# AUDIO-BASED MUSIC STRUCTURE ANALYSIS

**Jouni Paulus**[*]
Fraunhofer Institute for
Integrated Circuits IIS
Erlangen, Germany
jouni.paulus@iis.fraunhofer.de

**Meinard Müller**
Saarland University and
MPI Informatik
Saarbrücken, Germany
meinard@mpi-inf.mpg.de

**Anssi Klapuri**
Queen Mary Univ. of London
Centre for Digital Music
London, UK
anssi.klapuri@elec.qmul.ac.uk

## ABSTRACT

Humans tend to organize perceived information into hierarchies and structures, a principle that also applies to music. Even musically untrained listeners unconsciously analyze and segment music with regard to various musical aspects, for example, identifying recurrent themes or detecting temporal boundaries between contrasting musical parts. This paper gives an overview of state-of-the-art methods for computational music structure analysis, where the general goal is to divide an audio recording into temporal segments corresponding to musical parts and to group these segments into musically meaningful categories. There are many different criteria for segmenting and structuring music audio. In particular, one can identify three conceptually different approaches, which we refer to as repetition-based, novelty-based, and homogeneity-based approaches. Furthermore, one has to account for different musical dimensions such as melody, harmony, rhythm, and timbre. In our state-of-the-art report, we address these different issues in the context of music structure analysis, while discussing and categorizing the most relevant and recent articles in this field.

## 1. INTRODUCTION

The difference between arbitrary sound sequences and music is not well-defined: what is random noise for someone may be ingenious musical composition for somebody else. What can be generally agreed upon is that it is the structure, or the relationships between the sound events that create musical meaning. This structure starts from the level of individual notes, their timbral characteristics and pitch and time intervals. Notes form larger structures, phrases, chords, and chord progressions, and these again form larger constructs in a hierarchical manner. At the level of entire musical pieces the subdivision can be made

---

to musical sections, such as *intro*, *chorus*, and *verse* in popular music. Recovering a description of this structure, often referred to as *musical form*, is what is here meant by *music structure analysis*. In this paper, we mainly focus on Western popular music in terms of the musical structures and acoustic assumptions we make, even though many of the employed principles can be utilized to analyze other kinds of music as well. For a tutorial and a review of earlier methods for music structure analysis, we refer to the book chapter by Dannenberg and Goto [16]. Our objective is to give an updated overview on this important topic by discussing a number of new trends and recent research articles. Computational analysis of the structure of recorded music constitutes a very active research field within the area of music information retrieval. Here we focus on music structure analysis at the largest temporal scale, and assume that the musical form can be expressed as a sequence of musically meaningful parts at this level.[1] The musical form is of great importance for both understanding as well as processing music and is often characteristic to the particular genre.

Structure in music signals arises from certain relationships between the elements—notes, chords, and so forth—that make up the music. The principles used to create such relationships include *temporal order*, *repetition*, *contrast*, *variation*, and *homogeneity*. Obviously, the *temporal order* of events, as also emphasized by Casey and Slaney [11], is of crucial importance for building up musically and perceptually meaningful entities such as melodies or harmonic progressions. Also, the principle of *repetition* is central to music, as Middleton [51] states: *"It has often been observed that repetition plays a particularly important role in music—in virtually any sort of music one can think of, actually. [...] In most popular music, repetition processes are especially strong."* Recurrent patterns, which may be of rhythmic, harmonic, or melodic nature, evoke in the listener the feeling of familiarity and understanding of the music. The principle of *contrast* is introduced by having two successive musical parts of different character. For example, a quiet passage may be contrasted by a loud one, a slow section by a rapid one, or an orchestral part by a solo. A further principle is that of *variation*, where motives and parts are picked up again in a modified or transformed

---

form [39]. Finally, a section is often characterized by some sort of inherent *homogeneity*, for example, the instrumentation, the tempo, or the harmonic material being similar within the section.

In view of the various principles that crucially influence the musical structure, a large number of different approaches to music structure analysis have been developed. One can roughly distinguish three different classes of methods. Firstly, *repetition-based* methods are employed to identify recurring patterns. From a technical point of view, these methods are also often referred to as *sequence approaches*, see also Sec. 5. Secondly, *novelty-based* methods are used to detect transitions between contrasting parts. Thirdly, *homogeneity-based* methods are used to determine passages that are consistent with respect to some musical property. 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. From a technical point of view, the homogeneity-based approach has often been referred to as *state approach*, see also Sec. 5. Finally, in all the method categories, one has to account for different musical dimensions, such as melody, harmony, rhythm, or timbre. To this end, various feature representations have been suggested in the literature.

The remainder of this paper is organized as follows. In Sec. 2, we approach the structure analysis task from different angles and give a problem definition used in this paper. In Sec. 3, we discuss feature representations that account for different musical dimensions. In Sec. 4, we introduce the concept of a self-distance matrix often used in music structure analysis, and show how the various segmentation principles are reflected in this matrix. Then, in Sec. 5, we discuss the principles of repetition-based, novelty-based, and homogeneity-based structure analysis methods. Here, we also discuss and categorize the most relevant and recent articles in this field. In Sec. 6, we address the issue of evaluating analysis results, which in itself constitutes a non-trivial problem. Finally, in Sec. 7, we conclude with a discussion of open problems.

## 2. PROBLEM SPECIFICATION

As mentioned before, the task of *music structure analysis* refers to a range of problems, and different researchers have pursued slightly different goals in this context. A common theme, however, is that the temporal scale of the analysis has been approximately the same in all the cases. In the rest of the paper, we use the following terminology. A *part* is understood to be a musical concept that loosely refers to either a single instance or all the instances of a musical section, such as chorus or verse, whereas a *segment* is understood to be a technical concept that refers to the temporal range of a single occurrence of a musical part. The term *group* is used to denote one or more segments that represent all the occurrences of the same musical part.

The methods discussed in the following take an acoustic music signal as the input and produce some information about the structure. The output of the discussed methods varies from images created for visualization purposes to representations that specify the time range and musically meaningful label of each found part. In the simplest form, no explicit structural analysis is performed, but some transformation of the acoustic features of the piece are used to yield a visual representation of structural information, e.g., the self-similarity matrix visualization by Foote [24].

The next category of methods aim to specify points within a given audio recording where a human listener would recognize a change in instrumentation or some other characteristics. This problem, which is often referred to as *novelty detection*, constitutes an important subtask [25]. For example, as we explain later, having computed novelty points in a preprocessing step may significantly speed up further structure analysis [62].

Another and yet more complex task level involves grouping the sections that represent the same underlying musical part: sections that can be seen as repetitions of each other [59, 64, 56]. Finding and grouping all repeated sections provides already a fairly complete description of the musical form, by considering the non-repeated segments as separate and mutually unrelated parts.

Some structure analysis methods have been motivated by finding only one representative section for a piece of music, a "thumbnail" that provides a compact preview of the piece [31, 8, 23, 64]. For this purpose, the most often repeating section is typically suitable.

In this paper, we focus on the structure analysis problem where the objective is to determine 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 by letters $A, B, C, \ldots$ in the order of their first occurrence. Since some of the musical parts have distinct "roles" in Western music, some methods aim to automatically assign the groups with *labels*, such as *verse* or *chorus* [61].

## 3. FEATURE REPRESENTATION

Since the sampled waveform of an acoustic signal is relatively uninformative by itself, some feature extraction has to be employed. The first question to be addressed concerns the acoustic and musical features that humans observe when determining the musical form of a piece. Bruderer et al. [10] conducted experiments to find the perceptual cues that humans use to determine segmentation points in music. The results suggest that "global structure," "change in timbre," "change in level," "repetition," and "change in rhythm" indicated the presence of a structural boundary to the test subjects. We now summarize how some of these aspects can be accounted for by transforming the music signal into suitable feature representations.

### 3.1 Frame Blocking for Feature Extraction

The feature extraction in audio content analysis is normally done in relatively short, 10-100 ms frames. In music struc-

ture analysis each frame of a piece is usually compared to all other frames, which can be computationally intensive. Many of the proposed methods employ a larger frame length in the order of 0.1-1 s. Not only does this reduce the amount of data, but it also allows focusing on a musically more meaningful time scale [63]. The importance of the temporal resolution of feature extraction on the final structure analysis results has been emphasized in [52, 62].

