SYSTEMATIC EXPLORATION OF COMPUTATIONAL MUSIC STRUCTURE RESEARCH

Oriol Nieto
Pandora Media, Inc.
onieto@pandora.com

Juan Pablo Bello
Music and Audio Research Laboratory
New York University
jpbello@nyu.edu

ABSTRACT

In this work we present a framework containing open source implementations of multiple music structural segmentation algorithms and employ it to explore the hyper parameters of features, algorithms, evaluation metrics, datasets, and annotations of this MIR task. Besides testing and discussing the relative importance of the moving parts of the computational music structure eco-system, we also shed light on its current major challenges. Additionally, a new dataset containing multiple structural annotations for tracks that are particularly ambiguous to analyze is introduced, and used to quantify the impact of specific annotators when assessing automatic approaches to this task. Results suggest that more than one annotation per track is necessary to fully address the problem of ambiguity in music structure research.

1. INTRODUCTION

In recent years, numerous open source packages have been published to facilitate research in the field of music information retrieval. These publications tend to focus on a specific part of the standard methodology of MIR: audio feature extraction (e.g., Essentia [2], librosa [14]), datasets (e.g., SALAMI [22], MSD [1]), evaluation metrics (e.g., mir_eval [20]), and task-specific algorithm implementations (e.g., segment boundary detection [13], pattern discovery [16], beat tracking [4]). What is often missing are integrated frameworks where the choice of different moving blocks of the whole process (i.e., feature design, algorithm implementations, annotated datasets and evaluation metrics) can be interchanged in a seamless fashion, allowing the type of in-depth comparative studies on state of the art techniques that are virtually impossible in the context of MIREX [1]: e.g., what combination of features or preprocessing stages maximize results? What mixture of approaches should be used if highly accurate boundary localization is important? What implementations are more resilient to changes in data, features or prior information?

In this work we introduce an open source framework to facilitate reproducibility and encourage research in music structural segmentation. Building on top of existing open projects [9, 14, 20, 22], this framework combines feature computation, algorithm implementations, evaluation metrics, and annotated datasets in a standalone software focused on this area of MIR. Besides describing the architecture of this framework, we show its potential by compiling a new dataset composed of poly-annotated tracks carefully selected by the presented software, and conducting a series of experiments to systematically explore the impact of each moving part of this task. These new data and explorations reinforce the notion that this task is highly ambiguous [3], since we show that the ranking of computational approaches in terms of performance depends not only on what feature or dataset is employed, but on which annotation is used as reference.

The rest of this article is organized as follows: In Section 2 the framework is introduced. Section 3 discusses the creation of the new dataset. In Section 4 the explorations of the different moving parts of the structural segmentation eco-system are presented. Finally, in Section 5, the conclusions are drawn.

2. MUSIC STRUCTURE ANALYSIS FRAMEWORK

MSAF [2] is an open source framework written in Python that allows to thoroughly analyze the entire music structure segmentation eco-system. In this section we provide an overview of this MIR task and a description of the most relevant characteristics of this framework.

2.1 Structural Segmentation

This task, whose main goal is to automatically identify the large-scale, non-overlapping segments of a given audio signal (e.g., verse, chorus), has been investigated in MIR for over a decade [19], and nowadays it is still one of the most active in MIREX [23]. Potential applications to motivate its research are numerous, e.g., improve intra-track navigation, yield enhanced segment-level music rec-

[1] One exception would be MARSYAS [25], where feature extraction, algorithm implementations for a limited number of tasks, dataset annotations, and evaluations coexist in a single environment.

[2] https://github.com/urinieto/msaf

© Oriol Nieto, Juan Pablo Bello.
Licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0). **Attribution:** Oriol Nieto, Juan Pablo Bello. “Systematic Exploration of Computational Music Structure Research”, 17th International Society for Music Information Retrieval Conference, 2016.

ommendation systems, produce educational visualisation tools to better understand musical pieces. This task is often divided in two subproblems: *boundary detection* and *structural grouping*. The former identifies the beginning and end times of each music segment within a piece, and the latter labels these segments based on their acoustic similarity.

Several open source implementations to approach this problem have been published [10, 12, 13, 26], but given the differences in feature extraction, datasets, and evaluation metrics, it can be challenging to easily compare their results (e.g., Weiss' implementation [26] expects features computed with Ellis' code [5]; Levy's implementation [10] is only available in the form of a Vamp Plugin; McFee's publication [12] reports non-standard evaluation metrics with the first and last boundary removed). Our proposed, open-source MSAF seeks to address these issues by integrating these various components.

