MUSIC STRUCTURE SEGMENTATION ALGORITHM EVALUATION: EXPANDING ON MIREX 2010 ANALYSES AND DATASETS

Andreas F. Ehmann[1] **Mert Bay**[1] **J. Stephen Downie**[1] **Ichiro Fujinaga**[2] **David De Roure**[3]

[1]GSLIS
University of Illinois Urbana-Champaign
{aehmann, mertbay, jdownie}@illinois.edu

[2]Schulich School of Music
McGill University
ich@music.mcgill.ca

[3]Oxford e-Research Centre
University of Oxford
david.deroure@oerc.ox.ac.uk

ABSTRACT

Music audio structure segmentation has been a task in the Music Information Retrieval Evaluation eXchange (MIREX) since 2009. In 2010, five algorithms were evaluated against two datasets (297 and 100 songs) with an almost exclusive focus on western popular music. A new annotated dataset significantly larger in size and with a more diverse range of musical styles became available in 2011. This new dataset comprises over 1,300 songs spanning pop, jazz, classical, and world music styles. The algorithms from the 2010 iteration of MIREX are re-evaluated against this new dataset. This paper presents a detailed analysis of these evaluation results in order to gain a better understanding of the current state-of-the-art in automatic structure segmentation. These expanded analyses focus on the interaction of algorithm performance and rankings with datasets, musical styles, and annotation level. Because the new dataset contains multiple annotations for each song, we also introduce a baseline for expected human performance for this task.

1. INTRODUCTION

The structural, or formal, analysis of music is one of the most fundamental of analyses performed by musicologists. Very simply, the main goal of structural analysis is to segment music into sections that share similar characteristics, and apply labels to these sections. These segmentations take forms such as AABB, or ABAC, etc. With further analysis, certain descriptors can also be applied to these sections, such as verse, chorus, and so on [3].

In recent years, there has been increasing interest in developing methods for performing structural analyses automatically. For a good overview on the state of automatic music audio structural segmentation we refer the reader to [10]. The growing interest in structural segmentation algorithms is evidenced by the establishment of the structural segmentation task of the Music Information Retrieval Evaluation eXchange (MIREX) campaign [2]. Evaluations of structural segmentation algorithms were performed in 2009 and 2010. These evaluations were performed over collections with a strong bias towards western, popular music.

To perform a novel and potentially more thorough evaluation of the performance of structural segmentation algorithms, the set of algorithms submitted to MIREX 2010 in July 2010 was re-evaluated in May 2011 using a newly constructed dataset. For the purposes of this paper, we are calling this new test collection the MIREX 2010 Version 2 (MRX10V2) dataset. MRX10V2 is much larger in size than the datasets used in earlier MIREX evaluations. It also contains a much broader range of music styles. Moreover, the MRX10V2 database contains multiple annotations per piece. Having multiple annotations per song allows us, for the first time, to explore how well algorithms perform this task relative to human experts.

The main motivation for this work stems from an ongoing project called the Structural Analysis of Large Amounts of Music Information (SALAMI) [3]. The SALAMI project is an endeavor to use music structure algorithms to annotate and segment a large corpus of music (on the order of 300,000 songs). Its main goal is to test the feasibility and usefulness of current music information retrieval algorithms on a larger scale than has commonly been performed. As a pilot to the SALAMI project, the work presented in this paper aims to further our understanding of how current state-of-the-art algorithms perform at music segmentation.

The rest of this paper is formatted as follows. Section 2 gives a description of the dataset used in this mid-cycle MIREX evaluation. Section 3 briefly describes the algorithms. Section 4 presents the evaluation results. Section 5 offers some conclusions and suggests future work.

2. DATASETS: OLD AND NEW

The evaluation of structure segmentation algorithms on a large dataset requires the creation of a suitable ground truth. As in virtually all cases of MIREX-style evaluations, ground truth creation is carried out by human annotators. The use of human annotators brings up two significant challenges. First, there is a large labor cost involved in manually annotating music pieces. Second, and perhaps more importantly, is the notion that it is difficult to truly assert that any subjective interpretation of something as complex as musical form is "truth." Many considerations must be taken into account regarding such annotations. Both [1] and [11] lay out methodologies for annotating musical structure. In this work, the dataset, and subsequent annotation methodologies described in [14] are used.

