TISMIR

Schreiber, H., et al. (2020). Music Tempo Estimation: Are We Done Yet? *Transactions of the International Society for Music Information Retrieval,* 3(1), pp. 111–125. DOI: https://doi.org/10.5334/tismir.43

OVERVIEW ARTICLE

Music Tempo Estimation: Are We Done Yet?

Hendrik Schreiber*, Julián Urbano† and Meinard Müller*

With the advent of deep learning, global tempo estimation accuracy has reached a new peak, which presents a great opportunity to evaluate our evaluation practices. In this article, we discuss presumed and actual applications, the pros and cons of commonly used metrics, and the suitability of popular datasets. To guide future research, we present results of a survey among domain experts that investigates today's applications, their requirements, and the usefulness of currently employed metrics. To aid future evaluations, we present a public repository containing evaluation code as well as estimates by many different systems and different ground truths for popular datasets.

Keywords: Evaluation; Tempo Estimation; Metric; Dataset

1. Introduction

The estimation of a music recording'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 (Scheirer, 1998; Dixon, 2001). In contrast to beat-tracking (Allen and Dannenberg, 1990; Goto and Muraoka, 1994) or local tempo estimation (Peeters, 2005), successful global tempo estimation requires the existence of a stable tempo as often occurs in Rock, Pop, or Dance music. To conduct a basic evaluation of a global tempo estimation system one needs the system itself, test recordings with globally stable tempo, suitable annotations, and at least one metric. Starting with the work of Goto and Muraoka (1994) and Scheirer (1998), the MIR research community has been conducting such evaluations for 25 years. Acknowledging the importance of making results comparable, the first systematic evaluation with a defined set of metrics and datasets was conducted in 2004 (Gouyon et al., 2006). One year later, the 2005 Music Information Retrieval Evaluation eXchange (MIREX) (Downie, 2008) established an automatic tempo extraction task, which has been conducted almost every year ever since. Through both the datasets and metrics established in 2004 and for MIREX, we have seen global tempo estimation systems improve and have been able to track their performance. In the meantime, new datasets have been published and another large-scale evaluation has been conducted (Zapata and Gómez, 2011), but neither applications nor metrics have been fundamentally questioned or updated. This is why recent near-perfect MIREX results (Böck et al., 2015; Schreiber and Müller, 2018b) beg the question: are we done yet?

In this work, we critically discuss the evaluation of global tempo estimation systems. We do so based on the idea that *applications* lead to *use cases* that define who the users are, how they use the system, in what context and for what purpose (Schedl et al., 2013). The combination of these elements determines the *success criteria* to evaluate systems and judge whether the task is indeed *solved* (Sturm et al., 2014; Sturm, 2016). This kind of evaluation also allows us to acquire new knowledge and advance the field (Serra et al., 2013, p. 31), if experiments are followed by interpretation of results, learning, system improvement, and eventually re-evaluation or even re-definition of the task or the evaluation methodology (**Figure 1**). This is referred to as the *research cycle* (Urbano et al., 2013;

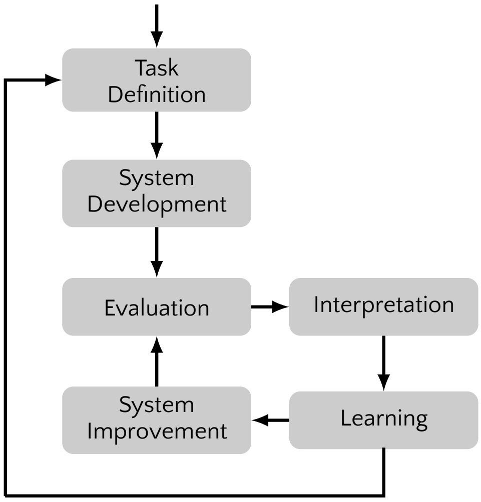


Figure 1: IR research cycle (Urbano et al., 2013).

