New Developments in Music Information Retrieval

Meinard Müller[1]

[1]*Saarland University and MPI Informatik, Campus E1.4, 66123 Saarbrücken, Germany*

Correspondence should be addressed to Meinard Müller (meinard@mpi-inf.mpg.de)

ABSTRACT
The digital revolution has brought about a massive increase in the availability and distribution of music-related documents of various modalities comprising textual, audio, as well as visual material. Therefore, the development of techniques and tools for organizing, structuring, retrieving, navigating, and presenting music-related data has become a major strand of research—the field is often referred to as Music Information Retrieval (MIR). Major challenges arise because of the richness and diversity of music in form and content leading to novel and exciting research problems. In this article, we give an overview of new developments in the MIR field with a focus on content-based music analysis tasks including audio retrieval, music synchronization, structure analysis, and performance analysis.

1. INTRODUCTION

As a result of massive digitization efforts, there is 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. For example, for classical music there often exists a large number of different acoustic representations (different performances) as well as visual representations (musical scores). As another example, users are generating an increasing number of home-made music videos with live performances of cover songs and remixes of popular songs. Additionally, for popular music, one can often find other representation types such as lyrics, tablatures, and chord sheets. In the last decade, great research efforts have been directed towards the development of technologies that address the challenges of organizing, understanding, and searching various types of music data in a robust, efficient and intelligent manner. Actually, a larger research community[1] systematically dealing with such issues has formed in the year 2000 having a first conference on Music Information Retrieval (MIR). Since then, rapid developments in music distribution and storage brought about by digital technology has fueled the importance of this young and vibrant research field. In this overview article, we report on new developments in the MIR field with a focus on various content-based audio analysis and retrieval tasks.

Because of the heterogeneity and complexity of music data, there are still many unsolved problems in content-based music analysis and retrieval. Here, *"content-based"* means that in the comparison of music data, one only makes use of the raw data itself, rather than relying on manually generated metadata such as keywords or other symbolic descriptions. While text-based retrieval of music documents using the composers name, the opus number, or lyrics can be handled by means of traditional database techniques, purely content-based music retrieval constitutes a difficult research problem. How should a retrieval system be designed, if the user's query consists of a whistled melody fragment or a short excerpt of some CD recording? How can (symbolic) score data be compared with the content of (waveform-based) CD recordings? What are suitable notions of similarity that capture certain (user-specified) musical aspects while disregarding admissible variations concerning, e. g., the instrumentation or articulation? How can the musical structure, reflected by repetitive and musically related patterns, be automatically derived from a CD recording? These questions only reflect a small fraction of current MIR research topics that are closely related to automatic music analysis [34].

In the following sections, we highlight some of these issues by means of four central MIR tasks. In Section 2,

[1] International Society for Music Information Retrieval (ISMIR), http://www.ismir.net/

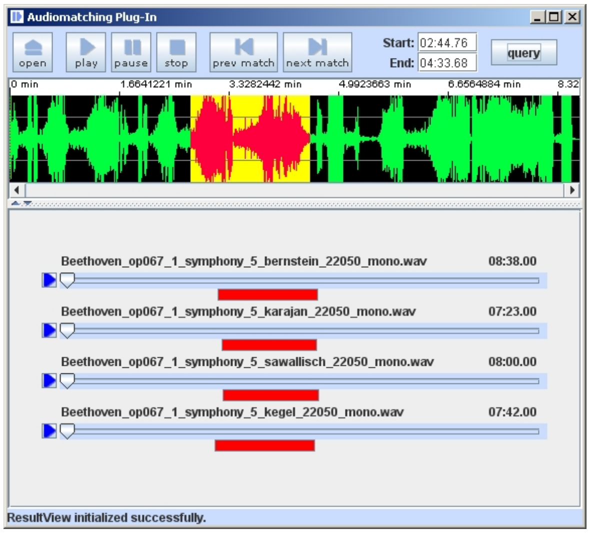
we start by discussing various types of music retrieval scenarios based on the query-by-example paradigm with the goal to access audio material in a content-based fashion. Then, in Section 3, we address the task of music synchronization, where the objective is to coordinate the multiple information sources related to a given musical work. While music synchronization generates relations between different versions of a piece of music, the goal of music structure analysis is to unfold relations within a given music representation. This topic is discussed in Section 4. Finally, in Section 5, we address the task of automated performances analysis, which can be regarded as being complementary to the previously discussed retrieval, synchronization, and structuring tasks.

2. AUDIO RETRIEVAL

Even though there is a rapidly growing corpus of audio material, there still is a lack of efficient systems for content-based audio retrieval, which allow users to explore and browse through large music collections without relying on manually generated annotations. In this context, the query-by-example paradigm has attracted a large amount of attention: given a fragment of an audio recording (used as *query*), the task is to automatically retrieve all documents from a given music database containing parts or aspects similar to the query. Here, the notion of similarity used to compare different audio fragments is of crucial importance and largely depends on the respective application as well as the user requirements.

In content-based retrieval, various levels of specificity can be considered. At the highest specificity level, the retrieval task consists in identifying a particular audio recording within a given music collection using a small audio fragment as query input [1, 6, 31, 58]. This task, which also aims at temporally locating the query fragments within the identified recording, is often referred to as *audio identification* or *audio fingerprinting*. Even though recent identification algorithms show a significant degree of robustness towards noise, MP3 compression artifacts, and uniform temporal distortions, the notion of similarity used in the scenario of audio identification is rather close to the identity. Existing algorithms for audio identification cannot deal with strong non-linear temporal distortions or with other musically motivated variations that concern, for example, the articulation or instrumentation.