The idea of a musically meaningful time scale has been taken even further in some methods that propose the use of event-synchronized feature extraction. In other words, instead of a fixed frame length and hop size, the division is defined by the temporal locations of sound events [36] or the occurrences of a metrical pulse, e.g., tatum or beat [47, 72, 42, 48, 14, 59]. Using a signal-adaptive frame division has two benefits compared to the use of a fixed frame length: tempo-invariance and sharper feature differences. Tempo-invariance is achieved by adjusting the frame rate according to the local tempo of the piece, which facilitates the comparison of parts performed in different tempi. Event-synchronized frame blocking also allocates consecutive sound events to different frames, which prevents them from blurring each others' acoustic features. In practice, one often calculates the features in short frames and then averages the values over the length of the event-synchronized frames [23, 60, 62, 50].
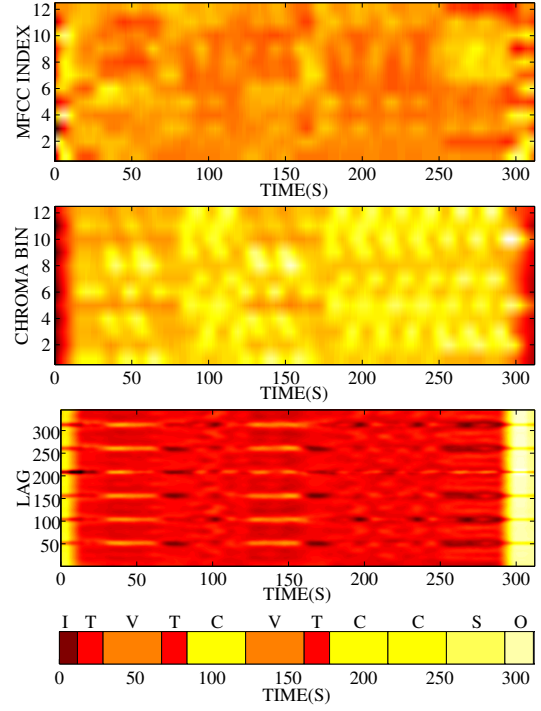
## 3.2 Features

The instrumentation and timbral characteristics are of great importance for the human perception of music structure [10]. Perceptually, timbre is closely related to the recognition of sound sources and depends on the relative levels of the sound at critical bands as well as their temporal evolution. Therefore, a majority of the timbre-based structure analysis methods use *mel-frequency cepstral coefficients* (MFCCs), which parametrize the rough shape of the spectral envelope and thus encode timbral properties of the signal [18]. MFCCs are obtained by discrete cosine transforming (DCT) log-power spectrum on the mel-frequency scale:

$$\mathrm{MFCC}(k) = \sum_{b=0}^{N-1} E(b) \cos\left(\frac{\pi(2b+1)k}{2N}\right), \quad (1)$$

where the subbands $b$ are uniformly distributed on the mel-frequency scale and $E(b)$ is the energy of band $b$. A generally accepted observation is that the lower MFCCs are closely related to the aspect of timbre [3, 74].

As an alternative to using MFCCs as a timbre parametrization, Maddage proposed replacing the mel-spaced filter bank with 4-12 triangular filters in each octave for the task [46]. Other parametrisations omit the DCT step and use some non-mel spacing in band definitions. For example, the MPEG-7 *AudioSpectrumEnvelope* descriptor [35] has been used [78, 41], or very similar constant-Q spectrograms [2, 11]. Aucouturier and Sandler [5] compared different parametrisations of timbral information in music structure analysis and found MFCCs



**Figure 1**: Acoustic features extracted from the piece "Tuonelan koivut" by Kotiteollisuus. The three feature matrices correspond to MFCCs (first panel), chroma (second panel), and rhythmogram (third panel). The annotated structure of the piece is given at the bottom panel, and the parts are indicated with: intro (I), theme (T), verse (V), chorus (C), solo (S), and outro (O).

to outperform other features such as linear prediction coefficients. MFCCs calculated from an example piece are illustrated in the top panel of Fig. 1.

Another important aspect of music is its pitched content on which harmonic and melodic sequences are built upon. In the context of music structure analysis, *chroma features* or *pitch class profiles* have turned out to be a powerful mid-level representation for describing harmonic content [8, 29, 52, 13, 48]. Assuming the equal-tempered scale, the chroma correspond to the set $\{C, C^\sharp, D, \ldots, B\}$ that contains the twelve pitch classes used in Western music notation. A normalized chroma vector describes how the signal's spectral energy is distributed among the 12 pitch classes (ignoring octave information), see Fig. 1 for an illustration.

Several methods for calculating chroma-based audio features have been proposed. Most approaches first compute a discrete Fourier transform (DFT) and then suitably pool the DFT coefficients into chroma bins [8, 29, 31]. Müller et al. [52, 56] propose to use a multirate filter bank consisting of time-domain band-pass filters that correspond to the semitone bands before the chroma projection. Ryynänen and Klapuri replace the DFT analysis by a multipitch estimation front-end [71]. Other chroma-like features are compared in a music structure analysis application by Ong et al. in [58]. Recently, Müller et al. [54] proposed a method to increase the timbre-robustness of chroma by removing some information correlating with the timbre before the octave folding. Some timbre-robustness

is also achieved by the spectral whitening as described in [71]. For an overview of other variants of chroma and pitch-based features, see Müller [52] and Gómez [29].

In contrast to timbral and harmonic content, there has been comparatively little effort in exploiting beat, tempo, and rhythmic information for music structure analysis. To extract such information from audio recordings, most approaches proceed in two steps. In the first step, a detection function, here called *onset accent curve*, is calculated, where high values correlate with the positions of note onsets in the music. The calculation typically relies on the fact that note onsets tend to cause a sudden change of the signal energy and spectrum [9, 80]. In the second step, the accent curves are analyzed with respect to quasiperiodic patterns. Important for the analysis is to obtain a shift-invariant representation that is immune to the exact temporal position of the pattern. Autocorrelation-based analysis allows for detecting periodic self-similarities by comparing an accent curve with time-shifted copies itself [19, 22, 65]. Alternatively, one can use a short-time Fourier transform and then omit the phase in order to derive a shift-invariant representation of the accent curve [65, 32]. Both methods reveal rhythmic properties, such as the tempo or beat structure. These properties typically change over time and are therefore often visualized by means of spectrogram-like representations referred to as *tempogram* [12], *rhythmogram* [38], or *beat spectrogram* [26].

Rhythmic features have not been used in music structure analysis very widely. For example, Jehan [36] used loudness curves, and Jensen [37, 38] included rhythmograms [2] for the structure analysis task. Paulus and Klapuri noted in [60] that the use of rhythmic information in addition to timbral and harmonic features provides useful information to structure analysis, see also Fig. 1. Finally, Peeters [63] has introduced *dynamic features* that aim to parametrize the rhythmic content by describing the temporal evolution of features.

Even though different features describe different musical properties, to date very few methods have utilized more than one feature at a time (except the methods with a large number of more simple features combined with feature vector concatenation [79, 57]). In some approaches MFCC and chroma features have been used to define a single, overlaid self-distance matrix [23, 64], see also Sec. 4. Levy et al. [40] combined information from timbral and harmony related features by feature vector concatenation. A similar approach was adopted by Cheng et al. [14]. Paulus and Klapuri [62] combine the information obtained from MFCCs, chroma features, and rhythmograms using a probabilistic framework.

## 4. SELF-DISTANCE MATRIX

As the musical structure is strongly implied by repetition, a useful strategy is to compare each point of a given audio recording with all the other points, in order to detect self-similarities. The general idea is to convert a given audio recording into a suitable feature sequence, say $(x_1, x_2, \ldots, x_N)$, and then to compare all elements of the sequence with each other in a pairwise fashion. More precisely, given a distance function $d$ that specifies the distance between two feature vectors $x_i$ and $x_j$, it is possible to compute a square *self-distance matrix* (SDM) $D(i, j) = d(x_i, x_j)$ for $i, j \in \{1, 2, \ldots, N\}$. Frequently used distance measures include the Euclidean distance $d_{\mathrm{E}}(x_i, x_j) = \| x_i - x_j \|$, and the cosine distance

$$d_{\mathrm{C}}(x_i, x_j) = 0.5 \left( 1 - \frac{\langle x_i, x_j \rangle}{\| x_i \| \| x_j \|} \right), \qquad (2)$$

where $\| \cdot \|$ denotes vector norm and $\langle \cdot, \cdot \rangle$ dot product. If the distance measure $d$ is symmetric, i.e., $d(x_i, x_j) = d(x_j, x_i)$, the resulting SDM is also symmetric along the main diagonal.

