MUSIC STRUCTURE BOUNDARY DETECTION AND LABELLING BY A DECONVOLUTION OF PATH-ENHANCED SELF-SIMILARITY MATRIX

Tian Cheng, Jordan B. L. Smith, Masataka Goto

National Institute of Advanced Industrial Science and Technology (AIST), Japan
{tian.cheng, jordan.smith, m.goto}@aist.go.jp

ABSTRACT

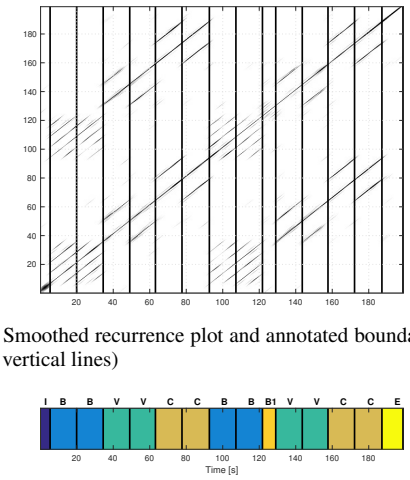
We propose a music structure analysis method that converts a path-enhanced self-similarity matrix (SSM) into a block-enhanced SSM using non-negative matrix factor 2-D deconvolution (NMF2D). With a non-negative constraint, the deconvolution intuitively corresponds to the repeated stripes in the path-enhanced SSM. Then the block-enhanced SSM is constructed without any clustering technique. We fuse block-enhanced SSMs obtained using different parameters, resulting in better and more robust results. Discussion shows that the proposed method can be a potential tool for analysing music structure at different scales.

Index Terms— Music structure analysis, NMF 2D deconvolution, path-enhanced self-similarity matrix, fusion

1. INTRODUCTION

A self-similarity matrix (SSM) is an important mid-level representation for music structure analysis, which is generated by computing frame-to-frame similarity. The repetition of segments typically leads to diagonal stripes in the SSM [1]. The stripes can be emphasised in the path-enhanced SSM by applying diagonal smoothing to the SSM [2] or to the recurrence plot [3]. As shown in Figure 1, the stripe patterns clearly illustrate the repetition of segments (e.g., verse or chorus) and the repetition within a segment (e.g., the bridge in Figure 1). The stripes of the same group can be generated by shifting a specific structure pattern. Based on this observation, we propose to decompose the path-enhanced SSM into structure patterns (as shown in Figure 2(a)) and shift activations corresponding to each pattern using non-negative matrix factor 2-D deconvolution (NMF2D) [4]. We then construct a new, block-enhanced SSM by computing the SSM of the normalised activations. In order to obtain a robust result, we fuse block-enhanced SSMs obtained with different parameters. Then we detect boundaries on the fused, block-enhanced SSM using a checkerboard kernel [5], and label the detected segments by comparing each pair of segments.

Several previous methods decompose the SSM to analyse music structures. Two methods, [6] and [7], apply non-negative matrix factorisation (NMF) to SSMs by assuming that columns within the same segment types have similar distributions. This assumption apparently works better on block-enhanced SSMs. In practice, to be effective, [6] only works on segment labelling by clustering NMF activations, and [7] roughly estimates the main segments by performing low-rank convex-NMF. In [8], the original SSM is decomposed using symmetric NMF by considering the symmetry of the SSM. Then the reconstruction of the symmetric NMF is taken as the block structure of the SSM. [1] decomposes the path-enhanced SSM by eigenvalue decompositions, obtaining different eigenvectors corresponding to different stripe structures. The clustered eigenvectors are used to compute an SSM in the form of a block structure. Then NMF is used to decompose the block-enhanced SSM for labelling.



(a) Smoothed recurrence plot and annotated boundaries (in vertical lines)

(b) Annotations: introduction (I), bridge (B), verse (V), chorus (C), bridge 1 (B1) and ending (E)

Fig. 1: An illustration of piece 'RWC-MDB-P-2001 No.23' in RWC Pop Database [9].

