THE HARMONIX SET: BEATS, DOWNBEATS, AND FUNCTIONAL SEGMENT ANNOTATIONS OF WESTERN POPULAR MUSIC

Oriol Nieto[1] **Matthew McCallum**[1] **Matthew E. P. Davies**[2]
Andrew Robertson[3] **Adam Stark**[4] **Eran Egozy**[5]
[1] Pandora Media, Inc., Oakland, CA, USA
[2] INESC TEC, Porto, Portugal
[3] Ableton AG, Berlin, Germany
[4] MI·MU, London, UK
[5] MIT, Cambridge, MA, USA
onieto@pandora.com

ABSTRACT

We introduce the Harmonix set: a collection of annotations of beats, downbeats, and functional segmentation for over 700 full tracks that covers a wide range of western popular music. Given the variety of annotated music information types in this set, and how strongly these three types of data are typically intertwined, we seek to foster research that focuses on multiple retrieval tasks at once. The dataset includes additional metadata such as MusicBrainz identifiers to support the linking of the dataset to third-party information or audio data when available. We describe the methodology employed in acquiring this set, including the annotation process and song selection. In addition, an initial data exploration of the annotations and actual dataset content is conducted. Finally, we provide a series of baselines of the Harmonix set with reference beat-trackers, downbeat estimation, and structural segmentation algorithms.

1. INTRODUCTION

The tasks of beat detection [8], downbeat estimation [2], and structural segmentation [34] constitute a fundamental part of the field of MIR. These three musical characteristics are often related: downbeats define the first beat of a given music measure, and long structural music segments tend to begin and end on specific beat locations – frequently on downbeats [10]. The automatic estimation of such information could result in better musical systems such as more accurate automatic DJ-ing, better intra- and inter-song navigation, further musicological insights of large collections, *etc*. While a few approaches exploiting more than one of these musical traits have been proposed [2, 11, 25], the amount of human annotated data containing the three of them for a single collection is scarce. This limits the training potential of certain methods, especially those that require large amounts of information (e.g., deep learning [18]).

In this paper we present the Harmonix set: human annotations of beats, downbeats, and functional segmentation for over 700 tracks of western popular music. These annotations were gathered with the aim of having a significant amount of data to train models to improve the prediction of such musical attributes, which would later be applied to various products offered by Harmonix, a videogame company that specializes in musically-inspired games. By releasing this set to the public, our aim is to let the research community explore and exploit these annotations to advance the tasks of beat tracking, downbeat estimation, and automatic functional structural segmentation. We discuss the methodology to acquire these data, including the song selection process, and the inclusion of standard identifiers (AcoustID and MusicBrainz) and a set of automatically extracted onset times for the first 30 seconds of the tracks to allow other researchers to more easily access and align, when needed, the actual audio content. Furthermore, we present a series of results with reference algorithmic approaches in the literature with the goal of having an initial public benchmark of this set.

The rest of this work is organized as follows: Section 2 contains a review of the most relevant public datasets of the tasks at hand; Section 3 discusses the Harmonix set, including the data gathering, their formatting, and various statistics; Section 4 presents numerous benchmarks in the set; and Section 5 draws some final conclusions and discusses future work.

2. BACKGROUND

Several datasets with beat, downbeat, and/or segment annotations have been previously published, and in this section we review the most relevant ones.

2.1 Beat and Downbeat Tracking Sets

Over the last 15 years, many annotated datasets for beat and downbeat tracking have appeared in the literature whose primary purpose has been to allow the comparison of newly proposed and existing algorithms. However, the well-known difficulties of sharing the audio component of large annotated datasets has led to a rather ad-hoc usage of different datasets within the literature, and to a lesser extent, the choice of which evaluation metrics are selected to report accuracy. Conversely, the MIREX evaluation campaign provides a more rigid model for evaluation, by withholding access to private test datasets, and instead relying on the submission of the competing algorithms in order to compare them under controlled conditions. To this end, MIREX can be a useful reference point to consider these two music analysis tasks from the perspective of annotated data.

The MIREX Audio Beat Tracking (ABT) task [1] first appeared in 2006 and ran on a single dataset [28,30] with the performance of the submitted algorithms determined using one evaluation metric, the P-Score. After a brief hiatus, the task reappeared in 2009 with the addition of a dataset of Chopin Mazurkas [36], and the inclusion of multiple evaluation metrics [5]. The task continued to run in this way until the incorporation of the SMC dataset [16] in 2012, from which point it has remained constant. In 2014, the Audio Downbeat Estimation (ADE) task [2] was launched which comprised six different datasets from diverse geographic and stylistic sources: The Beatles [24]; Hardcore, Jungle, Drum and Bass (HJDB) [15]; Turkish [41]; Ballroom [21]; Carnatic [42]; and Cretan [17], with the evaluation conducted using the F-measure. While the datasets contained with these two MIREX tasks are by no means exhaustive, they provide a useful window to explore both how the audio data is chosen and how the annotation is conducted for these MIR tasks. To this end, we provide the following breakdown of different properties including reference to both MIREX and non-MIREX datasets.