In the following subsections, the main parts of this framework are described. MSAF is written such that any of these parts could be easily extended.

2.2 Features

Most music structure algorithms accept different types of features in order to discover structural relations in harmony, timbre, loudness or a combination of them. Here we list the set of features that MSAF can compute by making use of librosa [14]: Pith Class Profiles (PCPs, representing harmony), Mel-Frequency Cepstral Coefficients (MFCCs, representing timbre), Tonal Centroids (or Tonnetz [7], representing harmony), and Constant-Q Transform (CQT, representing harmony, timbre and loudness).

Each of these features depend on additional analysis parameters such as sampling rate, FFT size, and hop size. Furthermore, a beat-tracker [4] (contained in librosa) is employed to aggregate all the features at a beat level, thus obtaining the so-called beat-synchronous representations. This process, which is common in structural segmentation, reduces the number of feature vectors while introducing tempo invariance. In this work we rely on this type of features exclusively, even though MSAF can operate both on beat- or frame-synchronous descriptors.

2.3 Algorithms

Algorithms of this task are commonly classified based on the subtask that they aim to solve. MSAF includes: seven algorithms that detect boundaries, and five that group structure (see Table 1).

The implementations in MSAF are either forked from the public repositories of their original publications [10, 12, 13, 26] or implemented from scratch when no access to the source code is available. Some differences in the results might arise given the difficulty of exactly recreating all implementation details, even though these differences appear to be minor.

Algorithm	Boundary	Grouping
2D-Fourier Magnitude Coeffs [15]	No	Yes
Checkerboard Kernel [6]	Yes	No
Constrained Cluster [10]	Yes	Yes
Convex NMF [18]	Yes	Yes
Laplacian Segmentation [12]	Yes	Yes
Ordinal LDA [13]	Yes	No
Shift Invariant PLCA [26]	Yes	Yes
Structural Features [21]	Yes	No

Table 1: Approaches included in MSAF and used in the experiments.

2.4 Evaluation Metrics

Structural segmentation employs multiple metrics to evaluate each of its two subproblems. For boundary detection, the Hit Rate is the most standard one, where the estimated boundaries are considered "hits" if they fall within a certain time window from the reference ones. This yields Precision (how many estimated boundaries are correct) and Recall (how many reference boundaries were estimated) scores, which are weighted with the standard F-measure. The time windows are typically 3 or 0.5 seconds. Moreover, sometimes the first and last boundaries are "trimmed" (i.e., not considered) given the fact that they should correspond to the beginning and end of the track, which should be trivial to detect. It has been shown that having a stronger weight on Precision than Recall tends to better align with perception [17], therefore this weight parameter is also part of MSAF. The other standard metric to report the quality of the boundaries is the Median Deviation [24], where the median deviation from each estimated boundary to its closest reference, and vice versa, are reported.

The most standard metric to assess the quality of the structural grouping subproblem is the Pairwise Frame Clustering [10]. This metric compares each pair of frames by checking whether they belong to the same label (or cluster), both for the estimation and reference. The ratio between the two sets of pairs over the number of similar pairs in the estimation yields the Precision metric, while Recall is the ratio between the two sets over the number of similar pairs in the reference. Again, the F-measure weights these two scores. Finally, an alternative metric named Normalized Conditional Entropy [11], based on the entropy of each frame between the estimation and reference, is also reported. This metric is formed by the under- and over-segmentation scores, which, again, can be compacted in a single score with the F-measure.

These metrics are reported in MIREX, and are transparently implemented in mir_eval [20], which MSAF employs.

2.5 Datasets

The following annotated datasets are the most common for assessing structural segmentation: Isophonics – 298 annotated tracks mostly of popular music [3]; SALAMI – two human references plus three levels of annotation per

[3] http://isophonics.net/datasets

track [22]. It contains 769 musical pieces ranging from western popular music to world music [4] ; The Beatles TUT – refined version of 174 annotations of The Beatles corrected and published by members of the Tampere University of Technology [5] .

Additionally, we make use of these uncommon and novel datasets: Cerulean – 104 songs collected by a company, subjectively deemed to be challenging tracks within a large collection. The genre of the songs varies from classical to heavy metal; Epiphyte – another industrial set of 1002 tracks composed mainly of pop music songs; Sargon – small set of 30 minutes of heavy metal tracks released under a Creative Commons license; SPAM – new dataset discussed in the next section.

