Segmenting Music through the Joint Estimation of Keys, Chords and Structural Boundaries

Johan Pauwels
STMS IRCAM-CNRS-UPMC
1, Place Stravinsky
75004 Paris, France
johan.pauwels@ircam.fr

Geoffroy Peeters
STMS IRCAM-CNRS-UPMC
1, Place Stravinsky
75004 Paris, France
geoffroy.peeters@ircam.fr

ABSTRACT

In this paper, we introduce a new approach to music structure segmentation that is based on the joint estimation of structural segments, keys and chords in one probabilistic framework. More precisely, the boundaries of a structure segment are determined by detecting key changes and by utilizing the difference in prior probability of chord transitions according to their position in a structural segment. In contrast to many of the recent approaches to structural segmentation, this system does not work with self-similarity matrices, although it has been designed to integrate this kind of approach into the framework at a later stage. However, just the current version of the system, using only the estimated harmony, is already producing encouraging results, especially with respect to the precise localization of the boundaries.

Categories and Subject Descriptors

I.5.4 [**Pattern recognition**]: Applications—*Signal processing, Waveform analysis*; J.5 [**Arts and humanities**]: Performing arts (e.g., dance, music)

Keywords

music information retrieval; structural segmentation; key and chord estimation

1. INTRODUCTION

Structural segmentation of music is the process in which an audio recording is divided into a number of non-overlapping sections that correspond to the macro-temporal organisation of the piece. These entities usually take the form of verses and choruses in popular music or of movements in classical music. The obtained sections can then be used for audio summarisation, synchronization or as an intermediate step in further content-based indexation.

Recent approaches in music structural segmentation have mostly focussed on processing so-called self-similarity matrices [1], obtained by comparing some low level acoustic feature, like MFCC's or chromas, to time-delayed copies of itself. An overview of these and other methods can be found in [8]. However, our proposed method takes a different approach. We present a framework for the joint estimation of structural boundaries, keys and chords that builds upon an existing system for key and chord estimation [9]. It is based on the premise that some chord combinations are more common around structural boundaries, especially when expressed as relative chords in a key, giving a musicologically richer representation that is made possible by the concurrent estimation of keys and chords. In that regard, it is more similar to previous work by Maddage [5] or by Lee [3], only do their systems work sequentially because the chord and key estimates are used as inputs to the structure estimation, whereas our model generates them concurrently.

In the remainder of this paper, we will first elaborate the assumption on which our system is based and provide a theoretical underpinning of its validity in Section 2. Afterwards, we will present our probabilistic framework that uses this information to concurrently estimate keys, chords and structural boundaries in Section 3. Its output will then be compared to a database of manual annotations in Section 4. Finally, some conclusions as well as possible directions for future work will be given in Section 5.

2. STRUCTURE DEPENDENT RELATIVE CHORD TRANSITION MODELS

The basic premise around which our approach is constructed, is that chord sequences have a different prior probability according to their position in a structural segment and more specifically, that the number of different chord combinations that occur around the structural boundaries is lower than those occurring in the middle of a segment. A supporting example for this assumption is that movements in music of the Classical period often end in a limited number of specific chord combination, the so-called *cadences.* In order to verify this statement in a more methodological way, we construct three different key-chord transition models depending on the position of the chords in a structural segment.

A key k is defined as the combination of a tonic t and a mode m ($m \in \{\text{major}, \text{(natural) minor}\}$). A chord c is defined as the combination of a root r and a type p ($p \in \{\text{maj}, \text{min}, \text{dim}, \text{aug}\}$). In the following we will consider pairs of keys and chords (k, c). We define a relative chord c' with respect to a key k by expressing the root r as the interval i between the tonic and the root $i = d(t, r)$.

MM'13, October 21–25, 2013, Barcelona, Spain.

ACM 978-1-4503-2404-5/13/10 ...$15.00.
http://dx.doi.org/10.1145/2502081.2502193.

Table 1: Perplexity of relative chord transition models per mode according to structural position

mode	intra	inter	final
major	6.10	4.10	2.88
minor	6.24	4.37	3.68

A key-chord pair can thus equivalently be written as a key-relative chord pair $(k, c) = (t, m, r, p) = (t, m, i, p) = (k, c')$. The latter representation is musicologically more informative and corresponds to the way scholars analyse harmony.

