Case Study "Beatles Songs" – What can be Learned from Unreliable Music Alignments?

Sebastian Ewert[1], Meinard Müller[2], Daniel Müllensiefen[3], Michael Clausen[1], Geraint Wiggins[3]

[1] Universität Bonn, Institut für Informatik III
Römerstr. 164, 53117 Bonn, Germany
ewerts@iai.uni-bonn.de, clausen@iai.uni-bonn.de
[2] Saarland University and MPI Informatik
Campus E1-4, 66123 Saarbrücken, Germany
meinard@mpi-inf.mpg.de
[3] Goldsmiths, University of London, Centre for Cognition, Computation and Culture
Lewisham Way, New Cross London SE14 6NW, England, UK
d.mullensiefen@gold.ac.uk, g.wiggins@gold.ac.uk

Abstract. As a result of massive digitization efforts and the world wide web, there is an exploding amount of available digital data describing and representing music at various semantic levels and in diverse formats. For example, in the case of the Beatles songs, there are numerous recordings including an increasing number of cover songs and arrangements as well as MIDI data and other symbolic music representations. The general goal of music synchronization is to align the multiple information sources related to a given piece of music. This becomes a difficult problem when the various representations reveal significant differences in structure and polyphony, while exhibiting various types of artifacts. In this paper, we address the issue of how music synchronization techniques are useful for automatically revealing critical passages with significant difference between the two versions to be aligned. Using the corpus of the Beatles songs as test bed, we analyze the kind of differences occurring in audio and MIDI versions available for the songs.

Keywords. MIDI, audio, music synchronization, multimodal, music collections, Beatles songs

1 Introduction

Modern digital music collections as well as the world wide web contain an increasing number of relevant digital documents for a single musical work comprising audio recordings, MIDI files, digitized sheet music, music videos, and various symbolic representations. The field of music information retrieval (MIR) aims at developing techniques and tools for organizing, understanding, and searching multimodal music collections in a robust, efficient, and intelligent manner. In this context, various alignment and synchronization procedures have been proposed with the common goal to automatically link several types of music

representations, thus coordinating the multiple information sources related to a given musical work [1,2,3,4,5,6,7,8,9,10].

In general terms, *music synchronization* denotes a procedure which, for a given position in one representation of a piece of music, determines the corresponding position within another representation [1,5]. 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 temporal and polyphonic structure. In real-world scenarios, however, this assumption is often violated. For example, for a popular song there often exists various structurally different album, radio, or extended versions. Live or cover versions may contain improvisations, additional solos, and other deviations from the original song [11]. Furthermore, poor recording conditions, interfering screams and applause, or distorted instruments may introduce additional serious degradations in the audio recordings. On the other side, MIDI or other symbolic descriptions often convey only a simplistic representation of a musical work, where, e. g., certain voices or drum patterns are missing. Often, symbolic data is also corrupted by transcription errors.

For matching and synchronizing structurally different versions, first strategies based on partial alignment techniques are described in the literature [11,12,13]. In general, the synchronization of two strongly deviating versions of a piece of music constitutes a hard task with many yet unsolved research problems. The general reason for the complexity of this task may be explained as follows. In the case that one can rely on the assumption of *global* correspondence between the versions to be synchronized, *relative* criteria are sufficient for establishing an accurate alignment. However, in the case that one may assume only *partial* correspondence, *absolute* criteria are needed to first decide of what is considered to be similar before one can compute the actual alignments. Here, without further model assumptions on the type of similarity, the synchronization task becomes infeasible and, in many cases, constitutes an ill-posed problem.

In this paper, we discuss various aspects of partial similarity using the corpus of the Beatles songs as test bed. In our case study, we revert to a collection of MIDI-audio pairs for more than 100 Beatles songs along with other meta data. The idea is to use conventional global and partial MIDI-audio synchronization techniques to temporally align MIDI events with the corresponding audio data. Then, based on an automated validation of the synchronization results, we segment the MIDI and audio representations into passages, which are then further classified as *reliable* and *critical*. Here, the reliable MIDI and audio passages have a high probability of being correctly aligned with a counterpart, whereas the critical passages are likely to contain variations and artifacts. In other words, we suggest to use music synchronization techniques as a means to analyze MIDI and audio material in a cross-modal fashion in order to reveal the differences and deviations. Actually, the critical passages often constitute the semantically interesting and surprising parts of a representation. For our Beatles collections, we

then further analyze the critical passages and discuss various types of differences and deviations.

The remainder of this paper is organized as follows. In Sect. 2, we give an overview over the underlying Beatles collection. In Sect. 3, we summarize the global and partial MIDI-audio synchronization techniques and describe how critical passages with significant difference in the audio and MIDI versions are revealed by deviating alignments. Then, in Sect. 4, we describe various types of differences within these passages. Conclusions and prospects on future work are given in Sect. 5. Further related work is discussed in the respective sections.

2 Beatles Collection

2.1 Historical Position of the Beatles

The Beatles have been one of the most successful and probably the most influential bands in the history of popular music. They wrote and recorded during 1962 and 1970 more than 200 songs which span a wide variety of popular styles including rock'n'roll, beat, ballads, rock, and psychodelic. Their impact on contemporary and subsequent popular musicians can hardly be overestimated. The corpus of Beatles songs as a whole is characterized by a great variety not only in terms of style but also regarding instrumentation, song structures, and approaches to recording and sound. While the sound of their early recordings was very much dominated by their provenance as a guitar-centered rock'n'roll and beat band, their later songs incorporated string quartets, symphonic instrumentation, backwards recording, and Indian instruments. Sophisticated polyphonic lead and background singing can also be regarded as one of the Beatles' trademarks. Similar to this increase in instrumental and sound sophistication, the variety and complexity of their song structures greatly augments during their career. Starting mainly from verse-chorus-bridge structures in their early works the Beatles took much more compositional freedom on their later albums including intricate intro and outro passages, long instrumental parts, and irregular repetitions of song sections while always maintaining the gist of the popular song genre. It is not surprising that this wealth of musical ideas and concepts that the Beatles produced over one decade had a great impact on the history of popular music. This impact is evidenced by the vast number of cover versions of Beatles songs that have been recorded. According to the internet database coverinfo.de[1] their most popular songs (e.g. 'Yesterday') have been recorded and published more than 200 times by other artists. Numerous transcriptions of their songs as printed sheet music and as MIDI files are also available.