All of these datasets are converted to the JAMS format [9], which is the default format that MSAF employs, and are publicly available in the MSAF repository (except Cerulean and Epiphyte, which are privately owned). This format is JSON-compatible and allows for multiple annotations in a single file for numerous tasks operating on a given audio track, making it ideal for the purposes of this work.

3. STRUCTURAL POLY-ANNOTATIONS OF MUSIC

SPAM is a new dataset composed of 50 tracks sampled from a large collection containing all the previously discussed sets (a total of 2,173 tracks). Following an approach inspired by [8], all MSAF algorithms were run on these 2,173 tracks. The tracks were then ranked based on the average Hit Rate F-Measure with 3 seconds window (i.e., the most standard metric for boundary detection) across all algorithms. Formally, the rank is computed using the mean ground-truth precision (MGP) score, defined as follows:

$$\mathrm{MGP}_i(B, g) = \frac{1}{M}\sum_{j=1}^{M} g(\mathbf{b}_{ij}) \qquad (1)$$

where $B \in \mathbb{R}^{N,M}$ is the matrix containing all the boundary estimations $\mathbf{b}_{ij} \in B$ for track $i \in [1, N]$ using algorithm $j \in [1, M]$, and g is the evaluation function (i.e., Hit Rate at 3 seconds). Ranking the tracks using this metric yields a list sorted by how *challenging* these tracks are for automatic segmentation.

The SPAM dataset is composed by the 45 most challenging tracks (i.e., the 45 at the bottom of the ranked list) plus the 5 least challenging (i.e., the top 5 tracks in the list). The number of tracks was kept small to facilitate the collection of five additional annotations using the same guidelines as in SALAMI. These five annotations were collected by music students (four graduates and one undergraduate) from the Steinhardt School at New York University, with an average number of years in musical training of 15.3 ± 4.9, and with at least 10 years of experience as players of a musical instrument. The goal was to create a set in which to explore the variability of structural annotations across subjects, focusing on the most challenging tracks (45) while still having a reduced control group (5). This split could foster further investigation on the differences between *easy* and *challenging* tracks.

The type of music ranges between jazz and blues, classical, world music, rock, western pop, and live recordings. Due to legal copyright issues, the audio of these tracks is not available, however, the features described in Section 2.2 are included along with the five annotations for each of the 50 tracks of SPAM.

4. EXPERIMENTS

In this section we report a series of experiments to further explore the task of structural segmentation carried out using MSAF, classified by the moving parts described previously. Each experiment can be subdivided based on the subproblems of boundary detection and structural grouping. For each experiment the default parameters are the following, unless otherwise specified: sampling rate is 11025Hz; FFT and hop sizes are 2048 and 512 samples, respectively; default feature type is beat-synchronous PCP; number of octaves and starting frequency for the PCPs are 7 and 27.5Hz, respectively; number of MFCCs is fixed to 14; number of CQT bins, starting at 27.5Hz, is 87; evaluation metrics are the F-measures of the Hit Rate with 3 seconds window and the Pairwise Frame Clustering for boundary detection and structural grouping, respectively; the boundaries used as input to the structural grouping algorithms are annotated; and the default dataset is The Beatles TUT. Code to reproduce the plots and results is available online [6] .

4.1 Features

We start by running all MSAF algorithms [7] using different types of features. In Figure 1 we can see, as expected, that the average scores of the boundary algorithms vary based on the feature types. This aligns with the results of a two-way ANOVA on the F-measure of the Hit Rate using the algorithms and features as factors, where the effect on the type of features is significant ($F(3, 3460) = 4.20, p < 0.01$). Also as expected, there is significant interaction between factors ($F(12, 3460) = 15.15, p < 0.01$), which can be seen in the plot when observing the poor performance of the Constrained Clustering algorithm for the Constant-Q features in comparison with the rest of the features.

A similar behavior occurs when analyzing the performance of the structural grouping algorithms, as can be seen in Figure 2. The two-way ANOVA confirms dependency of the type of features for these algorithms ($F(3, 2768) = 18.07, p < 0.01$), with significant interaction ($F(9, 2768) = 14.5, p < 0.01$) mostly due to the behavior, again, of the Constrained Clustering algorithm

[4] Only the first half of the full SALAMI annotations were used, since the authors did not have access to the rest of audio files.

[5] `http://www.cs.tut.fi/sgn/arg/paulus/beatles_sections_TUT.zip`

[6] `https://github.com/urinieto/msaf-experiments`

[7] Except Ordinal LDA and Laplacian Segmentation, since they only accept a specific combination of features as input.