Duration: Unlike the task of structural segmentation, beat and downbeat tracking datasets can be comprised of musical excerpts [14, 15, 21, 28] rather than full tracks [9,12,13,24]. **Number of annotators**: The initial MIREX beat tracking dataset [28] was unique in that it contained the annotations of 40 different people who tapped the beat to the music excepts. Conversely, other datasets used multiple annotators contributing across the dataset [16], a single annotator for all excerpts [14], or even deriving the annotations in a semi-automatic way from the output of an algorithm [24]. **Annotation post-processing**: Given some raw tap times or algorithm output, these can either be left unaltered [28] or, as is more common, iteratively adjusted until they are considered perceptually accurate by the annotator(s) [14–16]. **Style-specificity**: While some datasets are designed to have broad coverage across a range of musical styles [13, 14, 23], others target a particular group of styles [15, 21], a single style [9], the work of a given artist [12, 24] or even multiple versions of the same pieces [36]. **Western / Non-Western**: Similarly, the make up of the dataset can target underrepresented non-western music [33,41,42]. **Perceived difficulty**: Finally, the choice of musical material can be based upon the perceived difficulty of the musical excerpts, either from the perspective of musical or signal level properties [16].

2.2 Structural Segmentation Sets

The task of structural segmentation has been particularly active in the MIR community since the late 2000s. Similarly to the beat tracking task, several datasets have been published, and some of them have evolved over time. This task is often divided into two subtasks: segment boundary retrieval and segment labeling. All well-known published datasets contain both boundary and label information. One of the major challenges with structural segmentation is that this task is regarded as both *ambiguous* (i.e., there may be more than one valid annotation for a given track [26]) and *subjective* (i.e., two different listeners might perceive different sets of segment boundaries [4]). This has led to different methodologies when annotating and gathering structural datasets, thus having a diverse ecosystem of sets to choose from when evaluating automatic approaches.

The first time this task appeared on MIREX was in 2009, [3] where annotations from The Beatles dataset (which also includes beat and downbeat annotations, as previously described) and a subset of the Real World Computing Popular Music Database (RWC) [13] were employed. These sets contain several functional segment annotations for western (The Beatles) and Japanese (RWC) popular music. These segment functions describe the *purpose* of the segments, e.g.: "solo," "verse," "chorus." A single annotation per track is provided for these two sets. The Beatles dataset was further revised at the Tampere University of Technology, [4] and additional functional segment annotations for other bands were added to The Beatles set, which became known as the Isophonics Dataset [24]. No beat or downbeat annotations were provided to the rest of the tracks in Isophonics, and the final number of tracks with functional structural segment annotations is 300. The number of annotated tracks in RWC is 365.

To address the open problems of ambiguity and subjectivity, further annotations per track from several experts could be gathered. That is the case with the Structural Annotations for Large Amounts of Music Information (SALAMI) dataset [39], where most of its nearly 1,400 tracks have been annotated by at least 2 musical experts. Similarly, the Structural Poly Annotations of Music (SPAM) dataset [32] provides 5 different annotations for 50 tracks. These two sets not only contain functional levels of annotations, but also large and small scale segments where only single letters describing the similarity between segments are annotated. Thus, these can be seen as sets that contain *hierarchical* data, which pose significant chal-

[1] https://www.music-ir.org/mirex/wiki/2006:Audio_Beat_Tracking
[2] https://www.music-ir.org/mirex/wiki/2014:Audio_Downbeat_Estimation
[3] https://www.music-ir.org/mirex/wiki/2009:Structural_Segmentation
[4] http://www.cs.tut.fi/sgn/arg/paulus/beatles_sections_TUT.zip

lenges, since ambiguity and subjectivity span across multiple layers [26] and remain largely unexploited in the MIREX competition [7,40]. As opposed to Isophonics and RWC, these two sets contain highly diverse music in terms of genre: from world-music to rock, including jazz, blues, and live music.

The following properties typically define segmentation datasets: **Number of annotators**: This can help when trying to quantify the amount of disagreement among annotators [26, 32], or when developing approaches that may yield more than one potentially valid segmentation. **Hierarchy**: The levels of annotations contained in the set. It typically contains functional, large, and/or small segment annotations. When only one level of annotations is provided, these are typically called *flat* segment annotations.