While the problem of audio identification can be regarded as largely solved even for large scale music collections, semantically more advanced retrieval tasks are still mostly unsolved. The task of *audio matching* can be seen as an extension of audio identification. Here, given a short query audio fragment, the goal is to automatically retrieve all fragments that musically correspond to the query from all documents (e. g., audio recordings, video clips) within a given music collection, see also Fig. 1. Here, opposed to conventional audio identification, one allows semantically motivated variations as they typically occur in different performances and arrangements of a piece of music. For example, two performances may exhibit significant non-linear global and local differences in tempo, articulation, and phrasing as well as variations in executing ritardandi, accelerandi, fermatas, or ornamentations. Furthermore, one has to deal with considerable dynamical and spectral deviations, which are due to differences in instrumentation, loudness, tone color, accentuation, and so on. A first chroma-based audio matching procedure, which can deal with some of these variations, has been described in [38]. This procedure has been extended by Kurth and Müller [29] to scale to medium size datasets using indexing methods.



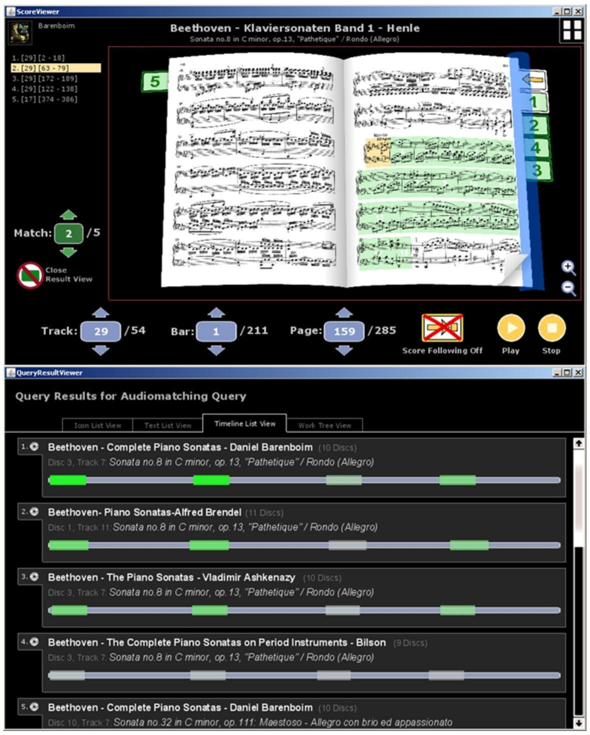
Start: 02:44.76
End: 04:33.68

Beethoven_op067_1_symphony_5_bernstein_22050_mono.wav 08:38.00
Beethoven_op067_1_symphony_5_karajan_22050_mono.wav 07:23.00
Beethoven_op067_1_symphony_5_sawallisch_22050_mono.wav 08:00.00
Beethoven_op067_1_symphony_5_kegel_22050_mono.wav 07:42.00

ResultView initialized successfully.

Fig. 1: Result of fragment-level content-based audio retrieval. Top: Query (yellow background) consisting of an audio fragment. Bottom: Retrieval result consisting of different performances.

Audio identification and audio matching are instances of *fragment-level* retrieval scenarios, where the goal is



1. [29] [2 - 18]
2. [29] [63 - 79]
3. [29] [172 - 189]
4. [29] [122 - 138]
5. [17] [374 - 386]

Match: 2 /5

Track: 29 /54 Bar: 1 /211 Page: 159 /285

Query Results for Audiomatching Query

1. Beethoven - Complete Piano Sonatas - Daniel Barenboim (10 Discs)
Disc 3, Track 7: *Sonata no.8 in C minor, op.13, "Pathetique" / Rondo (Allegro)*

2. Beethoven- Piano Sonatas-Alfred Brendel (11 Discs)
Disc 1, Track 11: *Sonata no.8 in C minor, op.13, "Pathetique" / Rondo (Allegro)*

3. Beethoven - The Piano Sonatas - Vladimir Ashkenazy (10 Discs)
Disc 3, Track 7: *Sonata no.8 in C minor, op.13, "Pathetique" / Rondo (Allegro)*

4. Beethoven - The Complete Piano Sonatas on Period Instruments - Bilson (9 Discs)
Disc 3, Track 7: *Sonata no.8 in C minor, op.13, "Pathetique" / Rondo (Allegro)*

5. Beethoven - Complete Piano Sonatas - Daniel Barenboim (10 Discs)
Disc 10, Track 7: *Sonata no.32 in C minor, op.111: Maestoso - Allegro con brio ed appassionato*

Fig. 2: Interface for simultaneous presentation of visual data (sheet music) and acoustic data (audio recording). The first measures of the third movement (Rondo) of Beethoven's Piano Sonata Op. 13 (Pathétique) are shown. Using a visual query (measures marked in green; theme of the Rondo), all audio documents that contain some matches are retrieved. Here, one audio recording may contain several matches (green rectangles; the theme occurs four times in the Rondo).

to retrieve all musically related fragments contained in the documents of a given music collection. To this end, time-sensitive similarity measures are needed to *locally* compare the query with subsections of a document. In contrast, in *document-level* retrieval, a single similarity measure is considered to globally compare entire documents. One recently studied instance of document-level retrieval is referred to as *cover song identification*, where the goal is to identify different versions of the same piece of music within a database [7, 18, 53]. A cover song may differ from the original song with respect to instrumentation and harmony, it may represent a different genre, or it may be a remix with a different musical structure. Different techniques have been suggested to deal with temporal variations including correlation and DTW-based methods [53], beat tracking methods [18], and audio shingles (small chunks of audio) [7]. Also, most procedures for general music classification tasks [44, 57] are based on document-level similarity.