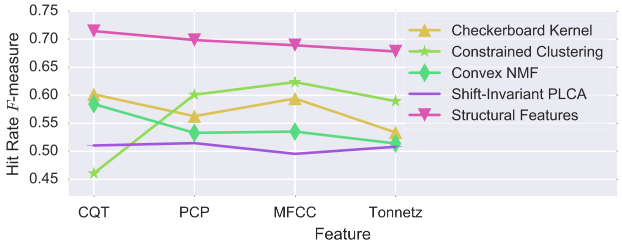


Figure 1: Boundary algorithms' performance depending on the type of features.

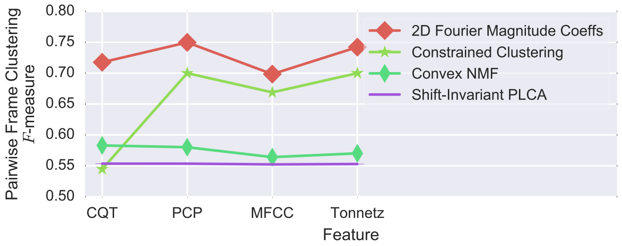


Figure 2: Structural algorithms' performance depending on the type of features.

when using Constant-Q features. Convex NMF still performs slightly better with this type of features, while the rest of the algorithms seem to be optimized to operate on the features suggested in their original publications.

This experiment yields two major points: (i) features describing timbre information (CQT, MFCCs) seem to be slightly better than those describing pitch information (PCPs, Tonnetz) for boundary detection, but the reverse seems to be true for clustering, and (ii) the Structural Features and Convex NMF methods obtain better results when using CQT, while in their original publications they recommend using harmonic features such as PCPs.

4.2 Algorithms

The quality of the segment boundaries can impact the results of the structural grouping subproblem [21]. MSAF lets us explore this by using the output of several boundary algorithms as input to the structural algorithms. Figure 3 shows average scores of the structural algorithms in MSAF. Additionally, the results with annotated boundaries are used and plotted in the first column. The boundary methods are sorted from left to right based on their performance on The Beatles TUT dataset. As expected, the quality of the boundary detection process affects the structural subproblem. A two-way ANOVA on the F-measure of the Pairwise Frame Cluster, with boundary and structural algorithms as factors, confirms this ($F(7, 6920) = 183.10, p < 0.01$). A significant interaction between the two factors is also present ($F(28, 6920) = 16.44, p < 0.01$), suggesting that the ranking of the algorithms will vary depending on the boundaries employed. This is confirmed by the Friedman test, which ranks the structural algorithms using Structural Features boundaries ($F(4) = 242.31, p < 0.01$) differently than when using Convex NMF boundaries ($F(4) = 225.05, p < 0.01$). For example, the 2D Fourier Magnitude Coefficients method becomes lower ranked than Convex NMF in the latter case, as can be seen in the plot.

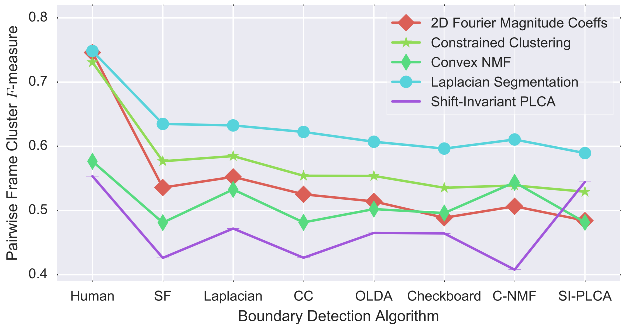
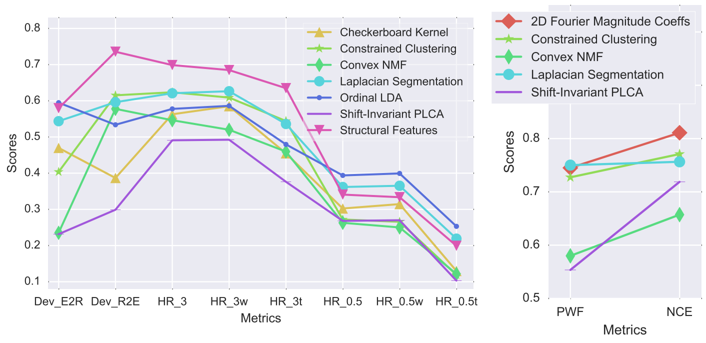


Figure 3: Performance of the structural algorithms contained in MSAF when using different types of previously estimated boundaries as input.

