A SEGMENT-BASED FITNESS MEASURE FOR CAPTURING REPETITIVE STRUCTURES OF MUSIC RECORDINGS

Meinard Müller, Peter Grosche, Nanzhu Jiang
Saarland University and MPI Informatik
{meinard,pgrosche,njiang}@mpi-inf.mpg.de

ABSTRACT

In this paper, we deal with the task of determining the audio segment that best represents a given music recording (similar to audio thumbnailing). Typically, such a segment has many (approximate) repetitions covering large parts of the music recording. As main contribution, we introduce a novel fitness measure that assigns to each segment a fitness value that expresses how much and how well the segment "explains" the repetitive structure of the recording. In combination with enhanced feature representations, we show that our fitness measure can cope even with strong variations in tempo, instrumentation, and modulations that may occur within and across related segments. We demonstrate the practicability of our approach by means of several challenging examples including field recordings of folk music and recordings of classical music.

1. INTRODUCTION

Music structure analysis constitutes a fundamental research topic within the field of music information retrieval. One major goal of structure analysis is to divide a music recording into temporal segments corresponding to musical parts and then to group these segments into musically meaningful categories [10]. Such segments may refer to chorus or verse sections of a popular piece of music, to stanzas of a folk song, or to the first theme, the second theme or the entire exposition of a symphony. Such important musical parts are often characterized by the property of being repeated several times throughout the piece. Therefore, finding the repetitive structure of a music recording is an important and well-studied subtask within structure analysis, see, e. g., [1, 2, 5, 6, 9] and the overview articles [3, 10]. Most of these approaches work well for music where the repetitions largely agree. However, in general, "repeating parts" are far from being simple repetitions. Actually, audio segments that refer to the same musical part may differ significantly in parameters such 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 yet unsolved problems.

In this paper, we address the problem of finding the most representative and repetitive segment of a given music recordings, a task often referred to as *audio thumbnailing*, see, e. g., [1]. Here, opposed to most of the previous approaches we want to admit even strong musical variations. As our main contribution, we introduce a fitness measure that assigns to each audio segment a fitness value that simultaneously captures two aspects. Firstly, it indicates *how well* the given segment explains other similar segments ("precisions") and, secondly, it indicates *how much* of the overall music recordings is covered by all these segments ("recall"). Furthermore, our fitness measure is normalized and disregards trivial self-explanations (reflexive relations). As a further contribution of this paper, we introduce a compact time-lag representation that yields a high-level view on the structural properties for the entire music recording. First experiments shows that our fitness measure, in combination with enhanced feature representations, can cope with even strong variations in tempo, instrumentation, and modulations that occur within and across the segments.

At this point, we want to note that our work has been inspired by Paulus and Klapuri [9], even though the task and concepts of this paper are fundamentally different to [9]. The fitness measure introduced in [9] expresses properties of an *entire structure*, whereas our fitness measure expresses properties of a *single segment*. In assigning a fitness value to a given segment, our idea is to simultaneously account for all its existing relations within the entire recording.

The remainder of this paper is organized as follows. In Section 2, we give a motivation of our approach, fix some notation, and quickly review the concept of self-similarity matrices. In Section 3, as our main contribution, we describe the technical details on the construction of our fitness measure. Finally, experimental results and an outlook on future work can be found in Section 4 and Section 5, respectively.

2. MOTIVATION AND NOTATION

In the following, we distinguish between a piece of music (in an abstract sense) and a particular audio recording (a concrete 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 [10]. Musical parts are often denoted by the letters $A, B, C, \ldots$ in the order of their first occurrence. For example, the sequence $A_1 A_2 B_1 A_3$ describes the *musical form* consisting of three repeating A-parts interleaved with one B-part. Then, for a given music recording of such a piece, the goal of the structure analysis problem as tackled in this paper would be to find the segments within the recording that correspond to the A-parts.

