PATH-CONSTRAINED PARTIAL MUSIC SYNCHRONIZATION

Meinard Müller

Max-Planck-Institut für Informatik
Campus E1 4, D-66123 Saarbrücken

Daniel Appelt

Universität Bonn, Informatik III
Römerstr. 164, D-53117 Bonn

ABSTRACT

Digital music collections often contain different versions and interpretations of a single musical work. In view of music retrieval and browsing applications, one important task, also referred to as audio synchronization, is to automatically time-align two given audio recordings of the same underlying piece. In this paper, we present a novel synchronization procedure, which can compute meaningful audio alignments even in the presence of structural variations. Such variations include the omission of repetitions, the insertion of additional parts (soli, cadenzas), or differences in the number of stanzas in popular, folk, or art songs. As one main contribution, we introduce the concept of path-constrained similarity matrices. This enables us to employ a flexible and efficiently computable partial matching procedure in the optimization step of our synchronization algorithm. Our overall strategy aims at aligning preferably long consecutive runs while avoiding an over-fragmentation of the audio material.

Index Terms— Music, Audio Recording, Alignment, Similarity Matrix, Partial Match

1. INTRODUCTION

Often, a large number of different versions and interpretations exist for a single musical work. The task of *music synchronization* aims at identifying and linking semantically corresponding events which are present in different versions. In particular, the task of *audio synchronization*, where the goal is to time-align two given audio recordings of the same underlying piece of music, has attracted a large amount of attention, see, e. g., [1, 2, 3] and the references therein. Even though recent synchronization algorithms can handle significant variations in tempo, dynamics, and instrumentation, most of them rely on the assumption that the two versions to be aligned correspond to each other with respect to their overall global structure. In real-world scenarios, however, this assumption is often violated. For a popular song, there may be various structurally different album, radio, or extended versions as well as cover versions. In classical music, audio recordings often show omissions of repetitions (e. g., in sonatas and symphonies) or significant differences in parts such as solo cadenzas of concertos. A further prominent example are recordings of popular, folk, or art songs. Here, different recordings of the same underlying song often exhibit a different number of stanzas.

Most previous approaches to music synchronization proceed in two steps. First, the two audio recordings to be aligned are transformed into sequences of (e. g., spectral, chroma, MFCC) features. Then, the two feature sequences are aligned using techniques based on dynamic time warping (DTW), see [2]. In classical DTW, all elements of one sequence are matched to elements in the other sequence (while respecting the temporal order). This is problematic when elements in one sequence do not have suitable counterparts in the other sequence. In the presence of global structural differences between the two sequences, this typically leads to misguided alignments, see Fig. 1a. Also, more flexible alignment strategies such as subsequence DTW or partial matching strategies as used in biological sequence analysis [4] do not properly account for such structural differences.

In this paper, we propose a novel synchronization procedure, which basically consists of three steps. In the first step, we construct a path-constrained similarity matrix, which encodes the common structure of the two audio recordings to be aligned (Sect. 2). In the second step, we compute an optimal path-constrained alignment using a standard partial matching procedure based on dynamic programming. Finally, in the third step, we improve the result by boosting the alignment of preferably long runs and eliminating the alignment of short audio fragments (Sect. 3). The main idea of the overall procedure is that constraining possible matches by a semantically motivated path structure automatically leads to a structurally meaningful global alignment, see Fig. 1.

In Sect. 4, we report on our experiments demonstrating the practicability of our algorithm. Further results and sonifications can be found at `http://www-mmdb.iai.uni-bonn.de/projects/partialSync/`. Conclusions and prospects on future work are given in Sect. 5. Further references to related work are given in the respective sections.

