A SIMPLE FUSION METHOD OF STATE AND SEQUENCE SEGMENTATION FOR MUSIC STRUCTURE DISCOVERY

Florian Kaiser
STMS IRCAM-CNRS-UPMC
1 Place Igor Stravinsky
75004 Paris
florian.kaiser@ircam.fr

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

ABSTRACT

Methods for music structure segmentation are based on strong assumptions on the acoustical properties of structural segments. These assumptions relate to the novelty, homogeneity, repetition and/or regularity of the content. Each of these assumptions provide a different perspective on the music piece. These assumptions are however often considered separately in the methods. In this paper we propose a method for estimating the music structure segmentation based on the fusion of the novelty and repetition assumptions. This combination of different perspectives on the music pieces allows to generate more coherent acoustic segments and strongly improves the final music structure segmentation's performance.

1. INTRODUCTION

Music structure segmentation (MSS) is the task of dividing a musical audio signal into its main structural parts. Examples of such main segments for popular music are the verse and the chorus. MSS allows for a large set of applications of interest in the context of digital music, such as automatic summarization or active listening. Because of this, the task emerged as an important challenge for the Music Information Retrieval (MIR) research and industrial communities.

A musical composition is a layered construction of quantifiable musical elements of various temporal scales, e.g. beats, notes, bars, etc. While these elements are qualified by strict musical definition, the higher temporal level music structure is perceptually audible but not qualified by any strict musical definition. Because of this, the MSS task raised questions about the definitions of the segments to be estimated. In order to cope with this lack of definition, MSS researchers have developed assumptions-driven methods. As described by Paulus et al. [13] methods can be categorized according to the used assumptions on the content: novelty, homogeneity and repetition. We can add to this list the Regularity hypothesis proposed by Sargent et al. [15].

1.1 Related Work

MSS algorithms usually rely on two successive steps: segmentation and grouping of segments. The temporal segmentation consists in estimating the borders of potential structural segments, therefore limiting the search space for the structural and fixing its temporal scale. This step is crucial to ensure the global performance of systems. This preliminary temporal segmentation is directly influenced and constrained by the above-mentioned assumptions. This is because the various assumptions give different perspectives of the music pieces content. We briefly review these assumptions here.

With the novelty hypothesis, boundaries between structural segments are considered as time points of high "acoustical contrast". This notion of contrast has been introduced by Foote [5] and extends previously proposed audio novelty segmentation techniques [3]. It considers jointly the homogeneity within segments as well as the dissimilarity between segments. A novelty function is computed by convoluting a Self-Similarity Matrix (SSM) with a kernel that reflects the novelty hypothesis. Peaks of the novelty function define potential structural boundaries. This method has been successfully applied to music structure segmentation and still produces state-of-the-art performances for the temporal segmentation step [4] [12] [9]. Slightly different, the homogeneity assumption only require strong inner acoustical homogeneity of the structural segments. A popular approach then consists in defining the structural segments as the states of a Hidden Markov Model (HMM) [1] [14] [11].

The novelty and homogeneity assumptions assume strong inner-homogeneity of segments, a property that Peeters formalized in [14] as the "state" representation of structural segments, and that is very often related to timbral properties of the music pieces. In contrast, repetition-based temporal segmentation aims at detecting repeated segments, homogeneous or not. In the case of non-homogeneity, repeated segments are visualized in a SSM as stripes on the diagonal and off-diagonals. The segmentation of these stripes is denoted by Peeters as the sequence approach and is usually done in a Time-Lag Matrix [6].

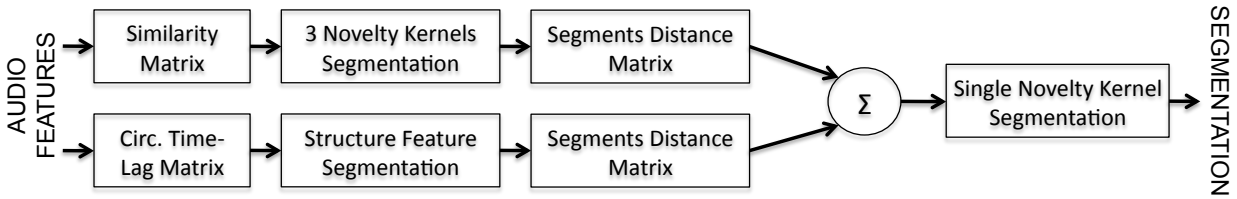

Figure 1. Fusion strategy between novelty and repetition temporal segmentation

Recently, an extension of this lag representation was proposed by Serra et al. [16] and allow for the detection of non-homogeneous as well as homogeneous repeated segments. This method is further described in the latter of this paper.