* International Audio Laboratories Erlangen, DE

† Multimedia Computing Group, Delft University of Technology, NL

Corresponding author: Hendrik Schreiber (hendrik.schreiber@audiolabs-erlangen.de)

Sturm, 2016). For it to succeed, we need to be able to conduct analyses of *all* parts of the evaluation process: task definition, data, metrics, systems, and analysis. As has been pointed out before (Urbano et al., 2013; Sturm, 2013a; Raffel et al., 2014; Salamon et al., 2014), this disqualifies evaluation campaigns with private or secret data and closed source evaluation code. Evaluation itself must follow the same cycle of learning. How we evaluate must be analyzed, questioned, and improved (Urbano et al., 2013; Serra et al., 2013, p. 33). Do datasets and metrics match current use cases? Are there recordings for which no system estimates the correct tempo, or recordings most systems estimate different tempi for? Does that mean the annotation is wrong, the tempo is hard to estimate, or the recording is not suitable for the task? To become aware of and address these issues, we need versioned annotations and publicly archived estimates. A step in this direction was taken by Böck et al. (2019), by publishing annotations and estimates as supplemental material.

We start our investigation in Section 2 with discussion of the relevance of tempo estimation in light of presumed and actual applications (in the general sense, not referring to a specific software program). In Section 3, we review popular metrics with emphasis on construct validity. Then, in Section 4, we present the results of a survey among domain experts, which aimed at finding out which applications are important to them and how they measure success. Based on these results, we propose the formal octave error as a complementary metric in Section 5. Then, in Section 6, we discuss size, quality, composition, and suitability of popular datasets. In Section 7, we propose a public repository for reference annotations, estimates, and evaluation code to help with future evaluations. Finally, in Section 8, we draw conclusions.

Throughout this article we will illustrate some observations with tempo estimates produced by three systems: 🟦 `perc` (Percival and Tzanetakis, 2014), 🟩 `böck` (Böck et al., 2015),[1] and 🟥 `schr` (Schreiber and Müller, 2018b). They were chosen for illustrative purposes, their conceptual differences, and availability, not because they necessarily represent the state of the art.

2. Applications

Even though tempo estimation is a well established MIR task, the existing research rarely discusses in depth why tempo estimation is *relevant* and what the *application requirements* are.

2.1 Research Justifications

Dixon (2001) identifies four main application types for his work on tempo and beat extraction: performance analysis, perceptual modeling, audio content analysis for retrieval, and performance synchronization. Most applications described in later work fall into these four broad categories. Alonso et al. (2003) mention automatic rhythmic alignment of audio, indexing for retrieval, and synchronized computer graphics. Peeters (2007) explicitly adds automatic playlist generation, DJ applications like beat-mixing and looping, and further beat-synchronous analysis (e.g., cover song identification, Ellis and Poliner (2007)). Tzanetakis and Percival (2013) list applications such as music similarity and recommendation, semi-automatic audio editing, automatic accompaniment, polyphonic transcription, beat-synchronous audio effects, and computer assisted DJ systems. Böck et al. (2015) add to this the contribution tempo estimation can make to beat-tracking, such that beats are aligned to a previously estimated tempo. Elowsson and Friberg (2015) consider tempo annotations useful for automated mixing, e.g., for beat-synchronous delay and compressor release settings. Similarly, Font and Serra (2016) mention remixing and browsing as potential applications.

In publications focused on new methods, most application descriptions serve a motivational purpose justifying the conducted research. In fact, even though some of the mentioned applications not only require tempo, but also phase information (e.g., beat-synchronous delay), they all stem from publications primarily (but not necessarily exclusively) about tempo estimation. To the best of our knowledge, no formal application survey for tempo estimation has ever been conducted. Therefore we simply do not know *how relevant* tempo estimation is for any of the mentioned applications and what *requirements* these applications have. Rephrased in terms of commercial engineering: for the past 25 years we have largely ignored the customer. As Salamon (2019) recently observed, 'There is a disconnect between MIR research and potential users of MIR technologies.' This is not to say that the MIR community has conducted the wrong kind of research. After all, it is the privilege of basic research to not require an immediate application, and prefacing each scientific project with a market study is not expedient. But as tempo estimation and MIR as a whole mature, one might want sound justifications as to why and for what research is conducted.