The origins of an SDM representation stems from recurrence plots proposed by Eckmann et al. [21] for the analysis of chaotic systems. The concept of a self-distance matrix [3] has been introduced to the music domain by Foote [24] in order to visualize the time structure of a given audio recording. Naturally, the properties of an SDM crucially depend on the chosen distance measure and the feature representation.

The distance measures are usually defined to compare single frames. Often, it is beneficial to also include the local temporal evolution of the features in order to enhance the structural properties of an SDM. To this end, Foote [24] proposed to average the distance values from a number of consecutive frames and to use that as the distance value. This results in a smoothing effect of the SDM. Müller and Kurth [55] extended these ideas by suggesting a contextual distance measure that allows for handling local tempo variations in the underlying audio recording. Instead of using sliding windows of several consecutive frames, other approaches calculate the average distance from the feature vectors within non-overlapping musically meaningful segments such as musical measure [72, 59]. Jehan [36] calculated SDMs at multiple levels of a temporal hierarchy, starting from individual frames to musical patterns. Each higher level in the hierarchy was calculated based on the SDM of the finer temporal structure.

Recurring patterns in the feature vector sequence $(x_1, x_2, \ldots, x_N)$ are visible in the SDM. The two most important patterns induced by the feature patterns are illustrated in an idealized SDM in Fig. 2. If the features capture musical properties (e.g., instrumentation) that stay somewhat constant over the duration of a musical part, *blocks* of low distance are formed. In case the features describe sequential properties instead of remaining constant within a

---

[2] Recently, Grosche et al. [33] suggested a cyclic variant of a tempogram, which may be a low-dimensional alternative in the structure analysis context. Similar to the concept of cyclic chroma features, where pitches differing by octaves are identified, the cyclic tempogram is obtained by identifying tempi that differ by a power of two.

[3] The dual of SDMs are self-similarity matrices in which each element describes the *similarity* between the frames instead of distance. Most of the following operations can be done with either representation, although here we discuss only SDMs.
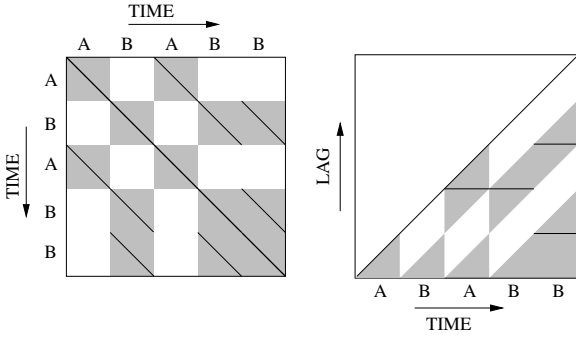
**Figure 2**: **Left:** An example of the patterns formed in SDMs. The sequence consists of two parts, A and B, repeating as indicated, and darker element denotes lower distance. **Right:** Corresponding time-lag matrix of the SDM The non-main diagonal stripes will be transformed into horizontal lines with the vertical position describing the interval (lag) between the occurrences.
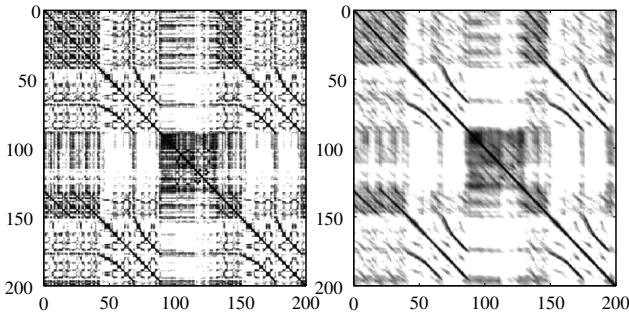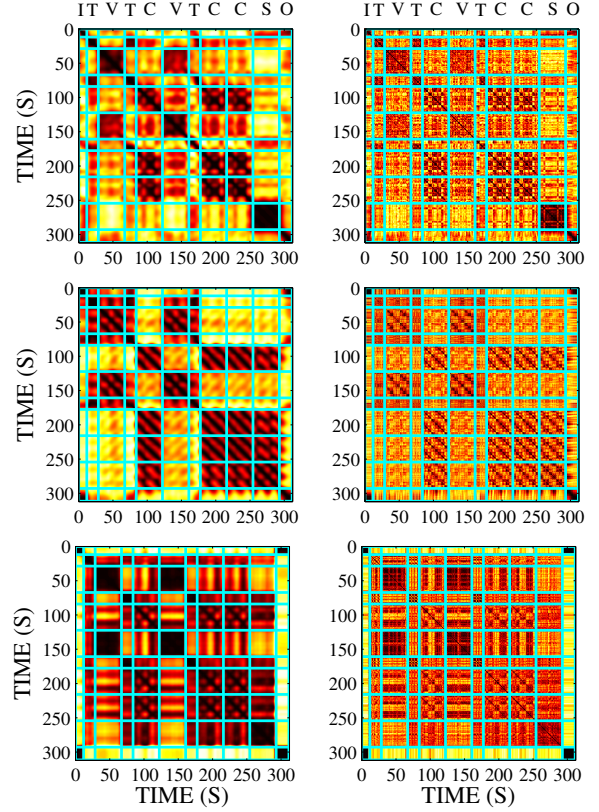


**Figure 3**: **Left:** Self-distance matrix of a piece with tempo variations. **Right:** Path-enhanced version. Darker pixels denote lower distances. Note that some of the stripes are curved expressing relative tempo differences in the repeating parts.



**Figure 4**: Example SDMs from features of Fig. 1 at a coarse (**Left**) and fine (**Right**) time scale. **Top:** MFCCs. **Middle:** Chroma features. **Bottom:** Rhythmogram. Darker pixels denote lower distances. The annotated structure of the piece is indicated by the overlay grid, and the part labels are indicated in the top of the figure with: intro (I), theme (T), verse (V), chorus (C), solo (S), and outro (O). The figure shows how different parts share some of the perceptual aspects, but not all, e.g., chorus and solo have similar harmonic but differring timbral content.

part, diagonal *stripes* of low distance are formed. If such a part is repeated, one finds stripes in the SDM that run parallel to the main diagonal. This is often the case when using chroma features, which then reveal repeated harmonic progressions within a piece. Locating and interpreting these patterns with various methods is the main approach employed in many of the structure analysis methods described in the literature.

As Peeters [63] noted, the features alone do not determine whether blocks or stripes are formed, but the temporal parameters of the feature extraction process are also important. In other words, the longer the temporal window is that the feature vector describes, the more likely it is that blocks are formed in the SDM. Therefore, working with low resolutions may not only be beneficial for computational, but also for structural reasons [56, 60]. The effect of the time scale parameter used in the feature computation on the resulting SDMs is also illustrated by Fig. 4.

Often a musical part is repeated in another key. Using chroma features, Goto [31] simulates transpositions by cyclically shifting the chroma. Adopting this idea, Müller and Clausen [53] introduced the concept of transposition-invariant SDMs, which reveals the repetitive structure even in the presence of key transpositions.

Another way to present repetitive information is to transform an SDM into a time-lag format [31]. In an SDM

$D$ both the axes represent absolute time, whereas in the time-lag matrix $R$ one axis is changed to represent time difference (lag) instead

$$R(i, i - j) = D(i, j), \quad \text{for } i - j > 0. \quad (3)$$

The ordinate transformation discards the duplicate information of a symmetric SDM, see Fig. 2. The diagonal stripes formed by repeated sequences appear as horizontal lines in the time-lag representation, and may be easier to extract. Even though a time-lag representation transforms the stripe information into a more easily interpretable form, the block information is transformed into parallelograms and may now be more difficult to extract. Furthermore, the time-lag representation only works when repeating parts occur in the same tempo, which is, in particular for classical music, often not the case. Structure analysis in the presence of temporal variations is discussed in [52, 56], see also Fig. 3 for an illustration.

