TWO DATA SETS FOR TEMPO ESTIMATION AND KEY DETECTION IN ELECTRONIC DANCE MUSIC ANNOTATED FROM USER CORRECTIONS

Peter Knees,[1] **Ángel Faraldo,**[2] **Perfecto Herrera,**[2] **Richard Vogl,**[1]
Sebastian Böck,[1] **Florian Hörschläger,**[1] **Mickael Le Goff**[3]
[1] Department of Computational Perception, Johannes Kepler University, Linz, Austria
[2] Music Technology Group, Universitat Pompeu Fabra, Barcelona, Spain
[3] Native Instruments GmbH, Berlin, Germany
peter.knees@jku.at

ABSTRACT

We present two new data sets for automatic evaluation of tempo estimation and key detection algorithms. In contrast to existing collections, both released data sets focus on electronic dance music (EDM). The data sets have been automatically created from user feedback and annotations extracted from web sources. More precisely, we utilize user corrections submitted to an online forum to report wrong tempo and key annotations on the *Beatport* website. *Beatport* is a digital record store targeted at DJs and focusing on EDM genres. For all annotated tracks in the data sets, samples of at least one-minute-length can be freely downloaded. For key detection, further ground truth is extracted from expert annotations manually assigned to *Beatport* tracks for benchmarking purposes. The set for tempo estimation comprises 664 tracks and the set for key detection 604 tracks. We detail the creation process of both data sets and perform extensive benchmarks using state-of-the-art algorithms from both academic research and commercial products.

1. INTRODUCTION

Electronic dance music (EDM) is one of the most important and influential music genres of our time. The genre has been defined as a broad category of popular music that, since the end of the 1990s, encompasses styles such as techno, house, trance, and dubstep, and, uniquely, utilizes electronic instruments such as synthesizers, drum machines, sequencers, and samplers. Traditionally, technologically-mediated live performances form an integral part of EDM [6, 8].

Historically, EDM evolved from and links genres from the 1950s to the 1980s such as soul, funk, disco, rap, and techno. After two decades of isolation as a genre, today, we are witnessing how it not only influences its legitimate forerunner genres, but also most generic and formulaic pop forms, including contemporary rock, r&b and rap music. In fact, given its spread over millions of followers, EDM is a central element in the 21st century's popular music — and therefore a major economical factor in the entertainment industry. [1] [2] [3]

Despite its popularity, in terms of musical sophistication, the reputation of EDM might not be the best: "simplistic," "too repetitive," "feasible with lack of talent," "fake music," or "button-pushing" are some of the criticisms we can find in press, social media, or even in academia. In contrast to such stereotyped views, for MIR research, EDM, in fact, presents an interesting area as some styles have inherent properties that may challenge or pose difficult problems for existing music description algorithms. These properties include complex rhythm patterns (as can be observed in IDM or breakbeat), tonal patterns beyond major-minor distinctions [41], structural development not using intro-verse-chorus, temporal developments simply based on reoccurring tension-relaxation patterns (such as "drops" [1, 43]), or, contrarily, developments that are not built on tension-relaxation schemes at all. This has been acknowledged by musicologists and theorists [8, 19, 40, 41, 44].

Although some work on topics pertinent to electronic music, e.g., regarding timbre, rhythm, segmentation, or individual sub genres, have been published in recent years [1, 10, 12, 17, 18, 26, 29–31, 33, 35, 42, 43], and there seems to be a trend towards tempo estimation, e.g., [20, 28], we still lack EDM-specific annotated collections and data sets. For instance, existing data sets for tempo (or beat) estimation comprise of ballroom dance genres [23], Beatles tracks [13, 25], classical, jazz, and (J-)pop [22], rock/pop, dance, classical, folk and jazz [24], or examples from classical music, romantic music, film soundtracks, blues, chanson, and solo guitar tracks selected for "difficulty" [27]. Similarly, for tonality-related tasks, existing data sets comprise of tracks by The Beatles and Queen [32], Robbie Williams [16], piano chords [2], and rock and pop mu-

© Peter Knees, Ángel Faraldo, Perfecto Herrera, Richard Vogl, Sebastian Böck, Florian Hörschläger, Mickael Le Goff.
Licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0). **Attribution:** Peter Knees, Ángel Faraldo, Perfecto Herrera, Richard Vogl, Sebastian Böck, Florian Hörschläger, Mickael Le Goff. "Two data sets for tempo estimation and key detection in electronic dance music annotated from user corrections", 16th International Society for Music Information Retrieval Conference, 2015.

[1] http://www.amsterdam-dance-event.nl/static/files/dance-onomics_economic-significance-edm-17102012.pdf
[2] http://www.thembj.org/2013/12/the-economics-of-the-electronic-dance-industry/
[3] https://smartasset.com/insights/the-economics-of-electronic-dance-music-festivals

sic [7, 15]. Other data sets used in MIR research that contain electronic dance music or other types of electronic music, such as the Million Song Dataset [3], the MediaEval 2014 Crowdsourcing Task data set, [4] or the art-oriented UbuWeb corpus [11], lack human annotations of tempo and key, among others.