2. PATH-CONSTRAINED SIMILARITY MATRIX

In this section, we introduce the concept of path-constrained similarity matrices. We start by reviewing the basic notions while fixing the notation. Given two audio record-

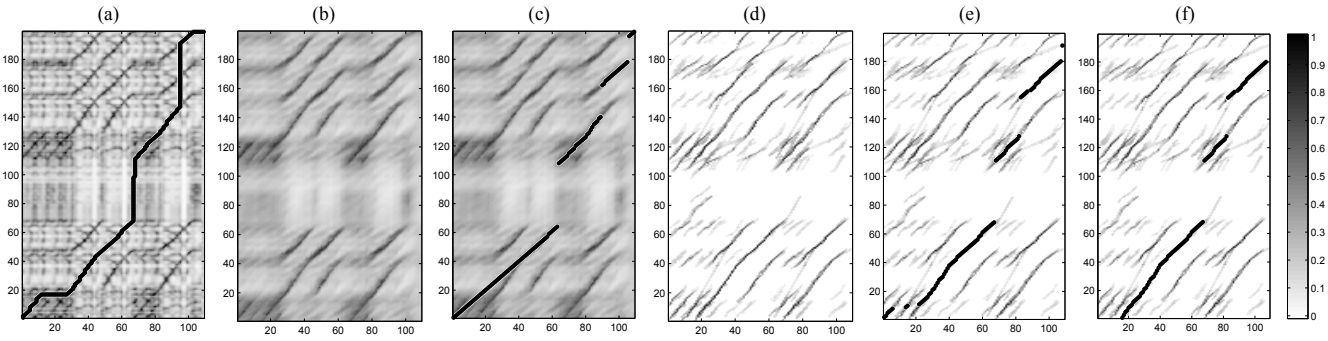

Fig. 1. **(a)** Similarity matrix $\mathcal{S}^{\mathrm{chroma}}$ of two different (structurally modified) audio recordings of Brahms' Hungarian Dance No. 5. The first recording (vertical axis) has the musical form $A_1^1 B_1^1 B_2^1 C^1 A_2^1 B_3^1 B_4^1 D^1$, whereas the second (horizontal axis) has the musical form $A_1^2 A_2^2 B_1^2 B_2^2 A_3^2 B_3^2 D^2$. **(b)** Smoothed similarity matrix $\mathcal{S}^{\mathrm{enh}}$. **(c)** $\mathcal{S}^{\mathrm{enh}}$ with score-maximizing match. **(d)** Path-constrained similarity matrix $\mathcal{S}^{\mathrm{pc}}$. **(e)** $\mathcal{S}^{\mathrm{pc}}$ with score-maximizing match. **(f)** Match after cleaning step. The three path components correctly match $A_1^1 B_1^1 B_2^1$ with $A_2^2 B_1^2 B_2^2$, A_2^1 with A_3^2, and $B_4^1 D^1$ with $B_3^2 D^2$, respectively.

ings, we transform them into suitable feature sequences $V := (v^1, v^2, \ldots, v^N)$ and $W := (w^1, w^2, \ldots, w^M)$, respectively. In the subsequent discussion, we employ smoothed normalized chroma features with a temporal resolution of 1 Hz as described in [2]. In this case, each 12-dimensional feature vector v^n, $n \in [1 : N]$, and w^m, $m \in [1 : M]$, expresses the audio's local energy distribution in the 12 chroma classes. Fixing a suitable local similarity measure— here, we use the inner vector product—the $(N \times M)$-*similarity matrix* $\mathcal{S}^{\mathrm{chroma}}$ is defined by $\mathcal{S}^{\mathrm{chroma}}(n, m) := \langle v^n, w^m \rangle$. Each tuple (n, m) is called a *cell* of the matrix. A *path* is a sequence $p = (p_1, \ldots, p_L)$ with $p_\ell = (n_\ell, m_\ell) \in [1 : N] \times [1 : M]$ for $\ell \in [1 : L]$ satisfying $1 \leq n_1 \leq n_2 \leq \ldots \leq n_L \leq N$ and $1 \leq m_1 \leq m_2 \leq \ldots \leq m_L \leq M$ (monotonicity condition) as well as $p_{\ell+1} - p_\ell \in \Sigma$, where Σ denotes a set of admissible step sizes. For example, in classical DTW one uses $\Sigma = \{(1, 0), (0, 1), (1, 1)\}$. The *score* of a path p is defined as $\sum_{\ell=1}^{L} \mathcal{S}^{\mathrm{chroma}}(n_\ell, m_\ell)$.

