A CROWDSOURCED EXPERIMENT FOR TEMPO ESTIMATION OF ELECTRONIC DANCE MUSIC

Hendrik Schreiber
tagtraum industries incorporated
hs@tagtraum.com

Meinard Müller
International Audio Laboratories Erlangen
meinard.mueller@audiolabs-erlangen.de

ABSTRACT

Relative to other datasets, state-of-the-art tempo estimation algorithms perform poorly on the GiantSteps Tempo dataset for electronic dance music (EDM). In order to investigate why, we conducted a large-scale, crowdsourced experiment involving 266 participants from two distinct groups. The quality of the collected data was evaluated with regard to the participants' input devices and background. In the data itself we observed significant tempo ambiguities, which we attribute to annotator subjectivity and tempo instability. As a further contribution, we then constructed new annotations consisting of tempo distributions for each track. Using these annotations, we re-evaluated two recent state-of-the-art tempo estimation systems achieving significantly improved results. The main conclusions of this investigation are that current tempo estimation systems perform better than previously thought and that evaluation quality needs to be improved. The new crowdsourced annotations will be released for evaluation purposes.

1. INTRODUCTION

Estimation of a music piece's *global tempo* is a classic *music information retrieval* (MIR) task. It is often defined as estimating the frequency with which humans tap along to the beat. A necessary precondition for successful global tempo estimation is the existence of a stable tempo as it often occurs in rock, pop, or dance music. To evaluate a tempo estimation system one needs the system itself, a dataset with suitable tempo annotations, and one or more metrics. One such dataset, named *GiantSteps Tempo*, has been released by Knees et al. in 2015 [6]. It was created by scraping a forum that let listeners discuss Beatport [1] songs with wrong tempo labels. Scraping was done via a script and 15% of the labels were manually verified. All 664 tracks in the dataset belong to the umbrella genre electronic dance music (EDM) with its subgenres trance, drum-and-bass, techno, etc. Since its release, several academic and commercial tempo estimation systems have been tested against the dataset (e.g. [12]). As is common for datasets annotated with only a single tempo per track, the two metrics *Accuracy1* and *Accuracy2* were used. *Accuracy1* is defined as the fraction of correct estimates while allowing a tolerance of 4%. *Accuracy2* additionally allows estimates to be wrong by a factor of 2, 3, 1/2 or 1/3 (so-called *octave errors*). The highest results reported for the *GiantSteps* dataset are 77.0% *Accuracy1* by the applications NI Traktor Pro 2 [2] (with octave bias $88 - 175$) and 90.2% *Accuracy2* by CrossDJ [3] (with octave bias $75 - 150$). [4] These results are surprisingly low—the highest reported *Accuracy2* values for other commonly used datasets like *ACM Mirum* [10], *Ballroom* [4], and *GTzan* [13] are greater than 95% [1]. Since EDM is often associated with repeating bass drum patterns and steady tempi [2, 7], it should be comparatively easy to estimate the tempo for this genre. We hypothesize that relatively low accuracy values were achieved for multiple possible reasons. Since the annotations were scraped off a forum for disputed tempo labels, the dataset may contain many tracks that are especially hard to annotate for humans. And if not difficult for humans to annotate, it is conceivable that the tracks are particularly hard for algorithms to analyze. Lastly, if neither humans nor algorithms fail, perhaps some of the scraped annotations are simply wrong.

In this paper we investigate why tempo estimation systems perform so poorly for *GiantSteps Tempo*. To this end, we conducted a large, crowdsourced experiment to collect new tempo data for *GiantSteps Tempo* from human participants. The experiment is described in detail in Section 2. The data is analyzed in Section 3 and used to create a new ground-truth. This ground-truth is then compared to the original ground-truth and used to evaluate two recent algorithms. The results are discussed in Section 4. Finally, in Section 5, we summarize our findings and draw conclusions.

