A Robust Fitness Measure for Capturing Repetitions in Music Recordings With Applications to Audio Thumbnailing

Meinard Müller, *Member, IEEE*, Nanzhu Jiang, and Peter Grosche, *Student Member, IEEE*

***Abstract*—The automatic extraction of structural information from music recordings constitutes a central research topic. In this paper, we deal with a subproblem of audio structure analysis called audio thumbnailing with the goal to determine the audio segment that best represents a given music recording. Typically, such a segment has many (approximate) repetitions covering large parts of the recording. As the main technical contribution, we introduce a novel fitness measure that assigns a fitness value to each segment that expresses how much and how well the segment "explains" the repetitive structure of the entire recording. The thumbnail is then defined to be the fitness-maximizing segment. To compute the fitness measure, we describe an optimization scheme that jointly performs two error-prone steps, path extraction and grouping, which are usually performed successively. As a result, our approach is even able to cope with strong musical and acoustic variations that may occur within and across related segments. As a further contribution, we introduce the concept of fitness scape plots that reveal global structural properties of an entire recording. Finally, to show the robustness and practicability of our thumbnailing approach, we present various experiments based on different audio collections that comprise popular music, classical music, and folk song field recordings.**

***Index Terms*—Structure analysis, audio, music, thumbnail, repetition, path, alignment, fitness.**

I. Introduction

MUSIC is highly structured and based on different principles such as repetition, contrast, variation, and homogeneity to create certain relationships between notes, melodies, chords, harmonies, or rhythms. The automated detection of such relations, a task closely related to audio structure analysis, constitutes a fundamental research topic within the area of music information retrieval. One major goal of structure analysis is to divide a music recording into temporal segments and to group these segments into musically meaningful categories [1]. Such segments may refer to chorus sections of a piece of popular music, to stanzas of a folk song, or to the first theme of a symphony. Such musical parts are often characterized by the fact that they are repeated several times throughout the piece. Actually, finding the repetitive structure of a music recording has been a well-studied problem over the last years, see, e.g., [2]–[11] and the overview articles [1], [12].

Most of these approaches work well for music in which the repetitions largely agree. However, in music performances, musical parts are rarely repeated in precisely the same way. For example, the repeated verses of a popular song typically share the same melody but differ in terms of the underlying lyrics. Furthermore, a verse may be repeated instrumentally, with the soloist deviating from the original verse by freely improvising the melody. Also, a verse may be repeated in a transposed form, with all notes being shifted, for example, one semitone upwards. The situation becomes even more complex for amateur recordings where non-professional singers often have severe intonation problems and deviate significantly from the expected pitches. For genres such as classical music, the main melody may be repeated by different instruments with changing accompaniment and in different keys. Also, repeated parts may show significant differences in tempo. In summary, audio segments that are considered as repetitions may differ significantly in such aspects as dynamics, instrumentation, articulation, and tempo, not to speak of pronounced musical variations. In such cases, structure analysis becomes a hard and ill-posed task with many problems still remaining unsolved.

In this paper, we focus on the problem of finding the most representative and repetitive segment of a given music recording, a task often referred to as *audio thumbnailing*, see, e.g., [2], [13]–[15]. Typically, such a segment should have many (approximate) repetitions, and these repetitions should cover large parts of the recording. As the main technical contribution of this paper, we introduce a fitness measure that assigns a fitness value to each audio segment that simultaneously captures two aspects. First, it indicates *how well* the given segment explains other related segments and, second, it indicates *how much* of the overall music recording is covered by all these related segments. Similar to [9], [14], the audio thumbnail is then defined to be the segment having maximal fitness. In the computation of the fitness measure, one important conceptual idea of our approach is to avoid hard decisions and error-prone steps in an early stage

Manuscript received February 07, 2012; revised July 11, 2012, October 19, 2012; accepted October 27, 2012. Date of publication November 15, 2012; date of current version December 31, 2012. This work was supported in part by the Cluster of Excellence on Multimodal Computing and Interaction at Saarland University, the International Max Planck Research School for Computer Science, and the German Research Foundation (DFG MU 2682/5-1). The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Nobutaka Ono.

M. Müller is with the International Audio Laboratories Erlangen, which is a joint institution of the University of Erlangen-Nuremberg and Fraunhofer IIS, 91058 Erlangen, Germany (e-mail: meinard.mueller@audiolabs-erlangen.de).

N. Jiang and P. Grosche are with Saarland University and the Max-Planck Institut für Informatik, 66123 Saarbrücken, Germany (e-mail: njiang@mpi-inf.mpg.de; pgrosche@mpi-inf.mpg.de).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASL.2012.2227732

of the algorithmic pipeline. In particular, as opposed to previous approaches, we introduce an optimization scheme that jointly performs path structure extraction and grouping—two error-prone steps that are usually performed successively. As a result, we obtain a robust procedure that can detect repetitive elements even in the presence of strong musical variations. We also describe an efficient algorithm based on dynamic programming for computing the fitness measure. As a further contribution of this paper, we introduce the concept of a scape plot representation that shows the fitness values over all possible audio segments. A visualization of this fitness scape plot yields a compact high-level view on the structural properties of the entire music recording. Finally, we present experiments based on different audio collections comprising popular music, classical music, and folk song field recordings. In combination with enhanced feature representations, we show that our fitness measure can even cope with strong variations in tempo, instrumentation, and transpositions that may occur within and across related segments. By discussing several explicit examples, we indicate the strengths as well as the limitations of our approach.

The remainder of this paper is organized as follows. In Section II, we discuss related work and specify some basic notation. Then, in Section III, we summarize the general concept of self-similarity matrices and describe various enhancement strategies. Section IV contains the main contributions of this article, where we give a detailed technical description of the fitness measure and the scape plot representation. Our experiments are then described in Section V. Finally, concluding remarks and an outlook on future work can be found in Section VI. Parts of this paper have been published in [16].

II. Background

As detailed in [1], musical structure depends on various principles such as temporal order, repetition, contrast, variation, and homogeneity. Therefore, a large number of different approaches to music structure analysis have been developed, whereas one can roughly distinguish between three different classes of methods. First, repetition-based methods are employed to identify recurring patterns. Second, novelty-based methods are used to detect transitions between contrasting parts. Third, homogeneity-based methods are used to determine passages that are consistent with respect to some musical property. From a more technical point of view, Peeters [9], [17] coined the first class as *sequence approaches*, and the second and third classes as *state approaches*. In all three cases, one has to account for different musical dimensions such as melody, harmony, rhythm, or timbre. In this paper, we contribute to repetition-based music structure analysis using chroma-based audio features that correlate to aspects of harmony and melody. In particular, we extend and improve on a sequence approach as introduced in [9]. In the remainder of this section, we summarize the main principles of repetition-based audio structure analysis, introduce some general notation, and discuss in more detail the relation of our approach to previous work.

In the following, we distinguish between a piece of music (in an abstract sense) and a particular audio recording (an actual performance) of the piece. The term *part* is used in the context of the abstract music domain, whereas the term *segment* is used for the audio domain [1]. Musical parts are often denoted by the letters $A, B, C, \ldots$ in the order of their first occurrence, where indices are used to indicate repetitions. For example, the sequence $A_1A_2B_1B_2CA_3B_3B_4$ describes the *musical form* of the Hungarian Dance No. 5 by Johannes Brahms. The structure of this piece, as given as a recording by Ormandy and serving as our running example, is shown in Fig. 1. The musical form consists of three repeating A-parts, four repeating B-parts, as well as a C-part. Hence, given the Ormandy recording, the goal of the structure analysis problem considered in this paper is to find the segments within the recording that correspond to the A-part or to the B-part.

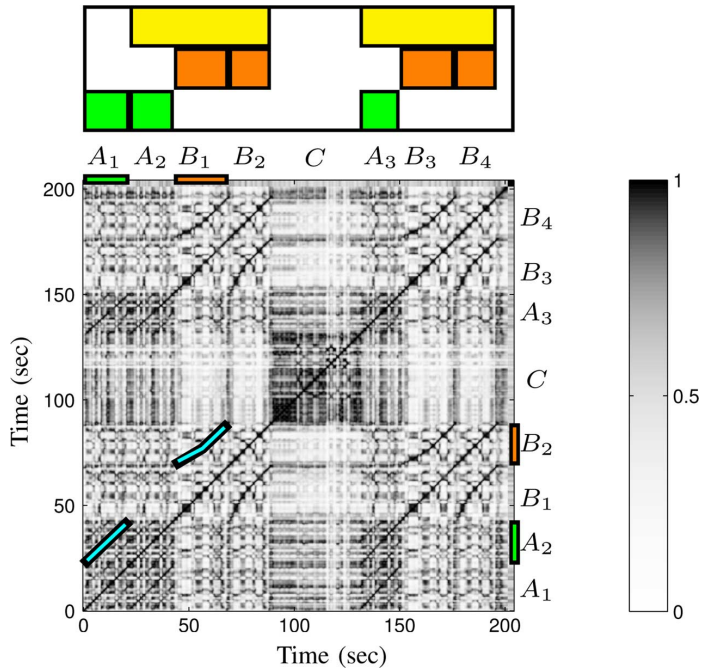


Fig. 1. Musical form and self-similarity matrix of a recording by Ormandy of Brahms' Hungarian Dance No. 5.