## 5. STRUCTURE ANALYSIS APPROACHES

As mentioned before, there are a variety of different methods proposed for music structure analysis. An overview of the operational entities of the proposed methods is shown
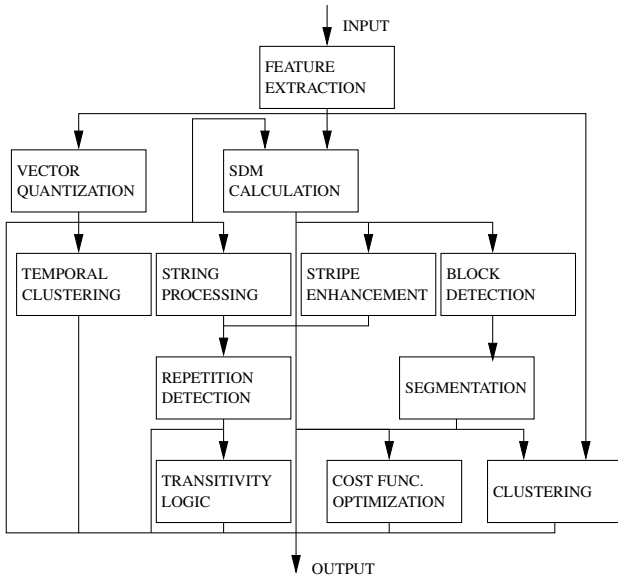
**Figure 5**: An overview block diagram of various operational entities employed in music structure analysis methods.

in Fig. 5. Furthermore, relevant literature along with a classification of the involved methods is summarized by Table 1. In this section, we describe the main approaches as well as the interconnections between the operational entities in more detail. The first categorization of music structure analysis methods was proposed by Peeters [63] dividing them into *sequence* and *state* approaches. The *sequence approaches* assume that there are sequences of events that are repeated several times in the given musical signal, thus forming diagonal stripes in the corresponding SDM. The *state approaches* in turn consider the piece to be produced by a finite state machine, where each state produces some part of the signal. Considering the SDM representation, the state approaches can be thought to form the blocks [4]. As mentioned in Sec. 1, we use the more semantically motivated term *repetition-based approach* instead of the more technically motivated term *sequence approach*. Similarly, we use the term *homogeneity-based approach* instead of the term *state approach*. Furthermore, we add a third category referred to as *novelty-based approach*. In the following, we describe some instantiations of each of the categories in more detail and then discuss some combined approaches.

## 5.1 Novelty-based Approaches

An important principle in music 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 [25] tries to identify segment boundaries by detecting 2D corner points in an SDM of size $N \times N$ using

---

[4] In principle a state is capable of emitting also a feature sequence forming stripes in SDM when repeated. However, the name "state approach" is more often used of methods that utilize principles of homogeneity.
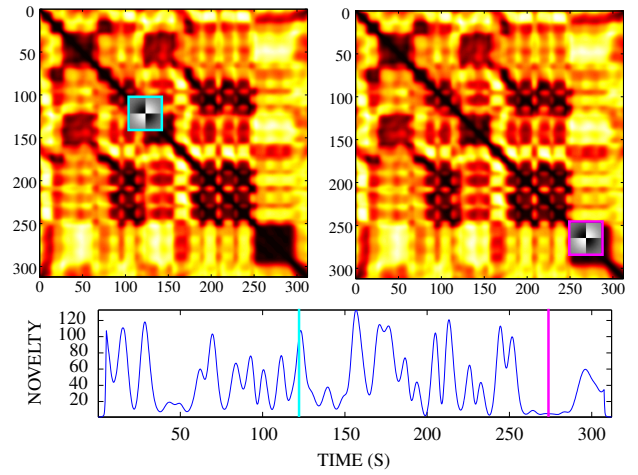


**Figure 6**: **Top:** Two instances of the SDM using MFCCs from Fig. 4. The checkerboard-like kernel that is correlated along the main diagonal is shown at two different positions on the left and right. **Bottom:** Resulting novelty curve.

a kernel matrix of a lower dimension. The kernel consists of an $M \times M$ matrix (with $M < N$) which has a $2 \times 2$ checkerboard-like structure and is possibly weighted by a Gaussian radial function. The kernel is illustrated within the small rectangles on top of the two SDMs in Fig. 6. The kernel is then correlated along the main diagonal of the SDM. 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. For an illustration, we refer to Fig. 6. Similarly, using other feature representation such as chroma features or rhythmograms, one obtains indicators for changes in harmony, rhythm, or tempo.

Jensen uses a different approach for locating the main diagonal blocks in an SDM [38] by formulating the segmentation as an optimization problem. The cost function to be optimized tries to minimize the average distance within blocks (defined by neighboring segment boundaries) of the SDM while keeping the number of segments small. Tzanetakis and Cook [76] propose to segment a signal by first extracting a set of features from the signal and then calculating a Mahalanobis distance between successive frames. Large differences in the distance values indicate possible segmentation points. For other methods to music segmentation, we refer to the publication by Turnbull et al. [75], in which several acoustic features and both supervised as well as unsupervised segmentation methods are evaluated.

## 5.2 Homogeneity-based Approaches

A direct continuation of the *novelty-based* procedure is to analyze the content of the created segments and to classify them building up *homogenous* clusters. Such an approach was introduced by Cooper and Foote in [15], where, after a novelty-based segmentation, the content of each segment is modeled by a normal distribution. Then, the similarity between two segments is computed using the Kullback-Leibler divergence between two multivariate normal distributions [28]. Having the distances for all segment pairs,

| Author / publication | Task | Acoustic features | Approach | Method |
|---|---|---|---|---|
| Aucouturier et al. [4] | full structure | spectral envelope | homogeneity | HMM |
| Barrington et al. [7] | full structure | MFCC / chroma | homogeneity | dynamic texture model |
| Bartsch & Wakefield [8] | thumbnailing | chroma | repetition | stripe detection |
| Chai [13] | full structure | chroma | repetition | stripe detection |
| Cooper & Foote [15] | summarisation | magnitude spectrum | homogeneity | segment clustering |
| Dannenberg & Hu [17] | repetitions | chroma | repetition | dynamic programming |
| Eronen [23] | chorus detection | MFCC+chroma | repetition | stripe detection |
| Foote [24] | visualization | MFCC | | self-similarity matrix |
| Foote [25] | segmentation | MFCC | novelty | novelty vector |
| Goto [31] | repetitions | chroma | repetition | stripe detection (*RefraiD*) |
| Jehan [36] | pattern learning | MFCC+chroma+loudness | homogeneity | hierarchical SDMs |
| Jensen [38] | segmentation | MFCC+chroma+rhythmogram | novelty | diagonal blocks |
| Levy & Sandler [41] | full structure | MPEG-7 timbre descriptor | homogeneity | temporal clustering |
| Logan & Chu [43] | key phrase | MFCC | homogeneity | HMM / clustering |
| Lu et al. [44] | thumbnailing | constant-Q spectrum | repetition | stripe detection |
| Maddage [46] | full structure | chroma | homogeneity | rule-based reasoning |
| Marolt [48] | thumbnailing | chroma | repetition | RefraiD |
| Mauch et al. [50] | full structure | chroma | repetition | greedy selection |
| Müller & Kurth [56] | multiple repetitions | chroma statistics | repetition | stripe search & clustering |
| Ong [57] | full structure | multiple | repetition | RefraiD |
| Paulus & Klapuri [59] | repeated parts | MFCC+chroma | repetition | cost function |
| Paulus & Klapuri [62] | full description | MFCC+chroma+rhythmogram | combined | fitness function |
| Peeters [63] | full structure | dynamic features | homogeneity | HMM, image filtering |
| Peeters [64] | repeated parts | MFCC+chroma+spec. contrast | repetition | stripe detection |
| Rhodes & Casey [70] | hierarchical structure | timbral features | repetition | string matching |
| Shiu et al. [72] | full structure | chroma | repetition | state model stripe detection |
| Turnbull et al. [75] | segmentation | various | novelty | various |
| Wellhausen & Höynck [78] | thumbnailing | MPEG-7 timbre descriptor | repetition | stripe detection |

**Table 1**: A summary of discussed methods for music structure analysis.

the segments are grouped with spectral clustering [77]. Logan and Chu [43] used a similar Gaussian parametrization on segments of fixed length and applied agglomerative hierarchical clustering. The method proposed by Goodwin and Laroche [30] performs the segmentation and clustering at the same time. The method itself resembles the optimization procedure described by Jensen [38], with the difference that the searched path can now return to a state defined earlier if it is globally more efficient for the structure description.