2. EXPERIMENT

In order to generate a new ground-truth for the *GiantSteps Tempo* dataset, we set up a web-based experiment in which

[1] http://www.beatport.com/, an online music store

[2] https://www.native-instruments.com/en/products/traktor/dj-software/traktor-pro-2/

[3] http://www.mixvibes.com/cross-dj-software-mac-pc/

[4] More benchmark results are available at http://www.cp.jku.at/datasets/giantsteps/

we asked participants to tap along to audio excerpts using their keyboard or touchscreen. The user interface for this experiment is depicted in Figure 1. Since most tracks from the dataset are 2 min long and tapping for the full duration is difficult, we split each track into half-overlapping 30 s segments. Out of the 664 tracks we created 4,640 such segments (in most cases 7 per track). To measure tempo, it is not important for tap and beat to occur at the same time. In contrast to experiments for beat tracking, phase shifts, input method latencies, or anticipatory early tapping—known as *negative mean asynchrony* (NMA)—are irrelevant, as long as they stay constant (see [11] for an overview of tapping and [3,5] for beat tracking). Therefore participants were asked to tap along to randomly chosen segments *as steadily as possible*, over the entire duration of 30 s without skipping beats. To encourage steady tapping, the user interface gave immediate feedback in the form of the mean tempo μ in BPM, the median tempo med in BPM, the standard deviation of the *inter-tap-intervals* (ITI) σ in milliseconds, as well as textual messages and emojis (Figure 1). When calculating the standard deviation, the first three taps were ignored, as those are typically of low quality (users have to find their way into the groove). When the standard deviation σ stayed very low, smilies, thumbs up and textual praise were shown. When σ climbed above a certain threshold, the user was shown sad faces and messages like "Did you miss a beat? Try to tap more steadily." To prevent low quality submissions, users were only allowed to proceed to the next track, once four conditions were met:

1. 20 or more taps
2. Taps cover at least 15 s
3. ITI standard deviation: $\sigma < 50$ ms
4. Median tempo: $50 \leq \text{med} \leq 210$ BPM

While the first three conditions were not explicitly communicated, the instructions made participants aware that the target tempo lies between 50 and 210 BPM. Once all four conditions were met, a large red bar turned green and the *Next* button became enabled. For situations in which the user was not able to fulfill all conditions, the user interface offered a *No Beat* checkbox. Once checked, it allowed users to bypass the quality check and proceed to the next song. It must be noted that there is a tradeoff between encouraging participants to tap well (i.e. steadily) and a bias towards stable tempi. We opted for this design for two reasons. 1) tempo in EDM is usually is very steady [2, 7]. 2) the bias is limited to individual tapping sessions at the segment level, i.e. we can still detect tempo stability problems on the track level by aggregating segment level annotations.

Participants were recruited from two distinct groups: Academics and people interested in the consumer-level music library management system beaTunes [5] . We refer to the former group as *academics* and the latter as *beaTunes*. While members of the *academics* group were asked to help

[5] https://www.beatunes.com/

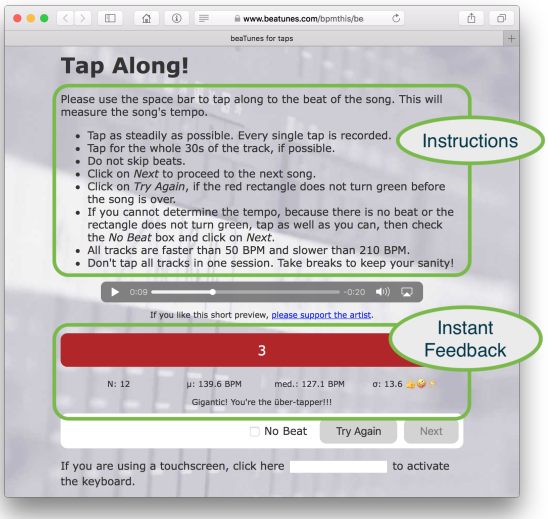

Figure 1: Illustration of the web-based interface used in our experimental user study.

in this experiment via relevant mailing lists without offering any benefits, members of the *beaTunes* group were incentivized by promising a reward license for the beaTunes software, if they submitted 110 valid annotations. While it was not explicitly specified what a "valid annotation" is, we attempted to steer people in the right direction using instructions and the instant feedback mechanisms described above (Figure 1).

3. DATA ANALYSIS

Over a period of $2^1/_2$ months 266 persons participated in the experiment, 217 (81.6%) belonging to *beaTunes* and 49 (18.4%) to *academics*. Together they submitted 18,684 segment annotations (avg = 4.03/segment). We made sure that all segments were annotated at least twice. Since some segments are harder to annotate than others, we monitored submissions and ensured that segments annotated by participants as very different from the original ground-truth—exceeding a tolerance of 4%—were presented to participants more often than others. The vast majority of annotations was submitted by the *beaTunes* group (95.1%). Overall 7.5% of all submissions were marked with *No Beat*. With 7.6% the *No Beat*-rate was slightly higher among members of the *beaTunes* group. Members of *academics* checked *No Beat* only for 5.2% of their submissions. Since the experiment was run in the participant's web-browser, the browser's user-agent for each submission was logged by the web-server. Among other information the user-agent contains the name of the participant's operating system. 17,012 (91.1%) of the submissions were sent from desktop operating systems that are typically connected to a physical keyboard. 1,672 (8.9%) were from mobile operating systems that are usually associated with touchscreens. Participants interested in a reward license, also had to enter name and email. Both datapoints have