In this paper, we want to address this lack of EDM data sets for MIR research. To this end, we propose two data sets – one for the task of tempo estimation and one for the task of key detection. In contrast to existing collections, both released data sets focus on electronic dance music. Since labeling a corpus manually is a labor-intense task, we follow another strategy to obtain human ground truth annotations for tracks from an digital online record store focusing on EDM, namely *Beatport*. [5] As tempo and key information given by the retailer are imperfect, users were encouraged to give feedback on spotted incorrect data using a dedicated online forum. We describe this forum in Section 2. We extract the contained information using regular expressions and knowledge-based filtering in order to obtain user-based annotations for the corresponding tracks (Section 3). In Section 4, we present some descriptive statistics on the extracted ground truth. Section 5 reports on benchmarking results obtained using a variety of academic and commercial algorithms on the two new data sets. We conclude this paper by discussing the modalities of making this data set available to the research community and by drawing conclusions in Section 6.

2. BEATPORT USER FORUM

Beatport is a US- and Germany-based online music store targeted at DJs and music producers. In comparison to standard music web stores, it emphasizes additional meta-data relevant for DJs, such as tempo, key, and style, as well as information on record label, release information, version, and remixing artists, making it an interesting source for MIR research. Meta-data associated with a track can be easily extracted in JSON format from the source code of the corresponding web page. This meta-data also contains links to the listening snippets of the tracks, which are typically between 60 and 120 seconds long.

An important observation is that tempo and key information provided on the website are determined algorithmically upon upload of the tracks by undisclosed algorithms. Thus, this information can not be considered a ground truth and is therefore useless for evaluation purposes. [6] However, apparently being aware of the imperfection of their automatic annotation algorithms, until late 2014, *Beatport* asked its customers to provide feedback on tempo and key information via a link ("Report Incorrect BPM/Key") pointing to a dedicated online forum. In this forum, users would post their corrections in free-form text using natural language, i.e., the feedback given is highly heterogeneous and in many cases incomplete (no information, reference to track missing, etc.) [7] Nonetheless, as other work has shown [37], online forums present a great opportunity to extract user-generated, music-related information. Table 1 shows typical comments posted into the forum.

"93 bpm not 111 or whatever it is!"
"bpm is 120 not 160. i should know, i made it ;)"
"173 bpm / g minor"
"key should be c# minor"
"wrong bpm"
"the bpm is fine... its the genre. it's progressive house, not tech house."

Table 1. Examples of correctional comments published on the online forum (links to tracks removed for readability)

We performed a complete web crawl of this user forum in May 2014. At the time of the crawl, there were 2,412 comments available, of which 1,857 contained a direct link to a track on the *Beatport* website. From the link to the track, we download the complete meta-data record in JSON format using web scraping techniques. From this, we also extract the associated style descriptor for statistical reasons, cf. Section 4.

3. GROUND TRUTH EXTRACTION

In this section we detail the process of extracting ground truth from the 1,857 comments that contained a link to a track. First, we describe the process of extracting BPM (beats-per-minute) information. Second, we describe the extraction of key information from the forum, as well as from expert sources available online. All steps were performed after case-folding the texts.

3.1 BPM Extraction

For BPM extraction, we retain all posts that contain the word 'bpm' and a two- or three-digit number, optionally followed by a decimal point and a one- to three-digit number. On the remaining posts, we apply several rule-based filter criteria to exclude unlikely or possibly unrelated numbers. This comprises of all numbers below 40 and above 250 as these represent tempo values with a low probability of occurrence in this context. Furthermore, we remove all two- or three-digit numbers (with optional decimal places) that are preceded by the word 'not' as well as the number representing the tempo given by the *Beatport* website (as this is obviously the wrong tempo). We then take the first matching number as ground truth for the linked track. Applying this restrictive filtering, we were able to extract 726 records of BPM tempo annotations that were made by humans rather than an algorithm.

[4] https://osf.io/h92g8/

[5] http://www.beatport.com

[6] The same holds for the associated genre/style information, which has to be set by the human uploading the tracks onto the platform and often results in rather arbitrary assignments, cf. [38, 39]

[7] The resulting difficulty in exploiting this information might be one of the reasons why none of the reported errors have led to a correction of the meta-data on the *Beatport* website, which has been also been negatively commented on by users, and could be a reason for discontinuing this form of feedback.