3. THE HARMONIX SET

In this section we present the Harmonix set, including the methodology of acquiring the data, its motivation, its contents, and a set of statistics of its annotations. The Harmonix set is publicly available on-line. [5]

3.1 Data Gathering

The primary motivation of this work is based on the need to create gameplay data for rhythm-action games (also known as beat matching games). Many such games exist, from early pioneers like Parappa The Rapper and Beatmania, to the rock simulation games Guitar Hero and Rock Band, as well as community-based games like OSU and more recently, VR games like Beat Saber. In most cases the gameplay data (also referred to as beatmaps), consisting of note locations in a song, are hand-authored. In certain games, additional control data may be desirable. For example, in the rock simulation games, where a 3D depiction of a rock concert is rendered, it can be desirable to simulate flashing lights (on the beat) or lighting color palette changes (on section boundaries). Again, these data tend to be hand-authored.

Harmonix's desire was to implement a suite of automatic music analysis tools that estimate certain musical attributes in order to expedite the process of hand-authoring gameplay data, or in some cases, to fully automate the process of creating these data. The songs of the Harmonix set were gathered and hand-annotated to create a ground-truth dataset for training and testing these algorithms.

The mix of genres in this corpus were chosen to be typical of ones used in the rhythm-action games, with a somewhat higher tendency towards EDM and popular songs suitable for dancing (see Figure 1 for the full genre distribution). As such, most tend to have a very stable tempo and a 4/4 time signature. However, we also added a selection of songs that may not be typical of dance or pop music to increase variety. Some of these (Classic Rock, Country, Metal) may have less stable tempo (where drums are played by actual musicians as opposed to drum-machines

[5] https://github.com/urinieto/harmonixset

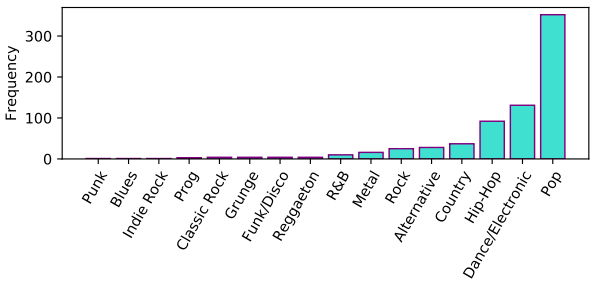

Figure 1. Genre distribution of the Harmonix set.

or DAW-based productions) and may deviate from a strict 4/4 meter.

All songs were annotated by trained professional musicians who regularly work in music production environments. As the project went on, the majority of annotation work fell to only a few individuals who became specialized in this task. Annotations were created in Digital Audio Workstation software (such as Reaper or Logic). First, a MIDI tempo track was established that corresponded to the song audio. Then beats, downbeats, and sections were coded into the MIDI track using note events and text events. MIDI files were then exported and automatically converted to a text-based representation of beats, downbeats, and named section boundaries. Every song was verified once by the original annotator.

3.2 Dataset Contents

The Harmonix set contains manual annotations for 715 western popular music tracks, thus being the largest published dataset to date containing beats, downbeats, and function structural segmentation information. The annotations and some of the song-level metadata are distributed via JAMS [19] files, one per track. This format is chosen given its simplicity when storing multi-task annotations plus song- and corpus-level metadata. Each JAMS file contains the beat, downbeat, and functional segmentation annotations, plus a set of estimated onsets for the first 30 seconds of the audio. These onsets are intended to help aligning the audio in case researchers obtain audio data with different compression formats that might include certain small temporal offsets. This onset information was computed using librosa [27], with their default parameters. [6]

For the sake of transparency and usability, we also publish the raw beats, downbeats, and segmentation data as space-separated text files, two per track: one for beats and downbeats, and the other for segments. We also distribute the code that converts these raw annotations into unified JAMS files. Furthermore, we provide other identifiers with the aim of easily retrieving additional metadata and/or audio content for each song. These identifiers include:

- **MusicBrainz** [7] : open music encyclopedia including

[6] librosa 0.6.3, using Core Audio on macOS 10.13.6.
[7] https://musicbrainz.org/

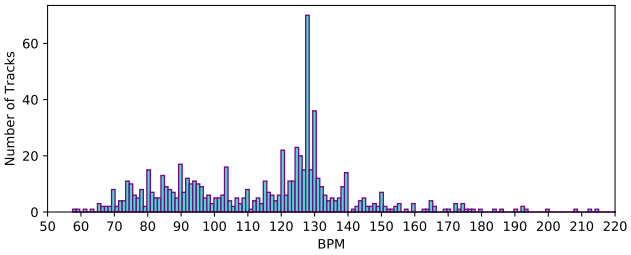

Figure 2. Tempo distribution of the tracks in the Harmonix set.