Most repetition-based approaches to audio structure analysis proceed as follows. In the first step, the music recording is transformed into a sequence $X := (x_1, x_2, \ldots, x_N)$ of feature vectors $x_n \in \mathcal{F}$, $1 \leq n \leq N$, where $\mathcal{F}$ denotes a suitable feature space. In the second step, based on a similarity measure $\mathbf{s} : \mathcal{F} \times \mathcal{F} \to \mathbb{R}$, one defines a *self-similarity matrix* $\mathcal{S} \in \mathbb{R}^{N \times N}$ by $\mathcal{S}(n, m) := \mathbf{s}(x_n, x_m)$, $1 \leq n, m \leq N$. In the following, a tuple $p = (n, m) \in [1 : N]^2$ is called a *cell* of $\mathcal{S}$, and the value $\mathcal{S}(n, m)$ is referred to as the *score* of the cell p. The crucial observation is that repeating patterns in the feature sequence X appear as diagonal "stripes" in $\mathcal{S}$ [2, 10]. More precisely, these stripes are paths of cells of high score running in parallel to the main diagonal. Therefore, in the third step, one extracts all such paths from $\mathcal{S}$, where each path encodes the similarity of a pair of segments. (These two segments are given by the two projections of the path onto the two axis of $\mathcal{S}$, see Figure 1.) In the fourth step, from the given pairwise relations of segments, one derives entire groups of segments, where each group comprises all segments of a given type of a musical part (e. g. all segments corresponding to A-parts). This step can be thought of forming some kind of transitive closure of the given path relations [3, 6]. However, this grouping process constitutes a main challenge when the extracted paths are erroneous and incomplete. In [5], a grouping process is described that balances out inconsistencies in the path relations by exploiting a constant tempo assumption. However, when dealing with music of varying tempo, the grouping process constitutes a challenging research problem.

As one main idea of our approach, we suggest to jointly perform the third and fourth step thus circumventing the separate grouping process. We realize this idea by assigning a fitness value to a given segment in such a way that all related segments simultaneously influence the fitness value. To express relations between segments, we will introduce the notion of a path family, see Section 3.1. Intuitively, instead of extracting individual paths, we extract entire groups of paths, where the consistency within a group is automatically enforced by the construction.

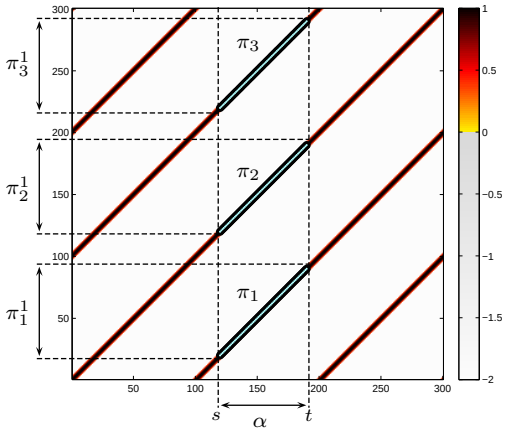

Figure 1. Idealized self-similarity matrix $\mathcal{S}$ for a recording of musical form $A_1 A_2 A_3$. The figures show an optimal path family $\mathcal{P} := \{\pi_1, \pi_2, \pi_3\}$ for the segment $\alpha = [s : t] = [120 : 190] = \pi_1^2 = \pi_2^2 = \pi_3^2$.

2.1 Desired Properties

We now motivate some basic properties that serve as a guideline for the construction of our fitness measure. Let $X = (x_1, x_2, \ldots, x_N)$ be the feature representation of the given audio recording. A *segment* α is defined to be a subset $\alpha = [s : t] \subseteq [1 : N]$ specified by its starting point s and its end point t (given in terms of feature indices). Let $|\alpha| := t - s + 1$ denote the length of α. In our approach, we introduce a *fitness measure* φ that assigns to each segment $\alpha \subseteq [1 : N]$ a fitness value $\varphi(\alpha) \in \mathbb{R}$. Intuitively, this fitness value should express to which extent the segment α "explains" the repetitive structure of X. In particular, the value $\varphi(\alpha)$ should be large in the case that the repetitions of α cover large portions of X, otherwise it should be small.