Most approaches to repetition-based structure analysis proceed as follows. First, the music recording is transformed into a sequence $X := (x_1, x_2, \ldots, x_N)$ of suitable feature vectors $x_n \in \mathcal{F}, n \in [1:N] := \{1, 2, \ldots, N\}$, where $\mathcal{F}$ denotes a suitable feature space. In the following, a *segment* α is understood to be a subset $\alpha = [s:t] \subseteq [1:N]$ specified by its start point s and its end point t. Based on a similarity measure $\mathbf{s} : \mathcal{F} \times \mathcal{F} \to \mathbb{R}$, one obtains a *self-similarity matrix* (SSM) denoted by $\mathcal{S} \in \mathbb{R}^{N \times N}$ and defined by $\mathcal{S}(n,m) := \mathbf{s}(x_n, x_m), 1 \leq n, m \leq N$. In the following, a tuple $(n,m) \in [1:N]^2$ is called a *cell* of $\mathcal{S}$, and the value $\mathcal{S}(n,m)$ is referred to as the *score* of the cell (n,m). The score value $\mathcal{S}(n,m)$ is high (dark color in Fig. 1) if the two feature vectors x_n and x_m are similar, otherwise $\mathcal{S}(n,m)$ is low (light color). The crucial observation is that repeating patterns in the feature sequence X appear as parallel "stripes" in $\mathcal{S}$, see [1], [3]. More precisely, these stripes are encoded by paths of cells of high score running roughly in parallel to the main diagonal. Each of the paths encodes the similarity of a pair of segments, which are given by the two projections of the path onto the vertical and horizontal axis of $\mathcal{S}$, respectively. This is also illustrated by Fig. 1, where two such paths are highlighted within

$\mathcal{S}$. One of the paths encodes the similarity between the audio segments corresponding to A_1 and A_2, and the other path encodes the similarity between the segments corresponding to B_1 and B_2. Note that the first path is exactly parallel to the main diagonal, indicating that the parts A_1 and A_2 are played in the same tempo, whereas the second path is curved, indicating that the parts B_1 and B_2 are played in different tempi. In fact, in the Ormandy interpretation, the B_2-part is played much faster than the B_1-part. This fact is also revealed by the gradient of the path, which encodes the relative tempo difference.

To identify repetitions, most approaches extract the path structure from an SSM and apply a clustering step to the pairwise relations obtained from the paths in order to derive entire groups of mutually similar segments. For example, one group contains all A-part segments, another all B-part segments. This step can be considered as forming some kind of transitive closure of the path relations [5], [8], [12]. However, note that strong musical and acoustic variations may lead to rather noisy and fragmented path structures, which makes both steps—path extraction as well as grouping—error-prone and fragile. In [4], a grouping process is described that balances out inconsistencies in the path relations by exploiting a constant tempo assumption. However, when dealing with varying tempo, the grouping process constitutes a challenging research problem [5], [7]. As opposed to previous approaches, the idea of our approach is to jointly perform the path extracting and grouping step. We realize this idea by assigning a fitness value to a given segment in such a way that all existing relations within the entire recording are simultaneously accounted for. In other words, instead of extracting individual paths, we extract entire groups of paths, whereby consistency properties within a group are automatically enforced by our construction. The general idea of assigning some kind of fitness value for each segment of the audio recording is not new and has already been formulated by Cooper and Foote [14]. In this early work, the authors calculate the fitness of a given segment as the normalized sum of the self-similarity between the segment and the entire recording, which can be thought of as some sort of "summary score." The thumbnail is then defined to be the fitness-maximizing segment. Also, a visualization of the fitness over all possible segments has been indicated in [14]. As one main limitation, the fitness measure does not take any path relations into account, thus yielding only limited information on the repetitiveness of a segment.

To capture the repetitive structure, Peeters [9] has introduced a fitness measure referred to as "likelihood." To compute this measure, a binary-valued diagonal path structure is extracted from an SSM. Then, for a given segment (called the "candidate mother segment"), the likelihood is defined as the sum of the lengths of all segments explained by diagonal paths over the candidate mother segment, where overlaps between repeating instances are prevented by suitable constraints. Finally, the thumbnail (referred to as the "mother segment") is defined as the candidate mother segment of maximal likelihood.

Our fitness measure builds upon and extends the pioneering work described in [9], [14] in various ways. First, instead of an explicit extraction of the binary-valued path structure, we avoid such a hard decision by working on a real-valued SSM

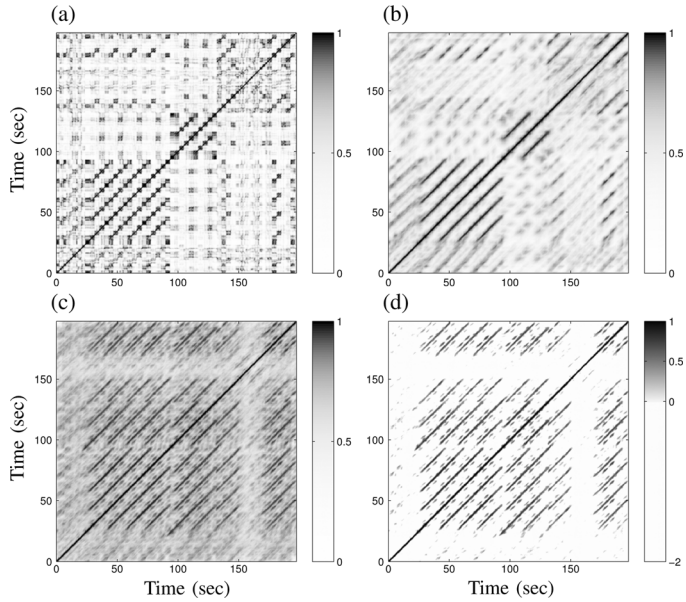


Fig. 2. Self-similarity matrices for the song "In the year 2525" by Zager and Evans, which has several transposed repetitions in its second half. **(a)** Initial self-similarity matrix. **(b)** Path-enhanced matrix. **(c)** Transposition-invariant matrix. **(d)** Thresholded matrix with $\delta = -2$.

(even though we also apply some thresholding for denoising purposes). Second, using a variant of dynamic time warping instead of looking for diagonal paths, our approach allows for handling tempo differences between repeating segments. Third, we combine a coverage criterion (which is similar to the likelihood in [9]) with a score criterion to define a fitness measure that balances out two contradicting principles (large coverage versus high average score). Fourth, introducing suitable normalization steps, where we disregard trivial self-explanations similar to [18], our fitness measure is well-suited to compare repetition properties of segments of different lengths. Finally, we introduce a unifying mathematical framework and an optimization scheme based on dynamic programming to efficiently compute the fitness measure.

We would like to point out that our work has also been inspired by Paulus and Klapuri [10], even though the task and concepts of this paper are fundamentally different from [10]. The fitness measure introduced in [10] expresses properties of an *entire structure*, whereas our fitness measure expresses properties of a *single segment*.

III. Self-Similarity Matrices

The degree of similarity between two repeating segments crucially depends on the feature type, the similarity measure, and on how the self-similarity matrix $\mathcal{S}$ is defined and post-processed. In this section, we describe the structure-enhanced and transposition-invariant SSM used in our experiments. Since the construction of the SSM is not the focus of this paper, we only give a brief description and revert to existing literature, see also Fig. 2 for an overview. From a technical point of view, our fitness measure is generic in the sense that it works with general self-similarity matrices that fulfill the normalization properties $\mathcal{S}(n,m) \leq 1$ for $n,m \in [1:N]$ and $\mathcal{S}(n,n) = 1$ for $n \in [1:N]$.

First of all, we convert the audio signal into twelve-dimensional chroma-based audio features, which closely correlate to the aspect of harmony and have become a widely used tool in processing and analyzing music data [2], [8], [9], [19]–[22]. In our experiments, we use a chroma variant referred to as CENS (Chroma Energy Normalized Statistics) ([8] Section 3.3)[1] with a feature rate of 2 Hz (two features per second). This resolution has turned out to be suitable not only for audio structure analysis [7], but also for related tasks such as cover song identification [22] and audio matching [23]. Normalizing the features, we use the inner product as a similarity measure to compute a self-similarity matrix $\mathcal{S}$, see Fig. 2(a).

To further enhance the path structure of $\mathcal{S}$, one typical procedure is to apply some kind of smoothing filter along the direction of the main diagonal, resulting in an emphasis of diagonal information in $\mathcal{S}$ and a denoising of other structures, see [2], [9], [21], [24] and Fig. 2(b) for an illustration. In our implementation, we use a smoothing variant as described in [21], which can deal with local tempo variations. Furthermore, to account for transpositions between related segments, we adopt the concept of *transposition-invariant* SSMs as introduced in [25]. The idea is to first compute the similarity between the original feature sequence and each of the twelve cyclically shifted versions of the chroma feature sequence [4], resulting in twelve similarity matrices. Then, the transposition-invariant SSM is calculated by taking the point-wise maximum over these twelve matrices, see Fig. 2(c).

In view of the subsequent application, we further process the SSM by suppressing all values that fall below a given threshold. Using a suitable threshold parameter $\tau > 0$ and a penalty parameter $\delta \leq 0$, we first set the score values of all cells with a score below τ to the value δ and then linearly scale the range $[\tau : 1]$ to [0:1], see Fig. 2(d). The thresholding introduces some kind of denoising, whereas the parameter δ imposes some penalty on all cells of low score. Intuitively, we want the relevant path structure to lie in the positive part of the resulting SSM, whereas all other cells are given a negative score. Finally, to enforce the normalization properties needed in our fitness construction, we set $\mathcal{S}(n,n) = 1$ for $n \in [1 : N]$ (this property may have been lost in the smoothing process due to boundary effects). Note that different methods can be used for thresholding [24]. In the following, we choose the threshold in a relative fashion by keeping $\rho \cdot 100\%$ of the cells having the highest score and set $\delta = -2$. The role of the parameters will be further explained and investigated in Section V.A3.

IV. Fitness Measure

In this section, we introduce and discuss our novel fitness measure. In assigning a fitness value to a given segment α, our idea is to simultaneously account for its relations to other segments of the audio recording. Extending the approach [9] based on diagonal paths, we introduce the concept of a path family over α, which allows for expressing repetitive relations even in the presence of tempo differences (Section IV.A). We then look for an optimal path family over α from among many possible path families. To compute such an optimal path family, we introduce an efficient algorithm based on dynamic programming (Section IV.B). Next, in Section IV.C, we explain how to assign a coverage value as well as an average score value to a given path family. The fitness of the segment α is then defined by the harmonic mean of coverage and score of the optimal path family over α. As for the thumbnail, we choose the segment having maximal fitness over all possible segments (Section IV.D). As another main contribution, we introduce a scape plot visualization that indicates the fitness values of all possible audio segments (Section IV.E). Finally, we discuss basic properties of our fitness concept and offer a number of illustrative examples (Section IV.F).

