A MUSIC STRUCTURE INFERENCE ALGORITHM BASED ON MORPHOLOGICAL ANALYSIS

Gabriel SARGENT
Université de Rennes 1
IRISA (UMR 6074)
gabriel.sargent@irisa.fr

Frédéric BIMBOT
CNRS
IRISA (UMR 6074)
frederic.bimbot@irisa.fr

Emmanuel VINCENT
INRIA
Centre INRIA Rennes
Bretagne Atlantique
emmanuel.vincent@inria.fr

ABSTRACT

Music structure refers to the description of the long term organization of a music piece through a sequence of *structural segments*. A structural segment can be defined by its *structural borders* (a start time, an end time) and a *label* reflecting the similarity of its music content compared to the other segments'. Its duration is typically around 16 s and more.

This document presents the music structure estimation system submitted to MIREX's structural segmentation task in 2012. It is composed of three steps : feature extraction, structural border estimation and segment labeling. First, the system produces a sequence of chroma vectors [6] expressed at the *snap* scale [1] (section 1). This sequence is used to calculate a segmentation criterion based on a morphological model of the structural segments [2] (section 2.1). The structural border estimation is performed by searching the segmentation with lowest cost, which combines this criterion and a regularity constraint (section 2.2). The segments are then labeled by clustering according to their similarity, through the minimization of an adaptive model selection criterion (section 3).

1. FEATURE EXTRACTION

The extraction of the sequence of chroma vectors of size 12 used to describe the music content of the piece is performed by means of the "Chroma Toolbox" by Muller and Ewert [6]. We use the CP features regularly and a hop of 0.1 s.

Then, they are expressed at the snap scale. The *snap* is here defined as the multiple of a beat whose period is closer to 1 s. The snap scale is synchronous to the downbeat scale, and they are often equal in practice. The beat and downbeat estimations are performed thanks to the MATLAB implementation by Davies *et al.* [4, 5]. The downbeat estimator is tuned so as to consider 4 beats per bar.

We associate to each snap the mean of the CP features contained in the window centered on the snap that lasts the duration of the snap period.

2. STRUCTURAL BORDER ESTIMATION

2.1 Morphological model

We assume that a structural segment can be characterized by its inner organization, according to its musical layers (timbre, harmony, melody ...). In this scope we consider the *system and contrast model* by Bimbot *et al.* [2]. It considers that each structural segment aimed is built from an group of typically four *morphological elements* of four snaps, we note $\{a_1, a_2, a_3, a_4\}$. The three first elements are related by simple transformations f and g so as $a_2 = f(a_1)$ and $a_3 = g(a_1)$. The fourth element can either follow the logic of the three elements and then form a *system* ($a_4 = f(g(a_1))$) or on the contrary *contrasts* with it ($a_4 = \delta(f(g(a_1)))$ where $\delta \neq id$). Note that we assume that the relevant layers for structural analysis can vary from one structural segment to another.

However, in much cases, either $f = id$ or $g = id$, or both. This leads to observe usual morphological motives like $aaaa, abab, aabb$ in the case of *systems* with no *contrast*, or $aaab, abac, aabc$ in the case of *systems* ending with a *contrast*. These motives can be extended to the case where the identity function id is replaced by "close to identity" functions id' ($aaa'b, aba'c, aa'bc, ...$). More information on this model can be found in [3].

2.2 Segmentation criterion

The aim is to evaluate for each time unit considered the likelihood that it corresponds to the beginning of a *system*. We assume that at least one of the relations (f or/and g) between the elements of a system equals the identity function. For each snap $t \in [1, T]$ of a music piece, we consider the analysis window of size $N = 16$ snaps so as to consider three morphological elements starting from t (a_1, a_2, a_3), and one morphological element before this snap (a_0) as represented in figure 1. We consider that the size of each morphological element is $N_m = 4$ snaps. The criterion Φ we consider in this work results from the linear combination of two quantities :

$$\Phi = \lambda_1 \sigma_{\text{System}} + \lambda_2 \sigma_{\text{Contrast}} \quad (1)$$

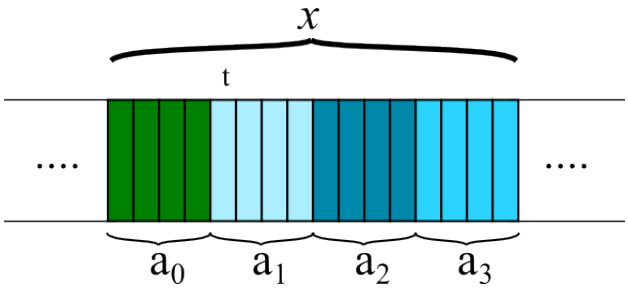

Figure 1. Analysis window used for segmentation criterion calculation, containing the sequence of features x. It is composed of 4 small windows, each one related to the position of a morphological element a_i of size 4 snaps, $i = \{0, 1, 2, 3\}$.

with λ_1 and $\lambda_2 \in \mathbb{R}^+$ learnt on a training database [1] .