Recall that a path with a high score reveals the similarity of two audio segments, which correspond to projections of the path onto the vertical (segment in the first audio recording) and the horizontal axis (segment in the second audio recording). For example, the path in Fig. 1f starting at cell $(1, 18)$ and ending at cell $(67, 69)$ reveals the similarity of the two audio segments that correspond to musical part $A_1^1 B_1^1 B_2^1$ in the first and musical part $A_2^2 B_1^2 B_2^2$ in the second recording. The extraction of the path structure from a similarity matrix is a difficult problem due to musical variations in audio recordings. To ease the extraction step, we further enhance the path structure of $\mathcal{S}^{\mathrm{chroma}}$ by using a contextual similarity measure as described in [5]. The enhancement procedure can be thought of a multiple filtering of $\mathcal{S}^{\mathrm{chroma}}$ along various directions given by gradients in a neighborhood of the gradient $(1, 1)$. We denote the enhanced similarity matrix by $\mathcal{S}^{\mathrm{enh}}$, see Fig. 1b for an illustration.

¿From $\mathcal{S}^{\mathrm{enh}}$, we extract paths with a high score using an iterative greedy strategy, see [2] for a similar strategy. The idea is to start a new path with a cell of maximal score, which is then successively extended to the upper right and lower left by cells that possess a score above a certain threshold, while respecting the step size condition. After removing such an extracted path along with a suitable neighborhood (setting the score to zero for all these cells), the process is iterated until all cells have a score below the threshold. Finally, the extracted path structure is converted into a similarity matrix $\mathcal{S}^{\mathrm{path}}$, where all cells that belong to extracted paths obtain a score of one and all other cells a score of zero, see Fig. 2c. To further improve and reinforce the extracted path structure, we additionally perform a joint structural analysis of the two audio recordings, see [2] for details. The resulting similarity clusters are translated back into a matrix representation denoted by $\mathcal{S}^{\mathrm{struct}}$. The idea is that the structural analysis comprises a transitivity step recovering missing path relations as well as complementing fragmented paths, cf. Fig. 2c and 2d.

We now combine the different similarity matrixes to form a single *path-constrained similarity matrix* denoted by $\mathcal{S}^{\mathrm{pc}}$:

$$\mathcal{S}^{\mathrm{pc}} := \frac{1}{6}(\mathcal{S}^{\mathrm{path}} + \mathcal{S}^{\mathrm{struct}}) * (\mathcal{S}^{\mathrm{chroma}} + \mathcal{S}^{\mathrm{enh}} + 1),$$

where $*$ denotes pointwise multiplication of matrix entries. Note that the entries of all involved matrices possess a value between 0 and 1. Because of the factor $(\mathcal{S}^{\mathrm{path}} + \mathcal{S}^{\mathrm{struct}})$, only cells that belong to the extracted or reinforced path structure can have a non-zero score in $\mathcal{S}^{\mathrm{pc}}$. The factor $(\mathcal{S}^{\mathrm{chroma}} + \mathcal{S}^{\mathrm{enh}} + 1)$ ensures that all these cells actually have a non-zero score and are additionally weighted through underlying score values given by $\mathcal{S}^{\mathrm{chroma}}$ and $\mathcal{S}^{\mathrm{enh}}$. The important point is that the resulting path-constrained similarity matrix explicitly incorporates structural information, thus constraining possible matches in a semantically meaningful way.