A. Path Family

Let $X = (x_1, x_2, \ldots, x_N)$ be a feature sequence and $\mathcal{S}$ a self-similarity matrix as introduced in Section III. Furthermore, let $\alpha = [s : t] \subseteq [1 : N]$ be a segment with $|\alpha| := t - s + 1$ denoting its length. For later usage, we define a *segment family* of size K to be a set

$$\mathcal{A} = \{\alpha_1, \alpha_2, \ldots, \alpha_K\} \tag{1}$$

of pairwise disjoint segments: $\alpha_k \cap \alpha_j = \emptyset$ for $k, j \in [1 : K], k \neq j$. Let $\gamma(\mathcal{A}) := \sum_{k=1}^{K} |\alpha_k|$ be the *coverage* of $\mathcal{A}$.

Next, a *path* over α of length L is a sequence $p = ((n_1, m_1), \ldots, (n_L, m_L))$ of cells $(n_\ell, m_\ell) \in [1 : N]^2$, $\ell \in [1 : L]$, satisfying $m_1 = s$ and $m_L = t$ (boundary condition) and $(n_{\ell+1}, m_{\ell+1}) - (n_\ell, m_\ell) \in \Omega$ (step size condition), where Ω denotes a set of admissible step sizes. In our setting, we use $\Omega = \{(1,2), (2,1), (1,1)\}$, which constrains the slope of a path within the bounds of 1/2 and 2, see ([8] Chapter 4). For a path p, we associate two segments defined by the projections $\pi_1(p) := [n_1 : n_L]$ and $\pi_2(p) := [m_1 : m_L]$, respectively. Note that the boundary condition enforces $\pi_2(p) = \alpha$. The other segment $\pi_1(p)$ is referred to as an *induced segment*, see Fig. 3 for examples. The *score* $\sigma(p)$ of p is defined as

$$\sigma(p) = \sum_{\ell=1}^{L} \mathcal{S}(n_\ell, m_\ell). \tag{2}$$

Note that each path over the segment α encodes a relation between α and an induced segment, where the score $\sigma(p)$ yields a kind of quality measure for this relation. Next, we introduce the concept of a *path family* over α, which is defined to be a set $\mathcal{P} := \{p_1, p_2, \ldots, p_K\}$ of size K, consisting of paths p_k over α with $k \in [1 : K]$. Furthermore, as an additional condition, we require the induced segments to be pairwise disjoint or, in other words, the set $\{\pi_1(p_1), \ldots, \pi_1(p_K)\}$ to be a segment family. This condition ensures that there are no overlaps between related segments as also required in the approach by Peeters [9], see also Fig. 3 for an illustration of this condition. Next, extending the definition in (2), the *score* $\sigma(\mathcal{P})$ of the path family $\mathcal{P}$ is defined as

$$\sigma(\mathcal{P}) := \sum_{k=1}^{K} \sigma(p_k). \tag{3}$$

[1]An implementation of these features is available at www.mpi-inf.mpg.de/resources/MIR/chromatoolbox/, see also [19].

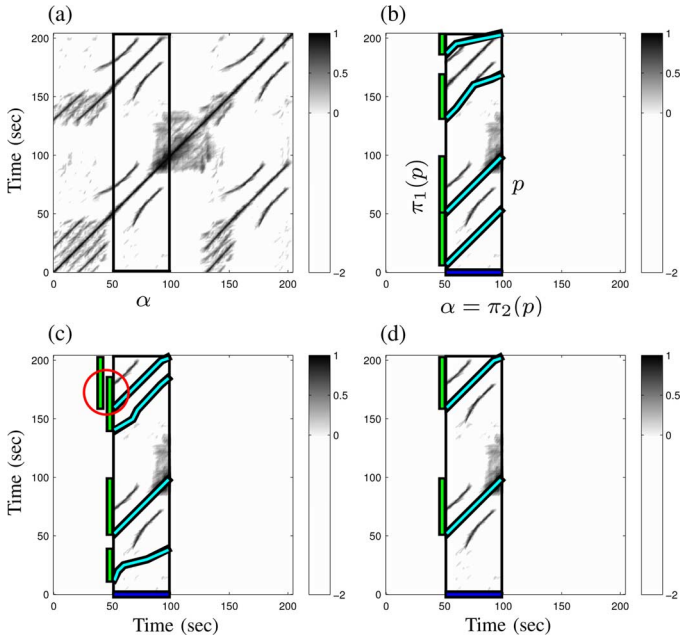


Fig. 3. SSM of our Brahms example with various paths over the segment $\alpha = [50 : 100]$. The induced segments are indicated on the vertical axis. **(a)** SSM. **(b)** Paths forming a path family. **(c)** Paths not forming a path family (induced segments overlap). **(d)** Paths forming an optimal path family.

As is also indicated by Fig. 3, there are in general a large number of possible path families over α. Among these path families, let

$$\mathcal{P}^* := \arg\max_{\mathcal{P}} \sigma(\mathcal{P}) \tag{4}$$

denote an optimal path family of maximal score. In the following, the family consisting of the segments induced by the paths of $\mathcal{P}^*$ will be referred to as the *induced segment family* (of $\mathcal{P}^*$ or of α). As will be shown in Section IV.B, the optimal path family $\mathcal{P}^*$ can be computed efficiently using dynamic programming. In Section IV.C, we explain how our fitness measure is derived from the score $\sigma(\mathcal{P}^*)$ and the induced segment family of $\mathcal{P}^*$.

B. Optimization Scheme

We now describe an efficient algorithm for computing an optimal path family for a given segment in a running time that is linear in the product of the length of the feature sequence and the length of the segment. Our algorithm is based on a modification of classical dynamic time warping (DTW) as originally developed for speech processing [26] and also extensively used in music processing [8], [27]. Given two sequences, say $X := (x_1, x_2, \ldots, x_N)$ and $Y := (y_1, y_2, \ldots, y_M)$, the objective of classical DTW is to compute an optimal path that *globally* aligns X and Y, where the first elements as well as the last elements of the two sequences are to be aligned. The step size condition as specified by the set Ω constraints the slope of the path. Furthermore, using $\Omega = \{(1,2), (2,1), (1,1)\}$, as in our case, each element of X is aligned to at most one element of Y (and vice versa).

Now, when computing an optimal path family over a given segment $\alpha = [s : t] \subseteq [1 : N]$, $M := |\alpha|$, the role of Y is taken over by α, and the conditions change compared to classical DTW. In particular, α can be simultaneously aligned to several

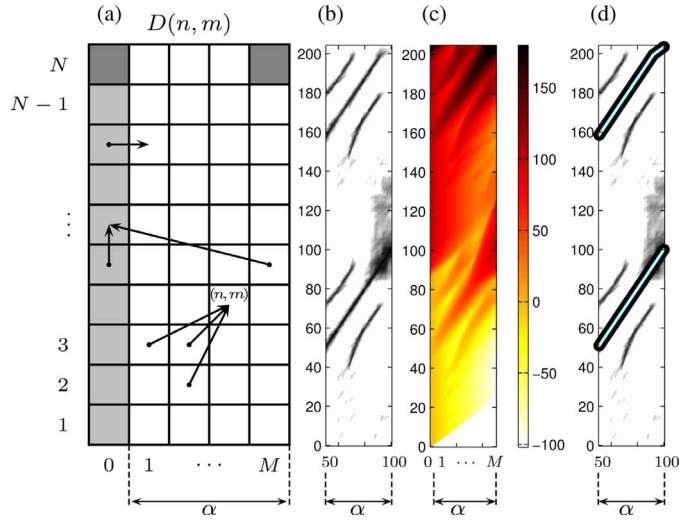


Fig. 4. **(a)** Illustration of the various predecessors in computing the accumulated score matrix. **(b)** Submatrix $\mathcal{S}^\alpha$ with $\alpha = [50 : 100]$ of the SSM shown in Fig. 3. **(c)** Accumulated score matrix D. **(d)** Optimal path family.

(non-overlapping) subsequences of X. However, for each such subsequence, the entire segment α is to be aligned. Furthermore, certain sections of X may be left completely unconsidered in the alignment. To account for these new constraints, we introduce additional steps that allow us to skip certain sections of X and to jump from the end to the beginning of the given segment α, see also Fig. 4 for an illustration. First, we define the $N \times M$ submatrix $\mathcal{S}^\alpha$ by taking the columns s to t of $\mathcal{S}$. Next, we define an accumulated score matrix $D \in \mathbb{R}^{N,M+1}$ (with rows indexed by $[1 : N]$ and columns indexed by $[0 : M]$), by the following recursion:

$$D(n,m) = \mathcal{S}^\alpha(n,m) + \max\{D(i,j) | (i,j) \in \Phi(n,m)\} \tag{5}$$

for $n \in [2 : N]$ and $m \in [2 : M]$, where $\Phi(n,m) = \{(n - i, m - j) | (i,j) \in \Omega\} \cap [1 : N] \times [1 : M]$ denotes the set of possible predecessors. So far, this is similar to the classical DTW algorithm. The constraint conditions and additional jump steps are realized by the definition of the values of D for the remaining index pairs (n,m) with $n = 1$ and $m \in \{0, 1\}$. The first column of D indexed by $m = 0$, which plays a special role, is recursively defined by $D(1,0) = 0$ and

$$D(n,0) = \max\{D(n-1,0), D(n-1,M)\} \tag{6}$$

for $n \in [2 : N]$. First, the term $D(n-1,0)$ enables the algorithm to move upwards without accumulating any (possibly negative) score, thus realizing the condition that sections of X may be skipped without penalty (negative score). Second, the term $D(n-1,M)$ closes up a path (ensuring that the entire segment α is aligned to a subsequence of X), while ensuring that the next possible segment does not overlap with the previous segment. The second column of D indexed by $m = 1$ is defined by

$$D(n,1) = D(n,0) + \mathcal{S}^\alpha(n,1) \tag{7}$$

for $n \in [1 : N]$, which starts a new path. Finally, to complete the initialization, we set $D(1,m) = -\infty$ for $m \in [2 : M]$,

which forces the first path to start with the first element of α. The score of an optimal path family is then given by

$$\sigma(\mathcal{P}^*) = \max\{D(N,0), D(N,M)\}. \tag{8}$$

The first term $D(N,0)$ reflects the case that the final section of X may be skipped, and the second term $D(N,M)$ ensures that in the other case the entire segment α is aligned to a suffix of X. The associated optimal path family $\mathcal{P}^*$ can be constructed from D using a back-tracking algorithm as in classical DTW, see ([8] Chapter 2) for details. As the only modification, the cells of $\mathcal{S}^\alpha$ that belong to the first auxiliary column (indexed by $m = 0$) are to be omitted to obtain the final path family. Obviously, the presented algorithm has a complexity (in terms of memory requirements and running time) of $O(MN)$.