In summary, it can be said that in the above mentioned problems one has to deal with a trade-off between efficiency and specificity. The more specific the search task is the more efficient it can be solved using indexing techniques. In the presence of significant spectral and temporal variations, the feature extraction as well as the matching steps become more delicate and cost-intensive requiring, e. g., local warping and alignment procedures. Here, the scalability to very large data collections consisting of millions of documents still poses many unsolved problems. Besides efficiency issues, future research also has to address the development of content-based retrieval strategies that allow a user to seamlessly adjust the specificity level in the search process ranging from high-specificity audio identification, over mid-specificity audio matching to low-specificity genre classification, while accounting for fragment-level as well as document-level retrieval. Another major challenge refers to cross-modal music retrieval scenarios, where the query as well as the retrieved documents can be of different modalities. As an example, Fig. 2 shows an interface for a cross-modal retrieval system that allows for bridging the visual and acoustic domain [11, 28]. Here, a user may formulate a query by marking certain musical measures within some sheet music. These measures are then used for retrieving semantically corresponding excerpts in audio recordings. In the future, comprehensive retrieval frameworks are to be developed that offer multi-faceted search functionalities in heterogenous and distributed music collections containing all sorts of music-related documents that vary in their formats (e. g. text, symbolic data, audio, image and video).

3. MUSIC SYNCHRONIZATION

As was already mentioned in the introduction, a musical work is far from simple or singular. In particular, there may exist various audio recordings, MIDI files, video clips, digitized sheet music, and other symbolic representations. In order to coordinate the multiple information sources related to a given musical work, various alignment and synchronization procedures have been proposed with the common goal to automatically link several types of music representations, see, e. g., [2, 14, 15, 17, 22, 26, 27, 30, 34, 39, 40, 41, 50, 52, 53, 54, 55, 56]. 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, see Fig. 3.

Depending upon the respective data formats, one distinguishes between various synchronization tasks [2, 34]. For example, *audio-audio* synchronization [17, 41, 56] refers to the task of time aligning two different audio recordings of a piece of music. These alignments can be used to jump freely between different performances, thus affording efficient and convenient audio browsing. The goal of *score-audio* and *MIDI-audio* synchronization [2, 14, 40, 50, 54] is to coordinate note and MIDI events with audio data. The result can be regarded as an automated annotation of the audio recording with available score and MIDI data. A recently studied problem is referred to as *scan-audio* synchronization [30], where the objective is to link regions (given as pixel coordinates) within the scanned images of given sheet music to semantically corresponding physical time positions within an audio recording. Such linking structures can be used to highlight the current position in the scanned score during playback of the recording. Similarly, the goal of *lyrics-audio* synchronization [22, 39, 27] is to align given lyrics to an audio recording of the underlying song. Finally, different music videos of the same underlying musical work can be linked by applying synchronization techniques to the videos' audio tracks [55].

In order to synchronize two different music representations, one typically proceeds in two steps. In the first step, the two music representations are transformed into sequences of suitable features. Here, on the one hand, the feature representations should show a large degree of robustness to variations that are to be left unconsidered in the comparison. On the other hand, the feature representations should capture characteristic information that suffice to accomplish the synchronization tasks. In this context, chroma-based music features have turned out to be a powerful tool for synchronizing harmony-based music, see [4, 24, 34]. Here, the chroma dimensions 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 content of a music representation in each of the 12 pitch classes, chroma features show a large degree of robustness to variations in timbre and dynamics, while keeping sufficient information to characterize harmony-based music. In the second step, the derived feature sequences have to be brought into temporal correspondence to account for temporal variations in the two music representations to be synchronized. Here alignment techniques such as *Dynamic Time Warping* (DTW) or *Hidden Markov Models* (HMM)—both techniques originally developed in speech processing [48]— are used to find optimal correspondences between two given (time-dependent) sequences under certain restrictions. Intuitively, the alignment can be thought of a linking structure as indicated by the red bidirectional arrows shown in Fig. 3. These arrows encode how the sequences are to be warped (in a non-linear fashion) to match each other.

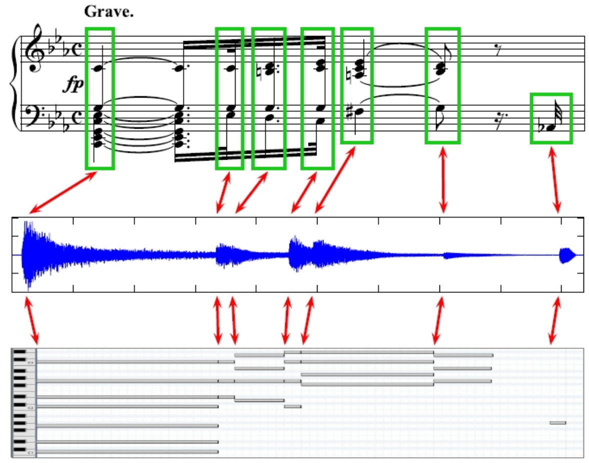


Fig. 3: Linking structure (red arrows) of various representations of different modalities (sheet music, audio, MIDI) corresponding to the same piece of music. These linking structures can be computed automatically using synchronization techniques. Here, the first measures of Beethoven's Piano Sonata Op. 13 (Pathétique) are shown (from [35]).