2.2 The Beatles Song as Testbed for MIR Tasks

Probably due to this impact on Western popular music in general and because of their structural and sound-related diversity, the songs written and recorded by

[1] http://www.coverinfo.de/

the Beatles have for a long time been favorite items for cultural, musicological, and music-technology studies. One example is the collection of chord label data for the audio recordings from 12 Beatles albums that was manually generated and compiled by Chris Harte [14]. This collection serves as publicly available ground truth data for the evaluation of chord labeling algorithms that take audio input. The Harte collection of chord labels was also an initial motivation for this present project. In [15], the authors proposed a Bayesian chord induction model that takes symbolic music data as input and annotates the music with chord labels. There are various ways to evaluate a chord labeling algorithm as explained in [16]. If the ambiguity of the harmonic content of a musical passage is not too high then comparing the algorithmically derived chord labels to a carefully crafted ground truth data set of chord labels of the same song is a good option. But in order to evaluate the symbolic chord labeling model by [15] against the ground truth data from the Harte collection that contains chord labels at points in real time in milliseconds, a good alignment has to be achieved that relates the audio recordings that formed the basis of Harte's manual chord annotations to the MIDI file from which the Bayesian model induced. Since the smallest metrical unit of the Bayesian chord model is the beat, the required precision of the MIDI-audio alignment has to be at least the beat (or crotchet) level.

2.3 The Goldsmiths Beatles Corpus

The MIDI files used in this project are part of a larger collection of transcriptions of popular music in MIDI format hosted at Goldsmiths College. This MIDI corpus was acquired in the context of the M^{4}S-project[2]. The corpus consists of a collection of 14067 transcriptions of pop songs from the 1950s to 2006. These files were produced by professional musicians for a distributing company[3], which sells its MIDI files to studio musicians, entertainers or karaoke enterprises where the files are used as backing tracks. The MIDI files may be considered accurate transcriptions of the original recordings and contain all the vocal and instrumental voices that are present in the original. In addition, the encoding conventions of the distributor allow to identify automatically the lyrics, main tune (vocal), bass line, drums, and harmony instruments. Apart from these conventions, the files are not otherwise annotated: there is no indication of song sections, harmonic labels, or abstracted representations of rhythmic and harmonic patterns. In short, the raw data of this corpus represent effectively faithful full-score transcriptions of pop songs. A corpus of symbolic music of this size and covering a large representative sample of a musical genre opens interesting perspective for musicological research which are discussed under the term *corpus-based musicology* in [17].

For obtaining pairs of MIDI and audio files of the same song, we used an approximate metadata match to relate MIDI files from the Goldsmiths corpus

[2] Modelling Music Memory and the Perception of Melodic Similarity, EPSRC-funded project, http://www.doc.gold.ac.uk/isms/mmm/

[3] Geerdes MIDI Music, http://www.midimusic.de/

to audio files in the Goldsmiths collection of recordings used in the OMRAS2-project[4]. For each MIDI transcription of a Beatles song in the Goldsmiths corpus, we computed the edit distance for song artist and title to all the songs in the OMRAS metadata database and checked all the matching audio files above a certain threshold manually. This resulted in a list of MIDI-audio pairs for more than 100 Beatles songs, which form the basis for the experiments to be described. The full list of songs is given in Table 1 in the Appendix. Even though the MIDI and audio representations of each pair generally show a high degree of correspondence, there are also a large number of local and even global differences. It is the object of the subsequent sections to reveal and describe some of these differences.

3 Music Synchronization

Given a MIDI-audio pair for a song, we now describe a procedure that allows for identifying the reliable and critical passages within the MIDI and the audio representation. Based on chroma-based music features (Sect. 3.1), the idea is to compute MIDI-audio correspondences using two conceptually different methods. On the one hand, we use classical dynamic time warping (DTW) to compute a global alignment path (Sect. 3.2). On the other hand, we use a matching procedure to compute a partial match (Sect. 3.3). Both strategies yield correspondences between the respective MIDI and the audio representation. The reliable passages within the two representations are then derived by essentially intersecting the correspondences encoded by the global alignment path and the ones encoded by the partial match (Sect. 3.4). The remaining passages are referred to as critical passages.

3.1 Chroma Features

For comparing a MIDI file and an audio recording of the same song, we convert both representations into a common mid-level representation. In the following, we revert to chroma-based music features, which have turned out to be a powerful tool for relating harmony-based music, see [18,4,5]. Here, the chroma refer to the 12 traditional pitch classes of the equal-tempered scale encoded by the attributes $\mathrm{C}, \mathrm{C}^\sharp, \mathrm{D}, \ldots, \mathrm{B}$. Representing the short-time energy content of the signal in each of the 12 pitch classes, chroma features show a high degree of robustness to variations in timbre and articulation [18]. Furthermore, normalizing the features makes them invariant to dynamic variations. It is straightforward to transform a MIDI file into a sequence of chroma features. Using the explicit MIDI pitch and timing information one basically identifies pitches that belong to the same chroma class within a sliding window of a fixed size, see [4]. Various techniques have been proposed for transforming an audio recording into a chroma sequence, either based on short-time Fourier transforms in combination