C. Definition of Fitness Measure

We now give a formal definition of our fitness measure. At this point, we only assume that the given SSM $\mathcal{S} \in \mathbb{R}^{N\times N}$ has the property that $\mathcal{S}(n,m) \leq 1$ for all cells $(n,m) \in [1:N]^2$, and $\mathcal{S}(n,n) = 1$ for $n \in [1:N]$.

For the segment α, let $\mathcal{P}^* = \{p_1, \ldots, p_K\}$ be an optimal path family. In view of our fitness measure, the score $\sigma(\mathcal{P}^*)$ does not yet have the desired properties, since it also depends on the lengths of the paths and captures trivial self-explanations. For example, the segment $\alpha = [1:N]$ explains the entire sequence X perfectly, which is a trivial fact. More generally, each segment α explains itself perfectly (this information is encoded by the main diagonal of a self-similarity matrix). Therefore, to disregard such trivial self-explanations, we simply subtract the length $|\alpha|$ from the score $\sigma(\mathcal{P}^*)$. Furthermore, we normalize the score with regard to the lengths L_k of the paths p_k (see Section IV.A) contained in the optimal path family $\mathcal{P}^*$. This yields the *normalized score* $\bar{\sigma}(\alpha)$ defined by

$$\bar{\sigma}(\alpha) := \frac{\sigma(\mathcal{P}^*) - |\alpha|}{\sum_{k=1}^{K} L_k}. \tag{9}$$

From the assumption $\mathcal{S}(n,n) = 1$, we obtain $\bar{\sigma}(\alpha) \geq 0$. Furthermore, note that, when using $\Omega = \{(1,2),(2,1),(1,1)\}$, we get $\sum_k L_k \leq N$. This together with $\mathcal{S}(n,m) \leq 1$ implies the property $\bar{\sigma}(\alpha) \leq 1 - |\alpha|/N$. Intuitively, the value $\bar{\sigma}(\alpha)$ expresses the *average score* of the optimal path family $\mathcal{P}^*$ (minus a proportion for the self-explanations).

Next, we define some kind of *coverage* measure for α. To this end, let $\mathcal{A}^* := \{\pi_1(p_1), \ldots, \pi_1(p_K)\}$ be the segment family induced by $\mathcal{P}^*$, and let $\gamma(\mathcal{A}^*)$ be its coverage as defined in Section IV.A. Similar to the normalized score, we define the *normalized coverage* $\bar{\gamma}(\alpha)$ by

$$\bar{\gamma}(\alpha) := \frac{\gamma(\mathcal{A}^*) - |\alpha|}{N}. \tag{10}$$

As above, the length $|\alpha|$ is subtracted to compensate for trivial coverage. Obviously, one has $\bar{\gamma}(\alpha) \leq 1 - |\alpha|/N$.

To combine the coverage and the score measure, we define the *fitness* $\varphi(\alpha)$ of the segment α to be the harmonic mean

$$\varphi(\alpha) := 2 \cdot \frac{\bar{\sigma}(\alpha) \cdot \bar{\gamma}(\alpha)}{\bar{\gamma}(\alpha) + \bar{\sigma}(\alpha)}. \tag{11}$$

In doing so, the fitness integrates the normalized score and coverage into a single measure, inheriting the property $\varphi(\alpha) \leq 1 - |\alpha|/N$ from $\bar{\sigma}(\alpha)$ and $\bar{\gamma}(\alpha)$.

In conclusion, note that our normalization neglects trivial self-explanation and allows for comparing segments of different length while slightly favoring shorter segments. To illustrate the last property, suppose that a piece has the musical form $A_1 A_2 \ldots A_6$. Then $\varphi(\alpha) = 5/6$ when α corresponds to A_1, $\varphi(\alpha) = 2/3$ when α corresponds to $A_1 A_2$, $\varphi(\alpha) = 1/2$ when α corresponds to $A_1 A_2 A_3$, and $\varphi(\alpha) = 0$ when α corresponds to the entire piece.

D. Thumbnail Selection

The fitness measure can be directly applied to audio thumbnailing. Similar to [9], [14], the basic idea is to define the thumbnail to be the segment of maximal fitness:

$$\alpha^* := \arg\max_{\alpha} \varphi(\alpha). \tag{12}$$

Note that the induced segment family of α^* reveals the repetitive structure of the thumbnail. To account for prior knowledge and to remove spurious estimates, one can impose additional requirements on the thumbnail solution. In particular, as also shown by our experiments in Section V, introducing a lower bound θ for the minimal possible thumbnail length allows us to reduce the effect of noise scattered in the underlying self-similarity matrix. Extending the above definition, we define

$$\alpha^*_\theta := \arg\max_{\alpha, |\alpha| \geq \theta} \varphi(\alpha). \tag{13}$$

In the next sections, we discuss and illustrate the properties of the fitness measure and the thumbnailing procedure in detail.

E. Fitness Scape Plot

We next introduce a compact fitness representation for the entire music recording, showing the fitness $\varphi(\alpha)$ for all possible segments α. Note that each segment $\alpha = [s:t] \subseteq [1:N]$ can be represented by its center $c(\alpha) := (s+t)/2$ and its length $|\alpha|$. Using the center to parameterize a horizontal axis and the length to parametrize the height, each segment corresponds to a point in some triangular representation, also referred to as a *scape plot*. Such scape plots were originally introduced by Sapp [28], [29] to represent harmony in musical scores in a hierarchical way. In our context, we define a scape plot Δ by setting $\Delta(c(\alpha), |\alpha|) := \varphi(\alpha)$ for segment α. Fig. 5(a) shows a color-coded representation of the scape plot for our Brahms example. Note that the maximal entry of Δ corresponds to the maximal fitness value, thus defining the thumbnail α^*. Furthermore, the segments with $|\alpha| \geq \theta$ correspond to all points in Δ that lie above a horizontal line with its height specified by θ.

F. Discussion of Properties

Before we give a quantitative evaluation of our thumbnailing procedure in Section V, we first discuss some explicit examples to illustrate the properties and potential of our fitness measure, the scape plot, and the derived segmentations.

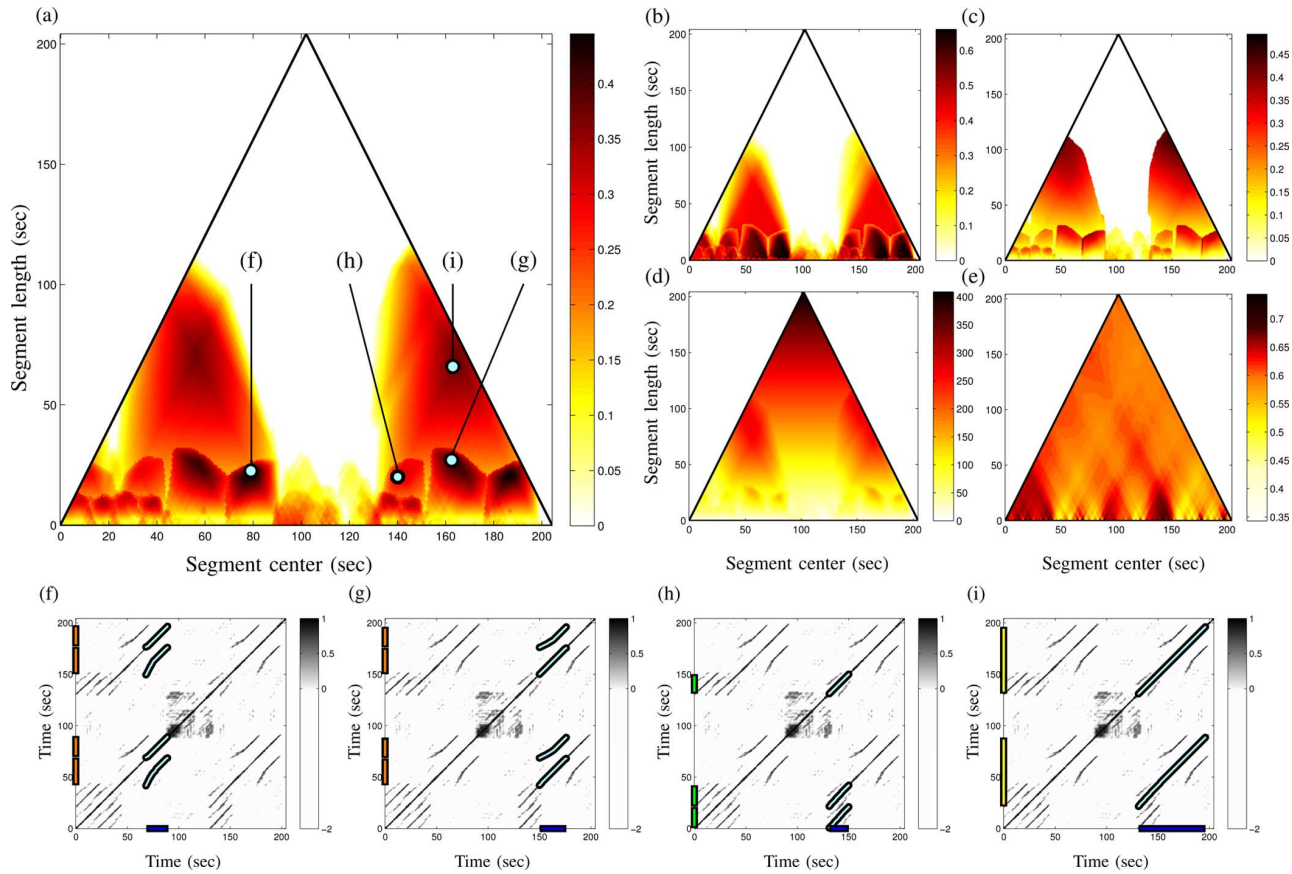


Fig. 5. Various scape plot representations as well as different optimal path families and induced segment families over different segments α for our Brahms example. **(a)** Fitness measure (harmonic mean). **(b)** Normalized score. **(c)** Normalized coverage. **(d)** Score. **(e)** Average measure as suggested in [14]. **(f)** $\alpha = \alpha^* = [68:89]$ (thumbnail, maximal fitness, corresponding to B_2). **(g)** $\alpha = [150:176]$ (corresponding to B_3). **(h)** $\alpha = [131:150]$ (corresponding to A_3). **(i)** $\alpha = [131:196]$ (corresponding to $A_3B_3B_4$).