We then look at all successions of 2 consecutive chords in a corpus that is annotated with keys, chords and structural segments. It is annotated such that all positions are indicated where at least one of key, chord or structural segment changes. In order to study the harmonic movement like musicologists do, both chords are interpreted in the same key. Because of the forward motion in music, this will be the key annotated for the first chord. Finally, to take the circularity of pitch perception into account, we ignore the tonic of the key and only keep the mode. The resulting representation of local harmony then consist of a mode and a pair of relative chords $(m_{n-1}, c'_{n-1}, c'_n)$.

These pairs of relative chords in a mode are divided into three categories according to their position in a structural segment. Examples for each of them have been indicated in Figure 1. We define following classes: *final*, when the two chords are the last ones in a structural segment; *inter*, meaning that the chord change is straddling a segment boundary; and *intra*, for the remaining chord transitions. These classes We then construct transition models for each structural position classes by counting the relative number of occurrences in a corpus. We use the part of the Isophonics set [6] that has been used for the MIREX 2010 chord estimation competition. It consists of 217 full songs, mostly by the Beatles (180 songs), the remainder by Queen and Zweieck. There are genres other than pop music that are a better fit for our restricted harmony model of 48 chords and 24 keys, but this corpus is one of the few that contains audio aligned to annotations of all our three considered aspects.

We can then quantify the difference in chord distribution between the various models by calculating the model perplexity $PP(C'_1, C'_2|m)$ per mode m for each of them. C'_1 and C'_2 represent the collection of all relative chords that appear as first, respectively second, element in the sequence. The model perplexity is defined as the exponential of the entropy $H(C'_1, C'_2|m)$ expressed in nats:

$$PP\left(C'_1, C'_2|m\right) = \exp\left(H\left(C'_1, C'_2|m\right)\right)$$
$$= \exp\left(-\sum_{c'_1} P\left(c'_1|m\right) \sum_{c'_2} P\left(c'_2|c'_1, m\right) \log P\left(c'_2|c'_1, m\right)\right)$$

This expresses the mean prior uncertainty of a bigram according to its position in the structural segmentation. A lower value means that the transition probability is concentrated into fewer combinations of two chords. As can be seen in Table 1, the values for the *intra*-model are indeed significantly higher than those for the *inter* and *final*-model, thereby confirming our hypothesis.

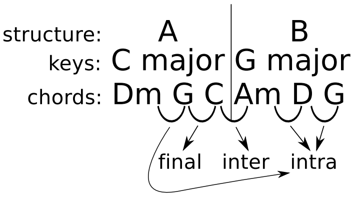

Figure 1: An example annotation with the three structural positions indicated

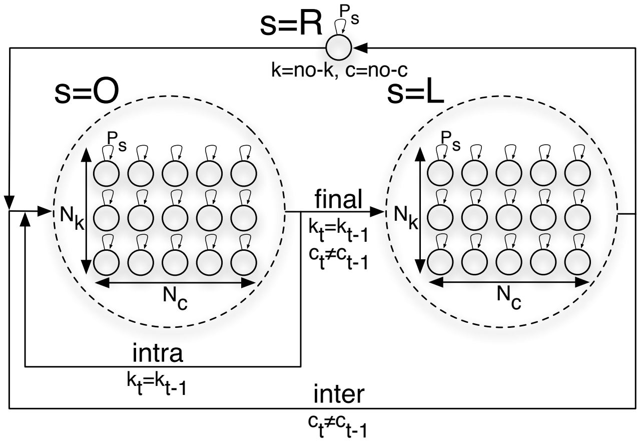

Figure 2: The state diagram

3. A PROBABILISTIC FRAMEWORK FOR THE JOINT ESTIMATION OF STRUCTURAL SEGMENTS, KEYS AND CHORDS

We will now describe a probabilistic framework in which the transition models as derived in the previous section will be used to determine a structural segmentation of a track, along with an estimation of the keys and chords. Our starting point is the system by Pauwels et al. [9] which concurrently estimates keys and chords, but does not estimate structural boundaries. It consists of an HMM in which each state represents a combination of a key and a chord. We extend it by letting each state q represent a structural position in addition to a key and a chord. A key k can take one of N_k values, a chord one of N_c values and the structural positions s can take one of two values: L which means that q is the last state of a structural segment or O which means that q is not the last state of a structural segment. In summary, $q = (k, c, s)$ with $k \in \{K_1, \ldots, K_{N_k}\}, c \in \{C_1, \ldots, C_{N_c}\}, s \in \{L, O\}$. Finally, we add a single state to handle the case when no chord is being played, notably at the beginning and end of a recording. In this state, the key will accordingly take a "no-key" value and the structural position will take a value of $s = R$. A simplified state diagram can be seen in Figure 2, in which states are grouped by their value of the structure variable. Only the transitions that change structure variable are drawn in order not to overload the picture, but the constraints on key and chord transitions are indicated next to the arrows. We are then looking for the state sequence $\hat{Q} = \{\hat{q}_1, \hat{q}_2, \ldots, \hat{q}_T\}$ that optimally explains the sequence of observations $X = \{x_1, x_2, \ldots, x_T\}$: $\hat{Q} = \arg\max P(Q|X)$. The resulting state sequence can then be split into separate sequences for keys, chords and structure states: $\hat{K}$, $\hat{C}$ and $\hat{S}$ respectively. The structural segmentation can be derived from the latter sequence simply by inserting a segment boundary for every transition from a state $s = L$ to one where $s = O$, or from or to $s = R$. Because the state variable q consists by definition of the combination of a chord, key and structure variable, these three optimal sequences will always be jointly decoded.