[4] Online Music Recognition and Searching, http://www.omras2.com

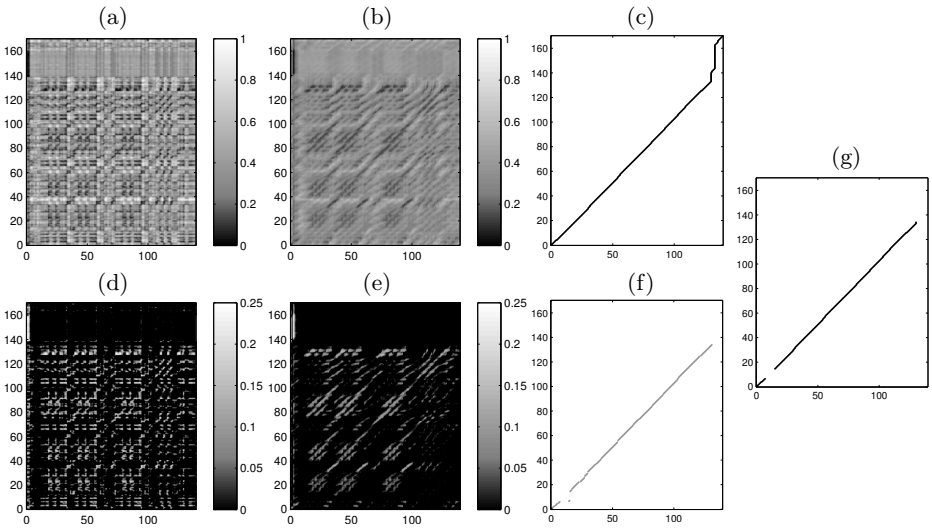

Fig. 1. Intersection technique for identifying reliable and critical passages within an audio recording (vertical axis) and a MIDI version (horizontal axis) of the song 'She loves you'. **(a)** Cost matrix. **(b)** Smoothed cost matrix. **(c)** Optimal warping path obtained via DTW based on matrix (b). **(d)** Score matrix. **(e)** Smoothed score matrix **(f)** Optimal match obtained via partial matching based on matrix (e). **(g)** Soft intersection of the warping path and the match.

with binning strategies [18] or based on suitable multirate filter banks [5]. For the technical details, we refer to the cited literature.

3.2 DTW Procedure

Let $V := (v^1, v^2, \ldots, v^N)$ and $W := (w^1, w^2, \ldots, w^M)$ be two feature sequences with $v(n)$, $w(m) \in \mathcal{F}$, $n \in [1:N]$, $m \in [1:M]$, where $\mathcal{F}$ denotes a feature space. In the subsequent discussion, we employ normalized 12-dimensional chroma features with a temporal resolution of 2 Hz (2 features per second), thus having $\mathcal{F} = [0,1]^{12}$. Furthermore, let $c : \mathcal{F} \times \mathcal{F} \to \mathbb{R}$ denote a *local cost measure* on $\mathcal{F}$. Here, we use the cosine measure defined by $c(v_n, w_m) = 1 - \langle v_n, w_m \rangle$ for normalized vectors. By comparing the features of the two sequences in a pairwise fashion, one obtains an $(N \times M)$-*cost matrix* C defined by $C(n,m) := c(v^n, w^m)$, see Fig.1a. Each tuple (n,m) is called a *cell* of the matrix.

Dynamic time warping is a standard technique for aligning two given (time-dependent) sequences under certain restrictions [19,5]. Given sequences V and W, the alignment is encoded by a *warping path*, which 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 the following three conditions:

Boundary condition: $p_1 = (1,1)$ and $p_L = (N,M)$,
Monotonicity condition: $n_1 \leq n_2 \leq \ldots \leq n_L$ and $m_1 \leq m_2 \leq \ldots \leq m_L$,
Step size condition: $p_{\ell+1} - p_\ell \in \Sigma$ for $\ell \in [1 : L-1]$.

Here, Σ denotes a set of admissible step sizes. For example, in classical DTW one uses $\Sigma = \{(1,0),(0,1),(1,1)\}$. The *cost* of a warping path p is defined as $\sum_{\ell=1}^{L} C(n_\ell, m_\ell)$. Now, let p^* denote a warping path having minimal cost among all possible warping paths. Then, the DTW distance $\mathrm{DTW}(V,W)$ between V and W is defined to be the cost of p^*. It is well-known that p^* and $\mathrm{DTW}(X,Y)$ can be computed in $O(NM)$ using dynamic programming. To this end, one recursively computes the accumulated $(N \times M)$-cost matrix D by

$$D(n,m) = \min\{D(n-1,m-1), D(n-1,m), D(n,m-1)\} + C(n,m), \quad (1)$$

with the starting values $D(n,1) = \sum_{k=1}^{n} C(k,1)$ for $n \in [1:N]$ and $D(1,m) = \sum_{k=1}^{m} C(1,k)$ for $m \in [1:M]$. The optimal cost is then given by $D(N,M)$ and the cost-minimizing warping path can be constructed by a simple backtrack algorithm. For details, we refer to [20,5,19].

To increase the robustness of the overall procedure, we further enhance the path structure of C by using a contextual similarity measure as described in [21]. The enhancement procedure can be thought of as a multiple filtering of C along various directions given by gradients in a neighborhood of the gradient $(1,1)$. We denote the smoothed cost matrix again by C, see Fig. 1b. An optimal warping path with respect to this smoothed cost matrix is shown in Fig. 1c.

3.3 Partial Matching Procedure

In classical DTW, all elements of one sequence are assigned 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. Particularly, in the presence of global structural differences between the two sequences, this typically leads to misguided alignments. For example, in the MIDI-audio pair underlying Fig. 1, the audio recording is overlayed and extended by additional applause, which is not present in the MIDI file. The end of the warping path shown in the upper right part of Fig. 1c is meaningless from a semantic point of view. To allow for a more flexible alignment, one can employ partial matching strategies as used in biological sequence analysis [20]. However, the additional degree of flexibility typically comes at the cost of a loss of robustness in the alignment. We now introduce such a partial matching strategy and then indicate in Sect. 3.4 how it can be combined with the DTW procedure.