We first continue with our Brahms example with the musical form $A_1A_2B_1B_2CA_3B_3B_4$, see Fig. 1. The fitness scape plot of the Ormandy recording of this piece, as shown in Fig. 5(a), reflects this structure in a hierarchical way. The fitness-maximizing segment is $\alpha^* = [68:89]$ ($c(\alpha) = 78.5, |\alpha| = 22$) and corresponds to B_2. Furthermore, the induced segment family consists of the four B-part segments, see Fig. 5(f). Note that all four B-part segments have almost the same fitness and lead to more or less the same segment family. For example, Fig. 5(g) shows the induced segment family, when considering the segment corresponding to B_3. This reflects the fact that each of the B-part segments may serve equally well as the thumbnail. Recall that our fitness measure slightly favors shorter segments. Therefore, since in this recording the B_2-part is played faster than the B_3-part, our fitness measure favors the B_2-part segment to the B_3-part segment. The scape plot also reveals other local maxima of musical relevance. For example, the local maximum corresponding to segment $\alpha = [131:150]$ ($c(\alpha) = 140.5, |\alpha| = 20$) reflects part A_3, and the induced segment family reveals the three A-parts, see Fig. 5(h). Furthermore, the local maximum corresponding to segment $\alpha = [131:196]$ ($c(\alpha) = 163.5$, $|\alpha| = 66$) reflects $A_3B_3B_4$, which is a repetition of $A_2B_1B_2$, see Fig. 5(i). Again, note that, because of the normalization, the fitness of $\alpha = [131:196]$ is well below the one of, e.g., the thumbnail $\alpha^* = [68:89]$.

Next, we illustrate that in the definition of the fitness measure, see Equation (11), the combination of the normalized score and coverage is of crucial importance. Fig. 5(b) shows the scape plot with only the normalized score. Since this measure only expresses the average score of a path family without expressing how much of the audio material is actually captured, many of the small segments have a relatively high score. Using such a measure would typically result in false positive segments of short length. In contrast, using only the normalized coverage would typically favor longer segments, see Fig. 5(c). Now, by combining score and coverage, our fitness measure balances out the two conflicting principles of having firm repetitions (high score) and of explaining possibly large portions of the recording (high coverage). Next, Fig. 5(d) illustrates the importance of normalization and subtraction of self-explanations. We obtain this scape plot by simply using the score $\sigma(\mathcal{P}^*)$ of the optimizing path family $\mathcal{P}^*$ over α. As a result, longer segments typically dominate the shorter segments, with the entire recording having maximal score. Finally, Fig. 5(e) shows a scape plot with the average score measure as proposed by [3], where each segment α is assigned the average score of all cells of the submatrix $\mathcal{S}^\alpha$ (the matrix $\mathcal{S}^\alpha$ is defined in Section IV.B).[2] This average score measure captures relatively coarse homogeneity proper-

[2]In the computation of the average score measure, we use the initial SSM without thresholding ($\rho = 1, \delta = 0$) as shown in Fig. 2(a). Actually, using enhanced matrices as shown in Fig. 2(c) yields similar results.

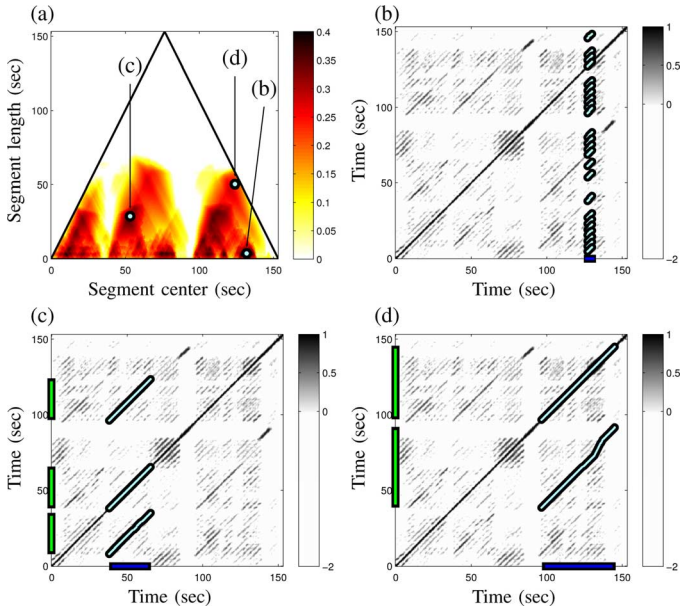


Fig. 6. Various optimal path families and induced segment families over different segments α for the Beatles song "Twist and Shout" having the musical form $IV_1V_2B_1V_3B_2O$. **(a)** Fitness scape plot. **(b)** $\alpha = \alpha^* = [127:130]$. **(c)** $\alpha = \alpha^*_\theta = [38:65]$ using $\theta = 10$ (corresponding to V_2). **(d)** $\alpha = \alpha^*_\theta = [97:145]$ using $\theta = 40$ (corresponding to V_3B_2).

TABLE I
OVERVIEW OF THE TWO EVALUATION DATASETS SHOWING THE RESPECTIVE NUMBER (NUM), AVERAGE DURATION (AV. DUR.) AND TOTAL DURATION (TOTAL DUR.) OF THE RECORDINGS. THE REMAINING NUMBERS ARE SPECIFIED IN THE TEXT

Dataset	Num [#]	Av. Dur. [sec]	Total Dur. [hh:mm]	$\lvert\mathcal{A}^{\mathrm{GT}}\rvert$ [#]	$\lvert\alpha^{\mathrm{GT}}\rvert$ [sec]	$\lvert\alpha^{\mathrm{GT}}\rvert$ [%]	$\gamma(\mathcal{A}^{\mathrm{GT}})$ [%]
BEATLES	175	162.1	7:52	4.0	22.1	14.4	54.1
MAZURKA	147	171.2	6:59	4.3	18.9	11.9	45.9

ties rather than repetitive structures.[3] As a result, the scape plot is less structured and only reveals high values for the A-parts that show a high degree of homogeneity with regard to the harmonic content (as captured by chroma-based audio features).

As a second example, Fig. 6 shows the scape plot and various induced segment families for the Beatles song "Twist and Shout." This song has the rough musical form $IV_1V_2B_1V_3B_2O$ consisting of a short intro (I-part), three verses (V-part), two bridges (B-part) and an outro (O-part). Interestingly, the fitness maximizing segment $\alpha^* = [127:130]$ is very short and leads to a large number of spurious induced segments, see Fig. 6(b). The reason is that the song contains a short harmonic phrase that is repeated over and over again. As a consequence, the self-similarity matrix contains many repeated spurious path fragments which, as a whole family, lead to a high score as well as to a high coverage value. To circumvent such problems, one can consider the segment α^*_θ as defined in (13) to enforce a minimal length for the thumbnail. For example, setting $\theta = 10$ (given in seconds) we obtain the segment $\alpha^*_\theta = [38:65]$ for our Beatles song, which corresponds to the verse V_2, see Fig. 6(c). This indeed yields a musically meaningful thumbnail. By further increasing the lower bound, we obtain superordinate repeating parts such as $\alpha^*_\theta = [97:145]$ corresponding to V_3B_2 (when using $\theta = 40$), see Fig. 6(d).

[3]The average score measure [3] was originally applied to timbre-related audio features, which tend to form homogeneity regions in an SSM.

V. APPLICATIONS AND EXPERIMENTS

In this section, we report on various experiments, which highlight the applicability and the performance of our fitness measure. In particular, we give a quantitative evaluation and discuss the benefits as well as the limitations of our approach in the context of an audio thumbnailing application for classical and popular music (Section V.A) and an audio segmentation application for folk song field recordings (Section V.B).

A. Thumbnailing

In a first series of experiments, we apply our fitness measure to determine the audio thumbnail. In the following, we describe the datasets (Section V.A1), introduce the evaluation measure (Section V.A2), investigate the role of various parameters (Section V.A3), compare our thumbnailing procedure to other approaches (Section V.A4), and finally discuss possible error sources (Section V.A5). In our evaluation, we rely on manually generated ground-truth (GT) annotations, which serve as references to compare against. Note that the term "ground-truth" as well as the usage of such annotations is problematic in the sense that different experts may disagree on how to segment and label the data. Therefore, it would be desirable to compare against annotations obtained by an entire panel of experts and to see if the results obtained by automated methods exhibit the same variability as the ones by the experts. Nevertheless, using annotations obtained by only one expert, the subsequent evaluation still indicates certain tendencies and illustrates the overall behavior of our procedure.

1) Datasets: We now describe two datasets consisting of popular and classical music recordings. The first dataset, denoted by BEATLES, consists of recordings from the 12 studio albums by "The Beatles" along with the structure annotations as described by [30]. Actually, for five of these songs, no clear repetitions were present in the annotations. Leaving out these songs in our experiments, BEATLES contains 175 recordings (instead of the original 180 songs). As a second dataset, denoted by MAZURKA, we use the complete recordings of the 49 Mazurkas composed by Frédéric Chopin in three different versions played by the pianists Rubinstein (1966), Cohen (1997), and Ezaki (2006), respectively. The resulting 147 files are a subset of the dataset of the Mazurka Project[4]. Furthermore, we use structure annotations generated by an expert on the basis of the musical scores. Then, using music synchronization techniques [31], these annotations were transferred automatically to the different recordings. Table I gives an overview of the two datasets, indicating the number of recordings, the average duration of the recordings, and the total duration. Furthermore, as explained in more detail in Section V.A2, the average number of GT thumbnails per song, the average duration (given in seconds and in percent) of GT thumbnails, and the average coverage (given in percent) of the induced segment families are specified.