In the above mentioned synchronization scenarios, the two data streams to be aligned are often entirely known prior to the actual synchronization. This assumption is exploited by alignment procedures such as DTW, which yield an optimal global match between the two complete data streams. Opposed to such an *offline* scenario, one often has to deal with scenarios where the data streams are to be processed *online*. One such prominent online scenario is known as *score following*, which can be regarded as a score-audio synchronization problem. While a musician is performing a piece according to a given musical score, the goal of score following is to identify the musical events depicted in the score with high accuracy and low latency [9, 15]. Note that such an online synchronization procedure inherently has a linear running time. As a main disadvantage, however, an online strategy is very sensitive to local tempo variations and deviations from the score—once the procedure is out of sync, it is very hard to recover and return to the right track. Similar to score following, Dixon et al. [17] describe a linear-time DTW approach to audio synchronization based on forward path estimation. Even though the proposed algorithm is very efficient, the risk of missing the optimal alignment path is still relatively high. A further synchronization problem, which involves score following, is known as *automatic accompaniment*. Here, one typically has a solo part played by a musician which is to be accompanied in real time by a computer system. The problem of real-time music accompaniment has first been studied by Dannenberg et al. [12]. Raphael [49] describes an accompaniment system based on Hidden Markov Models.

As can be seen from this overview, automated music synchronization and the related tasks constitute a challenging field of research, where one has to account for a multitude of aspects such as the data format, the genre, the instrumentation, or differences in parameters such as tempo, articulation and dynamics that result from expressiveness in performances. In the design of synchronization algorithms, one has to deal with a delicate tradeoff between robustness, temporal resolution, alignment quality, and computational complexity. The availability of linking information between different music representations is essential for many retrieval and analysis applications. For example, linking structures allow for navigating between different music documents (interdocument browsing), see also Fig. 4. Furthermore, synchronization techniques can help to bridge the gap between the demand of descriptive high-level features and the capability of existing feature extractors to automatically generate them. Here, note that the automated extraction of high-level metadata from audio such as score parameters, timbre, melodies, instrumentation, or lyrics constitutes an extremely difficult task with many unsolved problems. As a possible strategy to overcome some of the difficulties, one can exploit the fact that one and the same piece of music often exists in several versions on different semantic levels, e. g., one version on a semantically high level (score, lyrics, tablature, MIDI) and another version on a semantically low level (audio, CD recording, video clip). Then a possible strategy is to use the information given by a high-level version in order to support localization and extraction of corresponding events in the low-level version. In this context, synchronization strategies can be regarded as a kind of knowledge-based approach to generate various kinds of metadata annotations.

4. STRUCTURE ANALYSIS

While music synchronization can be used to establish relations across different versions of a piece of music, we now discuss the task of the structure analysis which reveals relations within a given music document. Fig. 4 shows how these relations can be used for inter-document as well as intra-document browsing. Generally speaking, the goal of *audio structure analysis* is to divide an audio recording into temporal segments and to group these segments into musically meaningful categories. Actually, there are different temporal levels as well as many different principles for segmenting and structuring music audio, see [46] for a recent overview article. As for the temporal dimension, structure starts from the level of individual notes, over musical motives, up to musical sections or musical parts that may last for several minutes. These hierarchically ordered structures express relationships between the individual sound events giving a piece some kind of musical meaning. To create these relationships, there are different principles that crucially influence the musical structure. In particular, the principles of *repetition*, *novelty*, and *homogeneity* are of fundamental structural importance and form the basis for many automated analysis methods, see, e. g., [4, 13, 25, 33, 37, 43, 45, 46, 47, 61]. Furthermore, the *temporal order* of events, as also emphasized in [8], is of crucial importance for building up musically and perceptually meaningful entities such as melodies or harmonic progressions. Following [46], we now give a brief overview of some of these methods.

Most frequently, *repetition-based* methods are employed where the goal is to identify recurring patterns. Actually, the repetitive structure of a piece often corresponds to a description that is close to the musical form of the underlying piece of music. Here, the description consists of a *segmentation* of the audio recording as well as of a *grouping* of the segments that are occurrences of the same musical part. The groups are often specified

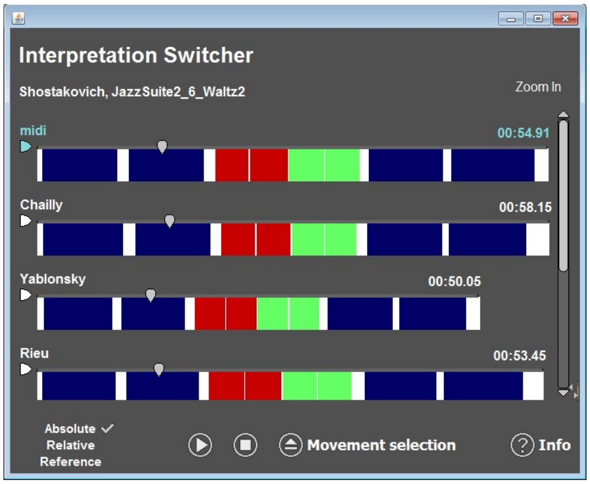


Fig. 4: User interface that facilitates navigation within an audio recording (intra-document browsing using the structure blocks) and across different performances (inter-document browsing switching between sliders). The four sliders correspond to four different performances of the second Waltz from the "Suite for Variety Orchestra No. 1" by Shostakovich. Above each slider, the musical form $A_1A_2B_1B_2C_1C_2A_3A_4$ is indicated by the color-coded blocks.