Instead of using the concept of a cost matrix with the goal to find a cost-minimizing alignment, we now use the concept of a *score matrix* with the goal to find a score-maximizing alignment. We start with the smoothed cost matrix C introduced in Sect. 3.2 and fix a cost threshold $\tau > 0$. We define a thresholded cost matrix C^τ by setting $C^\tau(n,m) := C(n,m)$ in case $C(n,m) < \tau$ and $C^\tau(n,m) = \tau$ otherwise. Then the $(N \times M)$-score matrix S is defined by

$$S(n,m) = \tau - C^\tau(n,m). \quad (2)$$

In other words, the cells of low cost (close to zero) in C have a high score in S (close to τ). Furthermore, all cells having a cost higher than τ in C have a score of zero in S. The idea of the thresholding is that in case a part in one music representation does not have a suitable counterpart in the other representation, we prefer to have no score at all rather than having small but meaningless score values, see (d) and (e) of Fig. 1.

We now introduce a more flexible notion of alignment that allows for arbitrary step sizes. A *match* is a sequence $\mu = (\mu_1, \ldots, \mu_K)$ with $\mu_k = (n_k, m_k) \in [1 : N] \times [1 : M]$ for $k \in [1 : K]$ satisfying the following condition:

Monotonicity condition: $n_1 < n_2 < \ldots < n_K$ and $m_1 < m_2 < \ldots < m_K$.

Opposed to a warping path, the monotonicity condition is strict, where an element in one sequence is assigned to at most one element in the other sequence. Furthermore there are no boundary or step size conditions imposed on a match. The *score* of a match μ with respect to a score matrix S is then defined as $\sum_{\ell=1}^{L} S(n_k, m_k)$.

This time, the objective is to compute an optimal match μ^* that maximizes the score over all possible matches with respect to S. Similarly to DTW, such an optimal match can be computed efficiently using dynamic programming. To this end, one recursively defines the accumulated score matrix T by

$$T(n, m) := \max\{T(n, m-1), T(n-1, m), T(n-1, m-1) + S(n, m)\} \quad (3)$$

with the starting values $T(n, 1) = S(n, 1)$ for $n \in [1 : N]$ and $T(1, m) = S(1, m)$ for $m \in [1 : M]$. The optimal score is then given by $T(N, M)$ and the optimal match can be constructed by a simple backtrack algorithm, see [20]. An optimal match for our running example is shown in Fig. 1f.

3.4 Reliable and Critical Passages

In general, when the two sequences to be aligned correspond to each other in terms of their global temporal structure and their feature similarity, the DTW procedure yields a robust alignment result. On the other hand, if structural differences are present, the more flexible partial matching procedure may yield more reasonable alignments than DTW. Now, when several strategies with different design goals yield similar alignment results, then the confidence of having good correspondences is high. Based on this simple idea, we present an automatic method towards finding passages in the MIDI and audio representations that can be synchronized in a reliable way. In contrast, this method can be applied for identifying critical passages, where the alignments disagree.

Now, given a MIDI-audio pair for a song, we first compute an optimal warping path as well as an optimal match. Next, the warping path and the match are intersected. To allow for small deviations between the warping path and the match, we admit some tolerance in the intersection. More precisely, a cell of the warping path is kept if and only if there is a cell of the match that lies in a specified neighborhood of the warping path cell. This can be regarded as a soft

intersection, where the warping path plays the role of the reference. Generally, the remaining cells neither form a warping path nor a match, but consists of a number of path fragments. As an example, Fig. 1g shows the soft intersection of the warping path shown in Fig. 1c and the match shown in Fig. 1f. In this case, the soft intersection contains a smaller and a larger path fragment. To avoid an over-fragmentation, two path fragments are combined to a larger one if the end of the first path fragment is sufficiently close to the start of the second one. Note that each path fragment describes an alignment between a segment (or passage) in the audio recording and a segment in the MIDI version. Intuitively, the audio (respective MIDI) segment is obtained by projecting the path fragment onto the vertical (respective horizontal) axis. In the following, we use $\mathrm{A}(t, u)$ to refer to the segment starting at t seconds and ending at u seconds in the audio recording. Similarly, $\mathrm{M}(t, u)$ refers to a MIDI segment. For example, the large path fragment shown in Fig. 1g describes an alignment between $\mathrm{A}(15, 134)$ and $\mathrm{M}(14, 130)$. In the following, each pair consisting of a MIDI and an audio segment aligned by a path fragment is referred to as *reliable pair*. The resulting MIDI and audio segments are called *reliable*. Furthermore, the complements of the reliable segments within the MIDI and audio representations are referred to as *critical* segments. In our example shown in Fig. 1g, the reliable audio and MIDI segments are $\mathrm{A}(0, 7)$ and $\mathrm{M}(0, 7)$ (first reliable pair) as well as $\mathrm{A}(15, 134)$ and $\mathrm{M}(14, 130)$ (second reliable pair). The critical segments are $\mathrm{A}(7, 15)$, $\mathrm{A}(134, 170)$, $\mathrm{M}(7, 14)$ and $\mathrm{M}(130, 140)$. In the following, we use our intersection technique for identifying reliable and critical segments in the MIDI and audio files of the Beatles songs.

4 Classification of Critical and Reliable Passages

Even though recent synchronization procedures can cope with some degree of variations in parameters such as tempo, dynamics, or timbre, most of these procedures rely on the assumption that the two versions to be aligned correspond to each other with respect to their overall global temporal and polyphonic structure. Generally, a violation of this assumption leads to unsatisfying or even useless synchronization results. We now examine which kinds of violations may occur in real-world music data and discuss how these violations may affect the synchronization methods. Based on some songs of the Goldsmiths Beatles corpus (Sect. 2.3), we use the intersection techniques described in Sect. 3 to identify critical and reliable passages. We then investigate some of the critical and reliable passages in more detail and give explicit statements about the underlying semantics behind this classification.