We consider our method to be most similar to [1]: both methods convert a path structure to a block structure. However, there are several differences. First, we compute the block-enhanced SSM without a clustering step because of the intuitive representation of the NMF2D. Second, we fuse the block-enhanced SSMs obtained using different parameter settings by taking the sum of them for a more robust result. Third, we directly use the average value from the block structure to indicate the similarity of two segments. Then segment labelling is achieved by simple thresholding, without any further post-processing steps. Our method may also be seen as a stripe (i.e., sequence)-oriented version of [7], which used NMF to decompose blocks (i.e., homogeneous segments) in SSMs. Focusing on stripes vs. blocks reflects two different interpretations of musical structure, and it is important to be able to analyse both.

The proposed method is evaluated using the RWC Pop Database [9], and compares favourably to some released results on the same database in the Music Information Retrieval Evaluation eXchange (MIREX) campaign.[1] Results show that the proposed method is among the top of the state-of-the-art.

This work was supported in part by JST ACCEL Grant Number JPMJAC1602, Japan.

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

2. METHOD

2.1. Generating the path-enhanced SSM

To generate a path-enhanced SSM, we adopt the method of computing recurrence plots in [3].[2] Firstly, we compute Harmonic Pitch Class Profile (HPCP) features [10] with 12 pitch classes, a window length of 743 ms and a hop size of 372 ms. Then we concatenate HPCP features of M adjacent frames to construct a new feature. We zero-pad the beginning and end of the HPCP features to keep the indices of new features the same as those of the original HPCP features. The concatenated features are denoted by $X = [\mathbf{x}_1, ..., \mathbf{x}_N]$, where $\mathbf{x}_i$, for $i = 1, ..., N$, is a column vector with the dimension of $12 \times M$, and N is the total number of feature frames.

Next, we compute the distance (the Euclidean norm) for each pair of frames. The recurrence plot R is obtained as follows. For each frame $\mathbf{x}_i$, $i = 1, ..., N$, we search for its K nearest neighbours in $\mathbf{x}_j$, $j = 1, ..., N$. If $\mathbf{x}_i$ is a neighbour of $\mathbf{x}_j$, and $\mathbf{x}_j$ is also a neighbour of $\mathbf{x}_i$, we set $R_{i,j} = 1$. K is set according to the frame length N by $K = kN$, where $k \in (0, 1]$.

We convolve the recurrence plot with a diagonal matrix covering 6 s to smooth the stripes and to eliminate short lines. The diagonal value of the matrix is a Gaussian function with a standard deviation of 3 s. A smoothed recurrence plot is shown in Figure 1(a).

2.2. Deconvolution of the path-enhanced SSM

In an NMF2D model, a pattern is convolved in both the x-axis and y-axis [4]. As shown in Figure 1, the columns within a repeated segment can be generated from a single column's pattern shifted along both the x- and y-axes. So we apply NMF2D to the smoothed recurrence plot to separate frames of different groups. The smoothed recurrence plot R is modelled as:

$$R \approx \sum_{\tau} \sum_{\Phi} \overset{\downarrow\Phi}{W^{\tau}} \overset{\rightarrow\tau}{H^{\Phi}}. \quad (1)$$