Dataset Split	+	-	p-value
±*academics*	0.0074	0.0090	3.11e−29
±*keyboard*	0.0088	0.0095	9.74e−7
beaTunes±*keyboard*	0.0089	0.0099	6.71e−10
academics±*keyboard*	0.0074	0.0073	8.68e−1

Table 1: Average coefficients of variation $\overline{c_v}$ for dataset splits, *academics* or not, *keyboard* or not, and *keyboard* or not for either *beaTunes* or *academics*. The low p-values indicate a significant difference between the dataset splits.

been removed from the collected data to ensure anonymity.

We analyzed the submitted data to find out whether we can find quality differences between submissions from different participant groups (Section 3.1). Section 3.2 introduces metrics for ambiguity and stability. In Section 3.3, we measure to which extent participants agree on one or multiple tempi for the same segment. Then, in Section 3.4, we take a look at segment annotations aggregated on the track-level. Finally, in Section 3.5, we investigate whether tempo ambiguity is a genre-dependent phenomenon.

3.1 Submission Quality

We wondered how steadily participants tapped and whether some groups of participants tapped more steadily than others. Specifically, are the *beaTunes* submissions as good as the *academics* submissions? We can use the coefficient of variation $c_v = \frac{\sigma}{\mu}$ of each submission's ITIs as a normalized indicator for how steadily a participant tapped. To remove tapping outliers within a segment, we sort each submission's ITIs and only keep the central 10 before calculating the c_v. This has the effect of reducing c_v for all submissions. The average c_v for all submissions is $\overline{c_v} = 0.0089$. Assuming a normal distribution, this means that on average 99.7% of all central 10 ITIs lie within $\pm 2.67\%$ ($\equiv 3\sigma$) of their submission's mean value. Using $\overline{c_v}$ as a measure for the submission quality of different dataset splits, we found that members of *academics* tapped significantly more steadily ($\overline{c_v} = 0.0074$) than members of *beaTunes* ($\overline{c_v} = 0.0090$) (Table 1). To test for significance we used Welch's t-test. Also, submissions from desktop operating systems that are typically installed on devices connected to a physical keyboard (i.e., no touchscreen) are of significantly higher quality ($\overline{c_v} = 0.0088$) than submissions from devices using iOS or Android as operating system ($\overline{c_v} = 0.0095$). Despite the differences, we found that even the ITIs from the group with the highest $\overline{c_v}$, i.e., *beaTunes* without keyboard, still lie within only $\pm 2.97\%$ ($\equiv 3\sigma$) of their mean value 99.7% of the time—again assuming a normal distribution. This is well below the tolerance of 4% allowed by *Accuracy1*.

We conclude that the data submitted by *academics* with keyboard is of the highest quality with regard to tempo stability, but find that the data submitted by members of *beaTunes* without keyboard is still acceptable, because the difference in $\overline{c_v}$ is not very large. This may be a direct result of the experiment's design which did not permit participants to submit highly irregular taps.

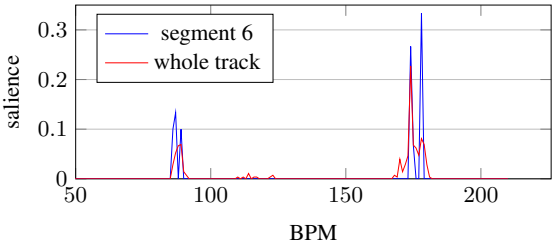

Figure 2: Tempo salience distribution for segment 6 of track 'Neoteric D&B Mix' by Polex (Beatport id 4397469). Measured values are: $P(T_{\text{track}}) = 4$, $P(T_{\text{seg6}}) = 2$, $A(T_{\text{track}}) = 0.30$, $A(T_{\text{seg6}}) = 0.40$, and $\text{JSD} = 0.24$.