Next, we impose some normalization constraints on φ. Note that the segment $\alpha = [1 : N]$ explains the entire sequence X perfectly. More generally, each segment α explains itself perfectly (this information is encoded by the main diagonal of a self-similarity matrix). We do not want such trivial, reflexive self-explanations to be captured by φ. Therefore, we require

$$0 \leq \varphi(\alpha) \leq \frac{N - |\alpha|}{N}. \tag{1}$$

In particular, one obtains $\varphi([1 : N]) = 0$. More generally, a value $\varphi(\alpha) = 0$ should mean that the segment α only explains itself but no other portions of X. As an illustrative example, we consider an "ideal" recording of a piece of music having the form $A_1 A_2 \ldots A_K$. Let α_k be the segment corresponding to A_k, $k \in [1 : K]$. Then our fitness measure should assume the value $\varphi(\alpha_k) = \frac{K-1}{K}$ for each segment α_k, see Figure 1 illustrating the case $K = 3$.

2.2 Self-Similarity Matrices

In general, repeating segments may differ significantly regarding tempo, instrumentation and other musical properties. The degree of the similarity between two repeating segments α and α' crucially depends on the used feature type, the similarity measure, and the resulting self-similarity matrix $\mathcal{S}$. Our fitness measure is generic in the sense that it can work with general self-similarity matrices that only fulfill some basic normalization properties. Actually, we only require the property $\mathcal{S}(n,m) \leq 1$ for $1 \leq n,m \leq N$ and $\mathcal{S}(n,n) = 1$ for $n \in [1:N]$. Since the construction of $\mathcal{S}$ is not in the focus of this paper, we only give a quick description of the type of self-similarity matrix as used in our experiments. Figure 2 illustrates the following steps. First of all, we use a variant of chroma-based audio features as described in [6, Section 3.3]. Normalizing these features, we simply use the inner product as similarity measure yielding a value between 0 and 1. To enhance structural properties, we apply temporal smoothing techniques that can deal with tempo variations, see [6, Section 7.2]. Furthermore, applying techniques as described in [7], we obtain a transposition-invariant matrix that can deal with modulation differences within and across repeating parts. Subsequently, using a suitable threshold parameter $\tau > 0$ and a penalty parameter $\delta \leq 0$, we post-process the matrix by first setting the score values of all cells with a score below τ to the value δ and then by linearly scaling the range $[\tau:1]$ to $[0:1]$. Finally, we set $\mathcal{S}(n,n) = 1$ for $n \in [1:N]$ (this property may have been lost by the smoothing step). In the following, we choose τ in a relative fashion by keeping 25% of the cells having the highest score and set $\delta = -2$.

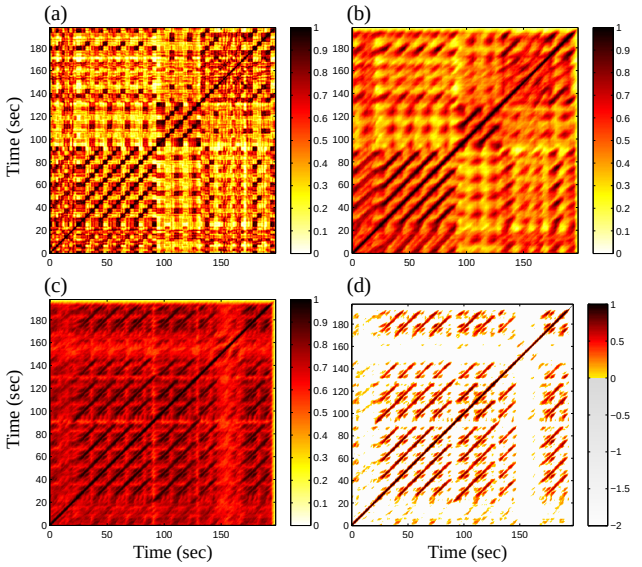

Figure 2. Self similarity matrices for the song "In the year 2525" by Zager and Evans. **(a)** Initial self-similarity matrix. **(b)** Path-enhanced matrix. **(c)** Transposition-invariant matrix. **(d)** Thresholded matrix with $\delta = -2$.