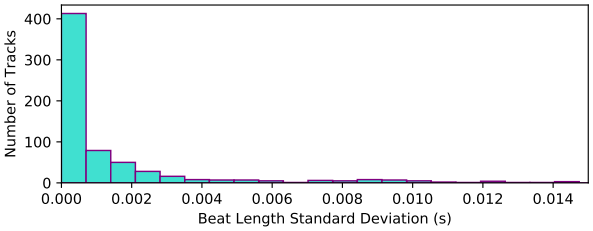

Figure 3. Standard deviation of the tempo distribution in the Harmonix set.

unique identifiers for recordings, releases, artists, etc.

- **AcoustID** [8] : open source fingerprinting service to easily match audio content, typically associated with MusicBrainz identifiers.

Finally, we provide a single CSV file including additional metadata information such as genre and proposed train/test splits. These splits are defined by retaining the genre distribution both in the training and test sets, with a 80% and 20% distributions of the tracks, respectively.

3.3 Data Statistics

In this subsection we provide several data insights obtained from the annotations to give an objective overview of the set. In Figure 2 we show the estimated tempo distribution in beats-per-minute (BPM) per track. These estimations were computed using the track-level median inter-beat-interval (IBI) for each of the annotated beats in a given track. There is a clear peak at 128 BPM, which could be explained by being the most common tempo in electronic dance music [29]. Furthermore, in Figure 3 we plot the standard deviation of the IBI. We can clearly see that the tempo is remarkably steady in this dataset, which is expected given the type of musical genres it spans.

In terms of segment statistics, we show data based on certain attributes described in a MIREX meta-analysis of the segmentation task [40]. In Figure 4 we plot track-level histograms for the number of segments, and the number of unique segments (i.e., those with the same associated label). Both distributions seem to be unimodal and centered around 10 and 11 for the number of segments per

[8] https://acoustid.org/

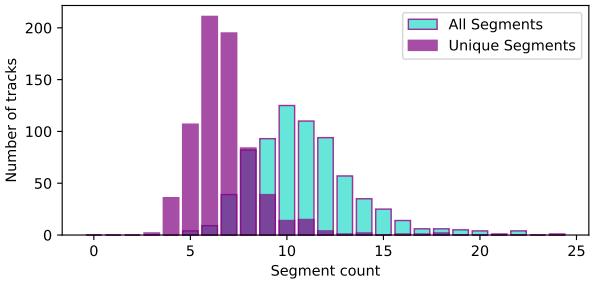

Figure 4. Number of segments per track, based on their segment labels.

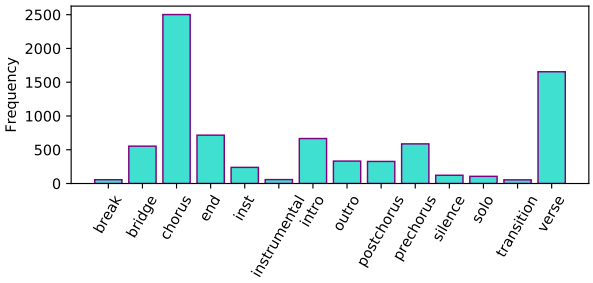

Figure 5. Most common segment labels.

tracks, and around 6 and 7 for the number of unique labels per track. This differs from the number of unique segments in The Beatles dataset, which is centered around 4 per track [31].

Figure 5 shows the frequency in which the most common segment labels appear in the set. The labels "chorus" and "verse" dominate the distribution, as these functional parts are common in western popular music. The plot also shows potentially repeated labels like "inst" and "instrumental." A further inter-song analysis of the labels might be necessary to potentially merge certain labels and thus unify the vocabulary of the set.

We plot in Figure 6 the distribution of the segment lengths, in seconds, across the entire dataset. As we showed in Figure 2, there is a majority of tracks at 128 BPM, for which a duration of 15 seconds would correspond to a segment of exactly 32 beats. This, in the common 4/4 time signature, would result in 8 bars per each 15-second segment in that tempo, and 8 bars are common in electronic dance music [29].

Finally, and thanks to having access to the annotated downbeats, we show in Figure 7 the number of segments starting at a specific beat within a given bar. We can see that the vast majority of segments (81.1%) start in a downbeat. Interestingly, several segments (10%) start in position 4, thus showing that 1-beat count-ins are more common than other types of count-ins on this dataset (a popular example of a 1-beat count-in song is Hey Jude by The Beatles, where the (1) is on the Jude and Hey is the (4) of the previous bar).