1.2 Paper overview

In this paper we study the combination of these different temporal segmentation approaches. In particular, we propose a generic fusion method that we apply to the fusion of the repetition and novelty-based approaches. The choice of these two methods is motivated by the strong antagonism that exists between them. Indeed, the detection of repetitions detection is global since it requires considering the whole signal (usually represented through a lag matrix). In contrast, the novelty approach is essentially local and has no global perspective on the music piece.

These two methods thus give different explanations of the music pieces' structure that are both true. Our claim is that these explanations are complementary and that their fusion enhances the global music structure segmentation performance. To this end, we propose a late fusion approach that maintains the acoustic coherency within the final estimated segments. Segments are first estimated under both hypotheses separately. A segments distance matrix is then computed for both segmentations. The sum of these two matrices then serves as the final representation for the temporal segmentation.

1.3 Paper organization

In Section 2 we introduce the algorithms we use for the temporal segmentation under both assumptions. In Section 3 we present the fusion procedure of the segmentations estimated with these two methods. The method is then illustrated in Section 4 on a real signal. In Section 5, we propose a comparative evaluation. Finally, conclusions are drawn in Section 6. The general architecture of the fusion method is illustrated in Figure 1.

2. TEMPORAL SEGMENTATION ALGORITHMS

2.1 Repetition-based Segmentation

The classical approach for repetition-based segmentation relies on the computation of a time-lag matrix representing chroma similarity. Recently, Serra et al. [16] proposed a novel method that extends this approach to a circular time-lag matrix representation. The latter incorporates both past and future samples. The audio signal is therefore represented by a multidimensional time series that contains, for each time frame instant, the chroma vector of the actual frame as well as knowledge of the recent past with the encapsulation of delay coordinates [10]. A recurrence plot R retaining only the nearest similar frames is then computed on this multidimensional time series. Circular shifting of the rows of R then allows to compute the circular time-lag matrix L:

$$L_{i,j} = R_{i,k+1} \tag{1}$$

with N the size of the feature vector, i = 1, ... , N and k = i + j -2 mod(N). An example of such a matrix is shown in Figure 4 (a). After smoothing with a bivariate gaussian kernel, the authors define the so-called "structure feature" that serves for the temporal segmentation as the rows of L. Structural changes are indeed detected by strong changes in the structure feature sequence and can be estimated in the difference between adjacent structure feature vectors. Evaluation at the 2012 MIREX [1] evaluation campaign for structural segmentation showed very convincing performances of this method.

2.2 Novelty-based Segmentation

The novelty-based segmentation was originally proposed by Foote in [5] and allows to detect transitions between homogenous segments of a musical signal. This is achieved by means of the correlation of 2×2 checkerboard novelty kernel along with the main diagonal of a SSM. [8] extends this method by introducing two new novelty kernels that allow for the detection of non-homogeneous to homogeneous segments transitions and vice versa. These kernels are illustrated in Figure 2. Three novelty curves are computed for all three kernels on a SSM computed on timbre-related features (MFCCs, Spectral Centroid, Spread, Skewness and Spectral Flatness). Adaptive peak tracking technique described in [7] allows for the estimation of boundary candidates in the three novelty functions and a final segmentation is obtained by merging the three boundary sets within a tolerance range of 2s. In this paper, we use the novelty method extended by [8].

2.3 Segmentations Agreement

In order to highlight the differences between the two assumptions, we compare the segmentation results obtained

[1] http://www.music-ir.org/mirex/wiki/2012:MIREX2012_Results

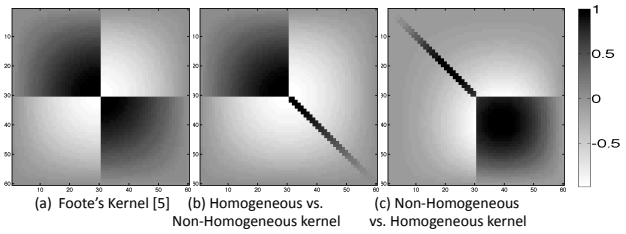

Figure 2. 60×60 Novelty detection kernels

with the repetition-based and novelty-based method presented above. We compare the results obtained on the Isophonics testset that consists of 297 popular music song. If we denote by T_R (T_N) one of the segment boundary obtained with the repetition-based segmentation (novelty-based segmentation), the boundaries T_R and T_N are said equivalent if within a tolerance window of 2 seconds.