If we assume that the first order Markov property holds and that acoustic observations are independent from state to state, then we can rewrite the probability that needs to be maximized using Bayes' theorem to

$$\hat{S}, \hat{K}, \hat{C} = \arg\max \prod_{t=1}^{T} P\left(s_t, k_t, c_t | s_{t-1}, k_{t-1}, c_{t-1}\right) \\ P\left(\mathbf{x_t} | s_t, k_t, c_t\right)$$

The transition probabilities $P(s_t, k_t, c_t|s_{t-1}, k_{t-1}, c_{t-1})$ are calculated by a prior musicological model that consists of a number of submodels. By introducing some musicologically motivated constraints to the transition probabilities, we want to enforce a number of relationships between the concepts of keys, chords and structural segments. These will ensure that our estimation always produces sensible results and have as an added benefit that this also speeds up the calculation. The first three constraints we impose are 1) a key change $k_t \neq k_{t-1}$ is only allowed to occur together with a chord change $c_t \neq c_{t-1}$, 2) there must be a change in chord or in key between segments, 3) a structural segment must contain at least two different chords (or a single no-chord). These three limitations can be easily enforced by ensuring that every state change implies a chord change. In other words, we let the chord level control the granularity of the key and structure estimations. This makes the state duration model effectively a chord duration model that we control by a single parameter P_s:

$$P\left(s_t, k_t, c_t | s_{t-1}, k_{t-1}, c_{t-1}\right) = \\ \begin{cases} P_s & s_t = s_{t-1} \wedge k_t = k_{t-1} \\ 0 & s_t \neq s_{t-1} \vee k_t \neq k_{t-1} \end{cases}, \forall c_t = c_{t-1}$$

The remaining probabilities $P\left(s_t, k_t, c_t | s_{t-1}, k_{t-1}, c_{t-1}\right)$, $\forall c_t \neq c_{t-1}$ of the chord changing transitions are calculated by a combination of three submodels. We further apply Bayes' theorem repeatedly to arrive at a decomposition into three terms

$$\begin{aligned} & P\left(s_t, k_t, c_t | s_{t-1}, k_{t-1}, c_{t-1}\right) \\ & \quad = P\left(c_t | s_t, s_{t-1}, k_{t-1}, c_{t-1}\right) P\left(k_t | s_t, s_{t-1}, k_{t-1}, c_t, c_{t-1}\right) \\ & \qquad P\left(s_t | s_{t-1}, k_{t-1}, c_{t-1}\right) \end{aligned}$$

We can already recognize the structure-dependent relative chord transition models of the previous section in the first term $P\left(c_t | s_t, s_{t-1}, k_{t-1}, c_{t-1}\right)$. Our three categories of chord transitions – *inter*, *intra* and *final* – each correspond to a certain combination of the state variables. The *intra* model will be used when $s_{t-1} = O$ and $s_t = O$, *inter* when $s_{t-1} = L$ and $s_t = O$ and *final* when $s_{t-1} = O$ and $s_t = L$. Finally, from our definition of L it follows that when $s_{t-1} = L$ and $s_t = L$, only the self probability P_s should be allowed, to account for the fact that the last chord of a structural segment can – and most likely will – last more than one time step. The other probabilities are set to zero.