The MIREX 2009 and the MIREX 2010 iterations of the MIREX structural segmentation task had an over bias toward popular music. The dataset known as MIREX 2009 contains 297 popular song annotations donated by Tampere University of Technology, Vienna University of Technology and Queen Mary, University of London. Music of *The Beatles* makes up a significant proportion of the MIREX 2009 dataset. The MIREX 2010 dataset consists of an annotated version of the RWC [4] database's popular music collection. Note that the published results to the MIREX 2010 dataset are evaluated against a ground truth donated by members of the QUAERO Project.[1] However, these annotations consist of only segment boundary annotations with no labeling. Hereafter, results pertaining to the MIREX 2010 dataset are evaluated against the original, labeled structural annotations as distributed with the RWC collection.

In order to compensate for the popular music bias exhibited by the older datasets, the new MRX10V2 dataset was deliberately created to include a much wider variety of musical styles. In addition to popular music, the new dataset contains classical, jazz, live, and world music. Table 1 presents the distribution of styles across the MRX10V2 dataset. While "live" may not truly be considered a musical style, live pieces are separated as they raise unique concerns such as applause sections, etc.

The "Double-keyed" pieces noted in Table 1 are those that have been annotated by two separate individuals. As Table 1 shows, the majority of pieces (1048 of 1383) have been annotated by two annotators. In addition, each annotation of a piece contains two levels of structural hierarchy. There is a fine-grained annotation and a coarse grained annotation, with each coarse-grained segment comprising one or more fine-grained segments. Therefore, a "fine" annotation may have form *abaabacdaba*, with equivalent "coarse" annotation of *AABA* where *A* represents an *aba* sequence and *B* represents a *cd* sequence. The new dataset contains 1,383 pieces which is over 4 times larger than earlier datasets used for evaluation.

Style	Double-keyed	Single-keyed	Total	Percentage
Classical	159	66	225	16%
Jazz	225	12	237	17%
Popular	205	117	322	23%
World	186	31	217	16%
Live	273	109	382	28%
Total	**1048**	**335**	**1383**	100%

Table 1. Breakdown of the MRX10V2 structure segmentation dataset by musical style.

3. ALGORITHMS

The algorithms used in this off-cycle MIREX evaluation are the same as the ones submitted to MIREX 2010. Five unique algorithms, including one with two distinct parameter settings (resulting in six overall algorithms), were run against the new 1,383 song dataset and evaluated. The algorithms are referred to in this paper using the code names assigned to them during MIREX 2010.[2]

Each of the algorithms under evaluation is composed of a unique combination of extracted features, segmentation methods, and labeling/grouping techniques. BV1-2 [13] uses beats, Mel Frequency Cepstral Coefficients (MFCCs) and chroma vectors as features, segments the song based on generalized likelihoods of three different criteria and gathers the segments using agglomerative hierarchical clustering. GP7 [12] uses MFCC, chroma vectors, spectral flatness and valley factors as features, calculates a weighted sum of 4 different distance matrices that is used to segment the signal. The segments are merged using hierarchical agglomerative clustering. MHRAF2 [8] uses chroma features and employs string matching techniques to identify strong harmonic redundancies using an iterative detection of major repetitions. MND1 [9] uses chroma vectors and calculates a similarity matrix using Pearson's correlation coefficient. MND1 searches the diagonals for repeated sequences and uses a greedy algorithm to decide on the segments. WB1 [16] uses beat synchronous chromagrams decomposed into basis patterns by shift-invariant probabilistic latent component analysis as features. Songs are segmented by computing the path of the basis patterns through a likelihood function that represents the structure of the song using the Viterbi algorithm.

[1] See http://www.quaero.org.

[2] See http://nema/mirex/wiki/2010:MIREX2010

On average, the runtimes for the algorithms is approximately two to six minutes per file. Table 2 presents the average runtimes per-file for each algorithm. We can see that most algorithms run roughly real-time. Therefore, any very large-scale effort to automatically segment music audio will require significant computational resources.

Algorithms	Average processing time (min. / piece)
WB1 [16]	2.28
GP7 [12]	2.64
BV1 & BV2 [13]	2.94
MND1 [9]	5.60
MHRAF2 [8]	6.38

Table 2. Algorithm names, corresponding references, and runtimes.