Results show that that 45.8% of the T_N were contained in the T_R. Moreover, 55.9% of the T_R were contained in the T_N. There is thus about 50% of the information estimated by the two methods that is very specific to the chosen assumption. While the two methods achieve rather comparative results when evaluated within MIREX, this experiment strengthens the assumption that fusion of both assumptions may increase the performance of the temporal segmentation.

3. SEGMENTATIONS FUSION

Repetition- and novelty- based segmentations are explicitly designed for different representations of the audio content. An early fusion approach of these representations would be therefore irrelevant. Instead, we choose a late fusion approach: the segmentations using both assumptions are first estimated, and then merged together. In the remaining of this section we propose two fusion strategies. We first introduce a baseline method (see part 3.1) that is the simplest way to merge the boundaries. We then present (see part 3.2) a method that merges the boundaries by explicitly considering the acoustical relevancy of the fused segments. This is done by using a segments distance matrix representation.

3.1 Baseline Method

The baseline fusion of boundaries simply consists in merging T_R and T_N if they are within a given tolerance window Δ. Since identical boundaries may be detected in both sets with a slight temporal deviation, the simple union of boundary sets is not precise enough for the fusion. We therefore merge boundaries within a tolerance window. As illustrated in Figure 3 we retain the earliest boundary of the two when a matching is found.

In our experiment, we set the tolerance window at $\Delta = 2$ seconds (1 measure @120bpm) to limit over-segmentation.

3.2 Segments Distance Based Fusion

Because the baseline method is only constrained by the heuristic rule of a tolerance range, it does not consider

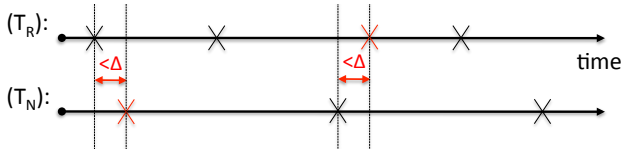

Figure 3. Illustration of the baseline fusion method.

the acoustic relevance of the newly formed segment that derived from excluding or keeping a boundary. The new segmentation might thus produce irrelevant segments and induce errors at the structure clustering step. We therefore propose here an alternative approach for the fusion of boundaries that includes knowledge of the acoustical coherence of the final estimated segments.

3.2.1 Segments distance matrix

We introduce the Segments Distance Matrix (SDM) that measures the acoustical consistency between segments formed by T_R and T_N. In this, the distance between two segments is calculated using the Mahalanobis distance between the features distributions within each pair of segments:

$$S(i,j) = d(X_i, X_j) = \sqrt{(X_i - X_j)^T \Sigma^{-1} (X_i - X_j)} \quad (2)$$

with X_i [m×n] and X_j [p×n] the feature vectors respectively within segments i and j and Σ their covariance matrix. We use the Mahalanobis distance since it provides a good compromise between the homogeneity and repetition assumptions for the segments comparison.

The size of the SDM is defined by the number of boundaries detected. We then temporally-scale this SDM to the original SSM size in order to reflect the music piece's structure. Two examples of temporally-scaled-SDMs are shown in Figure 4 (c) and (c') with the song One Vision by Queen. This example illustrates that SDMs give a perspective on the saliency of each boundary. Indeed, we easily distinguish adjacent segments with very strong acoustical dissimilarity as well as strong similarity.

3.2.2 Fusion and Boundary Selection

The fusion process should select the boundaries by taking into account the acoustic saliency of the newly formed segments. To provide a perspective on the acoustic contrasts given by the fusion of boundaries, we sum the SDMs computed corresponding to T_R and T_N. This summation allows displaying in a single representation the acoustic contrast brought by both assumptions. It gives an acoustical perspective on potentially merged boundaries. This is illustrated in Figure 4 (d). In this summed SDM, the acoustic contrasts of both segmentations are kept. Note that a similar fusion of different matrix data representations was used by Chen in [2] for structure labelling purposes. The summed SDM can be thought as a simplified SSM like representation of the music piece that gives a global acoustic description of the acoustic content.