We begin our discussion with the piece 'Hey Bulldog', see Fig. 2. From Fig. 2e, one can easily derive the critical passages. For example, the audio segment $\mathrm{A}(51, 64)$ and the MIDI segment $\mathrm{M}(56, 68)$ are classified as critical. Listening to these two segments, one may get the impression that the two segments correspond well from a semantic point of view. However, as indicated by the cost matrix shown in Fig. 2a, an alignment of these segment would results in very

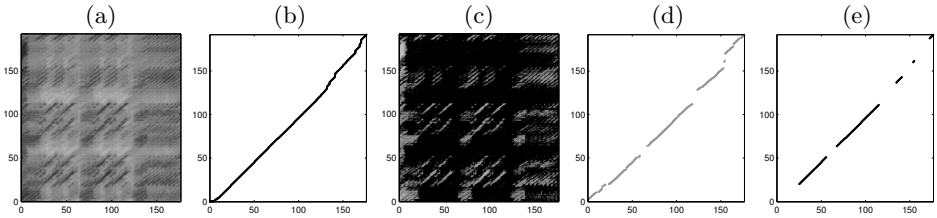

Fig. 2. Intersection technique applied to the MIDI-audio pair of the song 'Hey Bulldog', cf. Fig. 1. **(a)** Cost matrix. **(b)** Optimal warping path. **(c)** Score matrix. **(d)** Optimal match. **(e)** Soft intersection of the warping path and the match.

high costs. A closer investigation shows that in the audio segment A(51, 64) the singing voice sounds rather distorted and dominates the accompanying instruments. On the contrary, the MIDI segment M(56, 68) is dominated by the accompaniment instruments. Since the singing voice and the accompanying instruments are not strongly correlated with respect to harmony, these differences lead to rather different MIDI and audio chroma feature sequences in these segments, thus explaining the high costs. Similar findings hold for the critical segments A(111, 123) and M(115, 126).

Next, we look at the critical segments A(162, 188) and M(155, 173) located towards the end of the song. Looking at Fig. 2c, one recognizes a bundle of diagonal lines in the corresponding area of the score matrix. It turns out that the audio and the MIDI representations comprise a different number of local repetitions in this part of the song. Because of these structural differences, the alignment failed and the segments were correctly classified as critical.

Finally, we take a look at the pair given by the reliable segments A(64, 111) and M(68, 115). Even though this pair was classified as reliable, listening to the MIDI and audio segments reveals that they contain quite different improvisations. Nevertheless, the cost matrix assumes only lower values in the corresponding area. Here, the underlying reason is that, although different on the note level, the harmonizations underlying the improvisations are strongly related. Since the alignment is computed on the basis of chroma-based music features, the note level differences only have a marginal effect on the synchronization result. In other words, our classification as reliable is correct on the harmonic level but not necessarily on the note level. This example shows that the particular choice of feature is crucial in view of the final classification result.

Next, we study the classification result for the song 'The End', see Fig. 3. Here, the pair consisting of the audio segment A(0, 89) and the MIDI segment M(3, 94) has been classified as reliable, which makes perfect sense. Note that in the middle of the song, a short motive is repeated over and over again as indicated by the many parallel lines in the score matrix (Fig. 3c). Since the MIDI and audio version contain the same number of repetitions, no problems occur in the synchronization process.

On the contrary, towards the end of the song the audio segment A(91, 123) and the MIDI segment M(98, 127) are classified as critical. Indeed, the cost

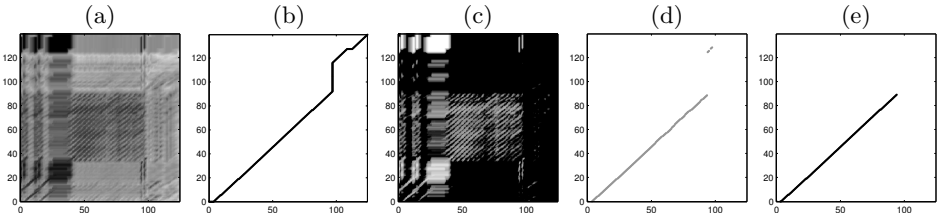

Fig. 3. Intersection technique applied to the MIDI-audio pair of the song 'The End', cf. Fig. 1.

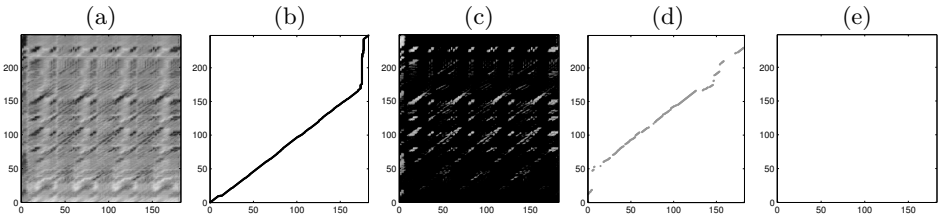

Fig. 4. Intersection technique applied to the MIDI-audio pair of the song 'Strawberry Fields Forever', cf. Fig. 1.

matrix shown in Fig. 3a reveals high cost values in the corresponding area, which is caused by various reasons. First, the piano accompaniment found in the audio is missing in the MIDI version. Second, the guitar in the audio is distorted resulting in a broad and diffuse spectrum. Furthermore, the audio comprises glissandi which are not reproduced from the MIDI version. As a consequence, the resulting audio and MIDI chroma features exhibit significant differences in the final part of the song.

For the song 'The End', one can make a further interesting observation when looking at the cost matrix (Fig. 3e). All cells within the entire area defined by the segments $A(19, 37)$ and $M(21, 41)$ are of constant low cost. As it turns out, these segments comprise a drum solo by Ringo Starr. In this case, the energy is more or less equally distributed over the twelve chroma bands resulting in chroma features with nearly identical entries. In other words, purely percussive music is not suitably represented by chroma features. In areas of constant cost, a warping path may take an arbitrary course resulting in a more or less random alignment. In our alignment strategy, however, we built in a bias towards the diagonal direction in case the choice of direction is arbitrary. As a result, our intersection technique classified these segments as reliable. Again, this example shows the necessity for using various feature types in view of a more multi-faceted classification.