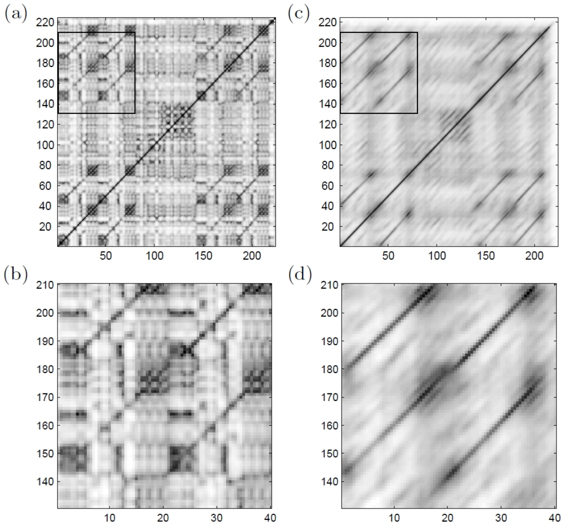


Fig. 5: Self-similarity matrices for the Shostakovich example (from [34]). Time is given in seconds and black/white encode high/low similarity. **(a),(b):** Similarity matrix and enlargement. **(c),(d):** Enhanced versions.

by letters $A, B, C, \ldots$ in the order of their first occurrence and correspond to musically meaningful sections such as *intro*, *chorus*, and *verse* a popular song is composed of. As more concrete example, we consider the second Waltz from the "Suite for Variety Orchestra No. 1" by Shostakovich, see Fig. 4. This piece has the musical form $A_1A_2B_1B_2C_1C_2A_3A_4$ consisting of four repeating A-parts (blue blocks), two recurring B-parts (red blocks) and two C-parts (green blocks). Based on such structures, various user interfaces as indicated by Fig. 4 have been suggested offering new navigation functionalities [21, 25]. Most repetition-based approaches proceed in the following fashion. First, the audio recording is converted into a sequence of suitable audio features, where often chroma-based features are used, see Section 3. Then, a self-similarity matrix is derived by comparing all elements of the feature sequence in a pairwise fashion based on a similarity measure. In this matrix, repetitive patterns are revealed by diagonal stripes parallel to the main diagonal, see Fig. 5 for an illustration. Even though it is often easy for humans to recognize these stripes, the automated extraction of such stripes constitutes a difficult problem due to significant distortions that are caused by variations in parameters such as dynamics, timbre, execution of note groups (e.g., grace notes, trills, arpeggios), modulation, articulation, or tempo progression [34]. To enhance the stripe structure, many approaches apply some sort of low-pass filtering or consider temporal context to smooth the self-similarity matrix along the diagonals, see Fig. 5c. For further details and pointers to the literature, we refer to [46].

A second important principle in music, referred to as *novelty*, is that of change and contrast introducing diversity and attracting the attention of a listener. The goal of *novelty-based* procedures is to automatically locate the points where these changes occur. A standard approach for *novelty detection* introduced by Foote [20] tries to identify segment boundaries by detecting 2D corner points in a self-similarity matrix. Here, a checkerboard-like kernel is locally correlated by shifting it along the main diagonal of the self-similarity matrix. This yields a *novelty function*, the peaks of which indicate corners of blocks of low distance. Using MFCCs, these peaks are good indicators for changes in timbre or instrumentation. Similarly, using other feature representations such as chroma features or rhythmograms, one obtains indicators for changes in harmony, rhythm, or tempo. Again we refer to [46] for details.

Finally, the principle of *homogeneity* is motivated by the observation that musically meaningful sections are often consistent with respect to some musical property such as instrumentation or the coarse harmonic context. Note that novelty-based and homogeneity-based approaches are two sides of a coin: novelty detection is based on observing some surprising event or change after a more homogenous segment. Therefore, homogeneity-based methods are often coupled with novelty-based procedures, where the created segments defined by novelty peaks are represented by statistical models (e. g. Gaussian distributions) and then suitably clustered [10]. Many of the recently introduced homogeneity-based methods employ some kind of state-based approach, where each musical part is represented by a state [3, 23, 33]. Then, Hidden Markov Models are employed to convert the feature sequence into some state sequence. Further constraints and post-processing methods are finally employed to alleviate the problem of temporal fragmentation, see [33, 46].

Most methods for music structure analysis described so far rely on a single strategy. The idea of Paulus and Klapuri [45] is to combine different segmentation principles by using a cost function for structural descriptions of a piece that considers all the desired properties. This cost function is then minimized over all possible structural descriptions for a given acoustic input. Methods that combine several musically motivated information sources and jointly account for various musical dimensions have shown first promising results and will set the trend for future research. Furthermore, to date, the research has mostly been focusing on Western popular music, in which the sectional form is relatively prominent. It would be both challenging and interesting to broaden the target data set to include classical and non Western music. Some of the principles employed by the current methods have been applied for these types of music too, but there is still a large need for research to cope with the complexity and diversity of general music data.

5. PERFORMANCE ANALYSIS

In the previously discussed retrieval, synchronization, and structuring tasks the goal was to detect musically meaningful relations even in the presence of significant variations in the underlying music material. The automated analysis of different versions or, more specifically, of different performances of a given piece of music can be regarded as a kind of complementary task. Here, given a number of performances, the goal is to capture the differences and commonalities between the different versions. In recent years, the task referred to as *performance analysis* has become an active subdiscipline within the field of Music Information Retrieval, see, e. g., [32, 51, 59, 60]. To better understand the starting point and goals of this task, note that musicians give a piece of music their personal touch by continuously varying tempo, dynamics, and articulation. Instead of playing mechanically they speed up at some places and slow down at others in order to shape a piece of music. Similarly, they continuously change the sound intensity and stress certain notes. Such performance issues are of fundamental importance for the understanding and perception of music. In principle, performance analysis addresses two different goals. One goal is to find commonalities between different interpretations, which allow for the derivation of general performance rules. A kind of orthogonal goal is to capture what is characteristic for the style of a particular interpreter [60].

