MUSIC STRUCTURE DISCOVERY: MEASURING THE "STATE-NESS" OF TIMES

Geoffroy Peeters
UMR STMS IRCAM-CNRS-UPMC / Sound Analysis/Synthesis Team

1. STATE AND SEQUENCE APPROACH

Music Structure Discovery (MSD) aims at estimating the underlying structure of a music track using observations of the audio signal. For this, a given time t_i of a music track is supposed to belong to one of the following categories:

• if t_i contains information similar to its adjacent times, t_i is said **homogeneous** [7] and belongs to a **"state"** [8],

• if t_i is similar to a foreign time t_j, and if the same is true for t_{i+l} $l \in [1, \lambda]$ (similar to t_{j+l}), we say that the corresponding segments are **repetitions** [7]. - If the corresponding times t_{i+l} are similar to their adjacent times than we have a "state repetition". - If this is not the case, we say that the times t_{i+l} and t_{j+l} belong to a **"sequence"** [8] of length λ which is **instantiated** at time t_i and at time t_j.

• if t_i is not similar to any other times, t_i is a **null-time**.

This subdivision has lead to two types of approaches to estimate the music structure: • the state approach, which is used to detect states (being repeated or not) and • the sequence approach, which is used to detect sequences (i.e. repetitions which are not states). See [8] for more details. This subdivision if summarized in the table below.

	Homogeneous	Non-Homog.
Repeated	State approach	Sequence approach
Non-Rep.	State approach	Null

Much more MSD systems have been proposed for the state approach. This is probably due to the fact that this approach can rely on well-established algorithms for segmentation (novelty measure of [1]), clustering [10] or hidden Markov models. In the state approach, there is no need to distinguish between repeated and non-repeated times since both will end up in states. The state approach is however not able to deal with non-homogeneous repeated times. This is the goal of the sequence approach.

The sequence approach first necessitates to distinguish the repeated times from the null times since only the repetitions will be used for the structure. Hence, the majority of the sequence approaches proceed in three successive separated stages: (1) extraction of audio observations, (2) detection of repetitions (sequence-instantiations) (3) connection of the detected sequence-instantiations to each others in order to estimate the sequences hence the structure.

1.1 Choice between state and sequence approach

To estimate the structure of a track, the choice between a state and a sequence approach depends on (A) the property of the music composition/production itself and (B) the audio observation we have from it. This second point can be subdivided into (B.1) the signal observations being used (B.2) the observation window length. Given a track and its observations, an automatic way to estimate the most appropriate approach to be used (among the state and sequence) would be beneficial. We propose here a measure which allows assigning each time of a track to one of the two approaches.

2. MEASURING THE STATE-NESS OF A TIME

As previously said in a MSD system, a given time t_i belongs to one of the following classes: - homogeneous/ states (repeated or not), - sequence (which are by definition repeated), - null. Corresponding to these classes are specific observations in the Self Similarity/Distance Matrix (SSM):

• homogenous/state: the local area around t_i in the main diagonal has continuous large values,

• sequence: the time corridor including t_i enclose at least one diagonal stripe,

• nul: neither the state or sequence conditions are observed.

Using this, we propose the "state-ness" coefficient $c(\tau)$ which represent the possibility to represent a time τ by a "state". For this, we first define the sub-matrix along the main diagonal of length L

$$\underline{\underline{E}}_{\tau}(t_i, t_j) = \underline{\underline{E}}(t_i \in [\tau, \tau + L], t_j \in [\tau, \tau + L]) \quad (1)$$

where $\underline{\underline{E}}(t_i, t_j)$ is the SSM provided by a specific MSD system and L is a fixed parameter set to 5s. We then compute the ratio of the mean value of the block $\underline{\underline{E}}_{\tau}$ over the mean value of its diagonal. If the block represents a "state" then

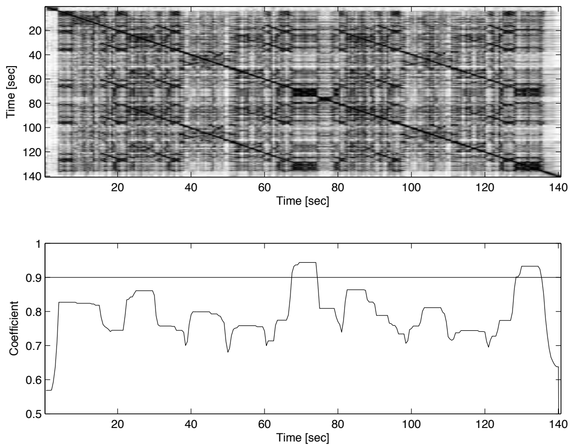

Figure 1. [Upper part]: Self Similarity Matrix [Lower Part]: $c(\tau)$ for $L = 5$s and a threshold at 0.9. On track: ”If I Needed Someone” from The Beatles ”Rubber Soul” album.

the mean value of the block will be close to the mean value along its diagonal. We also add a constraint related to the homogeneity of the block by subtracting to the mean value its standard deviation:

$$c(\tau) = \Big(\mu(\underline{\underline{E}}_{\tau}) - \sigma(\underline{\underline{E}}_{\tau})\Big) \,/\, \Big(\mu(diag(\underline{\underline{E}}_{\tau}))\Big) \qquad (2)$$

where μ denotes the mean value and σ the standard deviation. By experiments, we found that times for which $c(\tau) \geq 0.9$ correspond to ”states”. We illustrate this in Figure 1 where the values of $c(\tau)$ indicate two ”states” around times 70s and 130s. The remaining times of this track either belong to sequence-instantiations or are null-times.