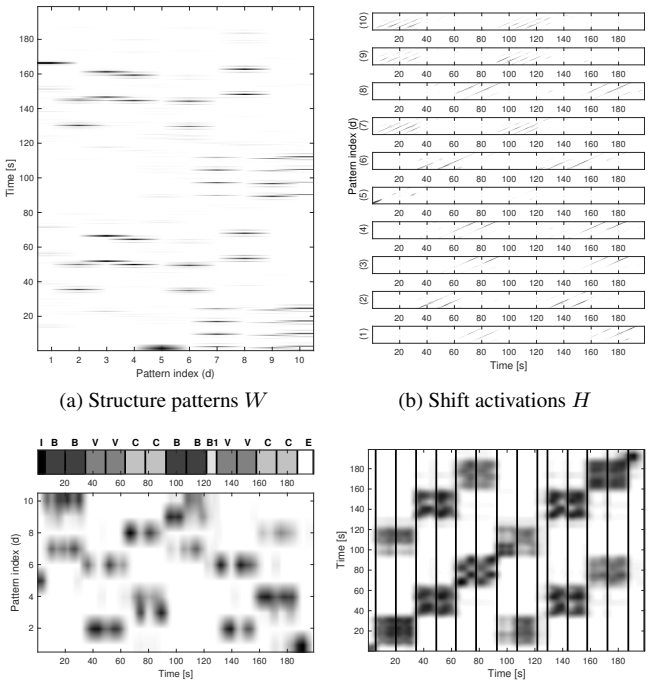
Each column W_d ($1 \leq d \leq D$) in Figure 2(a) represents a structure pattern, and its shifting activation $H_d \in \mathbb{R}^{\Phi_m \times N}$ is shown in the corresponding sub-figure in Figure 2(b), where Φ_m is the range of vertical shifting. We assume that the number of labels in a song is not larger than 10 ($D = 10$), and the length of a segment is not longer than 30 seconds (Φ_m equal to 30 seconds). Because the patterns are vectors, there is no need for horizontal shifting in this approach; the horizontal shift variable τ is thus set to 0. We use the implementation of [4][3] with a sparsity weight $\lambda = 2$ to enforce sparsity on H. For parameters not mentioned, we kept the default values.[4]

For the active frames in H_d, their structures in R are shifted from the same structure pattern W_d. Thus we sum H_d to indicate the frames corresponding to the structure pattern W_d: $H1_d(n) = \sum_{\Phi} H_d(\Phi, n)$. Then we normalise $H1_d$ to a maximum of 1 and smooth it using a 6-second window. The normalised activation NH is shown in Figure 2(c).

2.3. Generating the block-enhanced SSM

The structure patterns obtained via NMF2D are not guaranteed to correspond one to one to the structural groups. It is clear in Figure 2(c) that 3 patterns (W_7, W_9 and W_{10}) correspond to the bridge part; 2 patterns (W_2 and W_6) correspond to the verse part; 3 patterns (W_3, W_4 and W_8) correspond to the chorus part; and patterns W_5 and W_1 correspond to the introduction and the ending parts, respectively. Bridge 1 is not represented by any pattern.

Despite that, the active frames of patterns with the same label are more correlated than those of other patterns. Hence these frames have more similar distributions in NH. Based on this observation, we directly compute the SSM of NH to form a block-enhanced SSM, as shown in Figure 2(d).



(a) Structure patterns W (b) Shift activations H

(c) Normalised activations NH (d) Block-enhanced SSM

Fig. 2: The process of computing the block-enhanced SSM.

2.4. A simple fusion

Because of the random initialisation of the NMF2D, the deconvolution is different every time. In order to obtain more robust results, we run NMF2D 8 times with different parameters. When generating the recurrence plot in Section 2.1, we group $M \in \{5, 6, 7, 8\}$ frames and search for the top 2 or 3 percent nearest neighbours ($k \in \{0.02, 0.03\}$). To fuse the block-enhanced SSMs obtained from the 8 parameter sets $[M, k]$, we simply take their sum [11]. The preliminary test shows that the fusion step improves the results by over 6 percentage points on boundary detection in comparison to the result of each parameter set. Besides making the results more robust, the fused SSM keeps boundaries obtained in different block-enhanced SSMs. If a boundary is detected in more SSMs, it is easier to detect in the fused SSM, and more likely to be a true boundary.

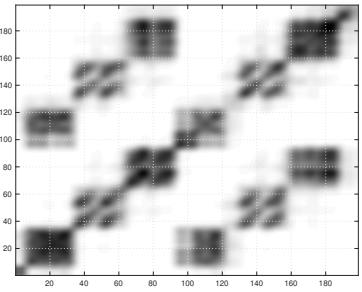
2.5. Boundary detection and labelling

After obtaining the fused block-enhanced SSM, we normalise it to a maximum of 1, denoted by M_B. We apply a checkerboard kernel [5] to M_B to generate a novelty curve. All peaks in the novelty curve are detected as boundaries. We add two other boundaries at the first and last frames. The fused SSM and detected boundaries are shown in Figure 3(a) and (b) respectively.