3. FITNESS MEASURE

Following the guidelines motivated in Section 2, we now introduce our novel fitness measure. In assigning a fitness value to a given segment α, our idea is to simultaneously account for all other segments that are related to α. To this end, in Section 3.1, we introduce the notation of a path family that allows for expressing these relations. Then, in Section 3.2, we explain how each path family can be assigned a coverage ("recall") as well as an average score measure ("precisions"). The fitness of the segment α is then determined by the path family that simultaneously maximizes coverage and score.

3.1 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 2.2. A *path* of length L is a sequence $\pi = (p_1, \ldots, p_L)$ of cells $p_\ell = (n_\ell, m_\ell)$ for $\ell \in [1:L]$ satisfying $p_{\ell+1} - p_\ell \in \Sigma$, where Σ denotes a set of admissible step sizes. In our setting, we use $\Sigma = \{(1,2),(2,1),(1,1)\}$, which constrains the slope of the admissible paths within the bounds of $1/2$ and 2, see [6, Chapter 4]. The *score* $\mu(\pi)$ of a path π is defined as

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

Considering the two projections, a path π defines two segments denoted by $\pi^1 := [n_1 : n_L]$ and $\pi^2 := [m_1 : m_L]$, see also Figure 1. Vice versa, given two segments α and α', a path π with $\pi^1 = \alpha$ and $\pi^2 = \alpha'$ is called an *alignment path* between the two segments. Given a segment α and a self-similarity matrix $\mathcal{S}$, we define a *path family over* α to be a set $\mathcal{P} := \{\pi_1, \pi_2, \ldots, \pi_K\}$ that consists of paths π_k and satisfies the following conditions. Firstly, $\pi_k^2 = \alpha$ for all $k \in [1:K]$. Secondly, the set $\{\pi_k^1 \,|\, k \in [1:K]\}$ consists of pairwise disjoint segments, i. e., $\pi_i^1 \cap \pi_j^1 = \emptyset$ for $i,j \in [1:K]$, $i \neq j$. Next, extending the definition in (2) in a straightforward way, the *score* $\mu(\mathcal{P})$ of the path family $\mathcal{P}$ is defined as

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

Finally, the *score* $\mu(\alpha)$ of a segment α is defined to be the score of a path family $\mathcal{P}^*$ having maximal score among all possible path families over α:

$$\mathcal{P}^* := \operatorname*{argmax}_{\mathcal{P}} \mu(\mathcal{P}) \tag{4}$$

$$\mu(\alpha) := \mu(\mathcal{P}^*). \tag{5}$$

Actually, the value $\mu(\alpha)$ is not yet the fitness value we are looking for since neither does it fulfill the basic properties

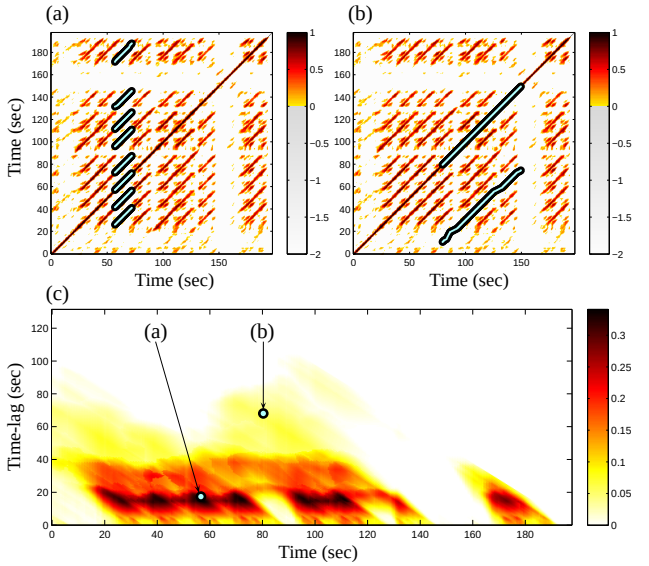

Figure 3. $\mathcal{S}$ and optimal path families $\mathcal{P}$ over different segments $\alpha = [s : t]$ for the song "In the year 2525" by Zager and Evans. **(a)** $\alpha = [57 : 72]$ (maximal fitness). **(b)** $\alpha = [80 : 150]$ **(c)** Fitness matrix.

formulated in Section 2 nor does it capture how much of the audio material is actually covered.