Interesting conclusions can be drawn: first, some structural algorithms are more robust to the quality of the boundaries than others (e.g., 2D-FMC sees a strong impact on its performance when not using annotated boundaries, especially when compared with the Laplacian method). Second, the best performing boundary algorithm will not necessarily make the results of a structural algorithm better, as can be seen in the structural results of C-NMF and SI-PLCA. To exemplify this, note how SF (which tends to outperform all other methods in terms of the standard metric, see Figure 1) produce, in fact, one of the lowest results in structural grouping for the C-NMF method. On the other hand, the Laplacian method (which outputs boundaries that are comparable to the ones by the Checkerboard kernel), obtains results on the structural part that are much better than those by SF. Finally, depending on the boundaries used, structural algorithms will be ranked differently in terms of performance (especially when using annotated boundaries as input). This is something that is not currently taken into account in the MIREX competition, and might be an interesting asset to add in the future for a deeper evaluation of the subtask of structural grouping.

4.3 Evaluation Metrics

In this section we explore the different results obtained by MSAF algorithms when assessed using the available metrics. For boundary detection, the metrics described in Section 2.4 are explored, which are depicted in Figure 4a as "Dev_E2R" for the median deviations from Estimations to References (R2E for the swapped version), and "HR_n" for the Hit Rate with a time window of n seconds (the w and t indicate the weighted and trimmed versions, respectively). The median deviations are divided by 4 in order to normalize the scores within the range of 0 to 1, and then inversed in order to indicate a better performance with a higher score. As expected, scores are significantly different depending on the metric used,



(a) Boundaries

(b) Structure

Figure 4: Scores of MSAF algorithms depending on evaluation metrics.

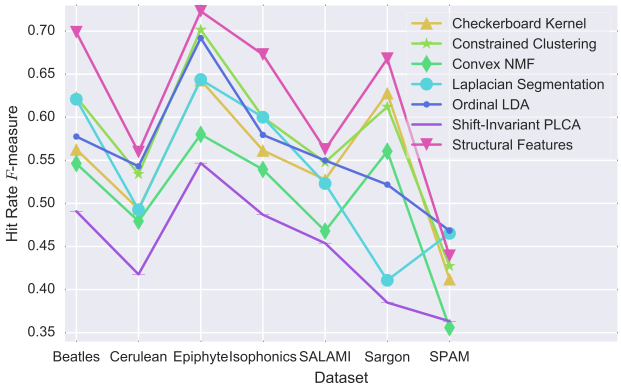


Figure 5: Boundary algorithms' performance depending on dataset.

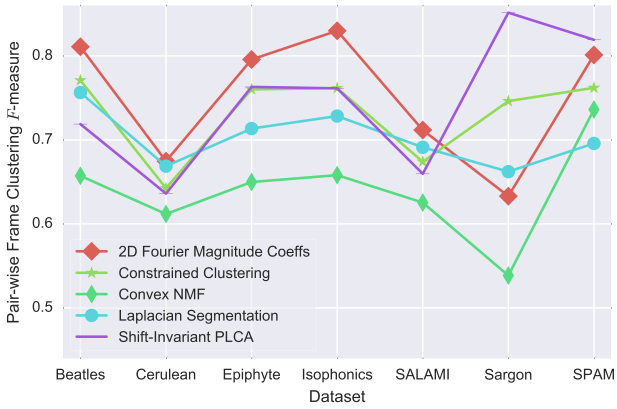


Figure 6: Structural algorithms' performance depending on dataset.

which is confirmed by the two-way ANOVA of the scores with metrics and algorithms as factors (the metric effect is $F(7, 9688) = 458, p < 0.01$). But perhaps more interesting is the fact that some algorithms perform better with some metrics than others (as suggested by the interaction effect of the two-way ANOVA: $F(42, 9688) = 11.24, p < 0.01$). For example, SF is the best algorithm in terms of the Hit Rate with a 3 seconds window, but it is surpassed by the Laplacian and OLDA algorithms when using a shorter window of 0.5, as the Friedman test confirms ($F(6) = 200.13, p < 0.01$) for the ranking of the Hit Rate with 3 seconds, which is different than the one for 0.5 seconds ($F(6) = 210.67, p < 0.01$). Therefore, we can state that, amongst these algorithms, SF is ideal if precise boundary localization is not necessary (HR_3), whereas Laplacian outperforms other methods when this localization has to be accurate (HR_0.5).