In the second term $P\left(k_t | s_t, s_{t-1}, k_{t-1}, c_t, c_{t-1}\right)$, we neglect the influence of the chords c_{t-1}, c_t in comparison to the other terms such that we end up with $P\left(k_t | s_t, s_{t-1}, k_{t-1}\right)$. We add the supplemental constraint that a key change can only occur between segments. This means that $s_{t-1} = O \vee s_t = L \Rightarrow P\left(k_t | s_t, s_{t-1}, k_{t-1}\right) = \delta_{k_t, k_{t-1}}$ with δ the Kronecker-delta. For the *inter* key transitions $P(k_t | s_t = O, s_{t-1} = L, k_{t-1})$, we reuse the model from [9], based on Lerdahl's theoretical distance [4] between keys.

The third term $P\left(s_t | s_{t-1}, k_{t-1}, c_{t-1}\right)$ will be used to control the ease of changing the structure variable s and thus to control the insertion of segment boundaries. We use a simple model that ignores the key and chord influence and consists of a single parameter ω that balances the probability of going to $s = O$ or $s = L$ after leaving $s = O$.

$$P\left(s_t | s_{t-1}\right) = \begin{cases} \omega & s_{t-1} = O \wedge s_t = O \\ 1 - \omega & s_{t-1} = O \wedge s_t = L \\ 1 & s_{t-1} = L \wedge s_t = O \\ 0 & s_{t-1} = L \wedge s_t = L \end{cases}$$

The result of adding the additional constraints is that the complete transition matrix will have a well-defined, sparse structure. The two upper quadrants consist of block diagonal matrices with N_k blocks of side N_c, the lower left quadrant is dense and the lower right quadrant is a diagonal matrix. In comparison to a system that only estimates keys and chords concurrently, the number of states gets doubled by repeating every key-chord state for $s = O$ and $s = L$. On the other hand, because of the sparsity of the transition matrix, the increase in the number of transitions stays limited. More specifically, the number of transitions is $(N_k N_c)^2 + 2N_k N_c^2 + N_k N_c + 1$, which corresponds in our configuration to an increase of 8% instead of the theoretically maximum of 400% that would be attained for a dense transition matrix. This sparsity will subsequently be used in the implementation of the Viterbi algorithm to limit the increase in computation time.

The features $\mathbf{x}$ that we use for the calculation of the acoustic probabilities $P\left(\mathbf{x_t} | s_t, k_t, c_t\right)$ are the Loudness Based Chromagrams as developed by Ni et al. [7]. These are 24-dimensional vectors that represent the loudness of each of the 12 pitch classes in both the treble and the bass spectrum. They are calculated with a step size of 23 ms and are afterwards averaged over interbeat segments as calculated by ircambeat [10].

We make the assumption that keys and chords can be independently tested for compliance with an observation and that the structure position is conditionally independent of the observations, such that $P\left(\mathbf{x_t} | s_t, k_t, c_t\right) = P\left(\mathbf{x_t} | c_t\right) P\left(\mathbf{x_t} | k_t\right)$. The key acoustic probability $P\left(\mathbf{x_t} | k_t\right)$ is then modelled as the cosine similarity between the observation vector x and Temperley's key templates [11]. These represent the stability of each of the 12 pitch classes relative to a given key. The chord acoustic probability $P\left(\mathbf{x_t} | c_t\right)$ is modelled by a multi-variate Gaussian with full covariance matrix. The models per chord are trained on the same data set from which the relative chord transition models have been derived.

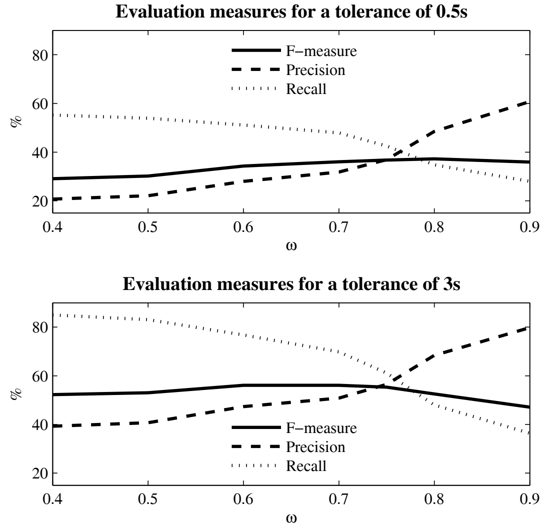

Figure 3: Evaluation measures for two tolerance settings

4. EXPERIMENTAL RESULTS

The performance of the generated structural boundaries will be evaluated by calculating a precision $\mathcal{P}(tol)$ and a recall $\mathcal{R}(tol)$ that is a function of a tolerance tol. The precision is defined as the number of estimated boundaries for which an annotated boundary falls within the window with length of tol centered around its estimated position. This number is expressed relatively to the total number of estimated boundaries. The recall on the other hand, is the relative number of annotated boundaries that have an estimated boundary within its tolerance window. Both measures are combined into an F-measure $\mathcal{F}(tol)$. These measures are calculated for every song of the previously mentioned data set and averaged.