From these 726 annotations, we identified duplicate BPM entries for the same IDs (e.g., when different users report a wrong tempo on the website for the same track or when the same user repeatedly urges to incorporate a suggestion made before). Furthermore, we use audio fingerprinting as well as manual inspection in order to map duplicate audio files with different IDs to one single ID (9 files). This joint information on duplicates is used to substantiate the tempo annotation: if there is more than one tempo correction available per track, we put all candidates with the same tempo (within $\pm 4\%$) into a bin and consider the mean of all tempo candidates within the bin that contains the absolute majority ($> 50\%$) as the correct annotation. If no such bin exists, the track is rejected. Since this was only the case for one file, we manually set the correct tempo for this file. This way 61 entries, including the 9 files with same audio but different IDs, were removed. In total, 42 resulting ground truth annotations are based on multiple sources. We further removed one file because the linked mp3 sample was no longer available. After this procedure, we obtain a human-annotated data set of 664 distinct electronic music tracks.

3.2 Key Extraction

A similar process was carried out on the same data, in order to extract user corrections on *Beatport*'s key tags. Additionally, we found three independently annotated sources that use the *Beatport* database for software benchmarking.

3.2.1 User Forum

In the 1,857 posts that contained a link to a track, we filter all that contain the sequences 'mixed-in-key,' 'mixed in key,' 'mik,' and 'melodyne' in order to exclude posts reporting on other algorithm's outputs. In the remaining posts, we search for occurrences of the regular expression `[a-g](\s*(#|b|sharp|flat))?\s*(min|maj)(or)?` where `\s` represents the class of whitespace characters. Additionally, all occurrences of this expression preceded by the word 'not' are excluded as well as matches that represent the same key as the key indicated in the *Beatport* meta-data (which, again, is obviously wrong).

After processing, we found a total of 404 key corrections which can be regarded as ground truth. In this group we found 15 duplicates and one track which is no longer available, leaving us with a total of 388 global-key annotations.

3.2.2 DJ Endo Labels

In order to compare different commercial key detection approaches, *DJ Endo* has published two online reports with different samples from *Beatport* that are built on his own ground truth annotations. For the first report (2011), [8] he annotates 100 songs, providing a (slightly truncated) GIF image file of the list. This image contains 99 items (one of which is a duplicate) with artist name, song title, his personal annotation, and the predictions of *Mixed-In-Key* and *Beatport*. We used OCR software to convert this list to a spreadsheet in order to obtain the human labels and access to the audio excerpts from the *Beatport* website. Using a simple script that queries the *Beatport* search page for artist and title, we retrieve the meta-data of candidate tracks. In case artist and title match perfectly, they are assigned, in case there are multiple candidates (e.g., different remix versions), a manual assignment to the correct version is done. Ultimately, this allowed us to obtain 92 out of the unique 98 tracks in the list image.

In the second report (2013), [9] DJ Endo makes a more exhaustive comparison between 7 different key estimation applications, including the *Beatport* database. The track list holds a total of 119 songs, 19 of which come from *YouTube* videos, while 7 tracks are listed without any link or *Beatport* key tag. We have excluded these 26 items, obtaining a batch of 93 songs with ground truth and links to the *Beatport* samples.

3.2.3 DJTechTools Labels

A third internet source (2014) [10] provides ground truth from human consensus for another 60 tracks. Besides the manual annotations and the *Beatport* key tags and links, 10 commercial products are evaluated.

Two of the annotations in this collection provide two key estimates per track. These have been checked and reduced to a single key manually by one of the authors, to fit with the rest of the collection.

3.2.4 Unification

With all these sources added together, we obtain a compound data set with 633 annotated tracks. However, we found a total of 29 duplicates among the different sources. In all cases, the different sources agree on the reported key, giving evidence that our approach is working (see also Section 6). This leaves us with a global-key detection data set of 604 EDM excerpts.

4. DATA SET CHARACTERISTICS

In this section we want to analyze the newly obtained data sets. To this end, we present descriptive statistics and also utilize the style information extracted from the *Beatport* meta-data. Please note that this style information does not represent a consistently annotated ground truth but merely serves as a broad reference to estimate the characteristics of the data sets.

4.1 Tempo Data Set Statistics

The Tempo data set contains tempo ground truth for 664 samples. Table 2 provides descriptive statistics for the samples within the different *Beatport* styles. The table contains the corresponding number of samples as well as the minimum, the maximum, the mean ($\bar{x}$), the median ($\tilde{x}$) and the standard deviation (σ) of the tempo annotations for each individual style. The extracted tempo ground truth