In terms of structural algorithms, two metrics (Pairwise Frame Clustering and Normalized Conditional Entropies) are depicted in Figure 4b. A similar behavior occurs here, where algorithms will be ranked differently depending on the metric of choice (Friedman test for the structural algorithms evaluated using the PWF yields $F(4) = 230.11, p < 0.01$ and different ranking than the one for NCE, which results in $F(4) = 215.12, p < 0.01$). Interestingly, all algorithms except Laplacian tend to yield better results when using the NCE scores. Given these results, it would be hard to firmly conclude what the best structural algorithm is for this dataset, since 2D-FMC outperforms Laplacian when evaluated using the NCE scores, which is the opposite behavior when using the PWF.

4.4 Datasets

In Figure 5, the average scores for all boundary algorithms in MSAF on different datasets are depicted. If a dataset contains more than one annotation per track, the first annotator in their JAMS files is used. As expected, different results are obtained depending on the dataset, as confirmed by the two-way ANOVA on the evaluation metric with dataset and algorithm as factors (dataset effect: $F(5, 16604) = 512.18, p < 0.01$). From the plot it can also be seen that some algorithms perform better than others depending on the dataset, which might indicate that they are tuned to solve this problem for a specific type of music. Overall, some datasets seem generally more challenging than others, the SPAM dataset being the one that obtains the worst results, which aligns with the method used to collect their data explained in Section 3.

In terms of the structural algorithms (Figure 6), the two-way ANOVA identifies significant variation, with a relevant effect on the dataset of $F(6, 11875) = 133.16, p < 0.01$. Contrasting with the boundary results, the scores for the SPAM dataset are, on average, one of the highest in terms of structural grouping. This, by itself, warrants discussion, since this dataset was chosen to be particularly challenging from a machine point of view, but only when taking the boundary detection subproblem into account. What these results suggest is that, (i) given the human reference boundaries (which are supposed to be difficult to detect), the structural algorithms perform well at clustering the predefined segments, and/or (ii) we might need a better evaluation metric for the structural subproblem.

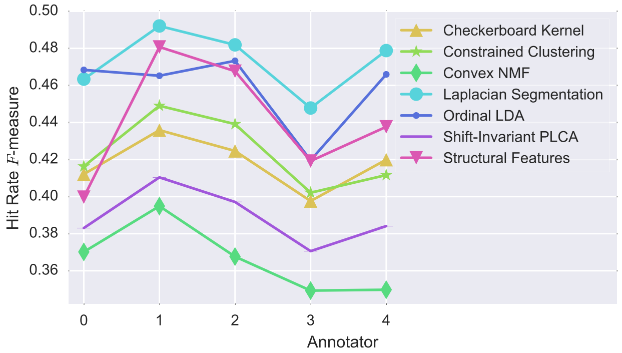


Figure 7: Scores of the boundary algorithms for each human reference in the SPAM dataset.

4.5 Human References

The last experiment focuses on analyzing the amount of variation of the MSAF algorithms depending on the annotator used. For this purpose, the five annotations per track of the SPAM dataset become particularly helpful. Starting with the boundaries, we can see in Figure 7 how variable the scores become when using different annotators for the same exact set of audio files. The two-way ANOVA of HR_3 with annotators and algorithms as factors validates this by reporting a significant annotator effect ($F(4, 1705) = 4.05, p < 0.01$). This suggests that subjectivity plays an important role for this subtask, and more than one set of boundaries would be actually valid from a human perspective. Therefore, the idea of a single "ground-truth" for boundary detection can potentially be misleading. Given this amount of variation depending on the annotator, it is interesting to see that the ranking also changes, making it difficult to compare algorithm behaviors. Even though the Laplacian algorithm performs the best for the majority of annotators, it is ranked as second when using annotator 0 by the Friedman test ($F(5) = 21.24, p < 0.01$), while it is ranked as first for the rest of annotators. These results suggest that, given the subjectivity effect in this task, it is indeed important to collect as many references as possible in order to better assess boundary algorithms.

Lastly, the results of the structural algorithms contrast with the previously discussed ones. In this case, there is little dependency on the human reference chosen, as there is no significant effect for the annotator factor in the two-way ANOVA ($F(4, 1225) = 1.08, p = 0.37$), without significant interaction ($F(16, 1225) = 0.93, p = 0.53$). This advocates that the structural grouping subproblem, when applied to a dataset where the grouping is not particularly challenging (as depicted in Figure 6), is not as affected by subjectivity as the boundary detection one, even though further analysis with larger and more challenging datasets —and perhaps with automatically estimated boundaries— should be performed in order to confirm this.