We continue our discussion with the song 'Strawberry Fields Forever'. As can be seen from Fig. 4e, the intersection of the warping path (Fig. 4b) and the match (Fig. 4d) resulted in no segments being classified as reliable. Listening to the audio and MIDI version reveals that the alignment described by the warping path between $A(0, 173)$ and $M(0, 170)$ is semantically reasonable. The same holds

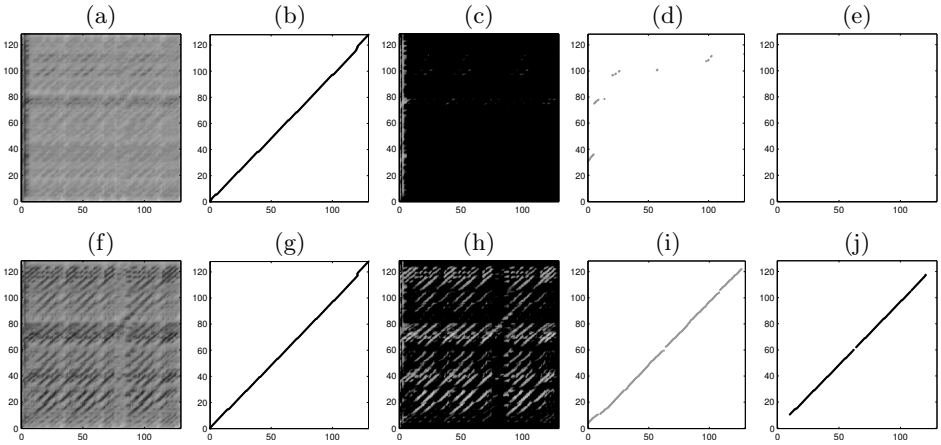

Fig. 5. Intersection technique applied to the MIDI-audio pair of the song 'I'll Get You', cf. Fig. 1. **Top:** Usage of standard audio chroma features. **Bottom**: Usage of pitch-shifted audio chroma features (shifted by a half semitone).

for the alignment described by the match between $A(60, 173)$ and $M(15, 123)$. However, the match takes a different route than the warping path, thus leaving the intersection of the warping path and the match empty. This is caused by the repetitive structure of the song in combination with structural differences between the MIDI and the audio, which results in many paths of similar costs (DTW) and matches of similar score (partial matching). This example reveals the effect when considering only one of possibly several alternative paths and matches in the intersection step. Anyway, our intersection techniques drastically revealed the global structural deviation in the MIDI-audio pair of the song.

Next, we look at the area in the cost matrix (Fig. 4a) defined by $A(0, 41)$ and $M(0, 40)$, which reveals rather high cost values. Listening to the corresponding parts showed that the audio version is slightly out of tune and that the dominating singing voice makes substantial use of glissandi not reflected well in the MIDI. Furthermore, a considerable use of rubato, or more generally the use of individual timings per voice can be found in the MIDI but not in the audio version. Altogether this causes the score matrix (Fig. 4c) to show almost no score in the corresponding area, thus causing the score maximizing match to take a route around this area (Fig. 4d).

As a final example, we consider the song 'I'll Get You'. A look at the upper part of Fig. 5 is quite deflating: the cost and score matrices do not reveal any notable structure resulting in a more or less random match and an empty intersection. Here, it turns out that in the audio recording the instruments have been tuned to play a half semitone below the standard Western pitch scale. However, so far, our procedure only accounts for semitone transpositions between the au-

dio and MIDI version. To account for global tuning differences around ±25 cents[5] we used a filter bank with all filters shifted by a half semitone for computing the audio chroma features. Using the resulting pitch-shifted chroma features to even out the detuning in the audio recording, one obtains that a reliable MIDI-audio synchronization basically throughout the entire song, see bottom part of Fig. 5.

5 Conclusions and Future Work

In this paper, we discussed how synchronization methods can be used for jointly analyzing different representations of the same piece of music. Given a MIDI and an audio version of the same piece, the basic idea was to apply two different, competing MIDI-audio synchronization strategies and then to look for consistencies in the established correspondences. From consistent correspondences, we obtained reliable pairs of MIDI and audio segments that allow for a robust synchronization. The remaining segments, referred to as critical segments, indicated local and global differences between the MIDI and audio versions. Based on a set of representative Beatles songs, we further investigated the kinds of differences occurring in real-world music scenarios. For the future, we plan to incorporate many more competing strategies by not only using different alignment strategies, but also by considering different feature representations and various local cost measures to account for different musical aspects. In combination with statistical methods, we expect to not only achieve a robust validation of music alignments, but also to obtain a semantically meaningful segmentation and classification of the music material.

Many MIR tasks are often performed either in the symbolic domain or in the audio domain. Each domain offers its assets and drawbacks. In a second research direction, we intend to apply MIDI-audio alignments for systematically comparing and studying MIR analysis results obtained from the two domains. In particular, we plan to compare various strategies for automatic chord detection that either work in the audio domain or work in the MIDI domain. By developing strategies for joint MIDI and audio analysis in a cross-domain fashion, we expect to achieve significant improvements for tasks such as automated chord recognition, which in turn may improve and stabilize the synchronization process. Such a feedback process could also help to locate musically stable respective unstable regions in different music representations, thus deepening the understanding of the underlying music.

References

1. Arifi, V., Clausen, M., Kurth, F., Müller, M.: Synchronization of music data in score-, MIDI- and PCM-format. Computing in Musicology **13** (2004)
2. Dannenberg, R., Hu, N.: Polyphonic audio matching for score following and intelligent audio editors. In: Proc. ICMC, San Francisco, USA. (2003) 27–34

[5] The *cent* is a logarithmic unit to measure musical intervals. The interval between two adjacent pitches or semitones of the equal-tempered scale equals 100 cents.