[4]mazurka.org.uk

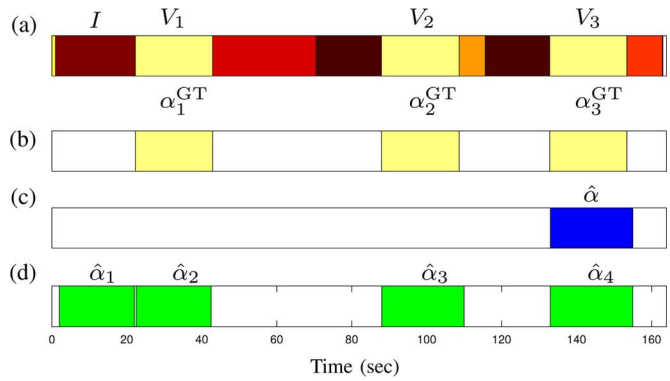


Fig. 7. Different segmentations for the Beatles song "Birthday." **(a)** GT structure annotation. **(b)** Induced segment family $\mathcal{A}^{\mathrm{GT}}$ of the GT thumbnail. **(c)** Estimated thumbnail $\hat{\alpha}$. **(d)** Induced segment family $\hat{\mathcal{A}}$ of $\hat{\alpha}$.

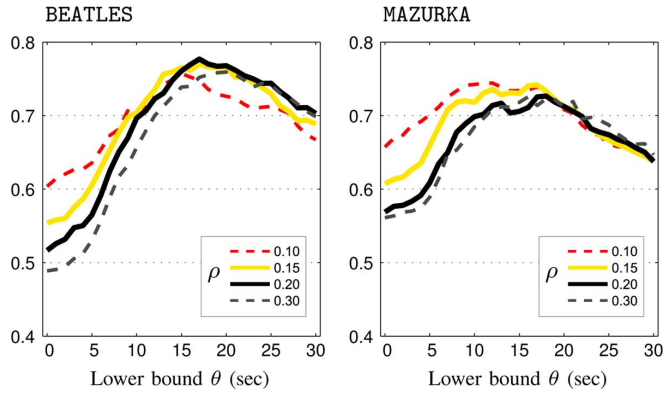


Fig. 8. Thumbnail F-measure values F^{α} in dependency of different parameters. The horizontal axis specifies the lower bound parameter θ (given in seconds) and the different colors correspond to different values for the relative threshold parameters ρ.

2) Evaluation Measures: For both datasets, there are structural annotations that consist of segmentations of the music recordings and a labeling of the segments by letters such as $A, B, C, \ldots$ representing musical parts. For each label, one can associate a segment family that consists of all segments marked with this label. In the thumbnailing scenario, we do not need the entire structure annotations, but only the label and associated segment family that represents the most repetitive musical part. This segment family then serves as the *GT thumbnail family*. To derive such a family, we compute for each labeled segment the normalized coverage for the associated segment family as in (10). Then, we take the labeled segment that maximizes the normalized coverage as the GT thumbail and the associated segment family as the GT thumbnail family. For example, in Fig. 7, the GT thumbnail family corresponds to the three verse segments labeled as V_1, V_2, and V_3. Note that each of these segments may serve equally well as the GT thumbnail. In general, let $\mathcal{A}^{\mathrm{GT}} := \{\alpha_1^{\mathrm{GT}}, \ldots, \alpha_K^{\mathrm{GT}}\}$ denote the GT thumbnail family representing the various possible GT thumbnails, see also Fig. 7(b).

Furthermore, let $\hat{\alpha}$ denote an estimated segment obtained from a given thumbnailing procedure. To measure how well the estimated thumbnail $\hat{\alpha}$ corresponds to the GT thumbnails, we compute the precision $\mathrm{P}_k^{\alpha} = (|\hat{\alpha} \cap \alpha_k^{\mathrm{GT}}|)/|\hat{\alpha}|$, the recall $\mathrm{R}_k^{\alpha} = (|\hat{\alpha} \cap \alpha_k^{\mathrm{GT}}|)/|\alpha_k^{\mathrm{GT}}|$, and the F-measure $\mathrm{F}_k^{\alpha} = 2\mathrm{P}_k^{\alpha}\mathrm{R}_k^{\alpha}/(\mathrm{P}_k^{\alpha} + \mathrm{R}_k^{\alpha})$ for each $k \in [1:K]$ and then define the *thumbnail F-measure* by

$$\mathrm{F}^{\alpha} = \max_{k \in [1:K]} \mathrm{F}_k^{\alpha}. \tag{14}$$

In other words, the thumbnail F-measure expresses to what extent $\hat{\alpha}$ agrees with one of the GT thumbnails contained in $\mathcal{A}^{\mathrm{GT}}$. As an example, Fig. 7 shows the case where the estimated thumbnail $\hat{\alpha}$ best agrees with α_3^{GT} (corresponding to V_3), yielding $\mathrm{F}^{\alpha} = 0.96$. Finally, let $\hat{\mathcal{A}} = \{\hat{\alpha}_1, \ldots, \hat{\alpha}_M\}$ be the induced segment family of $\hat{\alpha}$. For example, as shown by Fig. 7, the Beatles song "Birthday" has three segments annotated as verse, whereas the induced family of the estimated thumbnail consists of four segments. Actually, it turns out that the intro (segment labeled as I) is harmonically very similar to the verse segments[5].

3) Dependency on Parameters: We now examine the overall performance of our thumbnailing procedure. Note that it is not our intention to optimize and to advocate specific parameter settings. Instead, our main goal is to investigate the role and interdependencies of certain parameters as well as to indicate the conceptual benefits introduced by our fitness measure. In the following, we use the transposition-invariant SSM, as introduced in Section III and using the penalty parameter $\delta = -2$. Furthermore, there is the relative threshold parameter ρ (see Section III) used to suppress the small values in the SSM and the parameter θ (see Section IV.D) used to specify a minimum length for the chosen thumbnail.

Fig. 8 shows the thumbnail F-measure F^{α} for various choices of $\rho \in \{0.1, 0.15, 0.2, 0.3\}$ and different estimated thumbnails $\hat{\alpha} = \alpha_{\theta}^{*}$ using $\theta \in [0:30]$. Note that the results first improve with increasing θ (up to $\theta = 15$) and then deteriorate again when further increasing the lower bound. For example, in the case of the dataset BEATLES, the F^{α}-value is roughly 0.75 (or slightly above) when using $\theta = 17$, independent of the specific choice of ρ. As also illustrated by Fig. 6, using the lower bound θ allows us to disregard path families that consist of a large number of short spurious path fragments. This also explains why the role of ρ becomes more important for smaller θ: using smaller values for ρ removes more of the noise-like artifacts in the SSM that typically lead to spurious path fragments. Overall, it can be observed that the thumbnail F-measure is similar for an entire range of different parameter settings and that it also shows a similar behavior for the two datasets BEATLES and MAZURKA. This indicates that there has been no overfitting and that our general procedure does not depend on a specific parameter setting.

4) Comparison of Thumbnailing Procedures: For the subsequent experiments, we exemplary choose $\rho = 0.2$ in combination with $\theta = 15$. By comparing results obtained from different procedures, we now demonstrate that our fitness measure outperforms other measures in the thumbnailing context. Table II

[5]Actually, the intro is an instrumental version of the verse.

TABLE II
THUMBNAIL F-MEASURE F^{α} FOR VARIOUS SETTINGS (USING $\rho = 0.2$ AND $\theta = 15$). THE LAST FOUR SETTINGS SERVE AS BASELINE, WHERE THE SEGMENT LENGTH IS SPECIFIED USING PRIOR KNOWLEDGE

	BEATLES	MAZURKA
Fitness φ	0.761	0.711
Normalized score $\bar{\sigma}$	0.631	0.659
Normalized coverage $\bar{\gamma}$	0.618	0.565
Average score measure [14]	0.476	0.436
Baseline (entire song)	0.263	0.240
Baseline (second sixth)	0.620	0.526
Average score measure [14] (GT length)	0.555	0.516
Fitness (GT length)	0.775	0.836

shows the thumbnail F-measure for various settings. In the first four settings, we apply the same thumbnail selection strategy (using $\hat{\alpha} = \alpha^*_\theta$ as described in Section IV.D) based on four different measures: the fitness measure φ, the normalized score $\bar{\sigma}$, the normalized coverage $\bar{\gamma}$, and the average score measure suggested in [14], see also Fig. 5. For example, for BEATLES, we obtain $F^{\alpha} = 0.761$ when using φ and $F^{\alpha} = 0.476$ when using the average score measure. For both datasets, we obtain the best results when using φ, whereas the average score measure does not work well in this context. Also, using $\bar{\sigma}$ and $\bar{\gamma}$ separately does not yield the same quality as with their combined usage in φ.

For comparison, we also conducted a number of additional baseline experiments. First, we computed F^{α} when using the entire song as the estimated thumbnail. Second, splitting up each recording into six segments of equal length, we used the second segment as the estimated thumbnail ("second sixth").[6] Third, using the actual length of the GT thumbnail for each recording, we applied the thumbnailing procedure restricted to the respective GT length, once using the average score measure and then using our fitness measure. The last four rows of Table II show the resulting thumbnail F-measures for each of these four baseline procedures. Again, our fitness measure yielded better results than the average score measure. Furthermore, using the second sixth of a song yielded seemingly good results, e.g., $F^{\alpha} = 0.620$ for BEATLES. This relatively high number was not only a consequence of the datasets' statistics, but also a consequence of the rather "soft" nature of our evaluation measure.