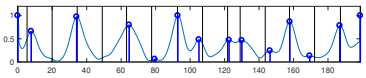
[2] Using SSM in [3] outperforms SSM in [2] in preliminary experiments.

[3] http://www2.imm.dtu.dk/pubdb/views/publication_details.php?id=4499

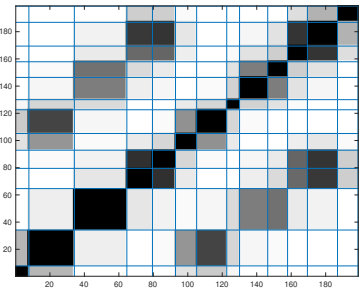
[4] It takes around 9 seconds to converge for a song of 200 seconds on a 2.8 GHz Intel Core i7 computer.



(a) Fused block-enhanced SSM



(b) Novelty curve and detected boundaries, with annotated boundaries indicated by black vertical lines



(c) Similarity of each pair of detected segments

Fig. 3: Boundary detection and labelling based on the block-enhanced SSM.

Having detected the boundaries, we compute the similarity of each pair of segments. The similarity is indicated by the average value of the block-enhanced SSM M_B. Given L detected boundaries, we have L-1 segments. The similarity between the i^{th} and j^{th} segments is computed as follows:

$$S_{i,j} = \frac{1}{N_i \times N_j} \sum_{I_i} \sum_{I_j} M_B(I_i, I_j),\ i, j = 1, ..., L-1. \quad (2)$$

Where, $(I_i,\ I_j)$ indicates the frame indices in the i^{th} and j^{th} segments, with lengths of N_i and N_j, respectively. The similarity of each segment to itself is forced to be $S_{i,i} = 1$. The similarity of the detected segments is shown in Figure 3(c). If the similarity value is larger than a threshold (0.5 in this paper), the pair of segments are detected as the same group. We enforce transitivity generously: if A and B belong to the same group, and A and C belong to the same group, then all A, B and C belong to the same group.

3. EXPERIMENT

3.1. Evaluation metrics and database

Two subtasks of music structure analysis are evaluated here: boundary detection and labelling. A boundary is considered to be correct if it lies within ± 3 s from a ground truth boundary [12]. We report boundary detection results of F-measure (F_B), precision (P_B) and recall (R_B). Labelling is evaluated in 0.1 s frames, with pairwise F-measure (F_L), precision (P_L) and recall (R_L).

We test the proposed method on the RWC Pop Database [9]. Two different annotations of this database are available, indicated by 'RWC-POP-A'[5] and 'RWC-POP-B'[6] respectively. We use both annotations to evaluate the proposed method for boundary detection and labelling.

3.2. Results

Among the results obtained by using checkerboards of different lengths, the best boundary detection result is achieved by the 2×14-second (28-second) checkerboard.[7] As shown in Table 1(a), for the RWC-POP-A annotations the proposed method achieves F-measure of 73.6% and 63.7% for boundary detection and labelling, respectively. For comparison, we select the 4 results published in MIREX from 2012 to 2017 with a boundary detection F-measure above 67% or a labelling F-measure above 60% (this ignores 33 competitors with lower scores). The SMGA method [3] achieved the best F-measure on both boundary detection (78.5%) and labelling (68%); see Table 1(a)). The proposed method has the second best F-measures, outperforming the rest of the methods on both tasks. Both the proposed and SMGA methods are based on the same recurrence plot. In Section 4, we discuss these two methods in detail.

Using the RWC-POP-B annotations (see Table 1(b)), SMGA [3] still has the best boundary detection F-measure of 79.7%, followed by GS1 and GS3 with F-measures of 79.3% and 72.9%, respectively. Both GS methods [13] train a Convolutional Neural Network (CNN) on the combined self-similarity lag matrix (of two different lags). The trained model works as a binary classifier to detect boundaries in a moving window. The proposed method achieves an F-measure of 72.1%, better than the other two methods (NB2 [14] and FK2 [15]) by 2.1 and 6.3 points, respectively.

