# Structural segmentation of music based on repeated harmonies

W. Bas de Haas*, Anja Volk*, Frans Wiering*
*Department of Information and Computing Sciences
Utrecht University
Email: {W.B.deHaas, A.Volk, F.Wiering}@uu.nl

***Abstract*—In this paper we present a simple, yet powerful method for deriving the structural segmentation of a musical piece based on repetitions in chord sequences, called FORM. Repetition in harmony is a fundamental factor in constituting musical form. However, repeated pattern discovery in music still remains an open problem, and it has not been addressed before in chord sequences. FORM relies on a suffix tree based algorithm to find repeated patterns in symbolic chord sequences that are either provided by machine transcriptions or musical experts. This novel approach complements other segmentation methods, which generally use a self-distance matrix based on other musical features describing timbre, instrumentation, rhythm, or melody. We evaluate the segmentation quality of FORM on 649 popular songs, and show that FORM outperforms two baseline approaches. With FORM we explore new ways of exploiting musical repetition for structural segmentation, yielding a flexible and practical algorithm, and a better understanding of musical repetition.**

***Keywords*-Structural Segmentation; Harmony; Repetition; Suffix Trees**

## I. INTRODUCTION

When humans perceive their surrounding environment, they tend to cluster similar percepts into larger cognitive structures. Also when perceiving music, human listeners group similar elements in time. In a melodic context this entails that notes form motifs, motifs form phrases, and phrases form sections that make up a piece. In multimedia systems, making use of this musical structure can be helpful when organising large quantities of musical data. Hence, segmenting pieces of music into smaller parts is a frequently used pre-processing step in various music information retrieval tasks. In this paper we present FORM[1], a novel approach to structural segmentation based on repeated patterns in musical harmony.

Repetition is considered a fundamental property of music: "Music-making is, to a large degree, the manipulation of structural elements through the use of repetition and change" [2]. Introducing changes to repeated patterns leads to what is called the *variation principle* in music [9], which is closely related to "musical parallelism", an important principle involved in the human segmentation of music as argued by Lehrdahl and Jackendoff [7, p. 52]. Hence, using repetition for inferring segment boundaries has been claimed to be an important aspect of modelling segments in music [3].

Though repetition and variation are considered fundamental for music, developing suitable formal models of what constitutes musically meaningful repeated patterns still remains a challenging task [9]. Many structural segmentation approaches in the audio domain detect repetition based on self-distance matrices, but modelling musical parallelism for automatically inferring a structural segmentation is still an unsolved problem, see e.g. [3], [8]. In this paper, we present a new approach that successfully models structural segmentation by exploiting suffix tree based repetition discovery.

*Harmony* is one of the most important components in Western tonal music. In musical practise, harmonic progressions are commonly represented as sequences of chords. Although chord sequences can be extracted from audio reasonably well, e.g. [5], they are not frequently used as mid-level features. Yet chord labels have some convenient properties: they are independent from tempo and rhythm, they generalise over sections by abstracting from the individually sounding notes, they are human interpretable, and they have been successfully applied to retrieval tasks [6].

In this paper, we make use of the following relationship between chord progressions and segments in music: harmony in popular music typically consists of short chord progressions that are repeated; these progressions function as structural building blocks that make up the piece. In pop songs, for instance, the verse and chorus are never played only once. Also, the main theme in a jazz piece, played at the beginning, is in general repeated at the end. Typically in these repetitions, melody, rhythm, and instrumentation are likely to vary, but the chord sequences hardly change. Hence, repetitions of chord progressions provide promising candidates for segments. Therefore, we assume to have a sequence of chord labels, which can, for example, be acquired by audio chord transcription techniques. Next, repeating sub-sequences are identified with suffix tree algorithms [4]. These repetitions form the basis for our structural segmentation approach.

The main contribution of this paper is FORM: a simple, yet powerful, method for deriving the structure of a musical piece that exploits repetitions in chord sequences. FORM combines symbolic sequence analysis and advanced chord transcription technologies, is fast, and intuitive to use. We evaluate the segmentation quality of FORM on 649 popular songs. For this, we use manually annotated chords as well as machine annotated chords, and show that FORM outperforms two baseline approaches. FORM demonstrates that musical repetition in the harmonic structure of a piece can be successfully applied to structural segmentation, and that chord labels are reliable

[1]Form Observation through Repeated Music

mid-level features in this context.