2.2 Presumed Applications

We would like to illustrate the issue with two presumed applications of tempo estimation: similarity and recommendation (Tzanetakis and Percival, 2013; Percival and Tzanetakis, 2014; Böck et al., 2015). By definition, two recordings with the same tempo are similar—at least in this respect. But since similarity has many facets, tempo cannot be the only feature used to predict it. It may not even be very important. In fact, in their introduction to music similarity Knees and Schedl (2016) briefly mention tempo, but do not deem it important enough to thoroughly discuss it. To quantify how important tempo estimation is for music similarity, we counted the number of MIREX submissions for the similarity task that used tempo as a feature.[2] Many submissions used low-level temporal or rhythmic features, but only 8 of 62 (13%) explicitly used a single beats per minute value. One team even removed tempo as feature in a subsequent submission.[3]

Music recommendation is another application mentioned when justifying tempo estimation research. But is tempo estimation really useful for recommendation? Content-based systems certainly *can* take advantage of tempo annotations (Vignoli and Pauws, 2005), but to the best of our knowledge this is not a common approach. Slaney (2011) points out that recommendation based on collaborative filtering usually outperforms content-based systems, if enough usage data is available. Merely in *cold-start* scenarios (e.g., lack of usage data) does content-based recommendation play a noteworthy role. Schedl et al. (2018) report that if content-based recommendation is attempted, 'almost all existing approaches rely on a number of predefined audio features that have been used over and over again, including spectral features, MFCCs, and a great number of derivatives.' This does certainly not exclude tempo, but in their report on current challenges for music recommender research tempo is never mentioned. Therefore, we conjecture that global tempo estimation is only of marginal importance for general similarity and recommendation. It may however still play a role when it comes to specific similarity or recommendation tasks, for example in the context of ballroom dances or physical exercise.

2.3 Actual Applications

On the positive side, there are plenty of existing applications that are very similar to those stated in the literature. Tempo estimation has been used in computational ethnomusicology (Cornelis et al., 2013). Life science researchers who study connections between exercise and music tempo (Waterhouse et al., 2010) and athletes who want to control the tempo of their workout naturally benefit from tempo estimation systems. Consumer applications like beaTunes (https://www.beatunes.com/) provide this information via offline analysis, and streaming services like Spotify (https://www.spotify.com/) or Deezer (https://www.deezer.com/) offer playlists with narrow BPM ranges made for runners. The music store BeatPort (https://www.beatport.com/) labels all its tracks with global BPM and key values to help DJs when shopping. And when performing, DJs can take advantage of tempo analysis and beat-tracking/matching features of their DJ software (e.g., Traktor, https://www.native-instruments.com/). Thus useful applications exist, even though they are typically not the result of user studies or other requirements gathering processes by the MIR community.

3. Metrics

The exemplary evaluation during the 2004 ISMIR conference effectively established the accuracy metrics ACC_1 and ACC_2 as standards. Few subsequent publications explicitly discuss the musical concept of global tempo. Instead, researchers seem to assume that measuring ACC_1 and ACC_2 is identical to measuring global tempo. *De facto*, the metrics have *become* the task definition (Salamon, 2019). The only popular alternative is the P-Score metric.

3.1 Accuracy 1 and 2

ACC_1 computes a 0 or 1 score per track, which indicates the correctness of an estimate, allowing a 4% tolerance. This tolerance is described as 'somewhat arbitrary' (Gouyon et al., 2006). It was not chosen because someone defined an application that required a certain precision, but because it was assumed that the test tracks have 'approximately constant tempi.' This may have been a good choice for traditionally produced music, but seems lenient for electronic music or music produced with modern production techniques like click tracks (Lamere, 2009), and strict for Romantic piano pieces. Attempting to justify the tolerance, Gouyon et al. (2006) argue that according to Friberg and Sundberg (1995) the Just-Noticeable Difference (JND) for music tempi is approximately 4% and therefore '4% is probably the highest precision level that should be considered.'