5) Error Sources: We now describe further experiments based on a second, "harder" evaluation measure that reveals typical error sources in audio thumbnailing. As mentioned before, one main problem in the evaluation is that the concept of a thumbnail is generally ambiguous. In the more general structure analysis context, typical error sources and evaluation measures have been discussed by Lukashevich [32], who specifically addressed the problems of *over-segmentation* and *under-segmentation*. Over-segmentation typically occurs in the presence of subordinate structures, as illustrated by the Beatles song "While My Guitar Gently Weeps" (Fig. 9(a)), where the verse has a subordinate structure basically consisting of two repeated subparts. Under-segmentation typically occurs when there are superordinate repeated parts that may also define

[6]The motivation for using six segments is that this results in a segment length close to the average ground-truth thumbnail length, see Table I. Furthermore, using the second segment for each recording is motivated by the fact that many songs or Mazurkas start with an intro and then continue with a verse or first theme corresponding to the thumbnail.

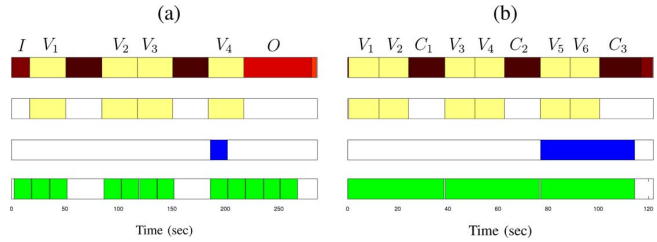


Fig. 9. GT segmentation and thumbnailing results as in Fig. 7 for (a) a song (BEATLES "While My Guitar . . .") with a subordinate structure and (b) a song (BEATLES "While My Guitar . . .") with superordinate structure.

TABLE III
ACCURACY RATES FOR FIVE DIFFERENT CASES (GT THUMBNAIL FAMILIES) AND DIFFERENT SETTINGS FOR DATASET BEATLES USING $\rho = 0.2, \theta = 15$, AND $F^{\alpha} \geq 0.8$. THE LAST COLUMN SHOWS THE TOTAL ACCURACY RATES SUMMED OVER THE FIVE CASES

	A	B	$A/2$	$B/2$	Sup	Total
Fitness φ	0.55	0.01	0.06	0.01	0.19	0.83
Normalized score $\bar{\sigma}$	0.34	0.10	0.25	0.02	0.01	0.72
Normalized coverage $\bar{\gamma}$	0.15	0.02	0.04	0.00	0.12	0.33
Average score measure [14]	0.18	0.06	0.18	0.01	0.01	0.45
Baseline (entire song)	0.00	0.00	0.00	0.00	0.00	0.00
Baseline (second sixth)	0.17	0.05	0.03	0.00	0.09	0.34

meaningful thumbnails (in terms of being the element maximizing the normalized coverage) as illustrated by the Beatles song "For No One" (Fig. 9(b)).

Using the dataset BEATLES, [7] we now give a quantitative evaluation that not only shows how often these phenomena actually occur, but also sheds a different light on the results presented in Table II. To this end, we introduce a different, "harder" evaluation method. Given a recording and a ground-truth family, we say that an estimated thumbnail $\hat{\alpha}$ is *correct* if the resulting thumbnail F-measure lies above a certain threshold, otherwise it is *incorrect*. In our experiments, we use the criterion $F^{\alpha} \geq 0.8$, basically saying that the estimated thumbnail must agree with a GT thumbnail by at least 80%. Then, as for the evaluation, we simply count the songs with correctly estimated thumbnails and divide this number by the total number of songs, leading to a proportion we refer to as *accuracy rate*.[8] We compute the accuracy rate using different segment families to serve as the ground-truth: the segment family maximizing the normalized coverage (this case was also considered before and is denoted as case A), the segment family having the second highest normalized coverage (case B), the segment families corresponding to the halves of the parts (subordinate structures, case $A/2$ and case $B/2$), and segment families corresponding to superordinate structures using suitable N-grams of labels (case Sup).

Table III shows the accuracy rates for the five different cases and for different settings. For example, using the thumbnailing procedure based on our fitness measure, the GT thumbnail A was identified in 55% of the songs, whereas a superordinate structure was identified in 19% of the songs. In total, the procedure delivered a thumbnail corresponding to one of the five cases in 83% of the songs. For the other settings, the results

[7]For MAZURKA the numbers are quite similar.

[8]Naturally, using a stricter (softer) criterion such as $F^{\alpha} \geq 0.85$ ($F^{\alpha} \geq 0.75$) leads to lower (higher) accuracy rates. The overall tendencies of the accuracy rates are similar for different choices.

get significantly worse. For example, the average score measure leads to a success rate of only 45%, when admitting all five cases. As mentioned above, using the normalized score tends to favor shorter segments (note the accuracy rate of 25% for the case $A/2$), whereas using the normalized coverage tends to favor longer segments (note the accuracy rate of 12% for the case Sup). Finally, in the last two rows of Table III one finds the accuracy rates for the two baseline methods as used in Table II. By no surprise, using the "harder" counting measure (instead of a "softer" thumbnail F-measure F^{α}), the accuracy rate is 0 when using the entire song as the thumbnail. Also, using the "second sixth" of a recording as the thumbnail leads to a rather poor accuracy rate of only 34%. In conclusion, Table III again shows that the fitness measure constitutes a valuable tool for audio thumbnailing.

B. Folk Song Segmentation

As a second application scenario of our fitness measure, we now introduce an automated procedure for segmenting a folk song field recording into its constituent stanzas. Such songs basically consist of a number of verses, yielding the musical form $A_1 A_2 \ldots A_K$. The main challenge is that the songs are recorded under poor recording conditions and are performed by elderly non-professional singers from memory. As a result, for some recordings, there are continuous intonation shifts of several semitones across the various stanzas as well as interruptions and significant temporal and melodic variations. By using this challenging audio material, we particularly want to demonstrate two things. First, the concept of transposition-invariance of the SSM (Section III) is essential for handling the intonation shifts. Second, the concept of path families (Section IV.A) is required to cope with the significant temporal variations across repeating patterns.

1) Dataset: In the Netherlands, folk songs have been extensively collected and studied. A long-term effort to record these songs was started by Will Scheepers in the early 1950s and was continued by Ate Doornbosch until the 1990s [33]. As a result, a collection known as *Onder de groene linde* (OGL)[9] was created, which not only represents a part of the Dutch cultural heritage but also documents textual and melodic variations resulting from oral transmission. The OGL collection consists of 7277 audio recordings. Nearly all of the field recordings are monophonic and comprise a large number of stanzas (up to 34). In our experiments, we use 47 of these recordings with a total runtime of 156 minutes and a total of 465 stanzas. Therefore, on average, each recording has a duration of roughly 199 seconds and consists of 10 stanzas, yielding an average stanza duration of roughly 20 seconds. For each of the songs, an expert user manually determined the segment boundaries of the stanzas, which serve as the ground-truth segmentation in our evaluation. This dataset was also used in [34].

2) Segmentation and Evaluation Measures: As opposed to the recordings of the `BEATLES` and `MAZURKA` datasets (Section V.A1), a folk song recording basically consists of a repetition of a single musical part. Therefore, instead of simply looking for the thumbnail (which would be one of the stanzas), we consider a more general segmentation problem with the objective to recover all segment boundaries. In [34], a reference-based segmentation algorithm is described that relies on the availability of a manually transcribed reference stanza. The segmentation is then achieved by locally comparing the field recording with the reference stanza. In this paper, we introduce a reference-free segmentation procedure by applying our audio thumbnailing approach, where the automatically computed thumbnail takes over the role of the reference. More precisely, we compute the fitness-maximizing segment α^* (or α^*_θ) and look at the induced segment family. The boundaries of the segments of this family then yield the segmentation result. For the evaluation, as in [34], we check to what extent the 465 manually annotated stanzas have been identified correctly by the segmentation procedure. We say that a computed starting boundary is a *true positive* if it coincides with a ground truth boundary up to a small tolerance of ± 2 seconds. Otherwise, the computed boundary is referred to as a *false positive*. A ground truth boundary that does not coincide with a computed boundary is referred to as a *false negative*. We then compute precision P and recall R values and define the F-measure as $\mathrm{F} := 2 \cdot \mathrm{P} \cdot \mathrm{R}/(\mathrm{P} + \mathrm{R})$.

3) Experiments: To account for the special characteristics of the field recordings, we apply in this experiment two enhancement strategies proposed in [34]. As a first enhancement strategy, we employ *F0-enhanced chroma features*, which capture only parts of the signal that correspond to the fundamental frequency (F0) of the monophonic recordings, leading to an increased robustness against recording artifacts and background noise. The second enhancement strategy allows us to deal with intonation shifts on a finer level than the twelve semitones shifts used in the transposition-invariant SSMs (Section III). The idea is to use generalized chroma representations with 36 bins (instead of the usual 12 bins) as suggested in [20], [22]. Taking the pointwise maximum over the resulting 36 matrices, we obtain a transposition-invariant SSM, which shows increased robustness to more continuous pitch fluctuations.

TABLE IV
F-MEASURE VALUES FOR THE REFERENCE-FREE FOLK SONG SEGMENTATION PROCEDURE (AND THE REFERENCE-BASED METHOD [34] FOR COMPARISON) WHEN USING TRANSPOSITION INVARIANCE (COLUMN "TRANS") OR NOT (ROW "NON") AS WELL AS USING THE CONCEPT OF PATH FAMILIES (CASE "DTW") OR NOT ("DIAG")

		Diag		DTW	
Method	ρ	non	trans	non	trans
Reference-free	0.4	0.674	0.715	0.750	**0.870**
Reference-free	0.3	0.611	0.678	0.754	0.828
Reference-free	0.2	0.529	0.631	0.664	0.758
Reference-based [34]		-	-	0.774	**0.926**

Table IV shows F-measure values obtained for our reference-free segmentation procedure using the lower bound $\theta = 10$ as well as for the reference-based method [34] for comparison. First, we show that one main benefit of our path family concept is that strong temporal variations, as they occur in the field recordings, can be handled. To this end, we compare the segmentation results obtained by using two versions of our fitness measure. In the case of "Diag", we simulate that only diagonal

[9]OGL is part of the *Nederlandse Liederenbank*, see www.liederenbank.nl

paths can be extracted by constraining the step size condition to $\Omega = \{(1,1)\}$. Similar to [9], this does not allow for handling tempo changes. The case "DTW" refers to our path family concept using the set $\Omega = \{(1,2),(2,1),(1,1)\}$. Table IV reveals how crucial this additional degree of flexibility is. For example, using a relative threshold $\rho = 0.4$ without accounting for transpositions (column "non"), the segmentation procedure yields an F-measure $\mathrm{F} = 0.674$ for "Diag", whereas it is $\mathrm{F} = 0.750$ for "DTW". We also obtain similar improvements for other parameter settings.