## II. Related work

The last decade, structural segmentation of music has been investigated widely. A musical piece can be divided in *parts*, where a part refers to a single instance or all instances of a musical section. For instance in pop music, a part would refer to a chorus, verse, introduction, bridge, etc. By *structure* we refer to a set of non-overlapping *segments*, where a segment is a temporal range that corresponds to a musical part. Generally, musical parts are repeated, and similar parts are grouped by assigning them the same label. Often capital letters, e.g. A, B, C, ..., etc., are used as labels, optionally augmented with primes to denote variations of the part.

Within the audio domain, most frequently *chroma* or *mel-frequency cepstral coefficients* (MFCCs) are used [8]. After feature extraction generally three different approaches can be discerned [8]: novelty based, homogeneity based, repetition based methods. All three approaches rely on a self-distance matrix, which can be created by comparing each frame every other frame using a Euclidean, cosine, or similar distances. The resulting square matrix is then used to identify similar sections in the recording. Novelty based approaches identify locations of maximal change on the main diagonal of the self-distance matrix. Depending on the features, peaks in such a *novelty curve* can be good predictors of changes in instrumentation, harmony, or rhythm. Homogeneity approaches continue this procedure, clustering created segments into homogeneous groups.

Repetition-based algorithms also use a self-distance matrix, but in a different way. These algorithms aim to identify the repetitive elements in an audio recording, and search for similar sections by tracking diagonal stripes parallel to the main diagonal in the self-distance matrix.

Within the symbolic domain few methods aim at segmenting a polyphonic score, and the majority of papers focusses on the segmentation of melodies, or on the separation of different voices. Gestalt-based knowledge-driven models and machine learning models are typical approaches to segmenting music in the symbolic domain. Still, modelling musical parallelism for inferring segment boundaries remains a challenge [3]. For a survey and comparisons of symbolic segmentation approaches we refer to [10].

## III. Repeated pattern analysis

Repeated patterns can be extracted from a sequence of symbols by identifying the left diverse nodes in the suffix tree of this sequence. Suffix trees and their applications originate from bio-informatics, but are found throughout many application areas, such as data compression, document clustering, music retrieval, and melodic repetition detection, but have not been used for structural segmentation. For an introduction into suffix trees and pattern discovery we refer to [4].

A sequence of symbols, or string, is denoted with $S$. In Section IV, these symbols are chord labels, but in this Section we use letters to keep it simple. The suffix tree of a $S$ we call

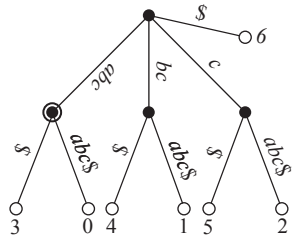


Figure 1. The suffix tree for a string $abcabc\$$. Leaves (white) are labelled with their starting index in the string. The internal node (black) that is marked with a circle, is left diverse.

$T$. The length of $S$ is referred to with $n$, and to point out a symbol at specific position $i$ in $S$, we use $S[i]$.

A suffix tree is a data structure that represents all suffixes of a string in a single tree.

**Definition:** For a string of $n$ symbols $S$, a suffix tree is defined as a rooted tree with exactly $n$ leaves, in which:

- the concatenation of the edge labels on the paths from root leaf spell out the suffixes of $S$,
- no two edges out of a node have edge labels starting with the same symbol,
- edge labels must be non-empty,
- all internal nodes, other than the root, have at least two children.

An example suffix tree of the string $abcabc\$$ is displayed in Figure 1. Here, we use the \$ symbol as termination unique termination symbol (see [4, Section 5.2, p. 90]).