We unfortunately see problems in this argument. First, Friberg and Sundberg's experiment measured whether participants were able to perceive the non-isochronous placement of the fourth tone in a sequence of six tones. But instead of 4%, they actually found an average JND of 2.5% for tracks with tempi between 60 and 250BPM. Secondly, and more importantly, it is not conclusively explained how this experiment relates to determining the tempo of a 30*s* sample, as was the task during the ISMIR 2004 contest. We therefore do not believe that the results of the experiment are suitable to derive the ACC_1 tolerance parameter. In fact, when plotting ACC_1 for the tempo estimation systems böck, schr, and perc with different tolerances (**Figure 2**), we see that all three systems are capable of estimating tempo for *Ballroom* (Gouyon et al., 2006) tracks with almost the same accuracy at 2% tolerance as they are at 4% tolerance. That said, for datasets with less stable tempi, 4% may be too strict.

This points to issues inherent to binary metrics. The threshold is usually arbitrary, because it cannot be derived in an indisputable, objective way. Furthermore, it hides information. ACC_1 does not tell us *how* wrong an estimate is, nor in which *direction*. This means that we cannot easily plot an error distribution or other descriptive statistics. ACC_1 is also blind to small systematic errors below the threshold. At the same time, it may overemphasize

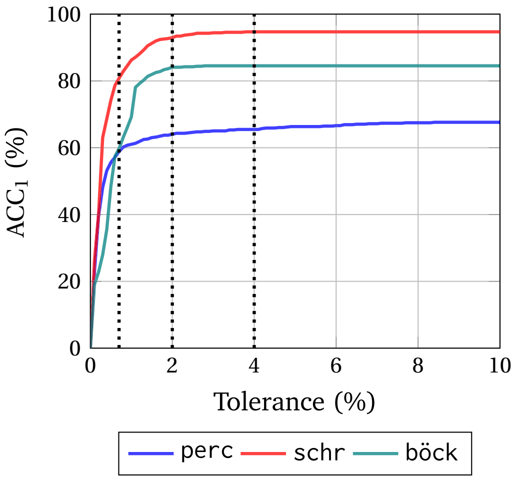


Figure 2: ACC_1 of several tempo estimation systems depending on tolerance measured on *Ballroom* with a ground truth based on beat annotations by Krebs et al. (2013).

differences between systems. As an extreme example, systematic errors of +4.01% and +3.99% may not differ much, but their ACC_1 scores could not be further apart. Specifying the tolerance for ACC_1 in percent may also be questioned. Assuming a fictional tolerance of 50%, a recording may be estimated half as fast, but not twice as fast. Contrary to that, estimating a triple meter recording at half its tempo is arguably less appropriate than at twice its tempo (Elowsson and Friberg, 2015).

ACC_2 additionally allows estimates to be wrong by the factors 2, 3, ½ or ⅓ (so-called *octave errors*). This metrical tolerance was not motivated by application requirements either, but by the realization that the used annotations may not match the perception of human listeners. Unfortunately, because the meter is not taken into account, ACC_2 counts some perceptually erroneous estimates as correct (Gouyon et al., 2006). Consequently, Elowsson and Friberg (2015) regard it as 'inappropriate.' Another limitation of ACC_2 is that it says nothing about a system's ability to help a user to distinguish between *slow* and *fast* tracks. This reduces this metric's usefulness for applications like playlist generation based on tempo continuity or when searching for slow music (Peeters and Flocon-Cholet, 2012). Gärtner (2013) states: 'From the perspective of the user of DJ software, it is absolutely mandatory that the tempo is annotated correctly. The so-called octave errors are unacceptable.' This mismatch between metric and usefulness illustrates that the *construct validity* (Urbano et al., 2013) of ACC_2, i.e., the correlation between use case, success criteria, and the employed metric, is far from perfect *for the mentioned use cases*.