[8] http://blog.dubspot.com/dubspot-lab-report-mixed-in-key-vs-beatport

[9] http://blog.dubspot.com/endo-harmonic-mixing-key-detection-analysis

[10] http://www.djtechtools.com/2014/01/14/key-detection-software-comparison-2014-edition

style	#	$\bar{x}$	$\tilde{x}$	σ	min	max
reggae-dub	2	70.0	70.0	0.0	70.0	70.0
chill-out	15	88.3	80.0	27.0	53.0	173.0
indie-dance-nu-dsc.	11	97.9	99.0	16.6	80.0	123.0
hip-hop	2	107.5	107.5	32.5	75.0	140.0
glitch-hop	17	109.9	110.0	26.9	80.0	174.0
deep-house	24	120.1	122.0	8.3	82.0	126.0
house	23	120.3	126.0	26.9	58.0	174.0
tech-house	22	123.8	126.0	5.3	107.0	130.0
techno	61	126.1	126.0	13.7	63.5	180.0
minimal	8	126.8	127.5	1.6	123.0	128.0
progressive-house	19	126.8	128.0	8.4	96.0	140.0
electronica	54	127.2	129.0	32.7	64.0	180.0
dj-tools	9	128.0	126.0	21.0	93.0	175.0
electro-house	22	129.4	128.0	21.7	63.0	175.0
funk-r-and-b	1	135.0	135.0	0.0	135.0	135.0
hard-dance	8	135.1	148.0	27.2	90.0	171.4
dubstep	76	135.2	140.0	23.7	70.0	180.0
breaks	26	138.9	140.0	14.4	83.5	170.0
trance	74	140.3	140.0	7.3	130.0	199.0
psy-trance	34	143.6	146.5	17.2	85.0	190.0
pop-rock	3	144.0	130.0	21.2	128.0	174.0
drum-and-bass	139	162.0	173.0	28.0	80.0	180.0
hardcore-hard-tech.	14	174.6	171.2	14.7	140.0	200.0
all	664	136.7	140.0	28.3	53.0	200.0

Table 2. Statistics for the *GiantSteps* Tempo data set per style (#...number of examples, $\bar{x}$...mean BPM value, $\tilde{x}$...median, σ...std.dev.).

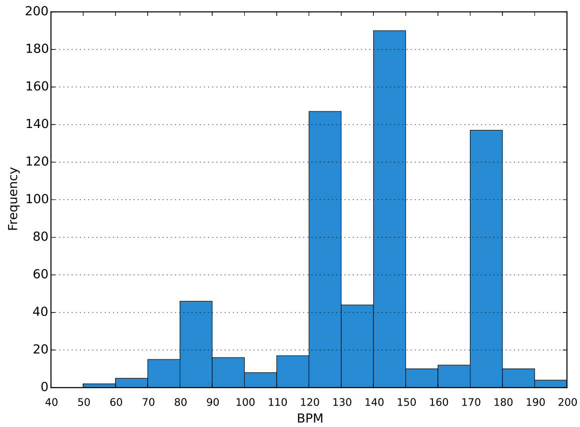

Figure 1. Distribution of BPM values in the Tempo set.

ranges from 53 to 200 BPM. Figure 1 contains a histogram of all BPM values in the data set. It reveals that most of the values are between 120 and 150 BPM, furthermore a peak between 170 and 180 BPM is apparent. This peak can most likely be attributed to the style *drum-and-bass* ($\bar{x} = 162$ BPM, $\tilde{x} = 173$ BPM), which makes up 20.9 % of all samples. This style is known for very high tempos (above 160 BPM) and seems to be a challenging and error prone task for beat and tempo estimation algorithms due to its syncopated beat structure. The evaluation of different tempo estimation approaches presented in Section 5 supports this theory. We argue that *Beatport*'s algorithmic issues with this genre, apart from the style's popularity, are the reason that many incorrect estimates were found by users and reported. Figure 2 visualizes the distribution of the different *Beatport* styles in the data set by means of a histogram.

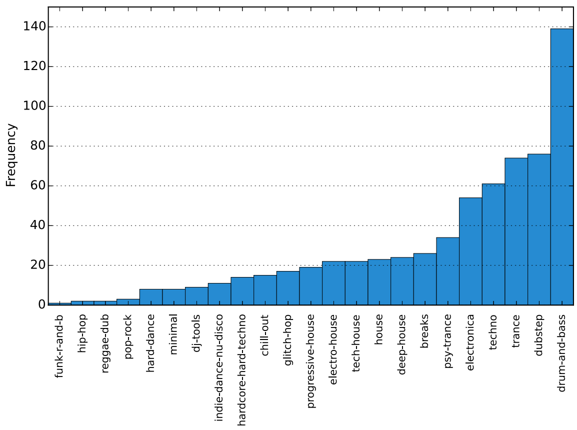

Figure 2. Histogram of tracks per style in the Tempo set.

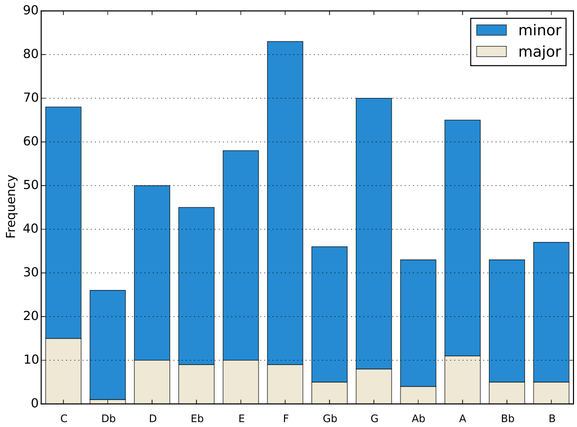

Figure 3. Distribution of keys in the Key data set.