4. EVALUATION AND RESULTS

4.1 Evaluation Methods

The same evaluation methods and metrics used in previous structural segmentation MIREX evaluations were used to evaluate the algorithms. The boundary retrieval metrics of [15] evaluate how close segment boundaries between algorithm results and ground truth are in time. This metric is label-agnostic and simply measures the segmentation of the piece and not whether similar sections are similarly labeled. The "hit rate" of the boundary retrieval measures if a returned segment boundary is within T seconds of a ground truth boundary. The hit rate is measured at two time-thresholds: T = 0.5 s and T = 3.0 s. The segment boundary hit rate measures encompass an F-measure (*SBR-F*), as well as a precision (*SBR-P*) and recall (*SBR-R*) measure. In addition, the median deviation, in seconds, between detected and ground truth boundaries is measured. *AB-2-RB* measures the median time difference between an annotated boundary and the nearest result boundary. Similarly, *RB-2-AB* measures the median time difference between a result boundary and the nearest annotated boundary.

Frame-pair clustering, as introduced in [6], divides the results and ground truth into short time frames (e.g. 100 ms). This metric then considers every possible pair of frames and their corresponding labels. Denoting the set of all frame-pairs that share the same label (i.e., same cluster) in the result as P_E, and likewise the set of all frame-pairs sharing the same label in the ground truth as P_A, we can define the pairwise precision, P, pairwise recall, R, and pairwise F-measure, F as

$$P = \frac{|P_E \cap P_A|}{|P_E|} \qquad R = \frac{|P_E \cap P_A|}{|P_A|} \qquad F = \frac{2PR}{P+R} \tag{1}$$

The frame-pair clustering F-measure, precision, and recall are to as *FPC-F*, *FPC-P*, and *FPC-R*, respectively.

The normalized conditional entropies introduced in [7] also represent structural annotations as sequences of short frames, similar to the frame-pair clustering metrics. Conditional entropies are calculated and normalized to yield a measure in [0, 1], the details of which are beyond the scope of this paper and can be found in the reference. The normalized conditional entropy measures are a dual measure with over-segmentation (*NCE-OSS*) and under-segmentation scores (*NCE-USS*). Because structure annotation can exist at multiple levels of granularity (as it does in the new ground truth), the two metrics will indicate if an algorithm tended to be too coarse (low under-segmentation score) or too fine (low over-segmentation scores). Finally, a random clustering index (RCI) measure is also calculated [5].

(a) Algorithm	NCE-OSS	NCE-USS	FPC-F	FPC-P	FPC-R	RCI	SBR-F@0.5s	SBR-P@0.5s	SBR-R@0.5s	SBR-F@3s	SBR-P@3s	SBR-R@3s	AB-2-RB	RB-2-AB
BV1	0.605	0.441	0.520	0.513	0.669	0.549	0.190	0.151	0.289	0.450	0.361	0.669	1.797	7.554
BV2	0.454	0.715	0.427	0.678	0.350	0.638	0.189	0.150	0.286	0.449	0.361	0.666	1.812	7.552
GP7	0.499	0.683	0.485	0.675	0.424	0.654	0.188	0.146	0.306	0.440	0.346	0.695	2.073	6.634
MHRAF2	0.546	0.591	0.559	0.617	0.583	0.659	0.195	0.218	0.197	0.435	0.485	0.440	7.262	5.338
MND1	0.624	0.625	0.556	0.649	0.586	0.662	0.291	0.302	0.326	0.470	0.479	0.534	8.565	5.389
WB1	0.609	0.540	0.546	0.583	0.608	0.630	0.237	0.240	0.272	0.393	0.395	0.446	10.780	3.881
(b) Algorithm	NCE-OSS	NCE-USS	FPC-F	FPC-P	FPC-R	RCI	SBR-F@0.5s	SBR-P@0.5s	SBR-R@0.5s	SBR-F@3s	SBR-P@3s	SBR-R@3s	AB-2-RB	RB-2-AB
BV1	0.643	0.323	0.384	0.321	0.680	0.505	0.179	0.236	0.159	0.567	0.744	0.499	2.905	2.007
BV2	0.521	0.567	0.373	0.452	0.386	0.712	0.177	0.234	0.157	0.565	0.741	0.497	2.937	1.980
GP7	0.584	0.557	0.432	0.467	0.482	0.720	0.163	0.208	0.153	0.472	0.605	0.436	4.946	2.300
MHRAF2	0.599	0.442	0.440	0.395	0.615	0.655	0.124	0.276	0.087	0.356	0.776	0.253	11.311	1.885
MND1	0.666	0.478	0.435	0.426	0.609	0.635	0.200	0.376	0.150	0.415	0.749	0.314	13.944	1.835
WB1	0.675	0.420	0.442	0.382	0.653	0.632	0.148	0.277	0.112	0.317	0.588	0.239	16.031	1.975

Table 3. Evaluations against coarse (a) and fine (b) ground truth annotations.