The concept of *state* is taken more explicitly in methods employing hidden Markov models (HMMs) for the analysis, see, e.g., [5, 27]. Here, the basic assumption is that each musical part can be represented by a state in an HMM, and the states produce observations from the underlying probability distribution. In an HMM, the probability of a state sequence $\mathbf{q} = (q_1, q_2, \ldots, q_N)$ given the observation sequence $\mathbf{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N)$ can be calculated by

$$P(\mathbf{q}|\mathbf{X}) \propto P(\boldsymbol{x}_1|q_1) \prod_{n=2}^{N} P(\boldsymbol{x}_n|q_n)p(q_n|q_{n-1}), \quad (4)$$

where $P(\boldsymbol{x}_n|q_n)$ is the likelihood of observing $\boldsymbol{x}_n$ if the state is $q_n$, and $p(q_n|q_{n-1})$ is the transition probability from state $q_{n-1}$ to state $q_n$. The analysis operates by training the HMM with the piece to be analyzed, and then by decoding (finding the most probable state sequence) the same signal with the model. Effectively this implements vector quantization of the feature vectors with some temporal dependency modeling expressed by the state transition probabilities. Though this model has a certain appeal, it does not work very well in practice because the result is often temporally fragmented, as noted by Peeters et al. [68]. The fragmentation is due to the fact that the in-

dividual states tend to model individual sound events rather than longer musical parts.

To alleviate the problem of temporal fragmentation, several post-processing methods have been proposed. Here, the state sequence produced by an HMM is only used as a mid-level representation for further analysis, where each state represents a certain context-dependent short sound event [41]. Fig. 7 shows the resulting state sequences of an example piece after analyzing it with fully connected HMMs with 8 and 40 states, respectively. The state sequence representation is included also for general audio parametrization in the MPEG-7 standard as the *SoundModelStatePathType* descriptor [35]. Abdallah et al. [1] proposed to calculate histograms of the states with a sliding window over the entire sequence and then to use the resulting histogram vectors as new feature representation. Based on these state histograms, probabilistic clustering is applied. This method was extended to include statistical modeling of the cluster durations [2]. Levy et al. [42] increased the amount of the contextual knowledge using a variant of a fuzzy clustering approach applied on the histograms. This approach was formalized by Levy and Sandler [41] using a probabilistic framework. Despite the relatively simple approach, the temporal clustering method [42] has proven to work quite well.

A slightly different approach to reduce the resulting fragmentation was proposed by Peeters [68]. He performed initial segmentation based on an SDM and then used the average feature value over each individual segment as initial cluster centroids that he further updated using k-means clustering. The obtained cluster centroids were then used to initialize the training of an HMM which produced the final clustering. In a recent publication Barrington et al. [7] propose to use dynamic texture mixture
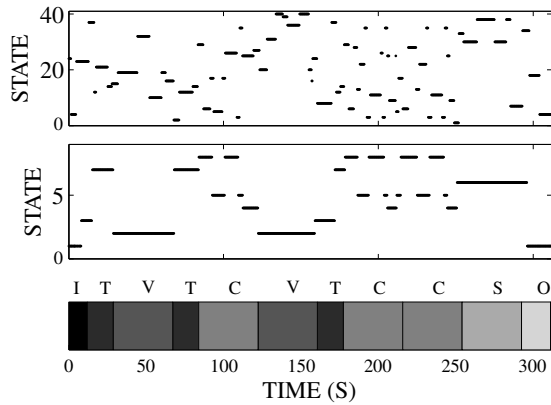
**Figure 7**: State sequences resulting from a fully connected HMM using 40 (**Top**) and 8 (**Middle**) states applied to the MFCC feature sequence of Fig. 1. The bottom panel shows the annotated ground truth structure.

models (DTM) for the structure analysis. DTM is basically a state model, where each (hidden) state produces observations that have a temporal structure. The main novelty of the method compared to the HMM-based state methods is that the observation model itself takes the temporal behavior of the produced observations into account, and there will be less need for heuristic post-processing.

### 5.3 Repetition-based Approaches

The repetition of musical entities, as already noted in Sec. 1, is an important element in imposing structure on a sequence of musical sounds. Here, the temporal order in which the sound events occur is crucial to form musically meaningful entities such as melodies or chord progressions. Therefore, the task of extracting the repetitive structure of a given audio recording of a piece of music amounts to first transform the audio into a suitable feature sequence and then to find repeating subsequences in it.

As was explained in Sec. 4, one possible approach is to compute an SDM and to search for diagonal stripes parallel to the main diagonal. 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 [56, 52]. To enhance the stripe structure, many approaches apply some sort of low-pass filtering to smooth the SDM along the diagonals [78, 8]. A similar effect can be achieved by averaging the distance values from a number of consecutive frames and to use that as the distance value [24]. Marolt [48] proposed to enhance the stripes by calculating multiple SDMs with different sliding window lengths and then by combining them with element-wise multiplication. Lu et al. [44] employed multiple iterations of erosion and dilation filtering along the diagonals to enhance the stripes by filling small breaks and removing too short line segments. Ong [57] extended the erosion and dilation filtering into two-dimensional filter to enhance the entire SDM. Goto [31] employed a two-dimensional lo-

cal filter to enhance the stripes; similar enhancement was later utilized by Eronen [23]. Peeters [64] proposed to low-pass filter along the diagonal direction, and high-pass filter along the anti-diagonal direction to enhance the stripes.

Most of the above approaches assume that the repeating parts are played in the same tempo, resulting in stripes that run exactly in parallel to the main diagonal. However, this assumption may not hold in general. For example, in classical music there are many recordings where certain parts are repeated in different tempi or where significant tempo changes (e.g. riterdando, accelerando, rubato) are realized differently in repeating parts. Here, the stripes may be even curved paths as indicate by Fig. 3. Müller et al. [55, 52] introduced smoothing techniques that can handle such situations by incorporating contextual information at various tempo levels into a single distance measure.

After enhancing the stripe structure, the stripe segments can be found, e.g., by thresholding. The *RefraiD* approach proposed by Goto [31] has later been employed by several studies [48, 57]. It uses the time-lag version of SDM to select the lags that are more likely to contain repeats, and then detect the line segments along the horizontal direction of the lags. Each of the found stripes specifies two occurrences of a sequence: the original one and a repeat. For chorus detection, or simple one-clip thumbnailing, selecting a sequence that has been repeated most often has proven to be an effective approach. In the case that a more comprehensive structural description is wanted, multiple stripes have to be detected as well as some logical reasoning to deduce the underlying structure as proposed by Dannenberg [17].

Similar to the dynamic programming approaches used for segmentation [30, 38], some of the stripes can be found by a path search. Shiu et al. [73] interpret the self-similarity values as probabilities and define a local transition cost to prefer diagonal movement. Then, Viterbi search is employed to locate the optimal path through the lower (or upper) triangle of the SDM. The stripes have large similarity values, thus the probability values are also large and the path is likely to go through the stripe locations. Another method to locate stripe segments by growing them in a greedy manner was proposed by Müller and Kurth [56]. These approaches are advantageous in that they are able to handle tempo differences in the repeats.