4.2 Key Data Set Statistics

The Key data set contains 604 tracks with ground truth. Table 3 presents some simple statistics, including number of excerpts per subgenre, percentage of major and minor keys and most frequent key. 84.8% of the data set is in minor. Figure 3 shows the distribution of the corpus by tonal centers. The most frequent key is *Fm*, closely followed by *Cm* and *Gm*. Overall the distribution of tonics is relatively balanced. This is possibly related to the modes of production of these styles of music.

Figure 4 presents a histogram of tracks arranged by *Beatport* genre tags. We observe that these are unevenly distributed, with 344 excerpts (57%) pertaining to different "house" styles, whereas other subgenres and categories are underrepresented, with only 3 to 6 tracks each (funk and r&b, glitch-hop, hard-dance, hardcore, hip-hop, psy-trance, reggae/dub, and dj-tools).

5. BENCHMARKING

In this section we provide benchmarking results for both academic and commercial approaches on both data sets to estimate the performance of current methods as well as getting an impression of the "difficulty" of the data sets.

style	#	maj (%)	min (%)	most freq. key (%)
breaks	14	28.6	71.4	C (21.0)
chill-out	11	36.3	63.6	Em, Dm, Ab (18.1)
deep-house	77	5.2	94.8	Cm (13.0)
dj-tools	3	33.3	66.7	—
drum-and-bass	38	18.4	81.6	Gm (28.9)
dubstep	22	9.1	90.9	Fm (22.7)
electro-house	51	9.8	90.2	Fm (25.5)
electronica	20	20.0	80.0	Fm (20.0)
funk-r-and-b	3	0.0	100.0	—
glitch-hop	6	20.0	80.0	Gm (15.0)
hard-dance	4	0.0	100.0	Gbm (50.0)
hardcore-hard-tech.	3	33.3	66.7	—
hip-hop	4	0.0	100.0	Em (50.0)
house	47	17.0	83.0	Gm,Cm (12.8)
indie-dance-nu-dsc.	14	21.4	78.6	—
minimal	11	0.0	100.0	Em,Am (27.3)
pop-rock	7	57.1	42.9	Gm (42.9)
progressive-house	88	21	67	Am (12)
psy-trance	5	0.0	100.0	Fm (40.0)
reggae-dub	3	33.3	66.7	—
tech-house	81	12.4	87.6	Dm (14.9)
techno	34	17.6	82.4	Cm (17.6)
trance	58	12.0	88.0	Fm (24.1)
all	604	15.2	84.8	Fm (12.0)

Table 3. Statistics for the *GiantSteps* Key data set per style (number of examples, percentage of major and minor keys, most frequent key)

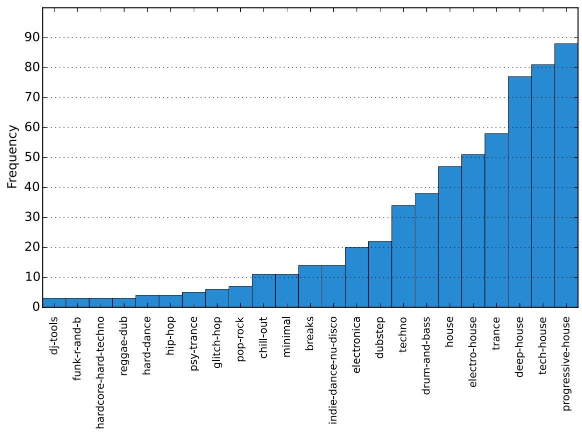

Figure 4. Histogram of tracks per style within the *GiantSteps* Key data set.

5.1 *GiantSteps* Tempo Data Set

Common in tempo estimation tasks, results are provided as accuracies within a $\pm4\%$ tolerance window. *Accuracy1* considers an estimate to be correct if it is within $\pm4\%$ of the true tempo. *Accuracy2* also considers an estimate to be correct if it is within $\pm4\%$ of either a third, half, double or triple of the true tempo, thus being permissive of so-called octave errors.

5.1.1 Algorithms

As a *baseline* we evaluate the annotations created by *Beatport*'s undisclosed algorithm, obtained by evaluating the tempo annotations that were initially reported as incorrect. We expect very low values from this strategy, as the data set consists only of cases where *Beatport* has given wrong estimates. However, the results are not trivially all zero, as the tolerance window allows for correct results if only minor deviations were corrected, as well as for corrections of octave errors.