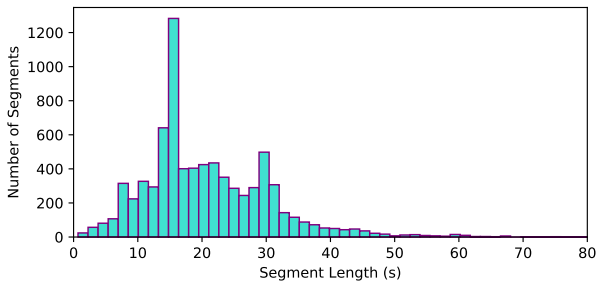

Figure 6. Segment length distribution.

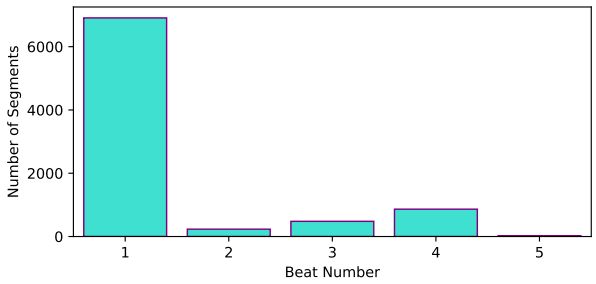

Figure 7. Number of segments based on their starting beat position within a bar.

4. RESULTS

4.1 Beat Results

In order to establish performance baselines over the dataset for the task of beat-tracking, we have evaluated a number of openly available beat tracking algorithms on the dataset [3,8,20,22]. Each of these algorithms can be found in either the madmom [1] [9] or librosa python libraries. This allows a comparison between datasets for which these algorithms have been previously evaluated, with respect to their affect on these algorithm's performance. The results are also provided with the dataset in CSV format. This is intendended as a convenience for any future work that wishes to evaluate novel algorithms against these benchmarks.

The beat tracking results for the aforementioned algorithms are displayed in Figure 8. There they are evaluated across two metrics, F-Measure, and Max F-Measure, where the latter refers to the maximum F-Measure obtained per track when evaluated across double and half-time metrical variations in the annotated beats provided with this dataset. In all experiments a tolerance window of ± 70 ms was employed in order to compute the F-Measure. For half-time metrical variations, both the downbeat and upbeat alignments were tested for a maximum F-Measure value. While [8] is the most computationally efficient of the algorithms, we see clear gains in the more recently developed methods. When investigating the types of errors present in the beat position estimates from [8], it was found the most common error was the alignment of beat

[9] We used madmom 0.16.1.

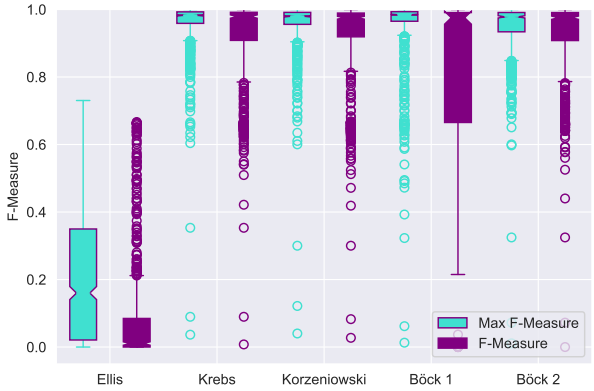

Figure 8. Beat tracking performance over the Harmonix set, for the algorithms Ellis [8], Krebs [22], Korzeniowski [20], Böck 1 - the "BeatDetector" technique from [3], and Böck 2 - the "BeatTracker" technique from [3].

phase. Often beat positions landed on the half beat or quarter beat, resulting in an F-Measure of 0 when this misalignment is consistent throughout the track. When comparing F-Measure and Max F-Measure metrics, it can be seen that with this dataset both [8] and the "BeatDetector" algorithm from [3] have a significant number of double-half time errors, compared to the other algorithms evaluated. Unlike the "BeatTracker" algorithm in [3], the "BeatDetector" algorithm assumes constant tempo.

4.2 Downbeat Results

Unfortunately the availability of open source downbeat estimation libraries is limited. In order to provide a baseline for downbeat detection performance with the Harmonix set specifically, results have been evaluated with the downbeat detection algorithms available in [1] in addition to Durand's algorithm [6] [10], making three algorithms in total. The algorithms from the Madmom python package [1] include the method proposed in [2] using the annotated beat positions as input, and the dynamic Bayesian bar tracking processor using the input from the RNN bar processor activation function. The results can be seen in Figure 9 in terms of F-Measure with a tolerance window of ± 70 ms. The superior performance of [2], which has oracle annotated beat information, highlights the importance of reliable beat tracking for downbeat estimation performance, and the interdependence between the beat tracking and downbeat estimation tasks.

4.3 Segmentation Results

There are several open source structural segmentation algorithms available in the Music Structure Analysis Framework (MSAF) [32]. [11] We run the best performing ones on the Harmonix set: (i) Structural Features [38] to identify boundaries, and (ii) 2D-Fourier Magnitude Coefficients

[10] Not open source, shared via private correspondence.
[11] MSAF version dev-0.1.8.