3.2 Definition of Fitness Measure

We now give a formal definition of our fitness measure, which has all the desired properties. Actually, at this point, we only need the assumption that the given self-similarity matrix $\mathcal{S} \in \mathbb{R}^{N\times N}$ 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 $n \in [1:N]$. We start by defining the *normalized score* $\bar{\mu}(\mathcal{P})$ of the path family $\mathcal{P}$ over α by

$$\bar{\mu}(\mathcal{P}) := \frac{\mu(\mathcal{P}) - |\alpha|}{\sum_{k=1}^{K} L_k}, \tag{6}$$

where L_k defines the length of path π_k. Here, the motivation for subtracting the length $|\alpha|$ of α is that the segment α trivially explains itself, see Section 2. It is not hard to see that the score $\bar{\mu}$ fulfills the conditions (1). From the assumption $\mathcal{S}(n,n) = 1$, one obtains $\bar{\mu}(\mathcal{P}) \geq 0$. Furthermore note that, when using $\Sigma = \{(1,2),(2,1),(1,1)\}$, one has $L_k \leq |\alpha|$ and $\sum_k L_k \leq N$. This together with $\mathcal{S}(n,m) \leq 1$ implies the property $\bar{\mu}(\mathcal{P}) \leq (N - |\alpha|)/N$. Intuitively, the value $\bar{\mu}(\mathcal{P})$ expresses the *average score* or precision of the given path family $\mathcal{P}$.

Next, we define some kind of *coverage* or recall measure for $\mathcal{P}$. To this end, let $\gamma(\mathcal{P}) := \cup_{k\in[1:K]} \pi_k^1 \subseteq [1:N]$ be the union of all segments defined by the first projection of the paths π_k. Then we define the *normalized coverage* $\bar{\gamma}(\mathcal{P})$ of $\mathcal{P}$ by

$$\bar{\gamma}(\mathcal{P}) := \frac{|\gamma(\mathcal{P})| - |\alpha|}{N}. \tag{7}$$

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

Inspired by the F-measure that combines precision and recall, we define the *fitness* $\varphi(\mathcal{P})$ of the path family $\mathcal{P}$ to be

$$\varphi(\mathcal{P}) := 2 \cdot \frac{\bar{\mu}(\mathcal{P}) \cdot \bar{\gamma}(\mathcal{P})}{\bar{\gamma}(\mathcal{P}) + \bar{\mu}(\mathcal{P})}. \tag{8}$$

In other words, the fitness integrates the normalized score and coverage into one measure. Finally, the *fitness* $\varphi(\alpha)$ of a segment α is defined to be the fitness value of the score-maximizing path family $\mathcal{P}^*$ defined in (4):

$$\varphi(\alpha) \quad := \quad \varphi(\mathcal{P}^*). \tag{9}$$

Note that the path family $\mathcal{P}^*$ defines in a natural way a set of disjoint segments revealing the repetitions of α within the sequence X, see Figure 1. An optimal path family $\mathcal{P}^*$ for a segment α can be computed efficiently with $O(|\alpha| \times N)$ operations using dynamic programming. Actually, the algorithm, which we do not describe in this paper due to space limitations, is an extension of classical dynamic time warping (DTW), see [4, 6].

When computing the fitness $\varphi(\alpha)$ for all possible segments $\alpha = [s : t] \subseteq [1 : N]$, one can obtain a compact fitness representation for the entire music recording. More precisely, we arrange all fitness values in some time-lag fitness matrix $\Phi \in \mathbb{R}^{N\times N}$ defined by $\Phi(s,\ell) := \varphi([s : s+\ell-1])$ for the starting point $s \in [1:N]$ and the segment length $\ell \in [1 : N - s + 1]$, whereas all other entries of Φ are set to zero, see Figure 3c for an example. Note that each cell (s,ℓ) of the fitness matrix Φ defines an optimal path family for the segment $\alpha = [s : s + \ell - 1]$. The maximal entry of Φ yields the segment with the highest fitness value, which can be regarded as the most representative segment of the recording. In this sense, a solution to our thumbnailing problem is given by

$$\alpha^* := \operatorname*{argmax}_{\alpha} \varphi(\alpha), \tag{10}$$

where the path family associated to α^* yields the structure analysis result.