Rhodes and Casey [70] employed a string matching method to the HMM state sequence representation to create a hierarchical description of the structure. Though the algorithm was presented to operate on a finite alphabet formed by the HMM states, the authors suggest that similar operations could be accomplished with feature vectors after modifying the matching algorithm to accept vector inputs. Aucouturier and Sandler [6] proposed another method for inspecting the HMM state sequences with image processing methods. The main idea is to calculate a binary co-occurrence matrix (resembling an SDM) based on the state sequence, which elements have the value 1, if the two frames have the same state assignment, and the value 0 otherwise. Then a diagonal smoothing kernel is
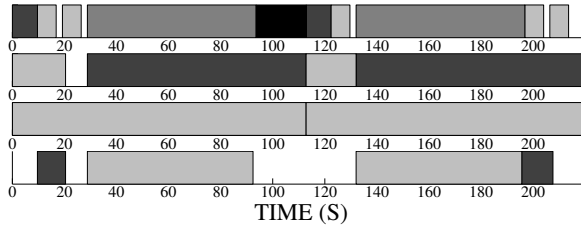
**Figure 8**: Effect of differently weighting the terms in the cost function of [59] on the final structure description. **Top:** Annotated ground truth. **Second row:** Analysis result with some reasonable values for the weights. **Third row:** Result with increased weight of the *complexity* term. **Bottom:** Result with a decreased weight for the term *amount unexplained*.
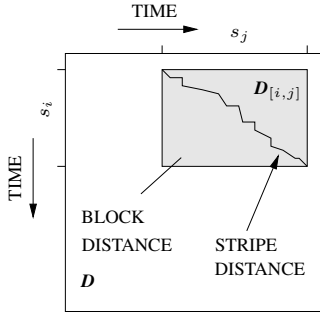


**Figure 9**: Illustration of the basic ideas behind the *stripe* and *block* distances between two segments $s_i$ and $s_j$ of a piece. The stripe distance is based on the path of least cost through the submatrix $D_{[i,j]}$ while the block distance is based on the average distance value within the submatrix.

applied on the matrix to smooth out small mismatches between sequences. Finally, stripes are searched from the resulting matrix with Hough transform, which is claimed to be relatively robust against bad or missing data points.

### 5.4 Combined Approaches

Most methods for music structure analysis described so far rely on a single strategy. For example, homogeneity-based approaches try to locate blocks of low distance on the SDM main diagonal and then to classify them. Or, repetition-based approaches try to extract stripes from the SDM and then to deduce the repetitive structure. An alternative approach is to focus on modeling the properties of a good structural description, and in doing so, to combine different segmentation principles. This is the idea of Paulus and Klapuri [59, 62], who proposed a cost function for structural descriptions of a piece that considers all the desired properties, and then, for a given acoustic input, minimized the cost function over all possible structural descriptions. A similar approach was also suggested by Peiszer [69]. In [59], the cost function included terms representing the *within-group dissimilarity* (repeats should be similar), the *amount unexplained* (the structural description would cover as much of the piece as possible), and the *complexity* (the structure should not be fragmented). The effect of the balancing of these three terms is illustrated in Fig. 8.

The main weakness of the cost function based method

described above—as well as with most of the other methods relying on locating individual stripes or blocks in the SDM—is that they operate only on parts of the SDM. In other words, when locating stripes, each of the stripes is handled separately without any contextual information. Considering structure analysis as a data clustering problem, each of the formed clusters should be compact (having small within-group distances), and the clusters should be well-separated (having large between-group distances). Paulus and Klapuri [62] formalized these ideas using a probabilistic framework. Here, replacing the cost function, a *fitness measure* is defined for jointly measuring within-group distance (which should be small) and between-group distance (which should be large). To this end, for each segment pair, two distances were calculated: a *stripe distance* that measures the distance of the feature sequences corresponding to the two segments (using dynamic time warping) and a *block distance* that measures the average distance over all frame pairs of the two segments, see also Fig. 9. Maximizing the fitness measure then resulted in a reasonable trade-off between these two types of complementary information. Multiple feature representations (MFCCs, chroma features, rhythmogram) were integrated into the fitness measure to account for the various musical dimensions, see Sec. 3.

In [62], the combinatorial optimization task over all descriptions was approximately solved by limiting the set of possible segments. To this end, a set of candidate segmentation points was created using a novelty-based method [25], and then a greedy algorithm over the remaining search space was applied. As a result, the method combines all the segmentation principles discussed in Sec. 5: a *novelty-based approach* was used to reduce the number segment candidates, and *homogeneity-based* and *repetition-based* approaches were integrated in the fitness measure. One drawback of the described approach is that the final structure description crucially depends on the first novelty detection step, which was found to be a bottle-neck in some cases.

## 6. EVALUATION

Music is multi-faceted and complex. Even though it is structured and obeys some general rules, music also lives from expanding and even breaking these rules. Therefore it can be problematic to give a concise and unique structural description for a piece of music. As a consequence, evaluating the performance of an automated structure analysis method is not as simple as it may initially seem. We now briefly discuss some of the evaluation metrics proposed in the literature.

To evaluate the accuracy of segmentation boundaries, most evaluation procedures involve some sort of recall rate, precision rate, and F-measure while accepting a small temporal deviation [75]. An alternative is to calculate the mean (or median) time between a claimed and annotated segmentation point [75]. The evaluation of music thumbnailing requires user studies, since the quality of the output is usually measured subjectively instead of an objective met-

ric, as described by Chai [13] and Ong [57].

Evaluating the result of a method producing a description of the full structure of a piece is less straightforward. Many of the evaluation metrics adopt an approach similar to evaluating clustering results: pairs of frames are inspected, and if they belong to any occurrence of the same musical part, they are considered to belong to the same cluster, denoted by the set $\mathbb{F}_A$ in case of ground truth and the set $\mathbb{F}_E$ in the case of analysis result. Based on these two sets, it is possible to calculate the pairwise precision rate $R_P = |\mathbb{F}_A \cap \mathbb{F}_E|/|\mathbb{F}_E|$, the pairwise recall rate $R_R = |\mathbb{F}_A \cap \mathbb{F}_E|/|\mathbb{F}_A|$, and the F-measure

$$F = \frac{2R_P R_R}{R_P + R_R}. \tag{5}$$

Using the above evaluation metric was proposed by Levy and Sandler [41]. Another closely related metric is the Rand index [34], used by Barrington et al. [7]. Abdallah et al. [1] proposed to match the segments in the analysis result and ground truth and to calculate a directional Hamming distance between frame sequences after the match. A similar approach with a differing background was proposed by Peeters [64]. A second evaluation metric proposed by Abdallah et al. [1] treats the structure descriptions as symbol sequences and calculates the mutual information between the analysis result and the ground truth. The mutual information concept was developed further by Lukashevich [45], who proposed an over- and under-segmentation measures based on the conditional entropies of the sequential representations of structures.

A property that can be considered to be a weakness in the metrics relying on pairs of frames, is that they disregard the order of the frames. In other words, they do not penalize hierarchical level differences between the computed parts such as splittings of segments into smaller parts. Chai [13], and Paulus and Klapuri [59] proposed heuristics finding a common hierarchical level for the computed structure result and the ground truth structure. However, the evaluation method is rather complicated, and the results are still subject for discussion.

Finally, it should be noted that most of the suggested evaluation metrics only consider one type of provided ground truth annotation. As the experiments by Bruderer et al. [10] suggest, the perception of musical structures is generally ambiguous. Thus the descriptions provided by two persons on the same piece may differ. A small-scale comparison of descriptions made by two annotators was presented by Paulus and Klapuri [62], and slight differences in the hierarchical levels as well as in the grouping were noted (using the F-measure (5) as the metric, human vs. human result was $89.4\%$ whereas the employed computational method reached $62.4\%$). Peeters and Deruty [67] proposed a more well-defined ground truth annotation scheme that allows annotating the structure of a piece from several different aspects and temporal scales at the same time. The annotation can then be transformed to focus on the aspect relevant to the current application, e.g., by reducing it to be a temporal segmentation and grouping, as with earlier data sets.

The first systematic evaluation of different structure analysis methods methods took place in the Music Structure Segmentation task at the Music Information Retrieval Evaluation eXchange (MIREX) 2009 [5]. MIREX itself is a framework for evaluating music information retrieval algorithms where the evaluation tasks are defined by the research community under the coordination of International Music Information Retrieval Systems Evaluation Laboratory at the University of Illinois at Urbana-Champaign [20]. The evaluation task was kept relatively straightforward: providing a temporal segmentation of an entire piece and grouping of segments to parts. The evaluation data was provided from the OMRAS2 metadata project [49], and it consisted of 297 songs, mostly by The Beatles (179 songs), and the remaining songs were from four other performers making the data rather homogenous. It should also be noted that a large part of the data was publicly available before the evaluation and may have been used in the development of some of the methods. The five submissions from four teams represent slightly different approaches: one searches diagonal stripes from SDM in a greedy manner [50] ($F = 60.0\%$), one aims at maximizing a fitness function from a combined approach [62] ($F = 53.0\%$), and one uses agglomerative hierarchical clustering on smaller segments [66] ($F = 53.3\%$). The details of the two other submissions ($F = 57.7\%$ and $F = 58.2\%$) were not published. Despite the differring approaches, there were no significant performance differences between the methods and depending on the evaluation metric the ranking order changed considerably (with the Rand index metric the ranking is almost reversed).