We can distinguish various kinds of repeated structures in sequences of symbols, but not all repetitions are equally useful. Therefore, to prevent a pattern explosion we focus on *maximal repeats*: pairs of repeated patterns that, if extended to either the left or right side, break their equality. Maximal repetitions can be defined with maximal pairs:

**Definition:** A *maximal pair* in a string $S$ is a pair of identical substrings $a$ and $b$ in S such that the symbol immediate left (or right) of $a$ is different from the character to the immediate left (or right) of $b$.

**Definition:** A *maximal repeat* is a substring of $S$ that occurs in a maximal pair in $S$.

Suffix trees can be used to locate maximal repeats in a string by exploiting the *left diverse* property of its nodes, which is based on the left symbol of the suffixes represented by that node. The left symbol of a suffix is the symbol immediate left of the starting symbol of the suffix. For instance, given the string $abcabc\$$, the left symbol of its suffix $cabc\$$ is $b$. This can be stated formally as:

**Definition:** Given a string $S$, the *left symbol* of a suffix starting at position $i$ in $S$ is the symbol at $S[i-1]$.

**Definition:** A node $v$ in a suffix tree is called *left diverse* if at least two leaves in $v$'s subtree have different left symbols.

In Figure 1 the circled internal node is the only left diverse node. The left diverse property propagates upwards: if a node is left diverse, all its ancestors are left diverse too.

Using the left diverse property, we can identify nodes in

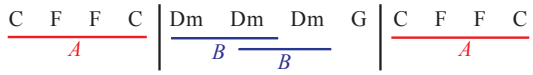


Figure 2. A chord sequence and its repetition-based segmentation. The horizontal lines represent the maximally repeated sequences, with subtree labels $A$ and $B$. The vertical bars denote the derived boundaries. The middle segment will be assigned label $B$.

the suffix tree that represent a maximal repeat. A suffix tree branches at a node, if for at least two similar suffixes the symbol on the right is different. If we also keep track of the symbol on the left of these suffixes (i.e. whether a node is left diverse or not), we can find all maximal repeated patterns of the string.

**Lemma:** Given a string $S$, a suffix of $S$ labelling the path to a node $v$ is a maximal repeat if and only if $v$ is left diverse.

For the proof of the lemma we refer to [4, p. 145].

If the suffix tree $T$ is pruned to contain only left-diverse nodes and their leafs, the suffixes represented by the leafs of $T$ contain a maximally repeated prefix. Next, the starting indices and lengths of these repetitions can be collected by traversing the tree.

## IV. Structural segmentation

Within FORM, a chord is represented as a single symbol consisting of a root and a quality (major/minor). All additional information that might be available in the labels, for example about inversions or chord additions, is ignored. This gives us 24 different chord labels, and a no-chord label for representing passages that lack harmonic information. Chord duration information is modelled by storing a chord label at every beat position. In general, we can make the repetition detection more specific, and use an equality function that takes also seventh chords etc. into account. However, we deliberately chose to use the rather broad major and minor categories because in chord label-based harmonic similarity estimation, a triadic representation gave better results than using more complex chords [6].

Given a chord sequence, we obtain a segmentation by constructing a suffix tree, and pruning the non-left diverse nodes. Only the root of the suffix tree has more than two subtrees, representing sequences that start with different chords. All sequences represented by such a subtree share a common prefix. However, they do not necessarily have to represent identical repeated patterns, they can have different suffixes. Furthermore, the left-diversity of these subtrees ensures that these subtrees do not share a suffix, unless this suffix is repeated. Because sequences extracted from the same subtree are *related* by common prefix, we assign a unique label to every subtree below the root, which we name *subtree label* and will become the part label used in the final segmentation.

We extract the repeated chord sequences from the suffix tree. Next, we obtain a segmentation by segmenting the chord sequence at every position where a repeated pattern starts and a subtree label changes. A brief statistical analysis of the repeated sequences and the data used in the experiment showed that segmentation boundaries are more likely to occur at repetition starts than at repetition endings. When a certain portion of the song is not part of a repeated pattern, the previous segment is extended up to the next pattern start. Musically this makes sense because, for instance, the last chord of a chorus is often varied to bridge to a verse. Figure 2 illustrates this segmentation strategy.