5. RESULTS & DISCUSSION

The evaluation results of the six algorithms using the new MRX10V2 dataset can be seen in Tables 3a (coarse-grained) and 3b (fine-grained). The figures in the tables represent weighted averages over the dataset, where the averaging was carried out as follows. All algorithms were evaluated over a single ground truth for the entire annotated dataset (~1300 pieces). Those pieces that were double-keyed were then used as a second ground truth and separately evaluated. These two separate evaluations were then weighted by the number of pieces in each set and averaged to produce the final results.

We take immediate note that the algorithms tend to perform better when evaluated using the coarser of the two human annotations (mostly evidenced by the *FPC-F* measure and low *NCE-USS* scores in Table 3b). With regard to the *FPC-F* data, the average performance for all algorithms using the coarse-grained ground truth is 0.520 versus 0.423 for the fine-grained. A Friedman's ANOVA test[1] run using the *FPC-F* measure data confirms that there exists a statistically significant difference in performance between the coarse and fine result sets (*p=0.01*). This result is not surprising, as the algorithms are designed for coarse annotation. We will talk about the relative performances of algorithms using only the coarse *FPC-F* scores later.

In comparing the MRX10V2 results with the previous MIREX datasets, we see that the evaluation results for all algorithms seem to be in the same general range. Using *FPC-F*-measure for comparison (as it provides a good balance between segmentation and labeling accuracy), Table 4 contains algorithm performances on the new dataset, the MIREX 2009 dataset, and the MIREX 2010 dataset. In general, average performance seems to be slightly worse on the new MRX10V2 dataset. Some algorithms seem to have been more strongly affected, with significant performance drops (e.g. BV2 and GP7). Some algorithms, however, also improved slightly on the MRX10V2 dataset over the MIREX 2009 dataset (e.g. MND1 and WB1). The smallest dataset, RWC, appears to generate the best performances. A Friedman's ANOVA test run against the Table 4 data indicated a statistically significant difference in performance among the three datasets (*p=0.02*). A subsequent Tukey-Kramer Honestly Significant Difference (TKHSD) test tells us that the MIREX 2010 collection results are significantly different than the other two collections. The same TKHSD also shows that MIREX 2009 and MRX10V2 are not different from each other. We suspect that the MIREX 2010 results are significantly better than the other two datasets because the RWC popular music database which makes up the MIREX 2010 set was artificially composed and performed to represent generic popular music and to overcome copyright problems.

Algorithm	MIREX09	MIREX10	MRX10V2	Ave.
BV1	0.502	0.520	0.520	0.514
BV2	0.493	0.531	0.427	0.484
GP7	0.536	0.592	0.485	0.538
MHRAF2	0.555	0.600	0.559	0.571
MND1	0.613	0.625	0.556	0.598
WB1	0.544	0.602	0.546	0.564
Ave.	0.541	0.578	0.516	0.545

Table 4. Comparison of algorithms over datasets

Recall that the earlier MIREX datasets have a strong bias toward western popular music. As mentioned in Section 2, MRX10V2 dataset was deliberately created to represent a wider range of musical styles to evaluate algorithmic performance across different genres. Table 5 presents a breakdown of algorithm performance across musical styles. Again, the *FPC-F* measure is used as a summary measure for comparison, and only the coarse annotations are considered. A Friedman's ANOVA test run against the Table 5 data indicates that there is no statistically significant differences in performance across musical styles (*p=0.90*). This is a promising result because it suggests that although, to date, most algorithms have been evaluated on popular music, they do seem to perform reasonably well on other styles. Such a claim is not meant to imply that individual algorithms do not perform significantly better on some musical styles than others. Rather, when all algorithms are looked at as a whole, musical style does not seem to have a large effect (i.e., individual idiosyncrasies average out).

Algorithm	Live	Classical	Jazz	Popular	World	Ave.
BV1	0.504	0.513	0.544	0.519	0.521	0.520
BV2	0.432	0.426	0.398	0.451	0.439	0.429
GP7	0.510	0.427	0.475	0.513	0.484	0.482
MND1	0.532	0.564	0.574	0.574	0.545	0.558
MHRAF2	0.557	0.590	0.556	0.543	0.555	0.560
WB1	0.560	0.524	0.547	0.548	0.537	0.543
Ave.	0.516	0.507	0.516	0.525	0.514	0.515

Table 5. Results by musical style considering only coarse annotations.