The final temporal segmentation is then obtained by applying the novelty segmentation on this fused representation. Since the matrix describes only segments of strong inner-homogeneity we solely employ Foote's kernel illustrated in Figure 2 (a).

4. CASE STUDY

In this section we illustrate on a real signal (the song "One Vision" by Queen) our method for segment detection based on late-fusion of repetition and novely-based segmentation. Figures 4 display - for the repetition method: the Circular Time-Lag Matrix (a), structure feature (b) and SDM (c) - for the novelty method: the SSM (a), novelty curve (b) and SDM (c). The SDMs calculated for both methods illustrate the different perspectives given on the song's temporal segmentation.

The sum of the SDMs and corresponding novelty curve with final estimated boundaries are displayed in Figure 4 (d) and (e). This clearly shows the compromise that is made in our method between repetition-based and novelty-based segments. Indeed, the different acoustic contrasts within the two segmentations can be corroborated by the final segmentation of the summed SDM but this not necessarily happening. For example boundaries that were detected within frames 180 and 420 by the repetition-based method are not all contained in the final segmentation because of insufficient acoustic contrast of the newly formed segment. Hence, the fusion method uses consistent acoustic clues to decide of the fusion of segmentations.

5. EVALUATION

We evaluate comparatively the performances of the various segmentation methods taken separately (repetition and novelty-based) and the proposed fusion methods (by baseline or distance-based fusion) as proposed in this paper. In order to investigate the impact of the method on the structure labelling of segments, an evaluation of the segments labelling is also proposed. We first introduce the segment labelling process, evaluation protocol and then present and discuss the results.

5.1 Segments Labeling

For all segmentation methods studied, the labelling of the segments is achieved using the method proposed by [9], i.e. a hierarchical clustering is applied on the basis vectors of the Non-Negative-Matrix-Factorization (NMF) of the SSM. We improve this method here by estimating automatically the optimal number k of clusters (hence of different segment labels) to be formed. For this, we use a method inspired by [17]. The method consists in varying the number k of clusters to be formed, and for each number, to compute the dispersion of the obtained partition. The dispersion D_k is defined as the average distance d_{xx} between all n_i elements x, x within each cluster C_i:

$$D_k = \sum_{i=1}^{k} \frac{1}{n_i} \sum_{x,x' \in C_i} d_{xx'} \qquad (3)$$

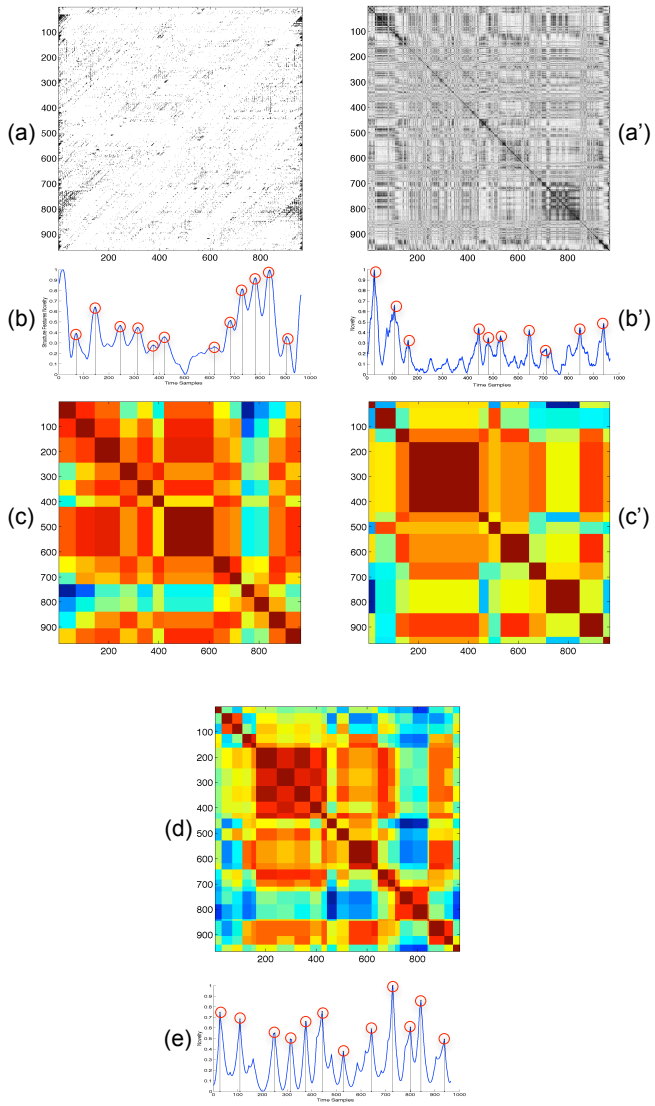

Figure 4. Example of the fusion method with the song example "One Vision" by Queen. (a) Circular Time Lag Matrix - (a') Timbre-related SSM - (b) Structure feature with estimated boundaries - (b') Novelty curve of the SSM with estimated boundaries - (c) Repetition segments SDM - (c') Novelty segments SDM - (d) Summed SDM - (e) Final novelty score with estimated boundaries

D_k monotonically decreases with the number of clusters and flattens for some k that is the ideal number of clusters. Differentiation of the D_k allows to estimate for each song the optimal number of labels.

5.2 Evaluation Protocol: Testset and Metrics

Testset: In order to allow the comparison between the results presented here and the ones obtained at the 2012 MIREX [2] evaluation for structural segmentation, we use the Isophonics testset [3], also known as the MIREX09 testset. This testset consists of 297 popular music songs (the

[2] http://www.music-ir.org/mirex/wiki/2012:MIREX2012_Results
[3] http://isophonics.net/

Method	Temp. Seg. Eval. @0,5s F	P	R	Temp. Seg. Eval. @3s F	P	R	Seg. Group. Eval. pF	pP	pR
$Repet$	22.8	22.4	24.1	64.3	63.6	67.6	59.9	**65.7**	58.3
$Novel$	**29.5**	26.7	34.7	61.8	55.9	72.5	60.0	62.5	63.2
$BaselineFusion$	28.8	24.6	**37.7**	63.4	52.8	**83.4**	59.6	64.2	59.6
$SDMFusion$	28.9	**29.5**	29.4	**65.2**	**66.7**	65.9	**62.1**	62.4	**66.7**

Table 1. Temporal segmentation and segment grouping evaluation on the Isophonics testset

Beatles, Queen, Michael Jackson...)..

Metrics: The temporal segmentation is, as in MIREX, evaluated using the precision P, recall R and F-Measure F. In order to compute the True Positives, False Positives and False Negatives, we used two tolerance windows: 0.5 and 3 seconds. The segment labelling is evaluated, as in MIREX, using of the pairwise Precision, Recall and F-Measure proposed by [11].

5.3 Results and Discussion

The results are indicated into Table 1. The repetition and novelty methods are respectively denoted by "*Repet*" and "*Novel*". The baseline fusion and segments distance based fusion are respectively denoted by "*Baseline Fusion*" and "*SDM Fusion*".

Repet versus Novel: The results obtained for temporal segmentation shows that both repetition- and novelty-based methods tend to over-segment the signal (recall $>$ precision). This is especially true for the novelty-based method. Evaluation with a 0.5s tolerance window shows better performances (F-measure) for the novelty-based method. Increasing the tolerance to 3s then turns to the advantage of the repetition-based method. The results obtained for segment labelling shows comparable performances (pairwise F-Measure) for both methods. The structural segmentations are however of different natures considering their differences in the pairwise recall and precision balance: - labelling using the Repet method tends to over-estimate the number of labels, hence inherently produce over-segmentation (pairwise Precision $>$ pairwise Recall). - the inverse phenomenon is observed using the Novel method.

Fusion methods: The performance evaluation of the baseline fusion method clearly shows a strong over-segmentation (R $>$ P for both tolerance window). Moreover, labelling of the segments for the baseline fusion method shows the worst performance. In contrast, the SDM based fusion method shows very convincing performances for both the temporal segmentation and segment labelling. Indeed, its performance for temporal segmentation (F-measure) is just behind the novelty- based method's performance at 0.5s and obtains the best score at 3s. It is also interesting to note that the temporal over-segmentation observed for both Repet and Novel segmentations is not observed in the SDM Fusion segmentation. This illustrates how the acoustic information is considered in the fusion. This is further validated by looking at the segmentations agreement. Conducting the same experiment as in Section 3.3 indeed shows that 61,8% of the repeated segments and 61,9% of the novelty segments are contained in the SDM Fusion segmentation.