3.2 P-Score

A metric that takes tempo ambiguity into account and treats it as an inherent property of music (Moelants and McKinney, 2004) is the P-Score proposed by Moelants and McKinney for the MIREX audio tempo extraction task in 2005.[4] The original metric incorporated two metrical levels as well as a phase estimate, and considered an estimation system's salience estimation. In 2006 it was simplified to:

$$P = ST1 * TT1 + (1 - ST1) * TT2 \tag{1}$$

Where each track is annotated with two reference tempi, T1 and T2, and T1's relative perceptual strength ST1 $\in [0, 1]$. T1, T2, and ST1 are the result of an expensive process involving many annotators per track. To calculate a P-Score, TT1 $\in \{0, 1\}$ encodes the ability of an estimation system to identify T1 with a tolerance of 8% as boolean value, 0 or 1. TT2 $\in \{0, 1\}$ is defined correspondingly.[5] In addition to the P-Score, 'One Correct' and 'Both Correct' percentages are published for systems participating in MIREX. Because P-Score accounts for ambiguity in human perception and does not reward perceptually erroneous estimates, it is an improvement compared to ACC_2, but still has shortcomings. We were unable to find any formal justification for the used 8% tolerance. According to McKinney, 'the tolerance was derived empirically through the evaluation of a number of excerpts, algorithms and studies. It is somewhat arbitrary […].'[6] Furthermore, since 2006 the metric does not require an estimation system to assign a salience value to its two estimates per track.[7] This means that an application using a system with a perfect P-Score still has to guess which of the two estimates is the more salient one. Just like ACC_2, P-Score does not test the ability of a system to distinguish between *slow* and *fast*. It also is not efficient in the sense that it is relatively expensive to create the necessary ground truth. This might explain why only one other suitable dataset (Schreiber and Müller, 2018a) has been created since the original MIREX dataset in 2005, which itself had been created for an experiment about the perception of tempo and not for MIREX.

4. Survey

To better answer some of the questions raised regarding applications and metrics, we have conducted a survey among domain experts who work or have worked on tempo estimation. In this section, we are highlighting the most important results. Details with graphical depictions are shown in Appendix A and the raw data is available as supplemental material.

Of the 24 individuals who filled out the questionnaire, 17 (71%) belonged to academia and 7 (29%) to the industry. Most participants identified themselves as researchers (92%), and a majority claimed to be involved in hands-on algorithm implementation (71%). We were surprised to learn that, according to participants, none of the usually mentioned applications is most important to them, but to produce 'input for other algorithms.' While 'other algorithms' may include 'recommendation' and 'similarity', neither of these two options was explicitly chosen by any participant. The second most important application is 'performance synchronization.'

Participants from the industry tend to focus their tempo estimation efforts much more on particular genres than those from academia. To industry, the danceable genres Ballroom, EDM/Disco, Hip Hop/Rap, and Reggae are most important (in that order). Classical is only ranked fifth. Contrary to this, those members from academia who target specific genres, ranked Classical first, followed by EDM/Disco, Pop/Rock, and Hip Hop/Rap. Ballroom and Folk were not ranked at all by academics. We speculate that this difference may be related to the respective group's motivation. Academia has already reached very good results for Ballroom music (Böck et al., 2015), which makes it uninteresting, while Classical music might still be seen as a challenge and may appear as more interesting from a musicological point of view. In contrast, the industry is not primarily driven by interestingness, and typical industry applications—like DJ software—focus on dance, not classical music.

Being able to distinguish *slow* from *fast* tracks is very important for the applications of most participants. This appears to be a central requirement. A strong majority of industry applications (71%) also seem to need a single

BPM value rather than a tempo distribution. In academia, this is only true for 57%. Both groups see ACC_1 as a very useful metric when it comes to measuring how well an application meets its requirements or how well a research objective is achieved. For ACC_2 the picture is less clear. While the industry leans toward 'useful,' members of academia gave answers covering almost the entire possible spectrum. This supports our criticism from Section 3.1 regarding the construct validity of ACC_2. When asked about the usefulness of P-Score, the two groups were of very different opinion. Most members of the industry tend to regard P-Score as 'not useful,' while many academics see it as 'essential.' This reveals a big divide between evaluation for industry applications and the scientific evaluation at MIREX.