In terms of academic algorithms we evaluate the tempo estimation approaches by Davies and Plumbley [14], Böck et al. [4], Gkiokas et al. [21], Percival and Tzanetakis [34], and Hörschläger et al. [28]. As a reference for non-academic algorithms we evaluate tempo estimators shipped with popular DJ tools, namely *Cross DJ Free*, [11] *Deckadance v2 (trial)*, [12] *Traktor 2 PRO*, [13] and *Rekordbox v3.2.2.* [14] We argue that those estimators are tailored to EDM and therefore should be able to perform well on this data set.

The commercial products typically enable (or require) the user to set an output range for BPM prediction, as a means of dealing with octave errors. *Deckadance* offers to choose among a predefined set of lower bounds of which we selected 80 BPM. In the *Traktor* option pane, the user can choose between a predefined set of tempo ranges. We decided to evaluate two ranges: 88-175 and 60-200 BPM. Similarly, for *CrossDJ*, we chose the 75-150 BPM setting as this is the best match for the given BPM distribution. The research algorithm by Böck et al. [4] also allows to set an arbitrary range. To compare to some of the range presets in commercial products, we evaluate the ranges 50-240, 95-190, and 88-175 BPM.

5.1.2 Results

Table 4 reports the obtained tempo accuracy values for all algorithms. As expected, commercial products outperform research algorithms, however none of the approaches exceeds 77% in terms of accuracy1. One important finding of more detailed investigations on a per-style level is that the proper choice of the output tempo range has a considerable influence on the accuracy for the style drum-and-bass. For instance, the algorithm by Böck et al. has a known deficiency when dealing with syncopated beats, thus, yielding only acceptable performance on drum-and-bass when being restricted to the 95-190 BPM range. Due to the fact that drum-and-bass makes up 20.9% of the collection, improvements in this style have a significant impact on the overall accuracy. This is further evidenced by [28], where performance is boosted through style-specific output ranges.

5.2 *GiantSteps* Key Data Set

The evaluation method follows the MIREX standard in key estimation tests. It assigns different weighting factors to different types of errors, depending of the proximity of the estimated key to the ground truth (fifth, relative, or parallel keys), and an overall weighted score. [15]

[11] http://www.mixvibes.com/products/cross
[12] http://www.image-line.com/deckadance/
[13] http://www.native-instruments.com/products/traktor/dj-software/traktor-pro-2/
[14] http://rekordbox.com
[15] http://www.music-ir.org/mirex/wiki/2015:Audio_Key_Detection

	accuracy1	accuracy2
Beatport	4.819	23.795
Davies, Plumbley [14]	29.367	48.042
Böck et al. [4] (50-240)	56.325	88.253
Böck et al. [4] (95-190)	76.506	86.597
Böck et al. [4] (88-175)	69.289	85.693
Gkiokas et al. [21]	58.886	82.380
Percival, Tzanetakis [34]	51.355	88.404
Hörschläger et al. [28]	75.000	82.831
Deckadance (80+)	57.681	81.627
CrossDJ (75-150)	63.404	90.211
Traktor (60-200)	64.608	88.705
Traktor (88-175)	76.958	88.705
Rekordbox	74.548	89.157

Table 4. Tempo estimation accuracies within a $\pm 4\%$ window for evaluated algorithms. BPM range restrictions in parentheses, if applicable.

5.2.1 *Algorithms*

On top of the *Beatport* annotations that serve as a baseline, we evaluate five different key estimation algorithms: two academic algorithms, namely *Queen Mary*'s Key Detector (QM-Key) [9] and *UPF's Essentia* key extractor [5], and three popular solutions, namely *KeyFinder*, [16] an open-source application by Sha'ath [36], the commercial software *Mixed-In-Key 7*, [17] and the online service/app *Rekordbox v3.2.2*. These applications are regarded trustworthy options for key detection within the EDM community.

KeyFinder is an application that allows the user to tweak the parameters of the algorithm, providing a single estimate per track. We use the default settings. On the other hand *Mixed-in-Key 7* and *Rekordbox* have a sealed approach and do not give the user any configuration option.

5.2.2 *Results*

Table 5 shows the results of the different algorithms on the key data set. If we look at the *Beatport* annotations, less than a third of the annotations match the ground truth (29.1%). However, it should be recalled that the majority of the collection (388 tracks) has been collected from reported mistakes in the *Beatport* forum, so the amount of correct keys is consequently very low. [18]

From the algorithms in the evaluation, we observe that the two academic algorithms perform poorly on this repertoire, very close to the baseline provided by the *Beatport* key tags, especially *Essentia*.

The two undisclosed approaches yield the best results, with *Rekordbox* providing 71.85% of correct estimations and a weighted score of 79.55 points. In any case, the experiment shows that there is room for improvement of the task in this specific repertoire.