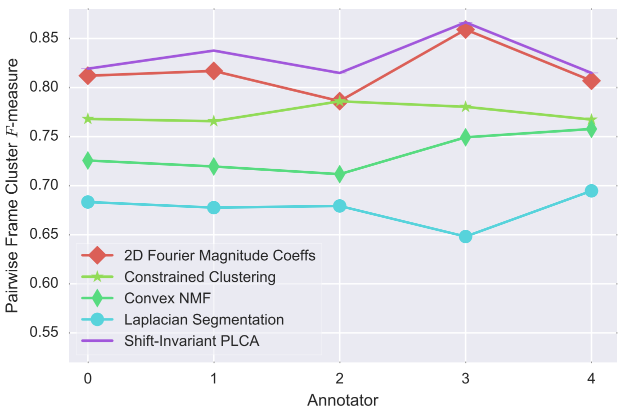


Figure 8: Scores of the structural algorithms for each human reference in the SPAM dataset.

5. CONCLUSIONS

We have presented an open-source framework that facilitates the task of analyzing, assessing, and comparing multiple implementations of structural segmentation algorithms and have employed it to compile a new poly-annotated dataset and to systematically explore the different moving parts of this MIR task. These experiments show that the relative rankings between algorithms tend to change depending on these parts, making it difficult to choose the "best" computational approach. Results also illustrate the problem of ambiguity in this task, and it is our hope that the new SPAM dataset will help researchers to further address this problem. In the future, we wish not only to include more algorithms in this open framework, but to have access to similar frameworks to encourage research on other areas of MIR.

6. REFERENCES

[1] Thierry Bertin-Mahieux, Daniel P. W. Ellis, Brian Whitman, and Paul Lamere. The Million Song Dataset. In *Proc of the 12th International Society of Music Information Retrieval*, pages 591–596, Miami, FL, USA, 2011.

[2] Dmitry Bogdanov, Nicolas Wack, Emilia Gómez, Sankalp Gulati, Perfecto Herrera, Oscar Mayor, Gerard Roma, Justin Salamon, José Zapata, and Xavier Serra. Essentia: An Audio Analysis Library for Music Information Retrieval. In *Proc. of the 14th International Society for Music Information Retrieval Conference*, pages 493–498, Curitiba, Brazil, 2013.

[3] Michael J. Bruderer. *Perception and Modeling of Segment Boundaries in Popular Music*. PhD thesis, Universiteitsdrukkerij Technische Universiteit Eindhoven, 2008.

[4] Daniel P. W. Ellis. Beat Tracking by Dynamic Programming. *Journal of New Music Research*, 36(1):51–60, 2007.

[5] Daniel P. W. Ellis and Graham E. Poliner. Identifying 'Cover Songs' with Chroma Features and Dynamic Programming Beat Tracking. In *Proc. of the 32nd IEEE International Conference on Acoustics Speech and Signal Processing*, pages 1429–1432, Honolulu, HI, USA, 2007.

[6] Jonathan Foote. Automatic Audio Segmentation Using a Measure Of Audio Novelty. In *Proc. of the IEEE International Conference of Multimedia and Expo*, pages 452–455, New York City, NY, USA, 2000.

[7] Christopher Harte, Mark Sandler, and Martin Gasser. Detecting Harmonic Change in Musical Audio. In *Proceedings of the 1st ACM Workshop on Audio and Music Computing Multimedia*, pages 21–26, Santa Barbara, CA, USA, 2006. ACM Press.

[8] André Holzapfel, Matthew E. P. Davies, José R. Zapata, J. Lobato Oliveira, and Fabien Gouyon. Selective Sampling for Beat Tracking Evaluation. *IEEE Transactions On Audio, Speech, And Language Processing*, 20(9):2539–2548, 2012.

[9] Eric J. Humphrey, Justin Salamon, Oriol Nieto, Jon Forsyth, Rachel M. Bittner, and Juan P. Bello. JAMS: A JSON Annotated Music Specification for Reproducible MIR Research. In *Proc. of the 15th International Society for Music Information Retrieval Conference*, pages 591–596, Taipei, Taiwan, 2014.

[10] Mark Levy and Mark Sandler. Structural Segmentation of Musical Audio by Constrained Clustering. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(2):318–326, feb 2008.