## 7. CONCLUSIONS

This paper has given an overview of the music structure analysis problem, and the methods proposed for solving it. The methods have been divided into three categories: novelty-based approaches, homogeneity-based approaches, and repetition-based approaches. The comparison of different methods has been problematic because of the differring goals, but an effort at this was made in MIREX2009. The results of the evaluations suggest that none of the approaches is clearly superior at this time, and that there is still room for considerable improvements.

Perhaps one of the largest problems in music structure analysis is not directly technical, but more conceptual: the ground truth for this task should be better defined. The need for this is indicated by the fact that the annotations made by two persons disagree to a certain degree [62]. Defining the ground truth better requires interdisciplinary work between engineers and musicologists. The current results suggest that the structure description should not only be on a single level, but include also the information of hierarchical recurrences—similar to human perception. Another major task consists in collecting and annotating a representative data set, which is free for use in research projects worldwide. Also, contrary to many earlier

data sets, it would be beneficial to have multiple parties involved to ensure data diversity and agreement on the target data. Having more accurate ground truth and a representative data set, the evaluation metrics can be defined more rigorously too: none of the current metrics corresponds to the perceived performance very accurately.

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. As has been discussed in this paper, it is not enough to only use a single musical aspect in the analysis—also humans typically utilize multiple cues simultaneously. Related to this, more perceptually (and musically) motivated features should be investigated, as well as the distance measures used to compare frame-level features. Methods combining several musically motivated information sources have shown promising results, and such trends should be pursued further.

## 8. REFERENCES

[1] S. Abdallah, K. Nolad, M. Sandler, M. Casey, and C. Rhodes. Theory and evaluation of a Bayesian music structure extractor. In *Proc. of 6th International Conference on Music Information Retrieval*, pages 420–425, London, England, UK, Sept. 2005.

[2] S. Abdallah, M. Sandler, C. Rhodes, and M. Casey. Using duration models to reduce fragmentation in audio segmentation. *Machine Learning*, 65(2–3):485–515, Dec. 2006.

[3] J.-J. Aucouturier and F. Pachet. Improving timbre similarity: How high's the sky. *Journal of Negative Results in Speech and Audio Sciences*, 1, 2004.

[4] J.-J. Aucouturier, F. Pachet, and M. Sandler. "The way it sounds": Timbre models for analysis and retrieval of music signals. *IEEE Transactions on Multimedia*, 7(6):1028–1035, Dec. 2005.

[5] J.-J. Aucouturier and M. Sandler. Segmentation of musical signals using hidden Markov models. In *Proc. of 110th Audio Engineering Society Convention*, Amsterdam, The Netherlands, May 2001.

[6] J.-J. Aucouturier and M. Sandler. Finding repeating patterns in acoustic musical signals: Applications for audio thumbnailing. In *Proc. of Audio Engineering Society 22nd International Conference on Virtual, Synthetic and Entertainment Audio*, pages 412–421, Espoo, Finland, 2002.

[7] L. Barrington, A. B. Chan, and G. Lanckriet. Modeling music as a dynamic texture. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(3):602–612, Mar. 2010.

[8] 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.

[9] 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, Sept. 2005.

[10] M. J. Bruderer, M. McKinney, and A. Kohlrausch. Structural boundary perception in popular music. In *Proc. of 7th International Conference on Music Information Retrieval*, pages 198–201, Victoria, B.C., Canada, Oct. 2006.

[11] M. Casey and M. Slaney. The importance of sequences in musical similarity. In *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 5–8, Toulouse, France, May 2006.

[12] A. T. Cemgil, B. Kappen, P. Desain, and H. Honing. On tempo tracking: Tempogram representation and kalman filtering. *Journal of New Music Research*, 28(4):259–273, 2001.

[13] W. Chai. *Automated Analysis of Musical Structure*. PhD thesis, Massachusetts Institute of Technology, Boston, Mass., USA, Sept. 2005.

[14] H.-T. Cheng, Y.-H. Yang, Y.-C. Lin, and H. H. Chen. Multimodal structure segmentation and analysis of music using audio and textual information. In *Proc. of IEEE International Symposium on Circuits and Systems*, pages 1677–1680, Taipei, Taiwan, May 2009.

[15] M. Cooper and J. Foote. Summarizing popular music via structural similarity analysis. In *Proc. of 2003 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 127–130, New Platz, N.Y., USA, Oct. 2003.

[16] 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, N.Y., USA, 2008.

[17] R. B. Dannenberg and N. Hu. Pattern discovery techniques for music audio. In *Proc. of 3rd International Conference on Music Information Retrieval*, pages 63–70, Paris, France, Oct. 2002.

[18] S. B. Davis and P. Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *Readings in Speech Recognition*, pages 65–74, 1990.

[19] S. Dixon, E. Pampalk, and G. Widmer. Classification of dance music by periodicity patterns. In *Proc. of 4th International Conference on Music Information Retrieval*, pages 159–165, Baltimore, Md., USA, Oct. 2003.

[20] J. S. Downie. The music information retrieval evaluation exchange (2005–2007): A window into music information retrieval research. *Acoustical Science and Technology*, 29(4):247–255, 2008.

[21] J.-P. Eckmann, S. O. Kamphorst, and D. Ruelle. Recurrence plots of dynamical systems. *Europhysics Letters*, 4(9):973–977, Nov. 1987.

[22] D. P. W. Ellis. Beat tracking by dynamic programming. *Journal of New Music Research*, 36(1):51–60, 2007.

[23] A. Eronen. Chorus detection with combined use of MFCC and chroma features and image processing filters. In *Proc. of 10th International Conference on Digital Audio Effects*, pages 229–236, Bordeaux, France, Sept. 2007.

[24] J. Foote. Visualizing music and audio using self-similarity. In *Proc. of ACM Multimedia*, pages 77–80, Orlando, Fla., USA, 1999.

[25] J. Foote. Automatic audio segmentation using a measure of audio novelty. In *Proc. of IEEE International Conference on Multimedia and Expo*, pages 452–455, New York, N.Y., USA, Aug. 2000.

[26] J. Foote and S. Uchihashi. The beat spectrum: A new approach to rhythm analysis. In *Proc. of IEEE International Conference on Multimedia and Expo*, pages 881–884, Tokyo, Japan, Aug. 2001.

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

[28] J. Goldberger, S. Gordon, and H. Greenspan. An efficient image similarity measure based on approximations of KL-divergence between two Gaussian mixtures. In *Proc. of Ninth IEEE International Conference on Computer Vision*, pages 487–493, 2003.

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

[30] M. M. Goodwin and J. Laroche. A dynamic programming approach to audio segmentation and music / speech discrimination. In *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 309–312, Montreal, Que., Canada, May 2004.

[31] M. Goto. A chorus-section detecting method for musical audio signals. In *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 437–440, Hong Kong, 2003.

[32] P. Grosche and M. Müller. A mid-level representation for capturing dominant tempo and pulse information in music recordings. In *Proc. of 10th International Conference on Music Information Retrieval*, pages 189–194, Kobe, Japan, Oct. 2009.

[33] P. Grosche, M. Müller, and F. Kurth. Cyclic tempogram a mid-level tempo representation for music signals. In *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing*, Dallas, Texas, USA, Mar. 2010.

[34] L. Hubert and P. Arabie. Comparing partitions. *Journal of Classification*, 2(1):193–218, Dec. 1985.

[35] International Organization for Standardization. *ISO/IEC 15938-4:2002 Information technology – Multimedia content description interface – Part 4: Audio*. Geneva, Switzerland, 2002.

[36] T. Jehan. *Creating Music by Listening*. PhD thesis, Massachusetts

Institute of Technology, Boston, Mass., USA, Sept. 2005.

[37] K. Jensen. A causal rhythm grouping. In *Computer Music Modeling and Retrieval*, volume 3310 of *Lecture Notes in Computer Science*, pages 83–95. Springer Berlin / Heidelberg, 2004.

[38] K. Jensen. Multiple scale music segmentation using rhythm, timbre, and harmony. *EURASIP Journal on Advances in Signal Processing*, 2007. Article ID 73205, 11 pages.

[39] F. Lerdahl and R. Jackendoff. *A Generative Theory of Tonal Music*. The MIT Press, Cambridge, Mass., USA, 1999.