Finally, the segment labelling evaluation shows a very positive impact of the SDM Fusion segmentation. We increase of about 2 percentage points the pairwise F-measure with very balanced pairwise precision and recall. Again, the SDM fusion of segments yield an original structural interpretation benefitting from both the repetition and novelty hypotheses.

6. SUMMARY AND CONCLUSION

In this paper, we proposed a method for the consistent fusion of repetition- and novelty- based temporal segmentations of music. We showed that this fused segmentation benefits from the temporal perspectives given by both hypotheses and is rather influenced by the acoustical consistency of the final segmentation than from one or the other original segmentation. Moreover, we showed that the fusion of the segmentations allows for a strong increase in the segment labelling performance. This paper thus illustrates the potential benefits of developing multiple hypotheses based structural segmentation algorithms. Moreover, we believe that the method is not restricted to the fusion of the repetition and novelty methods and could be applied to other temporal segmentation methods.

7. REFERENCES

[1] Jean-Julien Aucouturier, François 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, 2005.

[2] Ruofeng Chen and Ming Li. Music structural segmentation by combining harmonic and timbral information. In *Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR)*, 2011.

[3] Scott Shaobing Chen and P.S. Gopalakrishnan. Clustering via the bayesian information criterion with applications in speech recognition. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 2, pages 645–648 vol.2, May 1998.

[4] Matthew L. Cooper and Jonathan Foote. Automatic music summarization via similarity analysis. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, 2002.

[5] Jonathan Foote. Automatic audio segmentation using a measure of audio novelty. In *Proceedings of the IEEE International Conference on Multimedia and Expo*, 2000.

[6] Masataka Goto. Chorus-section detecting method for musical audio signals. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2003.

[7] A.L. Jacobson. Auto-threshold peak detection in physiological signals. In *Engineering in Medicine and Biology Society, 2001. Proceedings of the 23rd Annual International Conference of the IEEE*, volume 3, pages 2194–2195 vol.3, 2001.

[8] Florian Kaiser and Geoffroy Peeters. Multiple hypotheses at multiple scales for audio novelty computation in music. In *38th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Vancouver, Canada, 2013.

[9] Florian Kaiser and Thomas Sikora. Music structure discovery in popular music using non-negative matrix factorization. In *Proceedings of the 11th International Society for Music Information Retrieval Conference (ISMIR)*, Aug 2010.

[10] Holger Kantz and Thomas Schreiber. *Nonlinear Time Series Analysis*. Cambridge University Press, 2003.

[11] M. Levy and M. Sandler. Structural segmentation of musical audio by constrained clustering. *IEEE Transactions on Audio, Speech & Language Processing*, 16(2):318–326, 2008.

[12] Jouni Paulus and Anssi Klapuri. Music structure analysis using a probabilistic fitness measure and a greedy search algorithm. *IEEE Transactions on Audio, Speech & Language Processing*, 17(6):1159–1170, 2009.

[13] Jouni Paulus, Meinard Müller, and Anssi Klapuri. Audio-based music structure analysis. In *Proceedings of the 11th International Society for Music Information Retrieval Conference (ISMIR)*, 2010.

[14] Geoffroy Peeters. *Deriving Musical Structures from Signal Analysis for Music Audio Summary Generation: "Sequence" and "State" Approach*, volume 2771 of *Lecture notes in Computer Science*, pages 143–166. Springer, 2004.

[15] Gabriel Sargent, Frederic Bimbot, and Emmanuel Vincent. A regularity-constrained viterbi agorithm and its application to the structural segmentation of songs. *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, 2011.

[16] Joan Serra, Meinard Müller, Peter Grosche, and Josep Ll. Arcos. Unsupervised detection of music boundaries by time series structure features. In *AAAI International Conference on Artificial Intelligence*, 2012.

[17] Robert Tibshirani, Guenther Walther, and Trevor Hastie. Estimating the number of clusters in a data set via the gap statistics. *Journal of the Royal Statistical Society, series B*, 63:411–423, 2001.