gómez, P. Herrera, and X. Serra, "Chroma binary similarity and local alignment applied to cover song identification," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 16, no. 6, pp. 1138–1151, 2008.

[25] J. Serrà, X. Serra, and R. G. Andrzejak, "Cross recurrence quantification for cover song identification," *New Journal of Physics*, vol. 11, no. 9, p. 093017, 2009.

[26] H. Kantz and T. Schreiber, *Nonlinear time series analysis*. Cambridge, UK: Cambridge University Press, 2004.

[27] N. Marwan, M. C. Romano, M. Thiel, and J. Kurths, "Recurrence plots for the analysis of complex systems," *Physics Reports*, vol. 438, no. 5-6, pp. 237–329, 2007.

[28] J. S. Simonoff, *Smoothing methods in statistics*. Berlin, Germany: Springer, 1996.

[29] M. Müller, P. Grosche, and N. Jiang, "A segment-based fitness measure for capturing repetitive structures of music recordings," in *Proc. of the Int. Soc. for Music Information Retrieval Conf. (ISMIR)*, 2011, pp. 615–620.

[30] J. B. L. Smith, *A comparison and evaluation of approaches to the automatic formal analysis of musical audio*. MSc thesis, McGill University, Montreal, Canada, 2010.

[31] M. Goto, H. Hashiguichi, T. Nishimura, and R. Oka, “RWC music database: popular, classical, and jazz music databases,” in *Proc. of the Int. Conf. on Music Information Retrieval (ISMIR)*, 2002, pp. 287–288.

[32] S. Ewert, M. Müller, and P. Grosche, “High resolution audio synchronization using chroma onset features,” in *Proc. of the IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2009, pp. 1869–1872.

[33] B. S. Ong, “Structural analysis and segmentation of music signals,” Ph.D. dissertation, Universitat Pompeu Fabra, 2006.

[34] D. Turnbull, G. Lanckriet, E. Pampalk, and M. Goto, “A supervised approach for detecting boundaries in music using difference features and boosting,” in *Proc. of the Int. Conf. on Music Information Retrieval (ISMIR)*, 2007, pp. 51–54.

[35] H. Lukashevich, “Towards quantitative measures of evaluating song segmentation,” in *Proc. of the Int. Soc. for Music Information Retrieval Conf. (ISMIR)*, 2008, pp. 375–380.

[36] G. Peeters and E. Deruty, “Is music structure annotation multi-dimensional? A proposal for robust local music annotation,” in *Proc. of the Workshop on Learning the Semantics of Audio Signals (LSAS)*, 2009, pp. 75–90.

[37] E. Peiszer, *Automatic audio segmentation: segment boundary and structure detection in popular music*. MSc thesis, Vienna University of Technology, Vienna, Austria, 2007.

[38] R. Chen and M. Li, “Music structural segmentation by combining harmonic and timbral information,” in *Proc. of the Int. Soc. for Music Information Retrieval Conf. (ISMIR)*, 2011, pp. 477–482.

[39] M. Mauch, K. C. Noland, and S. Dixon, “Using musical structure to enhance automatic chord transcription,” in *Proc. of the Int. Soc. for Music Information Retrieval Conf. (ISMIR)*, 2009, pp. 231–236.

[40] R. J. Weiss and J. P. Bello, “Unsupervised discovery of temporal structure in music,” *Journal of Selected Topics in Signal Processing*, vol. 5, no. 6, pp. 1240–1251, 2011.

[41] F. Kaiser, T. Sikora, and G. Peeters, “MIREX 2012 - Music structural segmentation task: IRCAMSTRUCTURE submission,” Music Information Retrieval Evaluation eXchange (MIREX), 2012.

[42] G. Sargent, F. Bimbot, and E. Vincent, “A regularity-constrained Viterbi algorithm and its application to the structural segmentation of songs,” in *Proc. of the Int. Soc. for Music Information Retrieval Conf. (ISMIR)*, 2011, pp. 483–488.

[43] F. Bimbot, E. Deruty, G. Sargent, and E. Vincent, “Semiotic structure labeling of music pieces: concepts, methods and annotation conventions,” in *Proc. of the Int. Soc. for Music Information Retrieval Conf. (ISMIR)*, 2012, pp. 235–240.

[44] B. Martin, P. Hanna, M. Robine, and P. Ferraro, “Structural analysis of harmonic features using string matching techniques,” Music Information Retrieval Evaluation eXchange (MIREX), 2011.

Joan Serrà is a postdoctoral researcher with the Dept. of Learning Systems of IIIA-CSIC, the Artificial Intelligence Research Institute of the Spanish National Research Council in Bellaterra, Barcelona, Spain. He obtained both the degrees of Telecommunications and Electronics at Enginyeria La Salle, Universitat Ramón Llull, Barcelona, Spain, in 2002 and 2004, respectively. In 2006 he joined the Music Technology Group of Universitat Pompeu Fabra (UPF), Barcelona, where he received the MSc in Information and Communication Technologies in 2007 and the PhD in Computer Science in 2011. From 2006 to 2011 he was also an assistant and part-time associate professor with the Dept. of Information and Communication Technologies of the UPF. In 2010 he was a guest scientist at the Max Planck Institute for the Physics of Complex Systems in Dresden, and between 2011 and 2012 he did a research stay at the Max Planck Institute for Computer Science in Saarbrücken, both in Germany. His current research interests include machine learning, data mining, time series, computer audition, and music information retrieval.

Meinard Müller studied mathematics (Diplom) and computer science (Ph.D.) at the University of Bonn, Germany. In 2002/2003, he conducted postdoctoral research in combinatorics at the Mathematical Department of Keio University, Japan. In 2007, he finished his Habilitation at Bonn University in the field of multimedia retrieval writing a book titled *Information Retrieval for Music and Motion*, which appeared as Springer monograph. From 2007 to 2012, he was a member of the Saarland University and the Max-Planck Institut für Informatik leading the research group *Multimedia Information Retrieval & Music Processing* within the Cluster of Excellence on *Multimodal Computing and Interaction*. Since September 2012, Meinard Müller holds a professorship for *Semantic Audio Processing* at the International Audio Laboratories Erlangen, which is a joint institution of the University of Erlangen-Nuremberg and Fraunhofer IIS. His recent research interests include content-based multimedia retrieval, audio signal processing, music processing, music information retrieval, and motion processing.

Peter Grosche received the B.S. and M.Sc. degree in Electrical Egineering and Information Technology from Technical University of Munich (TUM) in 2006 and 2008, respectively. From 2008 to 2012, he pursued a Ph.D. in Multimedia Information Retrieval and Music Processing group at Saarland University and Max-Planck Institut für Informatik under the supervision of Meinard Müller. His thesis is titled “Signal Processing Methods for Beat Tracking, Music Segmentation, and Audio Retrieval”. In 2012, Peter joined Huawei European Research Center in Munich where he is working on topics related to 3D audio acquisition and rendering.

Josep Ll. Arcos is a Research Scientist of the Artificial Intelligence Research Institute of the Spanish National Research Council (IIIA-CSIC) where he is member of the Learning Systems Department. Received the MSc (1991) and PhD (1997) in Computer Science from the Technical University of Catalonia, Spain. He also received a MSc in Sound and Music Technology (1996) from the Pompeu Fabra University, Spain. He is co-author of more than 125 scientific publications with contributions on machine learning, case-based reasoning, multi-agent systems, self-organizing systems, or AI and music. He is co-recipient of several awards at case-based reasoning and computer music conferences. Presently, he is working on machine learning and on their applications to health care, Internet applications, and music. He acts as a reviewer of several international journals and of top international conferences.