[1] See [2] for an in-depth discussion of the applications of Friedman's ANOVA and the Tukey-Kramer Honestly Significant Difference (TKHSD) tests used in MIREX.

Algorithm	Fine	Coarse
BV1	0.392	0.525
BV2	0.371	0.434
GP7	0.433	0.485
MHRAF2	0.448	0.565
MND1	0.442	0.559
WB1	0.449	0.552
Human	0.629	0.721

Table 6. Fined-grained vs. coarse-grained FPC-F results.

5.1 Best Performances: Algorithms vs. Humans

In order to gain some general idea on how a human might perform relative to another human using the standard evaluation measures, the ground truths of all double-keyed files were compared. We performed the human-to-human comparison on both the coarse and fine-grained annotations. The double-keyed subset allows us to evaluate the human-generated annotations in the same manner as the algorithms. The Human results line in Table 6 was generated by declaring one human annotation set to be an "algorithm" while the other played the role of "ground truth." The algorithms were evaluated on only this subset of the data to allow for direct comparison of the results on the 794 double-keyed pieces that have both fine and coarse annotations.

Table 6 shows that structural annotation by music experts seems to be itself somewhat subjective. For example the average coarse-grained *FPC-F* score is 0.721. This indicates that some disagreement does exist amongst human experts. A higher degree of disagreement exists for the fine-grained annotations.

While algorithmic segmentations seem to perform similarly to each other, automatic segmentation has not reached human performance. We performed Friedman's ANOVA on the coarse-grained *FPC-F* scores for the algorithmic and human annotations across 794 tracks. At $p<0.01$, the Friedman's test indicates a statistically significant difference in performance among the annotation sources. The subsequent TKHSD multiple comparison tests show a set of four distinct performance groupings with each group being significantly different from the other groups (with no significant differences with each grouping). Figure 1 presents the results of the TKHSD test.

In the first performance group, we find, by itself, the results for the human annotations. These are noticeably better than any of the algorithmic results. This is to be expected given the relatively few years the community has be working on the structural segmentation problem. The second grouping (highlighted by an oval in Figure 1) consists of MHRAF2, MND1, and WB1. These three algorithms are not significantly different. BV1, GP7, and BV2 all have statistically significant performance differences. These results remind us of two important facts. First, the top performing algorithms are not significantly different in the MIREX 2010 Version 2 evaluations. We need to look at the stronger algorithms as a group to see what factors can be merged to build an improved segementation system. Second, notwithstanding human variations in strucutural annotations, we as a community still have a great way to go before our structural segmentation algorithms can be said to be acheiving human-like performances.

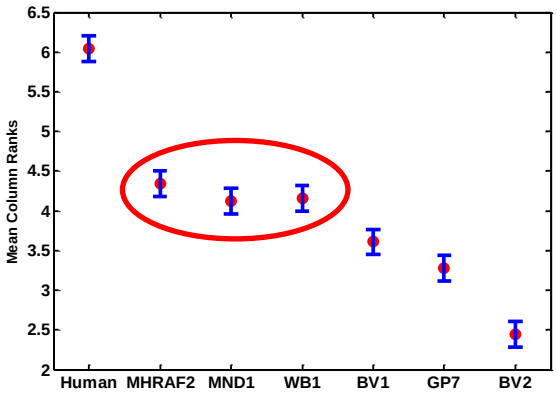

Figure 1. Tukey-Kramer HSD comparison plots of the human and algorithm mean performance ranks across 794 double-keyed tracks

6. CONCLUSIONS AND FUTURE WORK

In this paper we reported upon the most extensive evaluation of music structure segmentation algorithms to date. Our evaluation was performed on a new dataset spanning multiple musical styles. Top-ranked techniques for the automatic segmentation of music quantitatively perform similarly. Musical style does not seem to have an adverse affect on general performance, but individual algorithms have a nonuniform performances across styles. We can also conclude that the state of automatic segmentation is relatively immature. Even though we assert that structural or formal analysis is in itself a subjective endeavor, the comparison of two human annotators to one another far outperforms current algorithms. In summary, we have no current single technique that is clearly better than the others and none approach the capabilities of a music expert in this task.