3.2 Tempo Distribution Metrics

How steadily participants tapped does not say anything about whether they tapped along to the true tempo. But since the purpose of the experiment is to create a new ground-truth, we cannot easily verify submissions for correctness. What we can do though, is to measure annotator (dis)agreement both for a segment and for all segments belonging to the same track. To this end, we define some metrics based on tapped tempo distributions. To create such a tapped tempo distribution for a segment, we combine the 10 central ITIs from each of its submissions in a histogram T with a bin width of 1 BPM and then normalize so that $\sum_{i=1}^{n} T(x_i) = 1$, with n as the number of bins and x_i as the corresponding BPM values. For T we define local peaks as the highest non-zero $T(x_i)$ for all intervals $[x_i - 5, x_i + 5]$. This may include very small peaks. We interpret the BPM values x_i of the histogram's local peaks as the perceptually strongest tempi and their heights equivalent to their saliences. Per-track tempo distributions are created simply by averaging the 7 segment histograms belonging to a given track. For an example, please see Figure 2.

As a first, very simple indicator for annotator disagreement, we define $P(T)$ as the number of histogram peaks we find in a given tempo distribution T. A high peak count for a single segment $P(T_{\text{seg}})$ indicates annotator disagreement for that segment. This is not necessarily true for the peak count for a track $P(T_{\text{track}})$, since it may also be a sign of tempo instability, i.e., tempo changes or no-beat-sections. Because the peak count P does not say anything about the peaks' height or salience, it is a relatively crude measure. Therefore we define as second metric the salience ratio between the most salient and the second most salient peak as a measure for ambiguity. More formally, if s_1 is the salience of the highest peak and s_2 the salience of the second highest peak, then the ambiguity $A(T)$ is defined as:

$$A(T) := \begin{cases} 1, & \text{for } P(T) = 0 \\ 0, & \text{for } P(T) = 1 \\ s_2/s_1, & \text{for } P(T) > 1 \end{cases} \quad (1)$$

A value close to 0 indicates low and a value close to 1

high ambiguity. This definition is inspired by McKinney et al. [8] approach to ambiguity, but not identical. Just like P, we can use A for both segment and track tempo distributions. Again, for tracks we cannot be sure of the ambiguity's source.

Finally, we introduce a third metric that focuses more on tempo instability within tracks. Obvious indicators for instabilities are large differences between the tempo distributions of segments belonging to one track. Since we create tapped tempo distributions for each segment in a way that lets us interpret them as probability distributions, we can use the Jensen-Shannon Divergence (JSD) for this purpose, which is based on the Shannon entropy H. With the JSD we measure the difference between the tempo distribution's entropy for the whole track and the average of the the individual segment tempo distributions' entropies.

$$\mathrm{H}(T) := -\sum_{i=1}^{n} T(x_i)\log_b T(x_i) \tag{2}$$

$$\mathrm{JSD}(T_1, ..., T_m) := \mathrm{H}\left(\sum_{j=1}^{m} \frac{1}{m} T_j\right) - \sum_{j=1}^{m} \frac{1}{m}\mathrm{H}(T_j) \tag{3}$$

To allow an easy interpretation of JSD-values, we choose an unusual base for the entropy's logarithm. By setting $b = n$ in (2), we ensure that $0 \leq \mathrm{JSD} \leq 1$. This means, that a JSD-value near 0 indicates a small difference between the tempo distributions for a track's segments. Correspondingly, a JSD-value closer to 1 means that the tempo distributions of a track's segments are very different. To avoid detecting small tempo changes due to annotator disagreements, we convert the segment tempo distributions T to a bin width of 10 BPM before calculating JSD.

3.3 Segment Annotator Agreement

How much do participants agree on a tempo for a given segment? Recall that we have 4,640 segments (and 18,684 annotations for these segments) coming from 664 tracks. As depicted in Figure 3 top, the submissions for more than half the segments (2,500 or 53.9%) have just one peak, i.e., $P(T_{\text{seg}}) = 1$. For 1,514 or 32.6% of all segments we were able to find two peaks, indicating some ambiguity. For 432 segments (9.3%) we found 3 peaks and for 184 segments (4.0%) 4 peaks or more. 10 segments have no peak at all, because they have been marked as *No Beat* in all their submissions. When interpreting these numbers one has to keep in mind that some segments have been annotated by very few participants (Figure 3 bottom). To give an example, while the segments annotated with one peak are based on 3.64 submissions on average, the segments annotated with 6 peaks are annotated with 9.42 submissions per segment. This reflects the fact that we presented difficult segments to participants more often, but could also be caused by increased variability introduced by a higher number of submissions. Because submissions marked as *No Beat* do not show up in this overview unless all submissions for a segment were *No Beat*, we counted the segments for which a majority of submissions were marked with *No Beat*. That was the case for 118 segments (2.5%).