To investigate how robust FORM is to noisy chord sequences, we combine it with an automatic chord transcription algorithm. We decided to use MPTREE [5] which is freely available.[2] Automatically annotating the dataset used in our experiment with MPTREE yields an average correct overlap of 65% (the standard evaluation metric for automatic chord transcription used in MIREX[3] [5]). Using more noisy chord sequences is likely to result in shorter repeated patterns because inaccurate chords can break the equality between two related sub-sequences. Hence, we expect FORM to oversegment when using machine annotated chords. To reduce these effects, we can restrict FORM to only use repetitions greater than a specific length. Here, we use a length of 16 beats.

## V. Experiment

For our experiments we use the billboard dataset [1]. This dataset currently contains the audio and chord sequences of 649 popular songs randomly selected from the *Billboard* magazine's "Hot 100". All chord sequences are checked by at least three different musical experts, and contain structural segment annotations.

We evaluate the segmentation performance of FORM by comparing it to two baseline systems. To examine the robustness of FORM, we apply it to a near perfect chord sequence created by experts and an automatically transcribed chord sequence containing errors. The code of FORM is available to the research community on request. For the automatic transcription we relied on the system described in [5].

We compare FORM to two baseline systems: FIXED and RANDOM. FIXED is aimed to resemble a typical pop-song with the structure: ABBBBCCBBBBCCDCCE. In this, A, B, C, D, and E represent labels for Introduction, Verse, Chorus, Bridge, and Outro, respectively, where the verse is twice as long as the chorus, and the chorus is twice as long as the remaining sections. The sections are stretched to fit the length of the piece it is applied to. RANDOM generates random segments of 8 to 48 beats long with random labels in the range A...F. The first segment is shortened to fit the length of the song it is applied to.

We evaluate the quality of FORM by comparing its automatic segmentation to a ground truth segmentation, and calculate the pair-wise precision, recall and F-measure. Both the ground-truth and the computed segmentation are sampled at every 200 ms. Next, all pairs of frames with the same label are calculated for both segmentations. If $S_{gt}$ is the set of identically labelled pairs in the ground-truth annotation, and $S_m$ the set of identically labelled pairs in the machine annotation, then precision ($P$), recall ($R$) and F-measure ($F$) are defined as:

[2] http://hackage.haskell.org/package/HarmTrace
[3] http://www.music-ir.org/mirex/wiki/MIREX_HOME

| Method | $P$ | $R$ | $F$ |
|---|---|---|---|
| FORM expert chords | 0.54 | 0.72 | 0.58 |
| FORM automatic transcription *16 beats* | 0.50 | 0.62 | 0.52 |
| FORM automatic transcription | 0.55 | 0.39 | 0.43 |
| FIXED | 0.52 | 0.45 | 0.47 |
| RANDOM | 0.49 | 0.23 | 0.30 |

Table I
THE PAIRWISE PRECISION, RECALL AND F-MEASURE FOR THREE VARIANTS OF FORM AND TWO BASELINE SEGMENTATION APPROACHES.

$$P = \frac{|S_m \cap S_{gt}|}{|S_m|},\ R = \frac{|S_m \cap S_{gt}|}{|S_{gt}|},\ \text{and}\ F = \frac{2 \cdot P \cdot R}{P+R}.$$

A high pairwise precision reflects oversegmentation of the computed annotation. Similarly, a high pairwise recall indicates undersegmentation.

## VI. RESULTS

The results are displayed in Table I. FORM performs best when expert chord annotations are provided, outperforming the two baseline methods. When noisy automatically transcribed chords are used, the high precision and low recall indicate that FORM is oversegmenting considerably. However, when we remove the repetitions shorter than 16 beats, the performance improves substantially, and both baselines are outperformed. Eliminating repetition shorter than 16 beats is musically a reasonable assumption because general musical segments are longer than 4 measures (in $\frac{4}{4}$).

We tested if the differences in F-measure are statistically significant by performing a non-parametric Friedman test[4] with a significance level of $\alpha = 0.01$. The Friedman ANOVA is chosen because the underlying distribution of the F-Measure data is unknown, and the Friedman ANOVA does not assume a specific distribution of variance. There are significant differences between the five segmentation approaches, $\chi^2(4, N = 649) = 1647$, $p < 0.0001$. We determined that all pairwise differences are statistically significant by performing a post-hoc Tukey HSD test.