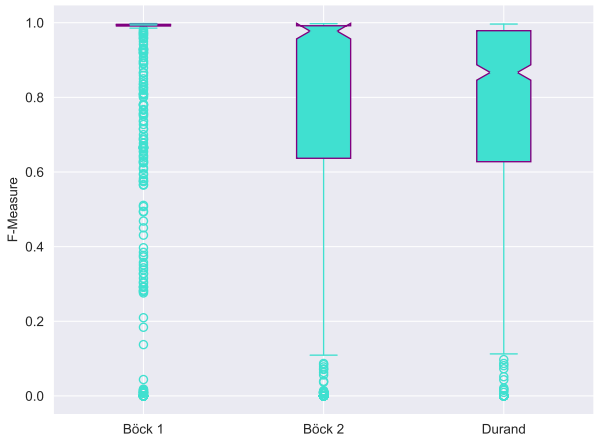

Figure 9. Downbeat tracking performance over the Harmonix set, for the algorithms BöckD 1 [2] and BöckD 2 - a dynamic Bayesian network provided within the Madmom package [1], and Durand [6].

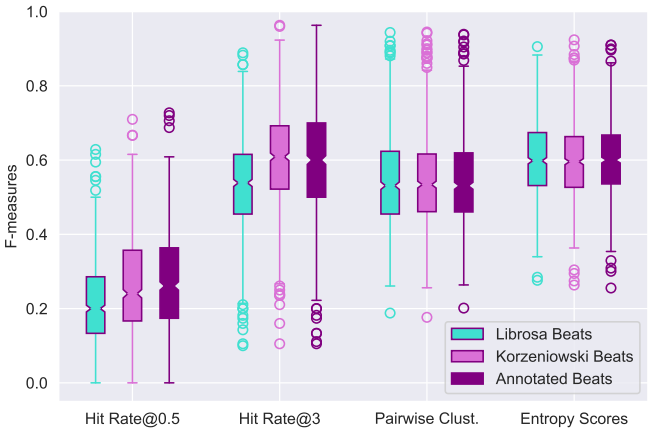

Figure 10. Segmentation results over the Harmonix set, using Structural Features for boundaries, 2D-FMC for the labeling process, and three types of beat information.

(2D-FMC) [31] to label the segments based on their acoustic similarity. Constant-Q Transforms [37] are the selected audio features given their ability to capture both timbral and harmonic content, and the default parameters in MSAF are the ones employed when computing these results. We use `mir_eval` [35] to evaluate these algorithms, and report the F-measures for the most common metrics: Hit Rate with 0.5 and 3 second windows for boundary retrieval, and Pairwise Frame Clustering and Entropy Scores for the labeling process. These algorithms can use beat-synchronized features, and we ran each algorithm three times, depending on the following beat information: (i) Librosa's estimations, (ii) Korzeniowski's estimations, and (iii) annotations from the Harmonix set. Thus, we are able to assess the segmentation results when employing the worst and best performing beat trackers from our previous study, plus those computed using human annotated beats. Song-level results for these three different runs are available as CSV files in the dataset repository disclosed above.

In Figure 10 all segmentation results are shown. The results in turquoise boxplots (on the left side) display the metrics of the algorithms when running on Librosa's estimated beat-synchronized features, those in light pink (in the middle) correspond to the results computed with Korzeniowski's beats, while the purple boxplots (on the right) show those using annotated beats instead. Given how related boundary retrieval is with respect to precise beat placement, it is not unexpected to see an improvement in the boundary metrics (Hit Rates) when using more accurate beat data. The boxplots further show that the smaller the time window used in the Hit Rate metrics the more accurate the beat information should ideally be. In other words, Korzeniowki's beats yield very similar results than those from human annotations when using a 3 second window, but there is clearly room for enhancement (in terms of beat tracking) when using 0.5 second windows, where the segmentation results using human annotated beats outperform any of the others that employ estimated ones. On the other hand, it is worth noting that the label results do not seem to depend as much on the quality of the beats in order to produce their outcomes, as the three different runs yield similar results for the Pairwise Frame Clustering and Entropy Scores metrics. As mentioned in Section 2.2, structural segmentation is a challenging task especially due to ambiguity, subjectivity, and hierarchy, and this is reflected in the overall results, which exhibit notable room for improvement.