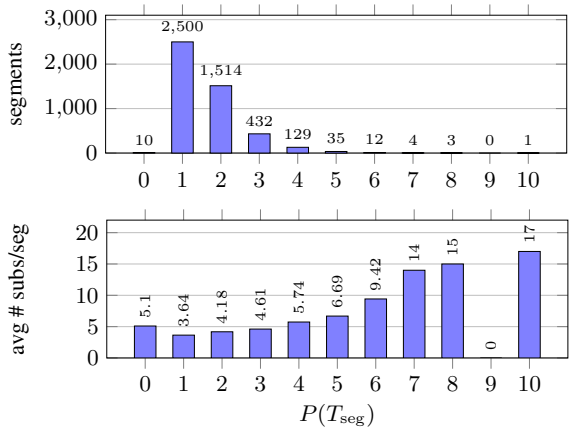

Figure 3: (top) Segments per peak count. (bottom) Average number of submissions per segment by peak count.

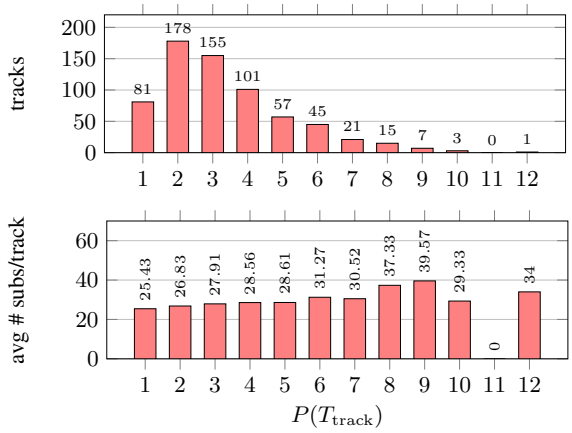

Figure 4: (top) Tracks per peak count. (bottom) Average number of submissions per track by peak count.

As mentioned in Section 3.2, the peak count does not say anything about the peaks' height or salience and is therefore a relatively crude measure. We found that the average ambiguity for all segments is $\overline{A(T_{\text{seg}})} = 0.25$ (with standard deviation $\sigma = 0.32$), meaning that on average the highest peak is 4 times more salient than the second highest peak. In other words, we can often observe a peak that is much more salient than others. At the same time, there may also be a second peak with considerable salience.

3.4 Track Annotator Agreement

Just like for the segments, we looked at the number of tracks per peak count. We found only 81 tracks (12.2%) with one peak and 582 tracks (87.8%) with two or more peaks (Figure 4 top). The largest group among the multi-peak tracks are tracks with two peaks (178 or 26.8%). These numbers are much more reliable than the segment peak counts as they are based on at least 25 submissions per track (Figure 4 bottom). Compared to the segments' peak counts we see a larger proportion of tracks with more than one peak. But this does not necessarily mean that the ambiguity A is much higher than for the segments, because

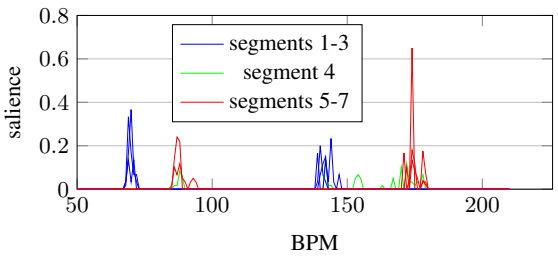

Figure 5: Tempo salience distributions for segments of the track 'Rude Boy feat. Omar LinX Union Vocal Mix' by Zeds Dead (Beatport id 1728723). The track's tempo changes in segment 4, leading to four distinct peaks. With JSD $= 0.44$ its Jensen-Shannon divergence is high.