In the RWC-POP-B annotations, the silent parts at the beginning and end of the music pieces are annotated as separate segments. This means that 4 correct boundaries can be detected only by detecting the silent parts. The results show that the improvement of 6.4 percentage points from GS3 to GS1 is achieved by identifying the silent segments. We do not think these boundaries are as important as others. In addition, the results may be biased, considering the high ratio of the number of silence-related boundaries to all boundaries. Thus, we also report the boundary detection results when ignoring all boundaries in the first 5 seconds and last 5 seconds [16] for the proposed and GS methods. The proposed method achieves 70.6% and 71.7% on F-measure with two annotations, respectively.[8] The F-measures of the GS method [13] are 70.9% and 75.2%, respectively, with no difference between GS1 and GS3.

The labelling F-measure of the proposed method is as high as 73.5% in Table 1(b). However, MIREX does not yet publish results with the RWC-POP-B labels, which were updated in 2014, so we cannot directly compare our work. Our results can serve as a baseline for future research.

4. DISCUSSION

When we inspected errors made by our algorithm, we found it was usually not due to poor modelling of the stripes, but a mismatch be-

[5] `https://staff.aist.go.jp/m.goto/RWC-MDB/AIST-Annotation/`

[6] `http://musicdata.gforge.inria.fr`. We use the reduced version of the annotation.

[7] Testing a reasonable range of checkerboard sizes (between 22 s to 30 s), the variance of the F-measures is less than 1 percentage.

[8] The smoothness in Section 2.2 reduces false alarms but blurs the boundaries. The F_Bs (within ± 0.5 s) of the proposed method are less than 25%.

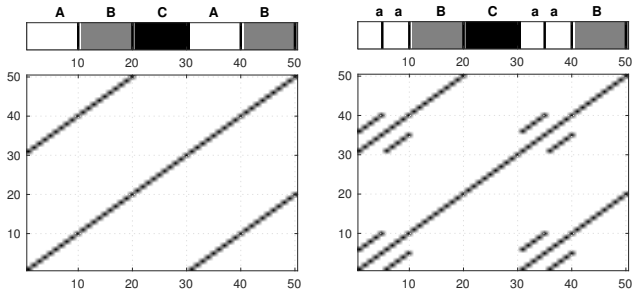
Method	F_B	P_B	R_B	F_L	P_L	R_L
Proposed (no fusion)	67.5	72.3	64.9	59.9	80.2	49.4
Proposed	73.6	80.2	69.3	63.7	81.2	54.3
SMGA^a [3]	**78.5**	81.7	77.3	**68.0**	72.8	66.5
GS1 (2015) [13]	71.5	80.6	65.6	54.2	77.9	43.1
GS3 (2015) [13]	73.6	89.4	64.0	54.2	77.9	43.1
NB2 (2014) [14]	70.3	76.6	66.9	55.5	52.5	61.7
FK2 (2013) [15]	65.7	81.6	56	63.5	79.6	61.2
Proposed^{-5}	70.6	79.8	64.9			
GS^{-5} (2015) [13]	70.9	88.6	61.0			

(a) Results in percentage with RWC-POP-A

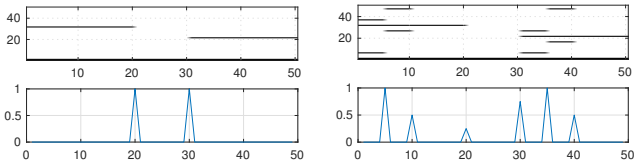
Method	F_B	P_B	R_B	F_L	P_L	R_L
Proposed (no fusion)	65.6	71.7	61.8	68 .0	77.7	63.5
Proposed	72.1	80.2	66.6	**73.5**	79.9	70.7
SMGA^a [3]	**79.7**	82.7	78.2			
GS1 (2015) [13]	79.3	91.9	71.1			
GS3 (2015) [13]	72.9	90.9	62.1			
NB2 (2014) [14]	70.0	78.1	64.6			
FK2 (2013) [15]	65.8	84.1	55.2			
Proposed^{-5}	71.7	76.9	69.0			
GS^{-5} (2015) [13]	75.2	89.4	67.2			