[11] Hanna Lukashevich. Towards Quantitative Measures of Evaluating Song Segmentation. In *Proc. of the 10th International Society of Music Information Retrieval*, pages 375–380, Philadelphia, PA, USA, 2008.

[12] Brian McFee and Daniel P. W. Ellis. Analyzing Song Structure with Spectral Clustering. In *Proc. of the 15th International Society for Music Information Retrieval Conference*, pages 405–410, Taipei, Taiwan, 2014.

[13] Brian McFee and Daniel P. W. Ellis. Learnign to Segment Songs With Ordinal Linear Discriminant Analysis. In *Proc. of the 39th IEEE International Conference on Acoustics Speech and Signal Processing*, pages 5197–5201, Florence, Italy, 2014.

[14] Brian Mcfee, Colin Raffel, Dawen Liang, Daniel P. W. Ellis, Matt Mcvicar, Eric Battenberg, and Oriol Nieto. librosa: Audio and Music Signal Analysis in Python. In *Proc. of the 14th Python in Science Conference*, pages 1–7, Austin, TX, USA, 2015.

[15] Oriol Nieto and Juan Pablo Bello. Music Segment Similarity Using 2D-Fourier Magnitude Coefficients. In *Proc. of the 39th IEEE International Conference on Acoustics Speech and Signal Processing*, pages 664–668, Florence, Italy, 2014.

[16] Oriol Nieto and Morwaread M. Farbood. Identifying Polyphonic Patterns From Audio Recordings Using Music Segmentation Techniques. In *Proc. of the 15th International Society for Music Information Retrieval Conference*, pages 411–416, Taipei, Taiwan, 2014.

[17] Oriol Nieto, Morwaread M. Farbood, Tristan Jehan, and Juan Pablo Bello. Perceptual Analysis of the F-measure for Evaluating Section Boundaries in Music. In *Proc. of the 15th International Society for Music Information Retrieval Conference*, pages 265–270, Taipei, Taiwan, 2014.

[18] Oriol Nieto and Tristan Jehan. Convex Non-Negative Matrix Factorization For Automatic Music Structure Identification. In *Proc. of the 38th IEEE International Conference on Acoustics Speech and Signal Processing*, pages 236–240, Vancouver, Canada, 2013.

[19] Jouni Paulus, Meinard Müller, and Anssi Klapuri. Audio-Based Music Structure Analysis. In *Proc of the 11th International Society of Music Information Retrieval*, pages 625–636, Utrecht, Netherlands, 2010.

[20] Colin Raffel, Brian Mcfee, Eric J. Humphrey, Justin Salamon, Oriol Nieto, Dawen Liang, and Daniel P. W. Ellis. mir_eval: A Transparent Implementation of Common MIR Metrics. In *Proc. of the 15th International Society for Music Information Retrieval Conference*, pages 367–372, Taipei, Taiwan, 2014.

[21] Joan Serrà, Meinard Müller, Peter Grosche, and Josep Lluís Arcos. Unsupervised Music Structure Annotation by Time Series Structure Features and Segment Similarity. *IEEE Transactions on Multimedia, Special Issue on Music Data Mining*, 16(5):1229 – 1240, 2014.

[22] Jordan B. Smith, J. Ashley Burgoyne, Ichiro Fujinaga, David De Roure, and J. Stephen Downie. Design and Creation of a Large-Scale Database of Structural Annotations. In *Proc. of the 12th International Society of Music Information Retrieval*, pages 555–560, Miami, FL, USA, 2011.

[23] Jordan B. L. Smith and Elaine Chew. A Meta-Analysis of the MIREX Structure Segmentation Task. In *Proc. of the 14th International Society for Music Information Retrieval Conference*, Curitiba, Brazil, 2013.

[24] Douglas Turnbull, Gert Lanckriet, Elias Pampalk, and Masataka Goto. A Supervised Approach for Detecting Boundaries in Music Using Difference Features and Boosting. In *Proc. of the 5th International Society of Music Information Retrieval*, pages 42–49, Vienna, Austria, 2007.

[25] George Tzanetakis and Perry Cook. MARSYAS: a framework for audio analysis. *Organised Sound*, 4(3):169–175, 2000.

[26] Ron Weiss and Juan Pablo Bello. Unsupervised Discovery of Temporal Structure in Music. *IEEE Journal of Selected Topics in Signal Processing*, 5(6):1240–1251, 2011.