Furthermore, handling intonation shifts is another crucial requirement for obtaining robust segmentations. Compared to the results obtained without transposition-invariant SSMs (e.g., $\mathrm{F} = 0.750$ for $\rho = 0.4$, "DTW", "non"), we observe a significant increase in the F-measure when using transposition-invariant SSMs (e.g., $\mathrm{F} = 0.870$ for $\rho = 0.4$, "DTW", "trans"). Actually, the last result is already very close to the results of the reference-based approach [34] ($\mathrm{F} = 0.926$). Finally, note that, in comparison to the thumbnailing scenario discussed in Section V.A, one requires larger values for the relative threshold parameter ρ. As a folk song recording basically consists of a large number of repetitions of a single musical part, the SSMs exhibit more relevant paths that need to be captured. In other words, the relevant path structure is "denser" requiring a larger threshold (e.g., $\rho = 0.4$).

VI. Conclusions and Future Work

In this paper, we have made several contributions to the field of music and audio structure analysis. First of all, we have introduced a novel fitness measure that expresses how representative a given segment is in terms of its repetitiveness. Our experiments have shown that the fitness-maximizing segment generally yields good estimates for musically meaningful thumbnails—even in the presence of acoustic and musical variations across repeated segments. The main idea was to jointly perform path extraction and grouping within a unifying optimization scheme, which yields a trade-off between quantity and length of paths (coverage) and quality of paths (score). Furthermore, we have shown how optimal path families (underlying the fitness measure) can be computed efficiently by suitably modifying the classical DTW algorithm. As a further contribution, we have introduced a scape plot representation that yields a compact and (so we think) aesthetically appealing visualization of the global repetitive structure of a given music recording.

Our techniques may be applicable not only to audio thumbnailing, but also to general structure analysis and segmentation tasks, as shown by our folk song scenario. For example, our fitness measure may be helpful for detecting and analyzing long-term repetitive structures that arise in large-scale works such as symphonies or sonatas. Further challenges regard efficiency issues when computing and analyzing the scape plot representation. For example, similar to [10], one may use additional cues such as points of novelty in order to restrict start and end points of thumbnail candidates. Then, the scape plot may only be evaluated on a grid or certain points. Furthermore, the scape plot computation may be accelerated by using multiscale approaches based on different feature resolutions. Particularly for regions within the scape plot corresponding to long segments or segments of poor fitness, a coarse feature resolution may suffice. Subsequently, only the regions of potentially high fitness need to be refined using a higher feature resolution.

References

[1] J. Paulus, M. Müller, and A. P. Klapuri, "Audio-based music structure analysis," in *Proc. 11th Int. Conf. Music Inf. Retrieval (ISMIR)*, Utrecht, The Netherlands, 2010, pp. 625–636.

[2] M. A. Bartsch and G. H. Wakefield, "Audio thumbnailing of popular music using chroma-based representations," *IEEE Trans. Multimedia*, vol. 7, no. 1, pp. 96–104, Feb. 2005.

[3] M. Cooper and J. Foote, "Summarizing popular music via structural similarity analysis," in *Proc. IEEE Workshop Applicat. Signal Process. Audio Acoust. (WASPAA)*, New Paltz, NY, 2003, pp. 127–130.

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

[5] R. B. Dannenberg and N. Hu, "Pattern discovery techniques for music audio," in *Proc. Int. Conf. Music Inf. Retrieval (ISMIR)*, Paris, France, 2002, pp. 63–70.

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

[7] M. Müller and F. Kurth, "Towards structural analysis of audio recordings in the presence of musical variations," *EURASIP J. Adv. Signal Process.*, vol. 2007, no. 1, 2007 [Online]. Available: http://asp.eurasipjournals.com/content/2007/1/089686

[8] M. Müller, *Information Retrieval for Music and Motion*. New York: Springer-Verlag, 2007.

[9] G. Peeters, "Sequence representation of music structure using higher-order similarity matrix and maximum-likelihood approach," in *Proc. Int. Conf. Music Inf. Retrieval (ISMIR)*, Vienna, Austria, 2007, pp. 35–40.

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

[11] C. Rhodes and M. Casey, "Algorithms for determining and labelling approximate hierarchical self-similarity," in *Proc. Int. Conf. Music Inf. Retrieval (ISMIR)*, Vienna, Austria, 2007, pp. 41–46.

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

[13] W. Chai and B. Vercoe, "Music thumbnailing via structural analysis," in *Proc. ACM Int. Conf. Multimedia*, Berkeley, CA, 2003, pp. 223–226.

[14] M. Cooper and J. Foote, "Automatic music summarization via similarity analysis," in *Proc. Int. Conf. Music Inf. Retrieval (ISMIR)*, Paris, France, 2002, pp. 81–85.

[15] M. Levy, M. Sandler, and M. Casey, "Extraction of high-level musical structure from audio data and its application to thumbnail generation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Toulouse, France, 2006, pp. 13–16.

[16] M. Müller, P. Grosche, and N. Jiang, "A segment-based fitness measure for capturing repetitive structures of music recordings," in *Proc. 12th Int. Conf. Music Inf. Retrieval (ISMIR)*, 2011, pp. 615–620.

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

[18] M. Mauch, "Automatic chord transcription from audio using computational models of musical context," Ph.D. dissertation, Queen Mary Univ. of London, London, U.K., 2010.

[19] M. Müller and S. Ewert, "Chroma toolbox: MATLAB implementations for extracting variants of chroma-based audio features," in *Proc. Int. Soc. for Music Inf. Retrieval Conf. (ISMIR)*, Miami, FL, 2011, pp. 215–220.

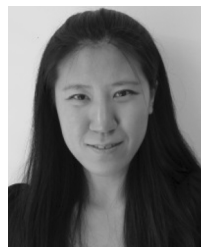
[20] E. Gómez, "Tonal description of music audio signals," Ph.D. dissertation, , UPF Barcelona, Barcelona, Spain, 2006.

[21] M. Müller and F. Kurth, "Enhancing similarity matrices for music audio analysis," in *Proc. Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Toulouse, France, 2006, pp. 437–440.

[22] J. Serrà, E. Gómez, P. Herrera, and X. Serra, "Chroma binary similarity and local alignment applied to cover song identification," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, pp. 1138–1151, Oct. 2008.

[23] F. Kurth and M. Müller, "Efficient index-based audio matching," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 2, pp. 382–395, Feb. 2008.
[24] J. Serrà, X. Serra, and R. G. Andrzejak, "Cross recurrence quantification for cover song identification," *New J. Phys.*, vol. 11, no. 9, p. 093017, 2009.
[25] M. Müller and M. Clausen, "Transposition-invariant self-similarity matrices," in *Proc. 8th Int. Conf. Music Inf. Retrieval (ISMIR)*, Vienna, Austria, Sep. 2007, pp. 47–50.
[26] L. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*, ser. Signal Processing Series. Englewood Cliffs, NJ: Prentice-Hall, 1993.
[27] N. Hu, R. B. Dannenberg, and G. Tzanetakis, "Polyphonic audio matching and alignment for music retrieval," in *Proc. IEEE Workshop Applicat. Signal Process. to Audio Acoust. (WASPAA)*, New Paltz, NY, Oct. 2003.
[28] C. S. Sapp, "Harmonic visualizations of tonal music," in *Proc. Int. Comput. Music Conf. (ICMC)*, 2001, pp. 423–430.
[29] C. S. Sapp, "Visual hierarchical key analysis," *Comput. Entertain.*, vol. 3, no. 4, pp. 1–19, 2005.
[30] M. Mauch, C. Cannam, M. E. Davies, S. Dixon, C. Harte, S. Kolozali, D. Tidhar, and M. Sandler, "OMRAS2 metadata project 2009," in *Proc. Int. Conf. Music Inf. Retrieval (ISMIR)*, Kobe, Japan, 2009.
[31] S. Ewert, M. Müller, and P. Grosche, "High resolution audio synchronization using chroma onset features," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Taipei, Taiwan, 2009, pp. 1869–1872.
[32] H. Lukashevich, "Towards quantitative measures of evaluating song segmentation," in *Proc. Int. Conf. Music Inf. Retrieval (ISMIR)*, Philadelphia, PA, 2008, pp. 375–380.
[33] L. P. Grijp and H. Roodenburg, Blues en Balladen. Alan Lomax en Ate Doornbosch, Twee Muzikale Veldwerkers. Amsterdam, The Netherlands, 2005, AUP.
[34] M. Müller, P. Grosche, and F. Wiering, "Robust segmentation and annotation of folk song recordings," in *Proc. 10th Int. Society Music Inf. Retrieval Conf. (ISMIR)*, Kobe, Japan, Oct. 2009, pp. 735–740.

Meinard Müller (M'09) 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.



Nanzhu Jiang received the M.Sc. degree in computer science from Saarland University in 2011. Since May 2011, she has been pursuing a Ph.D. in the Multimedia Information Retrieval and Music Processing group at Saarland University and Max-Planck Institut für Informatik under the supervision of Meinard Müller. Her main research interests focus on chord recognition, audio thumbnailing and music structure analysis.

Peter Grosche (S'09) received the B.S. and M.Sc. (Diplom) degree in electrical engineering and information technology from Technical University of Munich (TUM) in 2006 and 2008, respectively. Since June 2008, he is pursuing a Ph.D. in the Multimedia Information Retrieval and Music Processing group at Saarland University and Max-Planck Institut für Informatik under the supervision of Meinard Müller. Working in the field of music signal processing and music information retrieval, his research interests cover beat tracking, tempo estimation, music segmentation, and music transcription.