4. EXPERIMENTS

To investigate the behavior of our fitness measure, we have conducted various experiments using a number of challenging audio recordings that exhibit strong acoustic deformations and musical variations. We first discuss some representative examples and then report on an experiment conducted on a corpus of field recordings.

We start with the song "In the year 2525" by Zager and Evans, which already served as example in Figure 2 and Figure 3. This song has the musical form

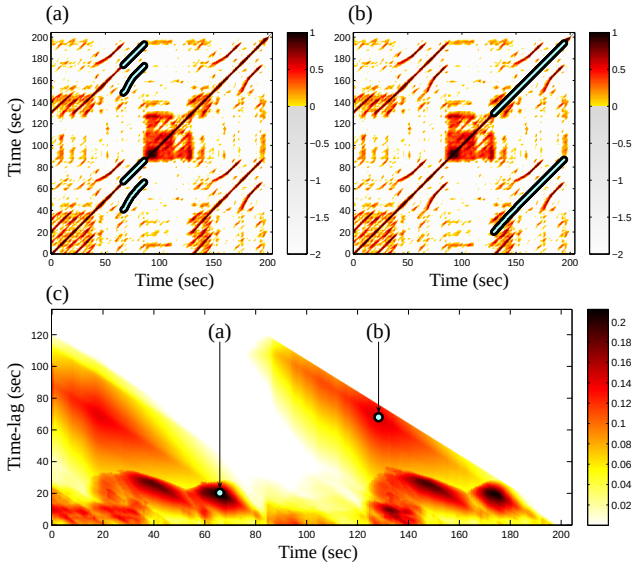

Figure 4. $\mathcal{S}$ and optimal path families $\mathcal{P}$ over different α for an Ormandy recording of Brahms' Hungarian Dance No. 5. **(a)** $\alpha = [67:87]$ (maximal fitness) **(b)** $\alpha = [130:195]$. **(c)** Fitness matrix.

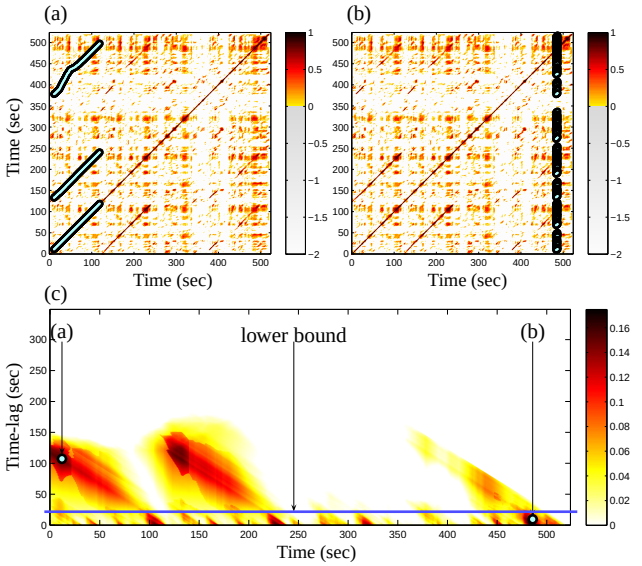

Figure 5. $\mathcal{S}$ and optimal path families $\mathcal{P}$ over different α for a Pollini recording of Beethoven's Op. 31, No. 2, first movement ("Tempest"). **(a)** $\alpha = [11:119]$ (maximal fitness when using the lower bound $\lambda = 20$ seconds.) **(b)** $\alpha = [483:487]$ (maximal fitness). **(c)** Fitness matrix.