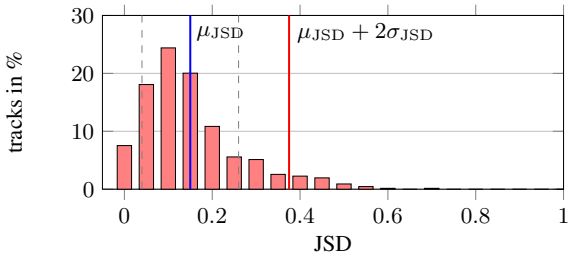

Figure 6: Distribution of tracks in the dataset per JSD interval with a bin width of 0.05. The blue line shows μ_{JSD} and the red line shows $\mu_{\mathrm{JSD}} + 2\sigma_{\mathrm{JSD}}$.

peak counts do not account for salience and even small local peaks are counted. In fact, we measured an average ambiguity of $\overline{A(T_{\mathrm{track}})} = 0.26$ (with standard deviation $\sigma = 0.27$)—almost the same average as for the segments. Therefore we attribute the shift towards more peaks to the much higher number of submissions per item and possible tempo instabilities in the tracks themselves. By tempo instability we mean for example a tempo change in the middle of the track, a quiet section, or no beat at all. Any of these cases inherently lead to more peaks. A typical example for a track with a tempo change is shown in Figure 5.

In an attempt to quantify tempo instabilities in the submissions we calculated the JSD introduced in Section 3.2. The histogram in Figure 6 shows the distribution of tracks per JSD interval with a bin width of 0.05. The average divergence for the whole dataset is $\mu_{\mathrm{JSD}} = 0.15$, the standard deviation is $\sigma_{\mathrm{JSD}} = 0.11$. To test whether a high JSD correlates with tempo instabilities, we considered all tracks with JSD $> \mu_{\mathrm{JSD}} + 2\sigma_{\mathrm{JSD}} = 0.375$, resulting in 39 tracks. Performing an informal listening test on these tracks revealed that 3 had no beat, 10 contained a tempo change (e.g. Figure 5), 7 had sections that felt half as fast as other sections (metrical ambiguity), 8 contained larger sections with no discernible beat, 9 were difficult to tap, and 2 had a stable tempo through the whole track. From this result one may conclude that a high JSD is connected to tempo instabilities, but it may also just indicate that a track is difficult to tap. Nevertheless, using JSD helped us find tracks in the *GiantSteps Tempo* dataset that exhibit tempo stability issues. Since 2.5% of the segments were annotated most often with *No Beat*, we wondered whether any tracks have a majority of segments that have predominantly been annotated with *No Beat*, hinting at the absence of not just a local beat (e.g., a sound effect or a silent section), but the lack of a global beat. This is true for 6 tracks, i.e., 0.9% of the dataset. All 6 of them are among the 39 tracks with very high JSD and either have no beat, are very difficult to tap or contain large sections without a beat.

3.5 Ambiguity by Genre

We wondered whether we can confirm findings by McKinney and Moelants [9] that the amount of tempo ambiguity depends on the genre or musical style. To ensure meaningful results, we considered only the 5 most often occurring genres in the dataset with 54 or more tracks each. We found that the genres techno and trance do not seem to be very affected by ambiguity. More than 65% of their segments are annotated with just one peak. In contrast to that, fewer than 38% of all segments in the genres drum-and-bass, dubstep, and electronica are annotated with just one peak (Figure 7 top). A similar picture presents itself when looking at the average segment ambiguity $\overline{A(T_{\mathrm{seg}})}$. As shown in Table 2, it is 0.12 for techno segments and thus much lower than the overall average of 0.25. The same is true for trance (0.17). Contrary to that, the ambiguity values for drum-and-bass (0.37), electronica (0.36) and dubstep (0.35) are all well above the average. We found similar relations for peak counts on the track level (Figure 7 bottom) and the average track ambiguity $\overline{A(T_{\mathrm{track}})}$ (Table 2). This strongly supports McKinney and Moelants' finding that tapped tempo ambiguity is genre-dependent. Perhaps it is even an inherent property.

Genre	$\overline{A(T_{\mathrm{seg}})}$	$\overline{A(T_{\mathrm{track}})}$
all	0.25	0.26
techno	0.12	0.10
trance	0.17	0.12
drum-and-bass	0.37	0.39
electronica	0.36	0.38
dubstep	0.35	0.43

Table 2: Average ambiguity for the top 5 genres.