[40] M. Levy, K. Noland, and M. Sandler. A comparison of timbral and harmonic music segmentation algorithms. In *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 1433–1436, Honolulu, Hawaii, USA, Apr. 2007.

[41] 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, Feb. 2008.

[42] M. Levy, M. Sandler, and M. Casey. Extraction of high-level musical structure from audio data and its application to thumbnail generation. In *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 13–16, Toulouse, France, May 2006.

[43] B. Logan and S. Chu. Music summarization using key phrases. In *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 749–752, Istanbul, Turkey, June 2000.

[44] L. Lu, M. Wang, and H.-J. Zhang. Repeating pattern discovery and structure analysis from acoustic music data. In *Proc. of Workshop on Multimedia Information Retrieval*, pages 275–282, New York, N.Y., USA, Oct. 2004.

[45] H. Lukashevich. Towards quantitative measures of evaluating song segmentation. In *Proc. of 9th International Conference on Music Information Retrieval*, pages 375–380, Philadelphia, Pa., USA, Sept. 2008.

[46] N. C. Maddage. Automatic structure detection for popular music. *IEEE Multimedia*, 13(1):65–77, Jan. 2006.

[47] N. C. Maddage, C. Xu, M. S. Kankanhalli, and X. Shao. Content-based music structure analysis with applications to music semantics understanding. In *Proc. of ACM Multimedia*, pages 112–119, New York, N.Y., USA, Oct. 2004.

[48] M. Marolt. A mid-level melody-based representation for calculating audio similarity. In *Proc. of 7th International Conference on Music Information Retrieval*, pages 280–285, Victoria, B.C., Canada, Oct. 2006.

[49] M. Mauch, C. Cannam, M. Davies, S. Dixon, C. Harte, S. Kolozali, D. Tidhar, and M. Sandler. OMRAS2 metadata project 2009. In *Proc. of 10th International Conference on Music Information Retrieval*, Kobe, Japan, Oct. 2009. Extended abstract.

[50] M. Mauch, K. Noland, and S. Dixon. Using musical structure to enhance automatic chord transcription. In *Proc. of 10th International Conference on Music Information Retrieval*, pages 231–236, Kobe, Japan, Oct. 2009.

[51] R. Middleton. Form. In B. Horner and T. Swiss, editors, *Key terms in popular music and culture*, pages 141–155. Wiley-Blackwell, Sept. 1999.

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

[53] M. Müller and M. Clausen. Transposition-invariant self-similarity matrices. In *Proc. of 8th International Conference on Music Information Retrieval*, pages 47–50, Vienna, Austria, Sept. 2007.

[54] M. Müller, S. Ewert, and S. Kreuzer. Making chroma features more robust to timbre changes. In *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 1877–1880, Taipei, Taiwan, Apr. 2009.

[55] M. Müller and F. Kurth. Enhancing similarity matrices for music audio analysis. In *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 437–440, Toulouse, France, May 2006.

[56] 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.

[57] B. S. Ong. *Structural analysis and segmentation of musical signals*. PhD thesis, Universitat Pompeu Fabra, Barcelona, Spain, 2006.

[58] B. S. Ong, E. Gómez, and S. Streich. Automatic extraction of musical structure using pitch class distribution features. In *Proc. of 1st Workshop on Learning the Semantics of Audio Signals*, Athens, Greece, Dec. 2006.

[59] J. Paulus and A. Klapuri. Music structure analysis by finding repeated parts. In *Proc. of 1st ACM Audio and Music Computing Multimedia Workshop*, pages 59–68, Santa Barbara, Calif., USA, Oct. 2006.

[60] J. Paulus and A. Klapuri. Acoustic features for music piece structure analysis. In *Proc. of 11th International Conference on Digital Audio Effects*, pages 309–312, Espoo, Finland, Sept. 2008.

[61] J. Paulus and A. Klapuri. Labelling the structural parts of a music piece with Markov models. In S. Ystad, R. Kronland-Martinet, and K. Jensen, editors, *Computer Music Modeling and Retrieval: Genesis of Meaning in Sound and Music - 5th International Symposium, CMMR 2008 Copenhagen, Denmark, May 19-23, 2008, Revised Papers*, volume 5493 of *Lecture Notes in Computer Science*, pages 166–176. Springer Berlin / Heidelberg, 2009.

[62] 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, Aug. 2009.

[63] G. Peeters. Deriving musical structure from signal analysis for music audio summary generation: ”sequence” and ”state” approach. In *Computer Music Modeling and Retrieval*, volume 2771 of *Lecture Notes in Computer Science*, pages 143–166. Springer Berlin / Heidelberg, 2004.

[64] G. Peeters. Sequence representation of music structure using higher-order similarity matrix and maximum-likelihood approach. In *Proc. of 8th International Conference on Music Information Retrieval*, pages 35–40, Vienna, Austria, Sept. 2007.

[65] G. Peeters. Template-based estimation of time-varying tempo. *EURASIP Journal on Advances in Signal Processing*, 2007(1):158–158, 2007.

[66] G. Peeters. MIREX 2009 “Music structure segmentation” task: Ircamsummary submission. In *Proc. of Fifth Annual Music Information Retrieval Evaluation eXchange*, Kobe, Japan, Oct. 2009. Extended abstract.

[67] G. Peeters and E. Deruty. Is music structure annotation multi-dimensional? A proposal for robust local music annotation. In *Proc. of 3rd Workshop on Learning the Semantics of Audio Signals*, pages 75–90, Graz, Austria, Dec. 2009.

[68] G. Peeters, A. La Burthe, and X. Rodet. Toward automatic music audio summary generation from signal analysis. In *Proc. of 3rd International Conference on Music Information Retrieval*, pages 94–100, Paris, France, Oct. 2002.

[69] E. Peiszer. Automatic audio segmentation: Segment boundary and structure detection in popular music. Master’s thesis, Vienna University of Technology, Vienna, Austria, Aug. 2007.

[70] C. Rhodes and M. Casey. Algorithms for determining and labelling approximate hierarchical self-similarity. In *Proc. of 8th International Conference on Music Information Retrieval*, pages 41–46, Vienna, Austria, Sept. 2007.

[71] M. P. Ryynänen and A. P. Klapuri. Automatic transcription of melody, bass line, and chords in polyphonic music. *Computer Music Journal*, 32(3):72–86, 2008.

[72] Y. Shiu, H. Jeong, and C.-C. J. Kuo. Musical structure analysis using similarity matrix and dynamic programming. In *Proc. of SPIE Vol. 6015 - Multimedia Systems and Applications VIII*, pages 398–409, 2005.

[73] Y. Shiu, H. Jeong, and C.-C. J. Kuo. Similar segment detection for music structure analysis via Viterbi algorithm. In *Proc. of IEEE International Conference on Multimedia and Expo*, pages 789–792, Toronto, Ont., Canada, July 2006.

[74] H. Terasawa, M. Slaney, and J. Berger. The thirteen colors of timbre. In *Proc. of 2005 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 323–326, New Platz, N.Y., USA, Oct. 2005.

[75] D. Turnbull, G. Lanckriet, E. Pampalk, and M. Goto. A supervised approach for detecting boundaries in music using difference features and boosting. In *Proc. of 8th International Conference on Music Information Retrieval*, pages 51–54, Vienna, Austria, Sept. 2007.

[76] G. Tzanetakis and P. Cook. Multifeature audio segmentation for browsing and annotation. In *Proc. of 1999 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 103–106, New Platz, N.Y., USA, Oct. 1999.

[77] Y. Weiss. Segmentation using eigenvectors: a unifying view. In *Proc. of Seventh IEEE International Conference on Computer Vision*, pages 975–982, Kerkyra, Greece, Sept. 1999.

[78] J. Wellhausen and M. Höynck. Audio thumbnailing using MPEG-7 low-level audio descriptors. In *Proc. of The SPIE Internet Multimedia Management Systems IV*, volume 5242, pages 65–73, Nov. 2003.

[79] C. Xu, X. Shao, N. C. Maddage, M. S. Kankanhalli, and T. Qi. Automatically summarize musical audio using adaptive clustering. In *Proc. of IEEE International Conference on Multimedia and Expo*, pages 2063–2066, Taipei, Taiwan, June 2004.

[80] R. Zhou, M. Mattavelli, and G. Zoia. Music onset detection based on resonator time frequency image. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(8):1685–1695, 2008.