The survey documents that many industry members are interested in more accurate tempo values than are tolerated by ACC_1 or ACC_2. Among them, the most often demanded accuracy was '2 decimal places.' This must be seen in the context of target genres, which for the industry are more oriented towards dance music, which typically has a very stable tempo. The most popular choice among academics was 'Other.' Here, free-form answers ranged from 'BPM with as small as possible tolerance' over 'no specific application yet' to 'depends on the dataset and the accuracy of the annotations.' The second most popular choices among academics were 'nearest integer' and '2% tolerance.' Regardless of affiliation, no one chose 8%—the tolerance traditionally used at MIREX.[8]

Lastly, while a strong majority (73%) of all participants still regard global tempo estimation as a relevant MIR task, only 57% of industry members believe so. Some of the stated doubts are: 'tempo estimation is good enough for most industrial use cases', 'local tempo estimation is a much more useful task', and 'beat tracking, as a more general task than tempo estimation, solves all problems.'

5. Formal Octave Error

We have argued in Section 3 that the tolerances of ACC_1, ACC_2, and P-Score are difficult to justify and that the binary nature of these metrics hides information. Furthermore, using a percentage as threshold is sub-optimal, and the survey results indicate that there is interest in metrics with lower tolerance, up to an 'as small as possible tolerance.' We therefore propose a complementary metric that measures how close and in which direction an estimate is to a reference value. Inherently, such a metric supports meaningful visual depiction of error distributions. Gouyon et al. (2006) and Peeters (2007) have used such a metric, by showing the $\log_2$ of the ratio between estimates and reference values in histograms. Following them, we formally define the octave error OE_1 as

$$OE_1(y, \hat{y}) = \log_2 \frac{\hat{y}}{y}, \tag{2}$$

with $y, \hat{y} \in \mathbb{R}_{>0}$ as ground truth and estimate. OE_1 is designed to highlight the most important error class, octave errors, in an intuitive way. Errors by factors k and $\frac{1}{k}$ have the same magnitude, which means that in an OE_1 visualization the octave errors 2, ½, 3, and ⅓, are easily identifiable as clusters around 1, −1, 1.58, and −1.58. **Figure 3a** shows examples for OE_1 distributions for *Ballroom* rendered as violin plots.[9] Clearly visible is the concentration around −1 tempo octaves (TO) for all systems but `böck`, `schreiber2017` (Schreiber and Müller, 2017), and `schr`. None of the systems suffer much from the relatively rare octave errors 3 or ⅓ (Peeters, 2007; Schreiber and Müller, 2017). The extent of the horizontal spread of the concentrations around 0 TO visualizes non-octave errors. OE_1 distributions can serve as indicators for the overall performance of a global tempo estimation system including the capability to help distinguish