4. EVALUATION

The tempo histograms for tracks can easily be turned into single tempo per track or two tempi+salience labels. This provides us the opportunity to evaluate the original ground-truth for the *GiantSteps Tempo* dataset by treating it like an algorithm. Since the original annotations are single tempo per track only, we are using *Accuracy1* and *Accuracy2* as metrics. To obtain one tempo value per track from a distribution, we are using just the tempo value with the highest salience. The three tracks without a beat have been removed. We refer to these new annotations as `GSNew` and to the original ones as `GSOrig`. Figure 8 shows the accuracy results for the comparison of `GSOrig` with `GSNew` and reveals a large discrepancy between the two. Only 81.5

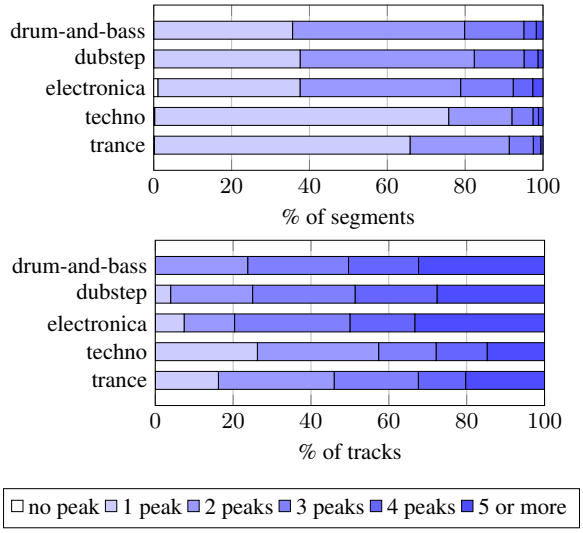

Figure 7: Percentage of segments (top) and tracks (bottom) with a given number of peaks by genre. Drum-and-bass, dubstep, and electronica suffer much more from tapped tempo ambiguity than techno and trance.

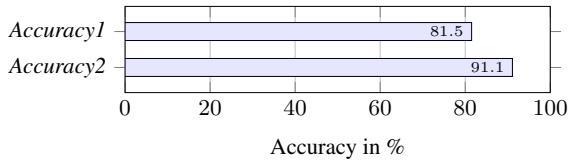

Figure 8: Accuracies measured when comparing `GSNew` with `GSOrig`.

of the labels match when using *Accuracy1*, and only 91.1% match when using *Accuracy2*.

Coming back to the original motivation for this paper—the poor performance of tempo estimation systems for *GiantSteps Tempo*—we evaluated the two state-of-the-art algorithms `schreiber` [12] and `böck` [1] with both the old and the new annotations. The algorithms were chosen for their proven performance and conceptual dissimilarity. While `schreiber` implements a conventional onset detection approach followed by an error correction procedure, `böck`'s core consists of a bidirectional long short-term memory (BLSTM) recurrent neural network. Despite their conceptual differences, both algorithms reach considerably higher accuracy values when tested against `GSNew` (Figure 9). *Accuracy1* increases for `böck` by 5.9 pp (58.9% to 64.8%) and for `schreiber` by 7.1 pp (63.1% to 70.2%). *Accuracy2* shows similar increases, 7.6 pp (86.4% to 94.0%) for `böck` and 6.5 pp (88.7% to 95.2%) for `schreiber`. Remarkably, both `böck` and `schreiber` reach higher *Accuracy2* values for `GSNew` than the original annotations reached, when compared with `GSNew`. The increased results for `GSNew` are much more in line with values reported for other tempo datasets. We therefore believe that this increase and the discrepancy between `GSOrig` and `GSNew` are hardly coincidences, but strong indicators for incorrect annotations in `GSOrig`.

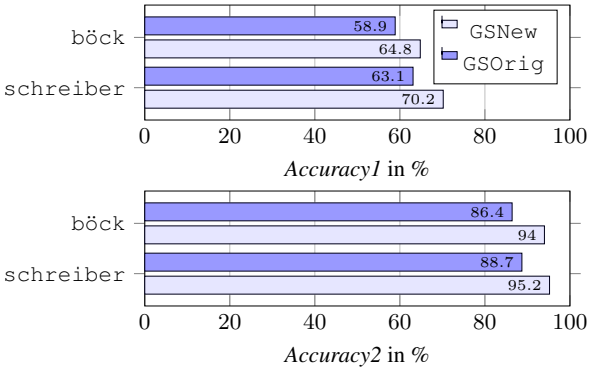

Figure 9: Accuracies for the algorithms `böck` and `schreiber` measured against both `GSOrig` and `GSNew`.