Before one can analyze a specific performance, one requires the information about when and how the notes of the underlying piece of music are actually played. Therefore, as the first step of performance analysis, one has to annotate the performance by means of suitable attributes that make explicit the exact timing and intensity of the various note events. The extraction of such performance attributes constitutes a challenging problem, in particular in the case of audio recordings. Closely following [35], we now describe aspects that concern the extraction step. Many researchers manually annotate the audio material by marking salient data points in the audio stream. However, being very labor-intensive, such a manual process is prohibitive in view of large audio collections. Another way to generate accurate annotations is to use a computer-monitored *player piano*. The advantage of this approach is that it produces precise annotations, where the symbolic note onsets perfectly align with the physical onset times. The obvious disadvantage is that special-purpose hardware is needed during the recording of the piece. In particular, conventional audio material taken from CD recordings cannot be annotated in this way. Therefore, the most preferable method is to automatically extract the necessary performance aspects directly from a given audio recording. Here, automated approaches such as *beat tracking* [16] and *onset detection* [5] are used to estimate the precise timings of note events within the recording. Even though great

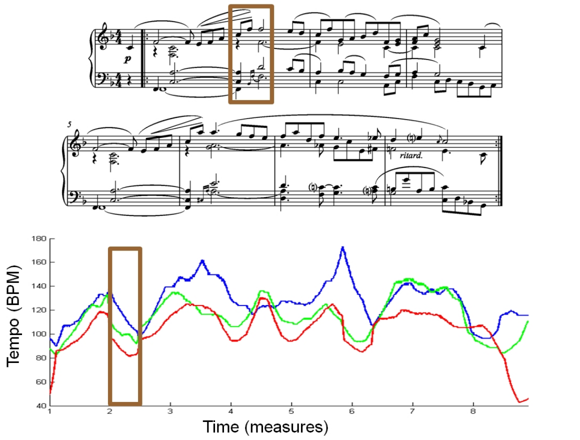


Fig. 6: Tempo curves for three different performances of Schumann's "Träumerei."

research efforts have been directed towards such tasks, the results are still unsatisfactory, in particular for music with weak onsets and strongly varying beat patterns.

Now, instead of trying to derive the tempo information only on the basis of a given music recording, one can exploit the fact that for many pieces there exists a kind of "neutral" representation, which can be used as a reference. Such a reference representation may be given in the form of a musical score or MIDI file, where the notes are played with a known constant tempo (measured in beats per minute or BPM) in a purely mechanical way. Using music synchronization techniques as described in Section 3, one can temporally align the MIDI note events with their corresponding physical occurrences in the music recording. From the synchronization result, it is possible to derive a *tempo curve* that reveals the relative tempo differences between the actual performance and the neutral MIDI reference representation. Assuming that the time signature of the piece is known, one can recover measure and beat positions from MIDI time positions. This information suffices to convert the relative values given by the tempo curve into musically meaningful absolute values. As a result, one obtains a tempo curve that describes for each musical position (given in beats and measures) the absolute tempo of the performance (given in BPM), see [36] for details.

As illustration, Fig. 6 shows the tempo curves for three performances of the first eight measures of the "Träumerei" by Robert Schumann. Despite of significant differences in the overall tempo, there are also noticeable similarities in the relative shaping of the curves. For example, at the beginning of the second measure (region marked by the box), all three pianists slow down, which can be explained by the ascending melodic line culminating in a local climax on the subdominant (B-flat major). After this climax, one can then notice a considerable speed up in all three performances.

In practice, it is difficult problem to determine whether a given change in the tempo curve is due to a synchronization error or whether it is the result of an actual tempo change in the performance. Therefore, to obtain reliable tempo information, one requires robust synchronization procedures of high temporal resolution [19, 42]. This is a particularly difficult research problem in itself for music with less pronounced onset information, smooth note transitions, and rhythmic fluctuation. The computer-based performance analysis, i. e., the automated interpretation of the extracted tempo parameters, is still in its infancy requiring interdisciplinary research efforts between computer science and musicology.

6. CONCLUSIONS

The interdisciplinary field of Music Information Retrieval has emerged over the last ten years into an independent and vibrant area of research that deals with a variety of music analysis and retrieval tasks. In this paper, we have only scratched the surface by discussing four central tasks with a focus on content-based audio analysis. Because of the complexity and diversity of music, one needs intelligent methods and tools that can detect the manifold relationships between the various modalities, versions, and interpretations of a given piece despite significant acoustic and musical variabilities. Therefore, a promising line of research is the development of multi-layered analysis methods that simultaneously account for different temporal resolution levels and various musical dimensions (e. g. rhythm, dynamics, harmony, timbre), while exploiting the availability of multiple versions and representations of a given musical work.

Acknowledgement. Meinard Müller is supported by the Cluster of Excellence on *Multimodal Computing and Interaction* at Saarland University. Special thanks go to the research group headed by Michael Clausen, Bonn University, for the long lasting collaborations.