$\sigma_{\text{System}}(t)$ quantifies the likelihood that t corresponds to the start time of a *system* through the analysis of the similarity of the morphological elements a_2 and a_3 with regard to a_1. Be $x = \{X_n\}_{1 \leq n \leq N}$ the sequence of features contained in the analysis window related to snap t. Let us define $X = \{X_0, X_1, X_2, X_3\}$, with vector $X_i = \{x_{1+iN_m}, ..., x_{N_m+iN_m}\}$ for $i = \{0, 1, 2, 3\}$, and $Y = [Y_1, Y_2] = [(X_2 - X_1)^2, (X_3 - X_1)^2]$. X_i contains the sequence of features of a_i, and Y_j is the squared distance between the features of a_{j+1} and a_1, for each of its dimensions. We have :

$$\sigma_{\text{System}}(t) = \frac{\sum_{j=1}^{N_m} \min(Y_1(j), Y_2(j))}{||X_1||^2} \quad (2)$$

where $||X_i||$ corresponds to the l_2 norm of vector X_i. $\sigma_{\text{System}}(t)$ allows to evaluate the contribution of X_1 to explain either X_2 or X_3 according to the various dimensions of the features. A high contribution implies that a system is likely to begin at snap t in the piece.

In the scope of the *system and contrast* model, a structural segment is likely to begin at t if a_2 and/or a_3 is similar to a_1. This is considered through $\sigma_{\text{System}}(t)$. If the preceding structural segment is different from the current one, the third and the fourth morphological elements differ from a_1. If the two segments are the same, then the third element may be similar to a_1, but the fourth one generally differs from it. We therefore introduce $\sigma_{\text{Contrast}}(t)$ which evaluates the dissimilarity between the morphological element preceding snap t we note a_0, and a_1. We choose to formulate it as follows :

$$\sigma_{\text{Contrast}}(t) = \text{cotan}(X_0, X_1) \quad (3)$$

where $\text{cotan}(X_i, X_j)$ corresponds to the cotangent of the angle between vectors X_i and X_j.

2.3 Regularity constraint

We consider a regularity constraint in the structural border estimation to favor segmentations with segments close to a typical segment size or *structural pulse* $\tau = 16$ snaps . Let m be the size of a structural segment :

$$\Psi_\alpha(m) = |\frac{m}{\tau} - 1|^\alpha \quad (4)$$

with $\alpha \in \mathbb{R}^+$ a factor which controls the convexity of this function. The use of a non-convex function will favor segmentations with a majority of segments of size equal to τ and few segments whose size is far from this structural pulse. On the opposite, a convex function loosen this constraint.

2.4 Performing the structural border estimation

The segmentation criterion and the regularity constraint are combined through a linear combination to form a segmentation cost C :

$$C = (1 - \lambda_3)\Phi + \lambda_3\Psi_\alpha \quad (5)$$

with $\lambda_3 \in [0, 1]$.

We use the Viterbi algorithm described in [8] to find the segmentation with the lowest segmentation cost.

3. SEGMENT LABELING

The labeling of the obtained segments is performed by the method described in [9]. Here, we transform the chroma sequence in a symbolic sequence by means of vector quantization, with the number of chroma clusters empirically fixed to 16. The Edit Distance on the symbolic features used to compare the content of the segments is replaced by a stripe distance [7] on the corresponding numeric chroma features. As we consider that the content of the fourth morphological element can be very variable as in section 2.1, we only consider the three fourth of the segment only when it lasts 16 snaps or more.

4. ACKNOWLEDGEMENTS

The authors would like to thank Matthew Davies for sharing his beat and downbeat estimation scripts with us. This work was partly supported by the Quaero project [2] funded by OSEO.

5. REFERENCES

[1] F. Bimbot, O. Le Blouch, G. Sargent and E. Vincent, " Decomposition into autonomous and comparable blocks : a structural description of music pieces", *Proceedings of the International Symposium on Music Information Retrieval*, pp. 189–194, 2010.

[2] F. Bimbot, E. Deruty, G. Sargent and E. Vincent, " Semiotic structure labeling of music pieces : concepts, methods and annotation conventions ", to appear in *Proceedings of the International Symposium on Music Information Retrieval*, 2012.

[1] RWC Popular database with structural annotations from [1] were used for parameter tuning.

[2] http://www.quaero.org/

[3] F. Bimbot, E. Deruty, G. Sargent and E. Vincent, " COMPLEMENTARY REPORT TO THE ARTICLE "Semiotic structure labeling of music pieces : concepts, methods and annotation conventions" (Proceedings ISMIR 2012) ", IRISA research report, 2012, http://hal.inria.fr/hal-00713196/

[4] M. E. P. Davies, "Towards Automatic Rhythmic Accompaniment" (Chapter 6), Ph.D. Thesis, Department of Electronic Engineering, Queen Mary University of London, 2007.

[5] A. M. Stark, M. E. P. Davies and M. D. Plumbley, "Real-Time Beat-Synchronous Analysis of Musical Audio" *Proceedings of the 12th Int. Conference on Digital Audio Effects*, Como, Italy, pp. 299–304, September 1-4, 2009.

[6] Meinard Müller and Sebastian Ewert, "Chroma Toolbox: MATLAB implementations for extracting variants of chroma-based audio features", *Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR)*, 2011.

[7] Jouni Paulus and Anssi Klapuri, "Music Structure Analysis Using a Probabilistic Fitness Measure and a Greedy Search Algorithm", *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 6, August 2009, pp. 1159-1170.

[8] G. Sargent, F. Bimbot and E. Vincent, "A regularity-constrained Viterbi algorithm and its application to the structural segmentation of songs", *Proceedings of the International Symposium on Music Information Retrieval*, 2011.

[9] G. Sargent, F. Bimbot, S. Raczynski, E. Vincent and S. Sagayama, "A music structure inference algorithm based on symbolic data analysis" *Music Information Retrieval Evaluation eXchange - ISMIR*, October 2011. http://hal.archives-ouvertes.fr/hal-00618141/PDF/Sargent_et_al_StructuralSegmentation_MIREX2011.pdf