(b) Results in percentage with RWC-POP-B

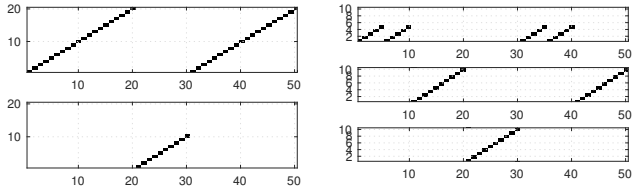
Table 1: Results with different annotations. Superscript ^{a} denotes results quoted from [3]. Superscript $^{-5}$ denotes boundary detection results by ignoring boundaries in the first and last 5 seconds.



(a) Annotations and path-enhanced SSMs



(b) Structure features and boundaries detected based on [3]



(c) Shift activations based on the proposed method

Fig. 4: Two examples: Left, 'ABCAB', right, 'aaBCaaB'.

tween our hypothesis and the song being analysed: the stripes that exist are found reliably, but they can be insufficient to model the structure. Methods based on path-enhanced SSMs (including [1, 3] and the proposed method) are sensitive to repetitions between segments and within segments. We illustrate this with two examples.

In Figure 4(a), we show two typical path-enhanced SSMs corresponding to segment sequences labelled 'ABCAB', but in the right example, segment 'A' consists of repeating sub-segments 'aa'. The simulated results of SMGA [3] and the proposed method are shown in Figure 4(b) and (c). We can see that in the first case 'ABCAB', both methods miss the boundaries between A and B; if both A and B are through-composed, this boundary is, in a sense, undiscoverable, and might only be found when considering musical novelty or homogeneity, or segment size uniformity.

In the second case 'aaBCaaB', both methods find the boundaries between A and B. SMGA also detects the boundary internal to 'aa', and this boundary is more noticeable than the one between A and B, whereas our method recognizes both as belonging to the same "repetition type". So, despite the gap in performance, we still recognize the special ability of our NMF2D approach to resolve intra-segment repetitions as being a kind of "homogeneously repetitive" state, which we hope to exploit better in future work.

Of course, there may be disagreement on the ground truth whether 'aa' should be labelled as one segment or two segments. So beyond this evaluation, we would like to provide meaningful representations for analysing music structure at different scales, such as in [17], [18] and the SALAMI database[9]. NMF2D shows promise as a multi-scale music structure analyser. In this paper, we find frames of the same group based on the summed shift activations. Then within the group, the repetition is clearly indicated by the diagonal stripes in the shift activations as shown in Figure 4(c) and also in the real example in Figure2(b).

Among the MIREX algorithms omitted from Table 1(b) are versions of [6] and [7], two NMF approaches directed at blocks instead of stripes. By targeting sequences instead of homogeneity, we improve F-measure from below 67 to 72.1, closer to the state-of-the-art performance of SMGA (F-measure = 79.7). We also found, like [11], that a simple late fusion approach led to increased robustness: both precision and recall increased in a balanced way to improve F-measure by 7 points. Based on these observations, we speculate that by fusing non-negative factorisation models of stripes and blocks, performance may improve even more.

5. CONCLUSION AND FUTURE WORK

In order to analyse music structure, we propose to deconvolve a path-enhanced SSM using NMF2D and generate a block-enhanced SSM based on the summed activations. We fuse block-enhanced SSMs obtained with different parameters. The analysis of the decomposition results shows that the summed activations intuitively separate frames belonging to different labels. The fusion step makes the results less sensitive to the parameters of computing the path-enhanced SSM, and also improves the performance.