We evaluate our output using two settings for the tolerance: 0.5 s and 3 s and the aforementioned Isophonics data set. The results in function of the structure change controlling parameter ω can be found in Figure 3. The movement in opposite directions of the precision and recall curves show that ω indeed is able to control the number of resulting structural segments. The optimal F-measures $\mathcal{F}(0.5s) = 36.04$ and $\mathcal{F}(3s) = 56.08$ are reached for $\omega = 0.7$.

Next, we will compare our results with a state-of-the-art segmentation algorithm by Kaiser & Peeters [2]. It calculates a novelty curve by taking the correlation between three types of kernels and the diagonal of an MFCC-based self-similarity matrix. Segment boundaries are then inserted at places where the novelty curve peaks. This system is a refinement of the one that ranked among the best segmentation algorithms during the MIREX 2012 contest. We obtained the output of their algorithm on the data set we use through personal correspondence with the authors.

From the results presented in Table 2, we can conclude that our approach is particularly strong in the fine localisation of the segment boundaries: we outperform Kaiser & Peeters' system for a tolerance of 0.5 s. However, increasing the tolerance to 3 s has a much larger beneficial effect on their performance than on ours. Consequently, their results for a tolerance of 3 s is significantly better, especially the precision. The latter is a consequence of the fact that our algorithm oversegments more than theirs (at least for the value of ω that gives the highest F-measure).

Table 2: A comparison between our proposed system and the state-of-the-art

	Kaiser & Peeters		Proposed ($\omega = 0.7$)	
tolerance	0.5s	3s	0.5s	3s
F-measure	32.94	64.53	36.04	56.08
precision	29.78	59.87	31.82	50.81
recall	39.28	73.50	47.91	69.85

5. CONCLUSION AND FURTHER WORK

We presented a new approach to structural segmentation by means of a probabilistic framework that concurrently estimates keys, chords and structural boundaries. It is based on the assumption that key changes indicate structural boundaries and that there is lesser variety in chords around structural boundaries than in the middle of structural segments.

Currently, our system is still in its proof-of-concept phase. In the future, we will first perform additional experiments to test the scaling of our approach to different types of music. The final goal however, is to integrate our method with a self-similarity matrix (SSM) based approach, aspiring to construct a synergistic system. For instance, just a change in instrumentation won't be detected as a structure boundary in our current system, whereas SSM methods are pretty efficient at spotting those.

6. ACKNOWLEDGEMENTS

This work was partly supported by the Quaero Program funded by Oseo French agency.

7. REFERENCES

[1] J. Foote. Visualizing music and audio using self-similarity. In *Proc. of ACM Multimedia*, 1999.

[2] F. Kaiser and G. Peeters. Adaptive temporal modeling of audio features in the context of music structure segmentation. In *Proc. AMR workshop*, 2012.

[3] K. Lee. *A system for acoustic chord transcription and key extraction from audio using hidden Markov models trained on synthesized audio.* PhD thesis, Stanford University, 2008.

[4] F. Lerdahl. *Tonal pitch space.* Oxford University Press, New York, 2001.

[5] N. C. Maddage. Automatic structure detection for popular music. *IEEE MultiMedia*, 13(1), 2006.

[6] M. Mauch, C. Cannam, M. Davies, S. Dixon, C. Harte, S. Kolozali, D. Tidhar, and M. Sandler. OMRAS2 metadata project 2009. In *Proc. ISMIR*, 2009.

[7] Y. Ni, M. McVicar, R. Santos-Rodríguez, and T. De Bie. An end-to-end machine learning system for harmonic analysis of music. *IEEE Trans. Audio, Speech and Language Proc.*, 20(6), 2012.

[8] J. Paulus, M. Müller, and A. Klapuri. Audio-based music structure analysis. In *Proc. ISMIR*, 2010.

[9] J. Pauwels, J.-P. Martens, and M. Leman. Improving the key extraction accuracy of a simultaneous key and chord estimation system. In *Proc. AdMIRe*, 2011.

[10] G. Peeters. Simultaneous beat and downbeat-tracking using a probabilistic framework: theory and large-scale evaluation. *IEEE Trans. Audio, Speech and Language Proc.*, 19(6), 2011.

[11] D. Temperley. *The cognition of basic musical structures.* MIT Press, 1999.