[16] http://www.ibrahimshaath.co.uk/keyfinder/

[17] http://www.mixedinkey.com/

[18] As a matter of fact, if we look at the performance of the *Beatport* algorithm on the different sources of ground truth separately, we find that the tracks from the user forum only contain 4.2% of correct predictions, while the manually-annotated expert sources result in about 66% of correct predictions each.

	corr.	5th	rel.	par.	other	weigh.
Beatport	29.14	21.52	8.77	19.20	21.36	46.37
QM-Key [9]	39.40	16.89	13.41	5.13	25.17	52.90
Essentia [5]	30.46	17.55	11.09	11.42	29.47	44.85
KeyFinder [36]	45.36	20.69	6.79	7.78	19.37	59.30
Mixed-In-Key	67.22	9.27	5.63	5.30	12.58	74.60
Rekordbox	71.85	10.10	3.97	7.28	6.79	79.55

Table 5. MIREX-style scores on the Key set with results from different algorithms.

6. CONCLUSIONS

We have presented two new data sets for tempo and key estimation in electronic dance music with 664 and 604 examples, respectively. The annotations have been automatically extracted from human feedback. In order to confirm the correctness of the labels, we have inspected randomly selected 15% of the annotations manually and found them all to be correct. In order to make this data set available to the community, we offer the annotations for download on a dedicated web page alongside scripts to retrieve the corresponding audio files from *Beatport* (and a backup location in case files change or are removed) and the original data including the crawl from the user forum and the code to extract the ground truth. [19] Since we performed rather restrictive filtering, a semi-automatic approach, for instance, would allow to extract even more ground truth labels for future work.

From the benchmarking results, we can see that there is still room for improvement for MIR algorithms. Although the data set is biased towards examples that are hard to classify specifically for the *Beatport* algorithms, these results challenge the stereotypical view on EDM as being "trivial cases". Commercial algorithms are ahead of research-oriented multi-purpose algorithms for both tempo and key estimation as they are likely optimized for EDM. We can conclude that academic algorithms still need to be improved in order to meet the characteristics of EDM, something we wish to contribute to with the publication of these new data sets.

7. ACKNOWLEDGMENTS

This work is supported by the European Union Seventh Framework Programme FP7 / 2007-2013 through the GiantSteps project (grant agreement no. 610591).

8. REFERENCES

[1] A. Aljanaki, M. Soleymani, F. Wiering, and R. Veltkamp. Mediaeval 2014: A multimodal approach to drop detection in electronic dance music. In *Proc MediaEval Workshop*, 2014.

[2] A. Barbancho, I. Barbancho, L. Tardón, and E. Molina. *Database of Piano Chords: An Engineering View of Harmony*. Springer, 2013.

[3] T. Bertin-Mahieux, D. Ellis, B. Whitman, and P. Lamere. The Million Song Dataset. In *Proc 12th ISMIR*, 2011.

[19] http://www.cp.jku.at/datasets/giantsteps/

[4] S. Böck, F. Krebs, and G. Widmer. Accurate tempo estimation based on recurrent neural networks and resonating comb filters. In *Proc 16th ISMIR*, 2015.

[5] D. Bogdanov, N. Wack, E. Gómez, S. Gulati, P. Herrera, O. Mayor, G. Roma, J. Salamon, J. Zapata, and X. Serra. Essentia: an open-source library for sound and music analysis. In *Proc 21st ACM Multimedia*, 2013.

[6] B. Brewster and F. Broughton. *Last Night a DJ Saved My Life: The History of the Disc Jockey*. Grove/Atlantic, 2007.

[7] J. Burgoyne, J. Wild, and I. Fujinaga. An expert ground-truth set for audio chord recognition and music analysis. In *Proc 12th ISMIR*, 2011.

[8] M. Butler. *Unlocking the Groove: Rhythm, Meter, and Musical Design in Electronic Dance Music*. Profiles in popular music. Indiana University Press, 2006.

[9] C. Cannam, M. Mauch, M. Davies, S. Dixon, C. Landone, K. Noland, M. Levy, M. Zanoni, D. Stowell, and L. Figueira. MIREX 2013 entry: Vamp plugins from the Centre for Digital Music, 2013.

[10] N. Collins. Influence in early electronic dance music: An audio content analysis investigation. In *Proc 13th ISMIR*, 2012.

[11] N. Collins. The UbuWeb electronic music corpus: an MIR investigation of a historical database. *Organised Sound*, 20(1), 2015.

[12] N. Collins and A. McLean. Algorave: A survey of the history, aesthetics and technology of live performance of algorithmic electronic dance music. In *Proc 14th NIME*, 2014.

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

[14] M. Davies and M. Plumbley. Context-dependent beat tracking of musical audio. *IEEE TASLP*, 15, 2007.

[15] T. de Clercq and D. Temperley. A corpus analysis of rock harmony. *Popular Music*, 30, 2011.

[16] B. di Giorgi, M. Zanoni, A. Sarti, and S. Tubaro. Automatic chord recognition based on the probabilistic modeling of diatonic modal harmony. In *Proc 8th Int'l Workshop on Multidimensional Systems (nDS)*, 2013.