$AB_1B_2B_3B_4C_1B_5B_6C_2B_7EB_8F$ starting with a slow intro (A-part) and continuing with eight repetitions of a chorus section (B-part), which are interleaved by two transitional C-parts and one E-part. The first four B-parts are rather similar, whereas the parts B_5 and B_6 are transposed by one and B_7 and B_8 by two semitones upwards. Using a transposition-invariant self-similarity matrix $\mathcal{S}$, all eight repeating B-parts are revealed by the path structure, see Figure 2. Figure 3 shows the time-lag fitness matrix Φ along with optimal path families for two different segments. The path family of the fitness-maximizing segment $\alpha^* = [57:72]$, which is shown in Figure 3a and corresponds to B_3, consists of eight paths. These paths correspond to the eight B-parts thus yielding the expected and desired result. Looking at other segments, one can notice that the fitness measure tries to balance out score and coverage. For example, for the long segment shown in Figure 3b, the lower path accepts even cells of negative score (as long as the accumulated score of the entire path is positive) for the sake of coverage. Here recall that, by definition, all paths of the family are forced to run over the entire segment α.

Next, we consider a recording by Ormandy of the Hungarian Dance No. 5 by Johannes Brahms, see Figure 4. This piece has the musical form $A_1A_2B_1B_2CA_3B_3B_4D$ consisting of three repeating A-parts, four repeating B-parts, as well as a C- and a D-part. As shown by the figure, the path structure of $\mathcal{S}$ again reflects this musical form. In particular, the curved paths reveal that the B-parts are played in different tempi. The fitness-maximizing segment is $\alpha^* = [67:87]$ and corresponds to B_2. As shown by Figure 4a, the path family consists of four paths, which correctly identify all four B-parts. The segment $\alpha = [130:195]$ shown in Figure 4b corresponds to $A_3B_3B_4$. Here note that because our fitness measure disregards self-explanations, the fitness of α is well below the one of α^*.

In our third example, we consider a Pollini recording of the first movement of Beethoven's piano sonata Op. 31, No. 2 ("Tempest"), see Figure 5. Being in the sonata form, the rough musical form of this movement is $A_1A_2BA_3C$ with A_1 being the exposition, A_2 the repetition of the exposition, B the development, A_3 the recapitulation, and C a short coda. Here, even though A_3 is some kind of repetition of A_1, there are significant musical differences. For example, the first theme in A_3 is extended by an additional section not present in A_1 and the second theme in A_3 is transposed five semitones upwards (and later transposed seven semitones downwards) relative to the second theme in A_1. Here note that the modulation does not apply to the entire A_3-part but only to the second theme within the A_3-part. Nevertheless, using transposition-invariance, our fitness measure can still identify the relation of the three A-parts when using $\alpha = [11:119]$, see Figure 5a. Interestingly, this is not the fitness-maximizing segment, which is actually given by $\alpha^* = [483:487]$, see Figure 5b. This example indicates a problem that occurs when the self-similarity matrix contains a lot of noise, i. e., scattered cells of relatively high score. Such cells may form numerous path fragments that, as a whole family, may yield significant average score as well as coverage values. To circumvent such problems, one may introduce a lower bound λ for the minimal possible segment length. For example, using a lower bound $\lambda = 20$ seconds, the fitness-maximizing segment is $\alpha = [11:119]$.

Finally, we report on an experiment using field recordings of the folk song collection *Onder de groene linde* (OGL), which is part of the *Nederlandse Liederenbank*.[1] Each song basically consists of a number of strophes yielding the musical form $A_1 A_2 \ldots A_K$. The main challenge is that the songs are performed by elderly non-professional singers with serious intonation problems, large tempo changes, and interruptions—not to speak of poor recording conditions and background noise. In [8], a reference-based segmentation algorithm, which reverts to an additional MIDI file used as stanza reference, is described and tested for 47 of these songs. As for evaluation, standard precision, recall and F-measures are used to measure the accuracy of the segmentation boundaries (with a tolerance of ± 2 seconds). The results of this reference-based method, which are shown in the last row of Table 1, serve as baseline.