The evidence that there is still a large room for improvement of current segmentation algorithms does not preclude them from being useful in their present form. Even

though it is understood that the algorithms have not yet met the sort of baseline that most researchers set for themselves (i.e., approaching human performance) it is important to note that these goals are far from being met in many facets of music information retrieval (MIR), be it chord estimation, multipitch detection, and so forth. The primary goal of the SALAMI project, and much of the future work that will stem from the evaluation performed here, is to assess just how useful current MIR algorithms can be.

For future work, we see the need to increase the size of our test collections. We would like to gather more annotations per song to augment our ability to explore the similarities and differences in human segmenting perceptions. We would also like to expand the number of styles and time periods represented in our test collections. Finally, we would like to perform a set of failure analyses on those songs that consistently scored poorly in order to discern what musical traits might be proving difficult for the annotators, both human and algorithmic, to process.

7. ACKNOWLEDGMENTS

We would like to thank the past music structure segmentation MIREX participants for allowing us access and use of their algorithms in evaluating them on this new structure dataset. This material is based upon work supported by the National Science Foundation under Grant No. IIS 10-42727.

8. 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 11th International Society for Music Information Retrieval Conference*, pp. 189–94, 2010.

[2] J.S. Downie. "The Music Informaton Retrieval Evaluation Exchange (2005-2007): A Window into Music Information Retrieval Research," *Acoustical Science and Technology*, 29 (4), pp. 247-55, 2008.

[3] A.F. Ehmann, M. Bay, J.S. Downie, I.Fujinaga, D. De Roure. "Exploiting Music Structures for Digital Libraries, " *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries*, pp. 479-80, 2011.

[4] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka. "RWC Music Database: Popular, classical, and jazz music databases," *Proceedings of the International Conference on Music Information Retrieval*, pp. 287–8, 2002.

[5] L. Hubert and R. Arabie. "Comparing Partitions," *Journal of Classification*, 2 (1), 193-218, 1985.

[6] C. Levy and M. Sandler. "Structural Segmentation of Musical Audio by Constrained Clustering," *IEEE Transaction on Audio, Speech, and Language Processing* , 16 (2), 318–26, 2008.

[7] H. Lukashevich. "Towards Quantitative Measures of Evaluating Song Segmentation," *Proceedings of the International Conference on Music Information Retrieval*, pp. 375–80, 2008.

[8] B. Martin, P. Hanna, M. Robine, and P.Ferraro. "Indexing Musical Pieces Using their Major Repetition," *ACM/IEEE Joint Conference on Digital Libraries*, Ottawa, Canada, 2011.

[9] M. Mauch, K. C. Noland, and S. Dixon. "Using musical structure to enhance automatic chord transcription," *Proceedings of the International Society for Music Information Retrieval Conference*, 231–6, 2009.

[10] J. Paulus, M. Mueller, and A. Klapuri: "State of the Art Report: Audio-Based Music Structure Analysis," *Proceedings of the 11th International Society for Music Information Retrieval Conference*, pp. 625–36, 2010.

[11] G. Peeters and E. Deruty. "Is music structure annotation multi-dimensional? A proposal for robust local music annotation," *Proceedings of the International Workshop on Learning the Semantics of Audio Signals*, pp. 75–90, 2009.

[12] G. Peeters. "Sequence representation of music structure using higher-order similarity matrix and maximum likelihood approach," *Proceedings of the International Conference on Music Information Retrieval,* pp. 35-40, 2007.

[13] G. Sargent, F. Bimbot, and E. Vincent. "Un système de détection de rupture de timbre pour la description de la structure des morceaux de musique," *Proceedings of Journées d'Informatique Musicale,* pp. 177–86, 2010.

[14] J.B.L. Smith, J.A. Burgoyne, I. Fujinaga, D. De Roure and J.S. Downie. "Design and creation of a large-scale database of structural annotations," *Proceedings of the 12th International Society for Music Information Retrieval Conference*, 2011.

[15] D. Turnbull, G. Lanckriet, E Pamalk, and M. Goto, "A Supervised Approach for Detecting Boundaries in Music Using Difference Features and Boosting," *Proceedings of the International Conference on Music Information Retrieval*, pp. 51-54, 2007.

[16] R. J. Weiss and J. P. Bello. "Identifying repeated patterns in music using sparse convolutive non-negative matrix factorization," *Proceedings of the International Society for Music Information Retrieval Conference*, pp. 123-8, 2010.