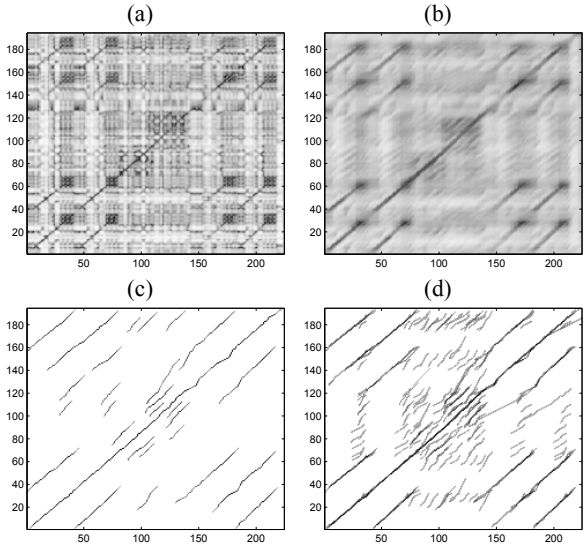

Fig. 2. Similarity matrices of two different audio recordings of a Waltz by Shostakovich having the musical form $A_1A_2BA_3A_4$. **(a)** $\mathcal{S}^{\mathrm{chroma}}$. **(b)** $\mathcal{S}^{\mathrm{enh}}$. **(c)** $\mathcal{S}^{\mathrm{path}}$. **(d)** $\mathcal{S}^{\mathrm{pc}}$.

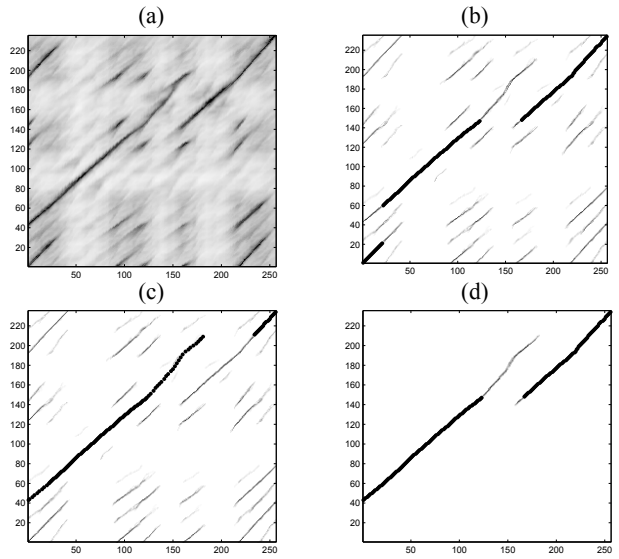

Fig. 3. Similarity matrices of two audio recordings with musical forms $A_1^1A_2^1B^1C^1A_3^1$ and $A_1^2B^2C_1^2C_2^2A_2^2$, respectively. **(a)** $\mathcal{S}^{\mathrm{enh}}$. **(b)** $\mathcal{S}^{\mathrm{pc}}$ with score-maximizing match μ. **(c)** Match μ' after substitution procedure. **(d)** Final match μ''. This match correctly aligns part $A_2^1B^1$ with $A_1^2B^2$ (first path component) and $C^1A_3^1$ with $C_2^2A_2^2$ (second path component).

3. PARTIAL MATCHING PROCEDURE

Our goal is to align similar and possibly long consecutive segments in the two audio recordings. In case, a part in one audio recording does not have a suitable counterpart in the other recording, we prefer to have no alignments rather than having bad alignments. In view of such requirements, we need a more flexible notion of alignment that allows for arbitrary step sizes. A *match* is a sequence $\mu = (\mu_1, \ldots, \mu_L)$ with $\mu_\ell = (n_\ell, m_\ell) \in [1:N] \times [1:M]$ for $\ell \in [1:L]$ satisfying $1 \leq n_1 < n_2 < \ldots < n_L \leq N$ and $1 \leq m_1 < m_2 < \ldots < m_L \leq M$. Note that a match induces a partial alignment, where an element in one sequence is assigned to at most one element in the other sequence. The *score* of a match μ with respect to a similarity matrix $\mathcal{S}$ is then defined as $\sum_{\ell=1}^{L} \mathcal{S}(n_\ell, m_\ell)$.