5. DISCUSSION AND CONCLUSIONS

In this paper we described a crowdsourced experiment for tempo estimation. We collected 18,684 tapped annotations from 266 participants for electronic dance music (EDM) tracks contained in the *GiantSteps Tempo* dataset. To analyze the data, we used multiple metrics and found that half of the annotated segments and more than half of the tracks exhibit some degree of tempo ambiguity, which may either stem from annotator disagreement or from intra-track tempo instability. This refutes the assumption that it is always easy to determine a single global tempo for EDM. We were able to identify tracks with no tempo at all, no-beat-sections or tempo changes, which raises questions about the suitability of parts of the dataset for the global tempo estimation task. Furthermore, we provided additional evidence for genre-dependent tempo ambiguity. Based on the user-submitted data we derived the new annotations `GSNew`. The relatively low agreement with the original annotations `GSOrig` indicates that one of the two ground-truths contains incorrect annotations for up to 8.9% of the tracks (ignoring octave errors). We re-evaluated two recent tempo estimation algorithms against both ground-truths and measured considerably higher accuracies when testing against `GSNew`. This leads us to the following conclusions: `GSOrig` contains incorrectly annotated tracks as well as tracks that are not suitable for the global tempo estimation task. The accuracy of state-of-the-art tempo estimations systems is considerably higher than previously thought. And last but not least, as a community, we have to get better at evaluating tempo algorithms in the sense that we need verified, high quality datasets that represent reality with tempo distributions instead of single value annotations. If we cannot accurately measure progress, we have no way of knowing when the task is done.

Datasets

All data is available at `http://www.tagtraum.com/tempo_estimation.html`.

Acknowledgments

The International Audio Laboratories Erlangen are a joint institution of the Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU) and Fraunhofer Institute for Integrated Circuits IIS. Meinard Müller is supported by the German Research Foundation (DFG MU 2686/11-1).

6. REFERENCES

[1] Sebastian Böck, Florian Krebs, and Gerhard Widmer. Accurate tempo estimation based on recurrent neural networks and resonating comb filters. In *Proceedings of the 16th International Society for Music Information Retrieval Conference (ISMIR)*, pages 625–631, Málaga, Spain, 2015.

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

[3] Olmo Cornelis, Joren Six, Andre Holzapfel, and Marc Leman. Evaluation and recommendation of pulse and tempo annotation in ethnic music. *Journal of New Music Research*, 42(2):131–149, 2013.

[4] Fabien Gouyon, Anssi P. Klapuri, Simon Dixon, Miguel Alonso, George Tzanetakis, Christian Uhle, and Pedro Cano. An experimental comparison of audio tempo induction algorithms. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(5):1832–1844, 2006.

[5] Andre Holzapfel, Matthew EP Davies, José R Zapata, João Lobato Oliveira, and Fabien Gouyon. Selective sampling for beat tracking evaluation. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(9):2539–2548, 2012.

[6] Peter Knees, Ángel Faraldo, Perfecto Herrera, Richard Vogl, Sebastian Böck, Florian Hörschläger, and Mickael Le Goff. Two data sets for tempo estimation and key detection in electronic dance music annotated from user corrections. In *Proceedings of the 16th International Society for Music Information Retrieval Conference (ISMIR)*, pages 364–370, Málaga, Spain, October 2015.

[7] Michaelangelo Matos. Electronic dance music. Encyclopædia Britannica `https://www.britannica.com/art/electronic-dance-music`, last checked 6/9/2018, August 2015.

[8] Martin F McKinney and Dirk Moelants. Deviations from the resonance theory of tempo induction. In *Proc. Conference on Interdisciplinary Musicology*, Graz, Austria, 2004.

[9] Martin F McKinney and Dirk Moelants. Ambiguity in tempo perception: What draws listeners to different metrical levels? *Music Perception: An Interdisciplinary Journal*, 24(2):155–166, 2006.

[10] Geoffroy Peeters and Joachim Flocon-Cholet. Perceptual tempo estimation using GMM-regression. In *Proceedings of the second international ACM workshop on Music information retrieval with user-centered and multimodal strategies (MIRUM)*, pages 45–50, New York, NY, USA, 2012. ACM.

[11] Bruno H. Repp. Sensorimotor synchronization: A review of the tapping literature. *Psychonomic Bulletin & Review*, 12(6):969–992, 2005.

[12] Hendrik Schreiber and Meinard Müller. A post-processing procedure for improving music tempo estimates using supervised learning. In *18th International Society for Music Information Retrieval Conference (ISMIR)*, pages 235–242, Suzhou, China, October 2017.

[13] George Tzanetakis and Perry Cook. Musical genre classification of audio signals. *IEEE Transactions on Speech and Audio Processing*, 10(5):293–302, 2002.