The proposed method sums the shift activations to indicate the frames of the same pattern. However, the stripes in the original shift activations can be potentially used to detect the repeated segments of the same label, providing an analysis at a different scale. Exploiting this remains a task for future work. We are also interested in the performance of the proposed method on other databases, especially the SALAMI database with different annotation scales.

[9] http://ddmal.music.mcgill.ca/research/salami.

6. REFERENCES

[1] H. Grohganz, M. Clausen, N. Jiang, and M. Müller, "Converting path structures into block structures using eigenvalue decompositions of self-similarity matrices," in *Proc. of the 14th International Society for Music Information Retrieval Conference (ISMIR)*, 2013, pp. 209–214.

[2] M. Müller, N. Jiang, and H. G. Grohganz, "SM Toolbox: MATLAB Implementations for Computing and Enhancing Similarity Matrices," in *Proc. of the 53rd Audio Engineering Society (AES)*, 2014.

[3] J. Serrà, M. Müller, P. Grosche, and J. L. Arcos, "Unsupervised music structure annotation by time series structure features and segment similarity," *IEEE Trans. on Multimedia, special Issue on Music Data Mining*, vol. 16, no. 5, pp. 1229–1240, 2014.

[4] M. N. Schmidt and M. Mørup, "Nonnegative matrix factor 2-D deconvolution for blind single channel source separation," in *Proc. of International Conference on Independent Component Analysis and Signal Separation*, 2006.

[5] J. Foote, "Automatic audio segmentation using a measure of audio novelty," in *Proc. of the IEEE International Conference on Multimedia and Expo(I)*, 2000, pp. 452–455.

[6] F. Kaiser and T. Sikora, "Music structure discovery in popular music using non-negative matrix factorization," in *Proc. of the 11st International Society for Music Information Retrieval Conference (ISMIR)*, 2010, pp. 429–434.

[7] O. Nieto and T. Jehan, "Convex non-negative matrix factorization for automatic music structure identification," in *Proc. of ICASSP*, 2013.

[8] J. Kauppinen, A. Klapuri, and T. Virtanen, "Music self-similarity modeling using augmented nonnegative matrix factorization of block and stripe patterns," in *Proc. of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2013.

[9] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka, "RWC Music Database: Popular, Classical, and Jazz Music Databases," in *Proc. of the 3rd International Conference on Music Information Retrieval (ISMIR)*, 2002, pp. 287–288.

[10] E. Gomez, *Tonal description of music audio signals*, Ph.D. thesis, Universitat Pompeu Fabra, Barcelona, Spain, 2006.

[11] F. Kaiser and G. Peeters, "A simple fusion method of state and sequence segmentation for music structure discovery," in *Proc. of the 14th International Society for Music Information Retrieval Conference (ISMIR)*, 2013.

[12] M. Levy and M. Sandler, "Structural segmentation of musical audio by constrained clustering," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 16, no. 2, pp. 318–326, 2008.

[13] T. Grill and J. Schlüter, "Music boundary detection using neural networks on combined features and two-level annotations," in *Proc. of the 16th International Society for Music Information Retrieval Conference (ISMIR)*, 2015.

[14] O. Nieto and J. P. Bello, "MIREX 2014 Entry: 2d Fourier Magnitude Coefficients," in *Proc. of MIREX*, 2014.

[15] F. Kaiser and G. Peeters, "MIREX 2013 - Music Structural Segmentation Task: IrcamStructure Submission," in *Proc. of MIREX*, 2013.

[16] J. B. L. Smith and E. Chew, "A meta-analysis of the MIREX structural segmentation task," in *Proc. of the 14th International Society for Music Information Retrieval Conference (ISMIR)*, 2013, pp. 251–256.

[17] M. Goto, "A Chorus-Section Detection Method for Musical Audio Signals and Its Application to a Music Listening Station," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 14, no. 5, pp. 1783–1794, 2006.

[18] B. McFee, O. Nieto, and Juan Bello, "Hierarchical evaluation of segment boundary detection," in *Proc. of the 16th International Society for Music Information Retrieval Conference (ISMIR)*, 2015.