5. CONCLUSIONS

We presented the Harmonix set, the largest dataset in terms of human annotations containing the following three types of music information: beats, downbeats, and function structural segments. This set contains mostly western popular music, with strong emphasis on Pop, EDM, and Hip-Hop. We provide metadata in terms of genre, song title, and artist information along with MusicBrainz and AcoustID identifiers plus predicted onset information to allow easier matching and alignment with audio data. We discussed a set of results using current algorithms in the literature in terms of beat tracking, downbeat estimation, and structural segmentation to disclose an initial public benchmark of the set. Given the rather large nature of the set and the three different types of music information contained in it, it is our hope that researchers employ these data not only to further advance one of these three MIR tasks individually, but also to potentially combine them to yield superior approaches in the near future.

6. ACKNOWLEDGMENTS

We would like to thank Simon Durand for sharing his downbeat estimation implementation. Matthew E.P. Davies is supported by Portuguese National Funds through the FCT-Foundation for Science and Technology, I.P., under the project IF/01566/2015.

7. REFERENCES

[1] S. Böck, F. Korzeniowski, J. Schlüter, F. Krebs, and G. Widmer. Madmom: A new python audio and music signal processing library. In *Proceedings of the 24th ACM international conference on Multimedia*, pages 1174–1178, 2016.

[2] S. Böck, F. Krebs, and G. Widmer. Joint Beat and Downbeat tracking with recurrent neural networks. *Proceedings of the International Society for Music Information Retrieval (ISMIR) Conference*, pages 255–261, 2016.

[3] S. Böck and M. Schedl. Enhanced beat tracking with context-aware neural networks. In *Proc. Int. Conf. Digital Audio Effects*, pages 135–139, 2011.

[4] M. J. Bruderer, M. F. McKinney, and A. Kohlrausch. The Perception of Structural Boundaries in Melody Lines of Western Popular Music. *Musicæ Scientiæ*, 13(2):273–313, 2009.

[5] M. E. P. Davies, N. Degara, and M. D. Plumbley. Evaluation methods for musical audio beat tracking algorithms. Technical Report C4DM-TR-09-06, Centre for Digital Music, Queen Mary University of London, 2009.

[6] S. Durand, J. P. Bello, B. David, and G. Richard. Feature adapted convolutional neural networks for downbeat tracking. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 296–300. IEEE, 2016.

[7] A. F. Ehmann, M. Bay, J. S. Downie, I. Fujinaga, and D. D. Roure. Music Structure Segmentation Algorithm Evaluation: Expanding on MIREX 2010 Analyses and Datasets. In *Proc. of the 13th International Society for Music Information Retrieval Conference*, pages 561–566, Miami, FL, USA, 2011.

[8] D. P. Ellis. Beat tracking by dynamic programming. *Journal of New Music Research*, 36(1):51–60, 2007.

[9] V. Eremenko, E. Demirel, B. Bozkurt, and X. Serra. Audio-aligned jazz harmony dataset for automatic chord transcription and corpus-based research. In *Proc. of the 16th Intl. Society for Music Information Retrieval Conf. (ISMIR)*, pages 483–490, 2018.

[10] J. T. Foote. Methods for the automatic analysis of music and audio. *FXPAL Technical Report FXPAL-TR-99-038*, 1999.

[11] M. Fuentes, B. McFee, H. C. Crayencour, S. Essid, and J. P. Bello. A Music Structure Informed Downbeat Tracking System Using Skip-Chain Conditional Random Fields and Deep Learning. In *Proc. of the 44th International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Brighton, UK, 2019.

[12] B. D. Giorgi, M. Zanoni, S. Böck, and A. Sarti. Multipath beat tracking. *Journal of the Audio Engineering Society*, 64(7/8):493–502, 2016.

[13] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka. RWC Music Database: Popular, Classical, and Jazz Music Databases. *International Conference on Music Information Retrieval*, (October):287–288, 2002.

[14] S. Hainsworth and M. Macleod. Particle filtering applied to musical tempo tracking. *EURASIP Journal on Applied Signal Processing*, 15:2385–2395, 2004.

[15] J. Hockman, M. E. P. Davies, and I. Fujinaga. One in the jungle: Downbeat detection in hardcore, jungle, and drum and bass. In *Proc. of the 13th Intl. Society for Music Information Retrieval Conf. (ISMIR)*, pages 169–174, 2012.

[16] A. Holzapfel, M. E. P. Davies, J. R. Zapata, J. L. Oliveira, and F. Gouyon. Selective sampling for beat tracking evaluation. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(9):2539–2548, 2012.

[17] A. Holzapfel, F. Krebs, and A. Srinivasamurthy. Tracking the "odd": Meter inference in a culturally diverse music corpus. In *Proc. of the 15th Intl. Society for Music Information Retrieval Conf. (ISMIR)*, pages 425–430, 2014.

[18] E. J. Humphrey, J. P. Bello, and Y. LeCun. Moving Beyond Feature Design: Deep Architecture and Automatic Feature Learning in Music Informatics. In *Proc. of the 13th International Society for Music Information Retrieval Conference*, pages 403–408, Porto, Portugal, 2012.