3. EXEMPLIFYING

We illustrate here the use of the ”state-ness” coefficient $c(\tau)$. For the computation of the SSM we use the system proposed in [9]: 13 MFCCs (excluding the 0th coefficient) combined with 12 Spectral Contrast Measures and Spectral Valley Measures [5] and 12 Pitch-Class-Profile coefficients [2]. Each dimension of the features is then modeled over time (texture window) by its mean value over a sliding window of length $P =$1s (or $P =$4s) with a 500ms hop size. We refer the reader to [9] for more details on the exact computation of the Self Similarity Matrix from these features. We demonstrates here the influence of the choice of P (using either short-term modeling $P = 1s$, or long-term modeling $P = 4$s) on $c(\tau)$ hence on the choice between a state of sequence approach. For each track of each test-set, we compute $c(\tau)$ for each frame of the track.

Using $P = 1s$, 6.2% of the frames of the Beatles test-set [6] have a value $c(\tau) > 0.9$. Hence, the ”sequence” representation is well-suited for 93.8% of the frames. Figure 2 illustrates the evolution of $c(\tau)$ over tracks (tracks are

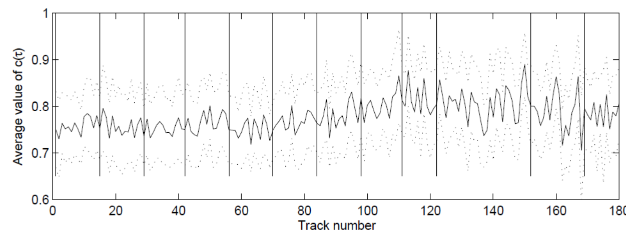

Figure 2. Average value of $c(\tau)$ over track number (dotted lines represent $\mu+\sigma$ and $\mu-\sigma$, vertical lines represent album separation) for the 180 tracks of the Beatles test-set.

arranged in recording date by album). It is interesting to note that the average-per-track $c(\tau)$ tends to increase over the years, which could be interpreted as a more important use of ”states” in the music structure process of The Beatles over times. The same applied to the RWC-Popular-Music test-set [4] [3] leads to 3.98% of the frames with $c(\tau) > 0.9$, hence belonging to states. **Using** $P = 4s$, the results change drastically: the state representation is now dominant among frames: 58.82% for the Beatles and 58.16% for the RWC test-set.

4. ACKNOWLEDGMENTS

This work was supported by the Oseo project “Quaero”.

5. REFERENCES

[1] J. Foote. Automatic audio segmentation using a measure of audio novelty. In *Proc. of ICME (IEEE Int. Conf. on Multimedia and Expo)*, pages 452–455, New York City, NY, USA, 2000.

[2] T. Fujishima. Realtime chord recognition of musical sound: a system using common lisp music. In *Proc. of ICMC*, pages 464–467, Bejing, China, 1999.

[3] M. Goto. Aist annotation for the rwc music database. In *Proc. of ISMIR*, pages pp.359–360, Victoria, Canada, 2006.

[4] M. Goto, H/ Hashiguchi, T. Nishimura, and R. Oka. Rwc music database: Popular, classical, and jazz music databases. In *Proc. of ISMIR*, pages pp. 287–288, Paris, France, 2002.

[5] D. Jiang, L. Lu, H.-J. Zhang, J.-H. Tao, and L.-H. Cai. Music type classification by spectral contrast. In *Proc. of ICME (IEEE Int. Conf. on Multimedia and Expo)*, Lausanne Switzerland, 2002.

[6] M. Mauch, C. Cannam, M. Davies, S. Dixon, C. Harte, S. Klozali, D. Tidhar, and M. Sandler. Omras2 metadata project 2009. In *Proc. of ISMIR*, Kobe, Japan, 2009.

[7] J. Paulus, M. Muller, and A. Klapuri. Audio-based music structure analysis. In *Proc. of ISMIR*, Utrecht, The Netherlands, 2010.

[8] G. Peeters. Deriving musical structures from signal analysis for music audio summary generation: Sequence and state approach. In U.K. Wiil, editor, *CMMR 2003 (LNCS 2771)*, Lecture Notes in Computer Science, pages 142–165. Springer-Verlag Berlin Heidelberg 2004, 2004.

[9] G. Peeters. Sequence representation of music structure using higher-order similarity matrix and maximum-likelihood approach. In *Proc. of ISMIR*, Vienna, Austria, 2007.

[10] G. Peeters, A. Laburthe, and X. Rodet. Toward automatic music audio summary generation from signal analysis. In *Proc. of ISMIR*, pages 94–100, Paris, France, 2002.