7. REFERENCES

[1] E. Allamanche, J. Herre, O. Hellmuth, B. Fröba, and M. Cremer. AudioID: Towards content-based identification of audio material. In *Proc. 110th AES Convention*, Amsterdam, NL, 2001.

[2] V. Arifi, M. Clausen, F. Kurth, and M. Müller. *Synchronization of Music Data in Score-, MIDI- and PCM-Format*, volume 13 of *Computing in Musicology*, pages 9–33. MIT Press, 2004.

[3] J.-J. Aucouturier and M. Sandler. Segmentation of musical signals using hidden Markov models. In *Proceedings of the 110th AES Convention*, Amsterdam, NL, 2001.

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

[5] J. P. Bello, L. Daudet, S. Abdallah, C. Duxbury, M. Davies, and M. B. Sandler. A tutorial on onset detection in music signals. *IEEE Transactions on Speech and Audio Processing*, 13(5):1035–1047, 2005.

[6] P. Cano, E. Batlle, T. Kalker, and J. Haitsma. A review of algorithms for audio fingerprinting. In *Proceedings of the IEEE International Workshop on Multimedia Signal Processing (MMSP)*, pages 169–173, St. Thomas, Virgin Islands, USA, 2002.

[7] M. Casey, C. Rhodes, and M. Slaney. Analysis of minimum distances in high-dimensional musical spaces. *IEEE Transactions on Audio, Speech & Language Processing*, 16(5), 2008.

[8] M. Casey and M. Slaney. The importance of sequences in musical similarity. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Toulouse, France, 2006.

[9] A. Cont. A coupled duration-focused architecture for real-time music-to-score alignment. *IEEE Trans. Pattern Anal. Mach. Intell.*, 32(6):974–987, 2010.

[10] M. Cooper and J. 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.

[11] D. Damm, F. Kurth, C. Fremerey, and M. Clausen. A concept for using combined multimodal queries in digital music libraries. In *Proceedings of the 13th European Conference on Digital Libraries (ECDL)*, 2009.

[12] R. B. Dannenberg. An on-line algorithm for real-time accompaniment. In *Proceedings of the International Computer Music Conference (ICMC)*, pages 193–198, 1984.

[13] R. B. Dannenberg and M. Goto. Music structure analysis from acoustic signals. In D. Havelock, S. Kuwano, and M. Vorländer, editors, *Handbook of Signal Processing in Acoustics*, volume 1, pages 305–331. Springer, New York, NY, USA, 2008.

[14] R. B. Dannenberg and N. Hu. Polyphonic audio matching for score following and intelligent audio editors. In *Proceedings of the International Computer Music Conference (ICMC)*, pages 27–34, San Francisco, USA, 2003.

[15] R. B. Dannenberg and C. Raphael. Music score alignment and computer accompaniment. *Communications of the ACM, Special Issue: Music information retrieval*, 49(8):38–43, 2006.

[16] S. Dixon. Evaluation of the audio beat tracking system beatroot. *Journal of New Music Research*, 36:39–50, 2007.

[17] S. Dixon and G. Widmer. Match: A music alignment tool chest. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, London, GB, 2005.

[18] D. P. W. Ellis and G. E. Poliner. Identifying 'cover songs' with chroma features and dynamic programming beat tracking. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 4, Honolulu, Hawaii, USA, Apr. 2007.

[19] S. Ewert, M. Müller, and P. Grosche. High resolution audio synchronization using chroma onset features. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 1869–1872, Taipei, Taiwan, Apr. 2009.

[20] J. Foote. Automatic audio segmentation using a measure of audio novelty. In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*, pages 452–455, New York, NY, USA, 2000.

[21] C. Fremerey, F. Kurth, M. Müller, and M. Clausen. A demonstration of the SyncPlayer system. In *Proceedings of the 8th International Conference on Music Information Retrieval (ISMIR)*, pages 131–132, Vienna, Austria, Sept. 2007.

[22] H. Fujihara, M. Goto, J. Ogata, K. Komatani, T. Ogata, and H. G. Okuno. Automatic synchronization between lyrics and music cd recordings based on viterbi alignment of segregated vocal signals. In *IEEE International Symposium on Multimedia (ISM)*, pages 257–264, Los Alamitos, CA, USA, 2006.

[23] S. Gao, N. C. Maddage, and C.-H. Lee. A hidden Markov model based approach to music segmentation and identification. In *Proceedings of the 4th Pacific Rim Conference on Multimedia (PCM)*, pages 1576–1580, Singapore, 2003.

[24] E. Gómez. *Tonal Description of Music Audio Signals*. PhD thesis, UPF Barcelona, 2006.

[25] M. 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.

[26] N. Hu, R. B. Dannenberg, and G. Tzanetakis. Polyphonic audio matching and alignment for music retrieval. In *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, US, October 2003.

[27] M.-Y. Kan, Y. Wang, D. Iskandar, T. L. Nwe, and A. Shenoy. LyricAlly: Automatic synchronization of textual lyrics to acoustic music signals. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(2):338–349, 2008.

[28] F. Kurth, D. Damm, C. Fremerey, M. Müller, and M. Clausen. A framework for managing multimodal digitized music collections. In *ECDL*, pages 334–345, 2008.

[29] F. Kurth and M. Müller. Efficient index-based audio matching. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(2):382–395, Feb. 2008.

[30] F. Kurth, M. Müller, C. Fremerey, Y. ha Chang, and M. Clausen. Automated synchronization of scanned sheet music with audio recordings. In *Proceedings of the 8th International Conference on Music Information Retrieval (ISMIR)*, pages 261–266, Vienna, Austria, Sept. 2007.