[19] E. J. Humphrey, J. Salamon, O. Nieto, J. Forsyth, R. M. Bittner, and J. 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.

[20] F. Korzeniowski, S. Böck, and G. Widmer. Probabilistic extraction of beat positions from a beat activation function. In *ISMIR*, pages 513–518, 2014.

[21] F. Krebs, S. Böck, and G. Widmer. Rhythmic pattern modeling for beat and downbeat tracking in musical audio. In *Proc. of the 14th Intl. Society for Music Information Retrieval Conf. (ISMIR)*, pages 227–232, 2013.

[22] F. Krebs, S. Böck, and G. Widmer. An efficient state-space model for joint tempo and meter tracking. In *ISMIR*, pages 72–78, 2015.

[23] U. Marchand and G. Peeters. Swing ratio estimation. In *Proc. of the 18th Intl. Conf. on Digital Audio Effects (DAFx)*, pages 423–428, 2015.

[24] M. Mauch, C. Cannam, M. Davies, S. Dixon, C. Harte, S. Kolozali, D. Tidhar, and M. Sandler. OMRAS2 Metadata Project 2009. In *Late Breaking Session of the 10th International Society of Music Information Retrieval*, Kobe, Japan, 2009.

[25] M. C. McCallum. Unsupervised Learning of Deep Features for Music Segmentation. In *Proc. of the 44th International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Brighton, UK, 2019.

[26] B. McFee, O. Nieto, M. M. Farbood, and J. P. Bello. Evaluating hierarchical structure in music annotations. *Frontiers in Psychology*, 8(1337), 2017.

[27] B. McFee, C. Raffel, D. Liang, D. P. W. Ellis, M. McVicar, E. Battenberg, and O. Nieto. librosa: Audio and Music Signal Analysis in Python. In *Proc. of the 14th Python in Science Conference*, pages 18–25, Austin, TX, USA, 2015.

[28] M. F. McKinney, D. Moelants, M. E. P. Davies, and A. Klapuri. Evaluation of audio beat tracking and music tempo extraction algorithms. *Journal of New Music Research*, 36(1):1–16, 2007.

[29] D. Moelants. Hype vs. Natural Tempo: a Long-term Study of Dance Music Tempi. In *Proc. of the 10th International Conference on Music Perception and Cognition*, Sapporo, Japan, 2008.

[30] D. Moelants and M. McKinney. Tempo perception and musical content: What makes a piece fast, slow or temporally ambiguous. In *Proc. of the 8th Intl. Conf. on Music Perception and Cognition*, pages 558–562, 2004.

[31] O. Nieto and J. P. 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.

[32] O. Nieto and J. P. Bello. Systematic Exploration of Computational Music Structure Research. In *Proc. of the 17th International Society for Music Information Retrieval Conference*, pages 547–553, New York City, NY, USA, 2016.

[33] L. O. Nunes, M. Rocamora, L. Jure, and L. W. Biscainho. Beat and Downbeat Tracking Based on Rhythmic Patterns Applied to the Uruguayan Candombe Drumming. In *Proc. of the 16th Intl. Society for Music Information Retrieval Conf. (ISMIR)*, pages 264–270, 2015.

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

[35] C. Raffel, B. Mcfee, E. J. Humphrey, J. Salamon, O. Nieto, D. Liang, and D. 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.

[36] C. Sapp. Comparative Analysis of Multiple Musical Performances. In *Proc. of the 8th Intl. Conf. on Music Information Retrieval, (ISMIR)*, pages 497–500, 2007.

[37] C. Schörkhuber and A. Klapuri. Constant-Q Transform Toolbox for Music Processing. In *Proc. of the 7th Sound and Music Computing Conference*, pages 56–64, Barcelona, Spain, 2010.

[38] J. Serrà, M. Müller, P. Grosche, and J. L. 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.

[39] J. B. Smith, J. A. Burgoyne, I. Fujinaga, D. De Roure, and J. S. 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.

[40] J. B. L. Smith and E. Chew. A meta-analysis of the MIREX Structure Segmentation task. In *Proceedings of the International Society for Music Information Retrieval Conference*, pages 251–256, Curitiba, Brazil, 2013.

[41] A. Srinivasamurthy, A. Holzapfel, and X. Serra. In search of automatic rhythm analysis methods for turkish and indian art music. *Journal of New Music Research*, 43(1):94–114, 2014.

[42] A. Srinivasamurthy and X. Serra. A Supervised Approach to Hierarchical Metrical Cycle Tracking from Audio Music Recordings. In *Proc. of the 39th IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5237–5241, 2014.