Our approach can be applied for accomplishing the same segmentation task without reverting to any reference. To this end, we determine the fitness-maximizing segment α^* as in (10) and derive the segmentation from the associated path family. Using the same evaluation measures as in [8], our reference-free method yields an F-measure value of $\mathrm{F} = 0.821$, see Table 1. Assuming some prior knowledge on the minimal length of a stanza, this result can be improved. For example, using the lower bound $\lambda = 10\,\mathrm{seconds}$ one obtains $\mathrm{F} = 0.855$, see Table 1. This result is still worse than the results obtained from the reference-based approach ($\mathrm{F} = 0.926$). Actually, a manual inspection showed that this degrade was mainly caused by four particular recordings, where the segmentation derived from α^* was "phase-shifted" compared to the ground truth. Employing a boundary-based evaluation measure resulted in an F-measure of $\mathrm{F} = 0$ for these four recordings. Furthermore, we found out that these phase shifts were caused by the fact that in all of these four recordings the singer completely failed in the first stanza (omitting and confusing entire verse lines). In a final experiment, we replaced the four recordings by a slightly shortened version by omitting the first stanzas, respectively. Repeating the previous experiment on this modified dataset produced an F-measure of $\mathrm{F} = 0.920$, which is already close to the quality obtained by baseline method. Overall, these results demonstrates that our fitness measure can cope even with strong temporal and spectral variations as occurring in field recordings.

5. CONCLUSIONS

In this paper, we introduced a novel fitness measure that expresses how representative a given segment is in terms of repetitiveness. Our experiments showed that the fitness-maximizing segment often yields a good candidate solution for the thumbnailing problem, even in the presence of strong acoustic and musical variations across repeating parts. We also introduced a time-lag fitness matrix that yields a high-level view on the structural properties for the entire music recording. For the future, we need to explore in more detail the role of the different parameter settings, including the role of the self-similarity matrix. We are convinced that our fitness matrix has great potential for visualizing and searching in hierarchical music structures in novel ways. Finally, efficiency issues need to be addressed as well as iterative approaches that allow for deriving the entire musical form.

[1] www.liederenbank.nl

Strategy	P	R	F
Maximal fitness	0.823	0.818	**0.821**
Maximal fitness ($\lambda = 10$)	0.863	0.847	**0.855**
Maximal fitness ($\lambda = 10$, modified dataset)	0.932	0.909	**0.920**
Reference-based method [8]	0.912	0.940	**0.926**

Table 1. Precision, recall, and F-measures for the reference-based segmentation method [8] and the three reference-free methods described in this paper.

Acknowledgement. This work has been supported by the Cluster of Excellence on Multimodal Computing and Interaction at Saarland University.

6. REFERENCES

[1] Mark A. Bartsch and Gregory H. Wakefield. Audio thumbnailing of popular music using chroma-based representations. *IEEE Transactions on Multimedia*, 7(1):96–104, February 2005.

[2] Matthew Cooper and Jonathan Foote. Summarizing popular music via structural similarity analysis. In *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 127–130, New Paltz, NY, US, 2003.

[3] Roger B. Dannenberg and Masataka Goto. Music structure analysis from acoustic signals. In David Havelock, Sonoko Kuwano, and Michael Vorländer, editors, *Handbook of Signal Processing in Acoustics*, volume 1, pages 305–331. Springer, New York, NY, USA, 2008.

[4] Roger B. Dannenberg and Ning Hu. Pattern discovery techniques for music audio. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, Paris, France, 2002.

[5] Masataka Goto. A chorus section detection method for musical audio signals and its application to a music listening station. *IEEE Transactions on Audio, Speech and Language Processing*, 14(5):1783–1794, 2006.

[6] Meinard Müller. *Information Retrieval for Music and Motion*. Springer Verlag, 2007.

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

[8] Meinard Müller, Peter Grosche, and Frans Wiering. Robust segmentation and annotation of folk song recordings. In *Proceedings of the 10th International Society for Music Information Retrieval Conference (ISMIR)*, pages 735–740, Kobe, Japan, October 2009.

[9] Jouni Paulus and Anssi Klapuri. Music structure analysis using a probabilistic fitness measure and a greedy search algorithm. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(6):1159–1170, 2009.

[10] Jouni Paulus, Meinard Müller, and Anssi Klapuri. Audio-based music structure analysis. In *Proceedings of the 11th International Conference on Music Information Retrieval (ISMIR)*, pages 625–636, Utrecht, The Netherlands, 2010.