FORM is fast. Runs with expert chords, calculated chords, and the minimal repetition limit finished in $2m39s$, $2m03s$, and $2m20s$, respectively. The runs were done on an Intel i7 3.4Ghz system, and exclude the chord transcription time.

## VII. DISCUSSION AND CONCLUSION

We have introduced a novel method for structural segmentation of music. FORM complements other methods because it obtains a segmentation by analysing repetitions in chord sequences, does not rely on a self-distance matrix, and can be used on audio and symbolic data. Therefore, it is likely that FORM uncovers segmentations that are overlooked by other methods. By focussing on chord sequences, we show that musically interpretable descriptors of harmony are useful in a structural segmentation context. Our results are similar to the F-measure scores reported at MIREX. However, for a fair comparison between FORM and the state-of-the-art in MIREX, FORM should be extended with other features that do not only account for harmony, like features describing, timbre, instrumentation, rhythm and melody.

[4]All statistical tests were performed in Matlab 2011b.

FORM can be tailored to specific data or user needs in many ways. For example by incorporating musical knowledge: choruses and verses often have regular lengths, and are likely to change at strong metrical positions. Moreover, FORM can easily be adapted to use other chord representations or features too, as long as a suitable equality function is provided. We expect repeated patterns in musical features describing timbre, melody, or rhythm can complement FORM and improve the overall performance. Moreover, there are other repeated patterns that are interesting for structural segmentation as well, like k-mismatch repeats that allow for inexact matching

By modelling segmentation based on harmonic repetition, we use a musical feature that possibly is more stable across different versions or renditions of the same song in jazz and pop than features associated with e.g. timbre or instrumentation. Hence, segmenting related songs into comparable segments based on harmonic structure is just one possible scenario where FORM complements existing segmentation methods. Moreover, by using repetitions of chords, we gain a better understanding of the important role of repetition for segmentation. To conclude, FORM provides a flexible, and musically grounded algorithm that can potentially be used in many practical multimedia tasks.

## VIII. ACKNOWLEDGEMENTS

A. Volk and W.B. de Haas are supported by the NWO-VIDI-grant 276-35-001 to A. Volk. F. Wiering is supported by the FES project COMMIT/.

## REFERENCES

[1] J. A. Burgoyne, J. Wild, and I. Fujinaga. An expert ground truth set for audio chord recognition and music analysis. In *Proc. ISMIR*, pages 633–638, 2011.

[2] G. Burns. A typology of 'hooks' in popular records. *Popular Music*, 6(1):1–20, 1987.

[3] E. Cambouropoulos. Musical parallelism and melodic segmentation. *Music Perception*, 23(3):249–268, 2006.

[4] D. Gusfield. *Algorithms on strings, trees and sequences: computer science and computational biology*. Cambridge University Press, 1997.

[5] W. B. de Haas, J. P. Magalhães, and F. Wiering. Improving audio chord transcription by exploiting harmonic and metric knowledge. In *Proc. ISMIR*, pages 295–300, 2012.

[6] W. B. de Haas, F. Wiering, and R. C. Veltkamp. A geometrical distance measure for determining the similarity of musical harmony. *International Journal of Multimedia Information Retrieval*, 2(3):189–202, 2013.

[7] F. Lerdahl and R. Jackendoff. *A Generative Theory of Tonal Music*. MIT Press, 1996.

[8] J. Paulus, M. Müller, and A. Klapuri. State of the art report: Audio-based music structure analysis. In *Proc. ISMIR*, pages 625–36, 2010.

[9] A. Volk, W. B. de Haas, and P. van Kranenburg. Towards modelling variation in music as foundation for similarity. In *Proc. ICMPC*, pages 1085–1094, 2012.

[10] F. Wiering, J. de Nooijer, A. Volk, and H. J. Tabachneck-Schijf. Cognition-based segmentation for music information retrieval systems. *Journal of New Music Research*, 38(2):139–154, 2009.