3. Dixon, S., Widmer, G.: Match: A music alignment tool chest. In: Proc. ISMIR, London, GB. (2005)
4. Hu, N., Dannenberg, R., Tzanetakis, G.: Polyphonic audio matching and alignment for music retrieval. In: Proc. IEEE WASPAA, New Paltz, NY. (2003)
5. Müller, M.: Information Retrieval for Music and Motion. Springer (2007)
6. Müller, M., Kurth, F., Röder, T.: Towards an efficient algorithm for automatic score-to-audio synchronization. In: Proc. ISMIR, Barcelona, Spain. (2004)
7. Müller, M., Mattes, H., Kurth, F.: An efficient multiscale approach to audio synchronization. In: Proc. ISMIR, Victoria, Canada. (2006) 192–197
8. Raphael, C.: A hybrid graphical model for aligning polyphonic audio with musical scores. In: Proc. ISMIR, Barcelona, Spain. (2004)
9. Soulez, F., Rodet, X., Schwarz, D.: Improving polyphonic and poly-instrumental music to score alignment. In: Proc. ISMIR, Baltimore, USA. (2003)
10. Turetsky, R.J., Ellis, D.P.: Force-Aligning MIDI Syntheses for Polyphonic Music Transcription Generation. In: Proc. ISMIR, Baltimore, USA. (2003)
11. J. Serrà, E. Gómez, P.H., Serra, X.: Chroma binary similarity and local alignment applied to cover song identification. IEEE Transactions on Audio, Speech and Language Processing **16** (2008) 1138–1151
12. Müller, M., Appelt, D.: Path-constrained partial music synchronization. In: Proc. IEEE ICASSP, Las Vegas, USA. (2008)
13. Müller, M., Ewert, S.: Joint structure analysis with applications to music annotation and synchronization. In: Proc. ISMIR, Philadelphia, USA. (2008)
14. Harte, C., Sandler, M., Abdallah, S., Gómez, E.: Symbolic representation of musical chords: A proposed syntax for text annotations. In Reiss, J.D., Wiggins, G.A., eds.: Proceedings of the 6th International Conference on Music Information Retrieval, London (2005) 66–71
15. Rhodes, C., Lewis, D., Müllensiefen, D.: Bayesian model selection for harmonic labelling. In: Program and Summaries of the First International Conference of the Society for Mathematics and Computation in Music, Berlin, May, 18-20, 2007. (2007)
16. Müllensiefen, D., Lewis, D., Rhodes, C., Wiggins, G.: Treating inherent ambiguity in ground truth data: Evaluating a chord labeling algorithm. In: Proceedings of the 8th International Conference on Music Information Retrieval. (2007)
17. Müllensiefen, D., Wiggins, G., Lewis, D.: High-level feature descriptors and corpus-based musicology: Techniques for modelling music cognition. In Schneider, A., ed.: Systematic and Comparative Musicology: Concepts, Methods, Findings. Volume 24 of Hamburger Jahrbuch für Musikwissenschaft. Peter Lang, Frankfurt (2008) 133–155
18. Bartsch, M.A., Wakefield, G.H.: Audio thumbnailing of popular music using chroma-based representations. IEEE Trans. on Multimedia **7** (2005) 96–104
19. Rabiner, L.R., Juang, B.H.: Fundamentals of Speech Recognition. Prentice Hall Signal Processing Series (1993)
20. Durbin, R., Eddy, S.R., Krogh, A., Mitchison, G.: Biological Sequence Analysis : Probabilistic Models of Proteins and Nucleic Acids. Cambridge University Press (1999)
21. Müller, M., Kurth, F.: Enhancing similarity matrices for music audio analysis. In: Proc. IEEE ICASSP, Toulouse, France. (2006)

Appendix: The Goldsmiths Beatles Corpus

ID	Title	Album	Year
123	I Saw Her Standing There	Anthology 1 [Disc 1]	1963
124	Ticket To Ride	Anthology 2 [Disc 1]	1996
125	Lady Madonna	Anthology 2 [Disc 2]	1968
126	Hey Jude	Anthology 3 (Disc 1)	1996
127	Day Tripper	1962-1966 [Disc 2]	1965
128	Here Comes The Sun	Abbey Road	1969
141	Yesterday	Anthology 2 [Disc 1]	1996
147	Hello Goodbye	1967-1970 [Disc 1]	1967
164	Eight Days A Week	Anthology 1 [Disc 2]	1964
165	Please Please Me	Anthology 1 [Disc 1]	1962
166	Something	Anthology 3 (Disc 2)	1996
167	Paperback Writer	1962-1966 [Disc 2]	1966
173	When I’m Sixty Four	Yellow Submarine	1999
192	Can’t Buy Me Love	Anthology 1 [Disc 2]	1964
254	In My Life	1962-1966 [Disc 2]	1965
277	I Want To Hold Your Hand	Anthology 1 [Disc 2]	1963
522	With A Little Help From My Friends	Sgt. Pepper’s Lonely Hearts Club Band	1967
525	Got To Get You Into My Life	Anthology 2 [Disc 1]	1996
531	And I Love Her	Anthology 1 [Disc 2]	1964
532	Norwegian Wood	Anthology 2 [Disc 1]	1996
550	Come Together	Anthology 3 (Disc 2)	1996
551	Because	Anthology 3 (Disc 2)	1996
552	Back In The USSR	1967-1970 [Disc 2]	1968
553	I Feel Fine	Anthology 2 [Disc 1]	1996
554	A Hard Day’s Night	Anthology 1 [Disc 2]	1964
556	In My Life	1962-1966 [Disc 2]	1965
565	Revolution	1967-1970 [Disc 1]	1968
585	You Never Give Me Your Money	Abbey Road	1969
586	The Fool On The Hill	1967-1970 [Disc 1]	1968
614	Good Day Sunshine	Revolver	1966
627	Taxman	Anthology 2 [Disc 1]	1996
635	Here, There And Everywhere	Revolver	1966
636	Twist And Shout	Anthology 1 [Disc 2]	1963
637	She Loves You	Anthology 1 [Disc 2]	1963
877	Drive My Car	1962-1966 [Disc 2]	1965
878	Money (That’s What I Want)	With The Beatles	1963
891	We Can Work It Out	1962-1966 [Disc 2]	1965
892	All My Loving	With The Beatles	1963
979	Your Mother Should Know	Anthology 2 [Disc 2]	1967
1046	Michelle	1962-1966 [Disc 2]	1965
1087	Penny Lane	Anthology 2 [Disc 2]	1967
1092	Let It Be	Let It Be... Naked	1969
1123	The Long And Winding Road	Let It Be	1970
1124	Eleanor Rigby	Yellow Submarine	1999
1125	For No One	Revolver	1966
1394	Oh! Darling	Abbey Road	1969
1408	Fixing A Hole	Sgt. Pepper’s Lonely Hearts Club Band	1967
1418	Roll Over Beethoven	Anthology 1 [Disc 1]	1963
1435	Ob-La-Di, Ob-La-Da	Anthology 3 (Disc 1)	1996
1497	Yellow Submarine	Yellow Submarine	1999
1828	Help!	Anthology 2 [Disc 1]	1996
2588	Ob-La-Di, Ob-La-Da	Anthology 3 (Disc 1)	1996
2716	I’ll Get You	Anthology 1 [Disc 1]	1963
2839	Rock And Roll Music	Anthology 2 [Disc 1]	1996
2961	Get Back	Let It Be	1970
3208	Norwegian Wood	Anthology 2 [Disc 1]	1996
3254	Ob-La-Di, Ob-La-Da	Anthology 3 (Disc 1)	1996
3794	All You Need Is Love	Yellow Submarine	1999
4016	I Wanna Be Your Man	With The Beatles	1963
4052	Across The Universe	Let It Be... Naked	1969
4053	Dear Prudence	The Beatles (White Album) [Disc 1]	1968
4083	Girl	1962-1966 [Disc 2]	1965
4084	Birthday	The Beatles (White Album) [Disc 2]	1968
4085	Honey Pie	The Beatles (White Album) [Disc 2]	1968
4284	Lucy In The Sky With Diamonds	Anthology 2 [Disc 2]	1967