[17] D. Diakopoulos, O. Vallis, J. Hochenbaum, J. Murphy, and A. Kapur. 21st century electronica: MIR techniques for classification and performance. In *Proc 10th ISMIR*, 2009.

[18] A. Eigenfeldt and P. Pasquier. Evolving structures for electronic dance music. In *Proc 15th Conf on Genetic and Evolutionary Computation (GECCO)*, 2013.

[19] L.-M. Garcia. On and on: Repetition as process and pleasure in electronic dance music. *Music Theory Online*, 11(4), 2005.

[20] D. Gärtner. Tempo detection of urban music using tatum grid non negative matrix factorization. In *Proc 14th ISMIR*, 2013.

[21] A. Gkiokas, V. Katsouros, G. Carayannis, and T. Stafylakis. Music tempo estimation and beat tracking by applying source separation and metrical relations. In *Proc 37th ICASSP*, 2012.

[22] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka. RWC music database: Popular, classical, and jazz music databases. In *Proc 3rd ISMIR*, 2002.

[23] F. Gouyon, A. Klapuri, S. Dixon, M. Alonso, G. Tzanetakis, C. Uhle, and P. Cano. An experimental comparison of audio tempo induction algorithms. *IEEE TASLP*, 14(5), 2006.

[24] S. Hainsworth and M. Macleod. Particle filtering applied to musical tempo tracking. *EURASIP JASP*, 2004:15, 2004.

[25] C. Harte. *Towards Automatic Extraction of Harmony Information from Music Signals*. PhD thesis, Queen Mary University of London, 2010.

[26] J. Hockman, M. Davies, and I. Fujinaga. One in the jungle: Downbeat detection in hardcore, jungle, and drum and bass. In *Proc 13th ISMIR*, 2012.

[27] A. Holzapfel, M. Davies, J. Zapata, J. Oliveira, and F. Gouyon. Selective sampling for beat tracking evaluation. *IEEE TASLP*, 20(9), 2012.

[28] F. Hörschläger, R. Vogl, S. Böck, and P. Knees. Addressing tempo estimation octave errors in electronic music by incorporating style information extracted from Wikipedia. In *Proc 12th SMC*, 2015.

[29] K. Jacobson, M. Davies, and M. Sandler. Toward textual annotation of rhythmic style in electronic dance music. In *AES Conv 123*, 2007.

[30] T. Kell and G. Tzanetakis. Empirical analysis of track selection and ordering in electronic dance music using audio feature extraction. In *Proc 14th ISMIR*, 2013.

[31] M. Leimeister, D. Gärtner, and C. Dittmar. Rhythmic classification of electronic dance music. In *AES 53rd Int'l Conf: Semantic Audio*, 2014.

[32] M. Mauch, C. Cannam, M. Davies, C. Harte, S. Kolozali, D. Tidhar, and M. Sandler. Omras2 metadata project 2009. 10th ISMIR Late-Breaking Session, 2009.

[33] M. Panteli, N. Bogaards, and A. Honingh. Modeling rhythm similarity for electronic dance music. In *Proc 15th ISMIR*, 2014.

[34] G. Percival and G. Tzanetakis. Streamlined tempo estimation based on autocorrelation and cross-correlation with pulses. *IEEE/ACM TASLP*, 22(12), 2014.

[35] B. Rocha, N. Bogaards, and A. Honingh. Segmentation and timbre- and rhythm similarity in electronic dance music. In *Proc 10th SMC*, 2013.

[36] I. Sha'ath. Estimation of key in digital music recordings. Tech Report, Birkbeck College, University of London, 2011.

[37] M. Sordo, J. Serrà, G. Koduri, and X. Serra. Extracting semantic information from an online carnatic music forum. In *Proc 13th ISMIR*, 2012.

[38] B. Sturm. Classification accuracy is not enough: On the evaluation of music genre recognition systems. *JIIS*, 41(3), 2013.

[39] B. Sturm. The GTZAN dataset: Its contents, its faults, their effects on evaluation, and its future use. *arXiv.org (e-prints)*, 2013.

[40] P. Tagg. Debate: From refrain to rave: the decline of figure and the rise of ground. *Popular Music*, 13, 1994.

[41] R. Wooller and A. Brown. A framework for discussing tonality in electronic dance music. In *Sound : Space - Proc Australasian Computer Music Conference*, 2008.

[42] R. Wooller and A. Brown. Note sequence morphing algorithms for performance of electronic dance music. *Digital Creativity*, 22(1), 2011.

[43] K. Yadati, M. Larson, C. Liem, and A. Hanjalic. Detecting drops in electronic dance music: Content-based approaches to a socially significant music event. In *Proc 15th ISMIR*, 2014.

[44] H. Zeiner-Henriksen. Moved by the groove: Bass drum sounds and body movements in electronic dance music. In Danielsen, ed, *Musical Rhythm in the Age of Digital Reproduction*. Ashgate, Farnham, Surrey, UK, 2010.