Similarly to DTW, one can use dynamic programming to compute a score-maximizing match with respect to $\mathcal{S}$. To this end, one recursively defines the accumulated similarity matrix D by $D(n,m) := \max\{D(n,m-1), D(n-1,m), D(n-1,m-1) + \mathcal{S}(n,m)\}$ and $D(n,0) := D(0,m) := 0$ for $n \in [0:N]$ and $m \in [0:M]$. The optimal score is then given by $D(N,M)$ and the score-maximizing match can be constructed by a simple backtrack algorithm, see [4]. Note that the flexibility of a match comes at the expense of loosing stability in the global alignment. For example, a match with respect to the unconstrained similarity matrices $\mathcal{S} = \mathcal{S}^{\mathrm{chroma}}$ or $\mathcal{S} = \mathcal{S}^{\mathrm{enh}}$ may lead to semantically useless alignments as illustrated by Fig. 1c. Our idea is to retain control over the final alignment by using the path-constrained similarity matrix $\mathcal{S} = \mathcal{S}^{\mathrm{pc}}$. This inherently leads to semantically more meaningful matches, where all cells of the match necessarily lie on the extracted path structure, cf. Fig. 1c and Fig. 1e.

Using $\mathcal{S} = \mathcal{S}^{\mathrm{pc}}$, the resulting match still may exhibit unnecessary "gaps", which induce an over-fragmentation in the alignment of the audio material. In particular in parts with consecutive repetitive segments (manifested as "striped regions" in the similarity matrices) the partial match may reveal more or less random gaps within such segments. To make this point clearer, we consider the example shown in Fig. 3. Here, the first recording has the musical form $A_1^1A_2^1B^1C^1A_3^1$ and the second one $A_1^2B^2C_1^2C_2^2A_2^2$. The match indicated by Fig. 3b aligns the beginning of A_1^2 with the beginning of A_1^1 and the end of A_1^2 with the end of A_2^1. ¿From a semantic point of view, however, A_1^2 should be entirely aligned either with A_1^1 or with A_2^1. In order to eliminate such gaps as well as to produce preferably long consecutive runs in the final alignment, we postprocess the match in a cleaning step. To this end, we first decompose match μ into pairwise disjoint path components of maximal length. Here, two consecutive cells of μ are considered to belong to the same path component if their corresponding indices differ by at most a given threshold τ. (In our implementation, we use $\tau = 3$.) For example, in Fig. 3b, the match consists of three path components. Next, we successively extend the longest path component of μ to the upper right and lower left, say by a cell (n,m), while eliminating a cell (a,b) that lies on one of the shorter path components in case the following three conditions are satisfied. First, the extension by (n,m) must comply with the step

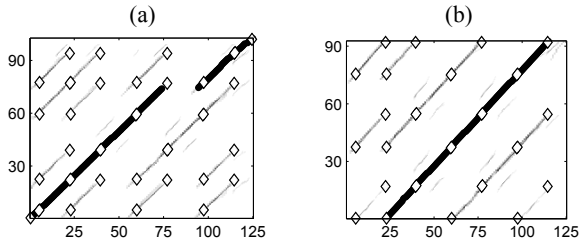

Fig. 4. Alignment of a solo recording of the song "Yesterday" by Paul McCartney with the corresponding original Beatles' album version. The solo version was modified by removing **(a)** a chorus section and **(b)** intro and outro sections. White diamonds indicate start and end points of musical sections.