[31] F. Kurth, A. Ribbrock, and M. Clausen. Identification of highly distorted audio material for querying large scale data bases. In *Proceedings of the 112th AES Convention*, 2002.

[32] J. Langner and W. Goebl. Visualizing expressive performance in tempo-loudness space. *Computer Music Journal*, 27(4):69–83, 2003.

[33] M. Levy and M. Sandler. Structural segmentation of musical audio by constrained clustering. *IEEE Transactions on Audio, Speech and Language Processing*, 16(2):318–326, 2008.

[34] M. Müller. *Information Retrieval for Music and Motion*. Springer Verlag, 2007.

[35] M. Müller, M. Clausen, V. Konz, S. Ewert, and C. Fremerey. A multimodal way of experiencing and exploring music. *Interdisciplinary Science Reviews (ISR)*, 35(2):138–153, 2010.

[36] M. Müller, V. Konz, A. Scharfstein, S. Ewert, and M. Clausen. Towards automated extraction of tempo parameters from expressive music recordings. In *Proceedings of the 10th International Society for Music Information Retrieval Conference (ISMIR)*, pages 69–74, Kobe, Japan, Oct. 2009.

[37] M. Müller and F. Kurth. Towards structural analysis of audio recordings in the presence of musical variations. *EURASIP Journal on Advances in Signal Processing*, 2007(1), 2007.

[38] M. Müller, F. Kurth, and M. Clausen. Audio matching via chroma-based statistical features. In *Proceedings of the 6th International Conference on Music Information Retrieval (ISMIR)*, pages 288–295, 2005.

[39] M. Müller, F. Kurth, D. Damm, C. Fremerey, and M. Clausen. Lyrics-based audio retrieval and multimodal navigation in music collections. In *Proceedings of the 11th European Conference on Digital Libraries (ECDL)*, pages 112–123, Budapest, Hungary, Sept. 2007.

[40] M. Müller, F. Kurth, and T. Röder. Towards an efficient algorithm for automatic score-to-audio synchronization. In *Proceedings of the 5th International Conference on Music Information Retrieval (ISMIR)*, pages 365–372, Barcelona, Spain, Oct. 2004.

[41] M. Müller, H. Mattes, and F. Kurth. An efficient multiscale approach to audio synchronization. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, pages 192–197, Victoria, Canada, 2006.

[42] B. Niedermayer and G. Widmer. A multi-pass algorithm for accurate audio-to-score alignment. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, pages 417–422, Utrecht, Netherlands, 2010.

[43] B. S. Ong. *Structural Analysis and Segmentation of Music Signals*. PhD thesis, University Pompeu Fabra, Barcelona, Spain, February 2007.

[44] E. Pampalk, A. Flexer, and G. Widmer. Improvements of audio-based music similarity and genre classification. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, pages 628–633, London, GB, 2005.

[45] J. Paulus and A. 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.

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

[47] G. Peeters. Sequence representation of music structure using higher-order similarity matrix and maximum-likelihood approach. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, pages 35–40, Vienna, Austra, 2007.

[48] L. Rabiner and B.-H. Juang. *Fundamentals of Speech Recognition*. Prentice Hall Signal Processing Series, 1993.

[49] C. Raphael. A probabilistic expert system for automatic musical accompaniment. *Journal of Computational and Graphical Statistics*, 10(3):487–512, 2001.

[50] C. Raphael. A hybrid graphical model for aligning polyphonic audio with musical scores. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, pages 387–394, Barcelona, Spain, 2004.

[51] C. S. Sapp. Comparative analysis of multiple musical performances. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, pages 497–500, Vienna, Austria, 2007.

[52] D. Schwarz, N. Orio, and N. Schnell. Robust polyphonic midi score following with hidden Markov models. In *International Computer Music Conference (ICMC)*, Miami, Florida, US, 2004.

[53] J. Serrà, E. Gómez, P. Herrera, and X. Serra. Chroma binary similarity and local alignment applied to cover song identification. *IEEE Transactions on Audio, Speech and Language Processing*, 16:1138–1151, Oct. 2008.

[54] F. Soulez, X. Rodet, and D. Schwarz. Improving polyphonic and poly-instrumental music to score alignment. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, pages 143–148, Baltimore, USA, 2003.

[55] V. Thomas, C. Fremerey, D. Damm, and M. Clausen. SLAVE: a Score-Lyrics-Audio-Video-Explorer. In *Proceedings of the 10th International Conference on Music Information Retrieval (ISMIR 2009)*, pages 717–722, October 2009.

[56] R. J. Turetsky and D. P. Ellis. Ground-truth transcriptions of real music from force-aligned MIDI syntheses. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, pages 135–141, Baltimore, USA, 2003.

[57] G. Tzanetakis and P. Cook. Musical genre classification of audio signals. *IEEE Transactions on Speech and Audio Processing*, 10(5):293–302, 2002.

[58] A. Wang. An industrial strength audio search algorithm. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, pages 7–13, Baltimore, USA, 2003.

[59] G. Widmer. Machine discoveries: A few simple, robust local expression principles. *Journal of New Music Research*, 31(1):37–50, 2003.

[60] G. Widmer, S. Dixon, W. Goebl, E. Pampalk, and A. Tobudic. In search of the Horowitz factor. *AI Magazine*, 24(3):111–130, 2003.

[61] C. Xu, N. C. Maddage, and X. Shao. Automatic music classification and summarization. *IEEE Transactions on Speech and Audio Processing*, 13(3):441–450, 2005.