ID	Title	Album	Year
4285	Nowhere Man	Yellow Submarine	1999
4286	She's A Woman	Anthology 2 [Disc 1]	1996
4287	Strawberry Fields Forever	1967-1970 [Disc 1]	1967
4433	Lovely Rita	Sgt. Pepper's Lonely Hearts Club Band	1967
4434	Maxwell's Silver Hammer	Anthology 3 (Disc 2)	1996
4435	You've Got To Hide Your Love Away	Anthology 2 [Disc 1]	1996
4630	Love Me Do	Anthology 1 [Disc 1]	1962
5212	Free As A Bird	Anthology 1 [Disc 1]	1977
5331	I'll Be Back	Anthology 1 [Disc 2]	1964
5334	It's Only Love	Anthology 2 [Disc 1]	1996
5335	No Reply	Anthology 1 [Disc 2]	1964
5337	Magical Mystery Tour	1967-1970 [Disc 1]	1967
5340	Hey Bulldog	Yellow Submarine	1999
5341	Getting Better	Sgt. Pepper's Lonely Hearts Club Band	1967
5342	You Can't Do That	Anthology 1 [Disc 2]	1964
5343	She's Leaving Home	Sgt. Pepper's Lonely Hearts Club Band	1967
5344	Sgt. Pepper's Lonely Hearts Club Band	Sgt. Pepper's Lonely Hearts Club Band	1967
5345	I Am The Walrus	Anthology 2 [Disc 2]	1967
5351	I Want You (She's So Heavy)	Abbey Road	1969
5352	Mother Nature's Son	Anthology 3 (Disc 1)	1996
5353	Helter Skelter	Anthology 3 (Disc 1)	1996
5543	Real Love	Anthology 2 [Disc 1]	1996
6688	My Bonnie	Anthology 1 [Disc 1]	1961
7852	Being For The Benefit Of Mr. Kite!	Sgt. Pepper's Lonely Hearts Club Band	1967
7879	Baby You're A Rich Man	Yellow Submarine	1999
8076	Savoy Truffle	The Beatles (White Album) [Disc 2]	1968
8092	What Goes On	Rubber Soul	1965
8101	This Boy	Anthology 1 [Disc 2]	1963
8103	Why Don't We Do It In The Road	Anthology 3 (Disc 1)	1996
8319	Don't Pass Me By	Anthology 3 (Disc 1)	1996
8320	Everybody's Got Something To Hide	The Beatles (White Album) [Disc 2]	1968
8362	From Me To You	1962-1966 [Disc 1]	1963
8535	It Won't Be Long	With The Beatles	1963
8536	You Won't See Me	Rubber Soul	1965
8552	Run For Your Life	Rubber Soul	1965
9828	Don't Let Me Down	Let It Be... Naked	1969
10570	All I've Got To Do	With The Beatles	1963
10571	And Your Bird Can Sing	Anthology 2 [Disc 1]	1996
10574	Boys	Anthology 1 [Disc 2]	1964
10575	Doctor Robert	Revolver	1966
10667	Long Tall Sally	Anthology 1 [Disc 2]	1964
10668	The Continuing Story Of Bungalow Bill	The Beatles (White Album) [Disc 1]	1968
10669	The End	Abbey Road	1969
11054	I'm Down	Anthology 2 [Disc 1]	1996
14209	The Ballad Of John And Yoko	1967-1970 [Disc 2]	1969
21577	While My Guitar Gently Weeps	Anthology 3 (Disc 1)	1996
22009	Dig A Pony	Let It Be... Naked	1969
22873	Everybody's Trying To Be My Baby	Anthology 2 [Disc 1]	1996
26541	Ain't She Sweet	Anthology 1 [Disc 1]	1961

Table 1: The Goldsmiths Beatles Corpus consisting of MIDI-audio pairs for more then 100 Beatles songs. The table indicates the song ID, the title of the song, the album the song was first published, and the publishing year.