size condition. Second, the substitution condition $a = n$ or $b = m$ must hold. (In our implementation, we allow some tolerance of up to τ indices, i.e., $n - \tau \leq a \leq n + \tau$ or $m - \tau \leq b \leq m + \tau$.) Third, the relative score condition $\mathcal{S}^{\mathrm{pc}}(n, m) > \rho \cdot \mathcal{S}^{\mathrm{pc}}(a, b)$ for some tolerance parameter $\rho \in (0, 1)$ must hold (we use $\rho = 0.6$). This process is iterated until the longest path component cannot be further extended. We then remove this component from the match and proceed in the same fashion with the remaining cells of the match. All resulting extended path components constitute a new match μ', see Fig. 3c. This procedure decreases or retains the number of path components. However, on the downside, the new match μ' has a lower overall score than μ. To partially compensate for this loss without again increasing the number of path components, we restrict $\mathcal{S}^{\mathrm{pc}}$ to all connected regions of positive score that contain at least one cell of μ' (the score of all other cells is set to zero). We then repeat the partial matching procedure to obtain an optimal match μ'' with respect to the so restricted similarity matrix, see Fig. 3d. This match constitutes our final alignment result.

4. EXPERIMENTS

To evaluate our synchronization procedure, we manually labelled musically meaningful sections of several audio recordings of various genres. The recordings we considered exhibit omissions and insertions of segments that have a duration of 20 seconds or more. In our evaluation, allowing a tolerance up to a few seconds, we compared the matches that are computed by our algorithm with musically meaningful matches.

In a first experiment, we randomly inserted and removed segments in a given recording and aligned it with the original one. In such a simple scenario, our synchronization procedure worked with nearly perfect precision. In a second experiment, we simulated a more realistic scenario. We formed synchronization pairs each consisting of two different interpretations of the same piece. We then modified the pairs by randomly removing some of the labelled sections. The match computed by our algorithm was analyzed by means of its path components. A path component is said to be *correct* if it aligns corresponding musical sections and *strongly correct* if, additionally, its start and end points appear at the labelled musical segment boundaries up to a certain tolerance, see Fig. 4. Similarly, a match is said to be *(strongly) correct* if it is maximal (up to a certain tolerance) and if all its path components are (strongly) correct. We tested our algorithm on 246 different synchronization pairs resulting in a total number of 565 path components. As a result, 91% of all paths are correct and 54% are even strongly correct (using a tolerance of 3 seconds). Furthermore, 86% (57%) of all matches were correct (strongly correct). Using a tolerance of 5 seconds, the number of correct (strongly correct) matches increased to 92% (72%). First experiments show that the correctness rates can be further improved by combining alignment results obtained from different temporal resolutions (e. g., 1 Hz and 2 Hz) and by integrating prior knowledge about the musical structure, e. g., obtained from a previous audio structure analysis [2]. For a detailed presentation of representative results, we refer to `http://www-mmdb.iai.uni-bonn.de/projects/partialSync/`. Here, one also finds sonifications of the alignment results.

5. CONCLUSIONS

In this paper, we have introduced a new synchronization procedure, which allows for partially aligning possibly long and connected portions of two given audio recordings in the presence of structural differences. Our contribution substantially extends recent synchronization procedures, which are based on the assumption of global correspondence. For future work, we will characterize the unaligned parts and extend our alignment scenario to also account for temporally reordered structures. As an important future application, our matching procedure may be applied for partially annotating audio recordings even in situations where one only has fragments of corresponding MIDI or score material (using a direct conversion of symbolic music into a chroma representation, see [1]).

6. REFERENCES

[1] Ning Hu, Roger Dannenberg, and George Tzanetakis, "Polyphonic audio matching and alignment for music retrieval," in *Proc. IEEE WASPAA, New Paltz, NY*, October 2003.

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

[3] Robert J. Turetsky and Daniel P.W. Ellis, "Force-Aligning MIDI Syntheses for Polyphonic Music Transcription Generation," in *Proc. ISMIR, Baltimore, USA*, 2003.

[4] Richard Durbin, Sean R. Eddy, Anders Krogh, and Graeme Mitchison, *Biological Sequence Analysis : Probabilistic Models of Proteins and Nucleic Acids*, Cambridge Univ. Press, 1999.

[5] Meinard Müller and Frank Kurth, "Enhancing similarity matrices for music audio analysis," in *Proc. IEEE ICASSP, Toulouse, France*, 2006.