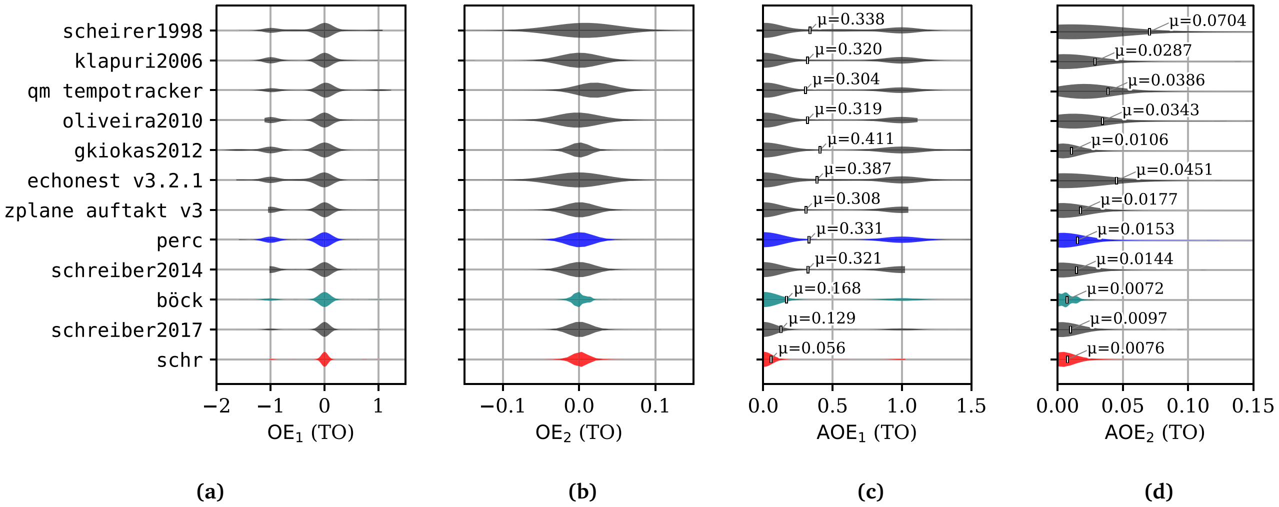


Figure 3: Empirical distributions of **(a)** OE_1, **(b)** OE_2, **(c)** AOE_1, and **(d)** AOE_2 using kernel density estimation (KDE). Based on values measured for *Ballroom* using a *median* ICBI-derived ground truth created from beat annotations by Krebs et al. (2013). Ordered by year of publication (Scheirer, 1998; Klapuri et al., 2006; Davies et al., 2009; Oliveira et al., 2010; Gkiokas et al., 2012; Percival and Tzanetakis, 2014; Schreiber and Müller, 2014; Böck et al., 2015; Schreiber and Müller, 2017, 2018b). Estimates for `zplane` and `echonest` stem from Percival and Tzanetakis (2014).

between *slow* and *fast*. Most importantly, one can see at a glance what kind of errors the tested systems are prone to.

We have seen in our discussion of P-Score that taking tempo ambiguity into account is desirable, but that suitable datasets are rare and new datasets are expensive to create. Furthermore, P-Score has not been adopted by the industry (Section 4). For these pragmatic reasons, we do not attempt to solve the metrical level problem, but define OE_2 similar to ACC_2 as

$$\begin{aligned}&\mathrm{OE}_2\left(y,\hat{y}\right)=\underset{x\in\Omega}{\arg\min}(|x|),\text{ with}\\&\Omega:=\left\{\mathrm{OE}_1\left(y,\hat{y}\right),\mathrm{OE}_1\left(y,2\hat{y}\right),\mathrm{OE}_1\left(y,\tfrac{1}{2}\hat{y}\right),\right.\\&\qquad\left.\mathrm{OE}_1\left(y,3\hat{y}\right),\mathrm{OE}_1\left(y,\tfrac{1}{3}\hat{y}\right)\right\}.\end{aligned}\tag{3}$$

OE_2 (**Figure 3b**) measures accuracy on a micro level, where the most common errors on the metrical level are ignored, i.e., it measures how close the estimate is to the nearest related tempo.[10] This is useful for genres with high tempo ambiguity, e.g., Dubstep (Schreiber and Müller, 2018a), and for applications that require errors to be as small as possible. The latter is a use case currently unsupported by ACC_1 and ACC_2, but desired by the industry (Section 4).

While the mean of OE_1 or OE_2 indicates whether an algorithm is expected to over- or underestimate the tempo, the *absolute* octave error (AOE = |OE|) can be used for system comparisons. To illustrate, **Figure 3c** shows annotated AOE_1-distributions. Most older systems have an average AOE_1 between 0.3 and 0.4TO, `böck` managed to halve this figure, and `schr` further reduced it to 0.056TO. When ignoring octave errors by using AOE_2 (**Figure 3d**), we can see that `böck` and `schr` perform on a similar level.

Note that though the mean AOE is informative, we recommend also reporting a distribution for a more complete picture.

6. Datasets

Evaluations of tempo estimation systems rely on datasets consisting of suitable recordings and annotations that model what we want to measure. Without claim to completeness, **Table 1** lists popular tempo datasets. Unfortunately, some of these datasets are relatively small, focus on a particular genre, are not freely available (any more), or have other flaws like duplicates, mislabelings, and distortions (Sturm, 2013b, 2014; Salamon, 2019).

6.1 Dataset Size

To reliably measure differences between systems, a dataset must be sufficiently large to minimize the effect of random variation due to the sampling of tracks it contains. Generalizability Theory (GT) offers a statistical tool to estimate the required size for performance assessments in general (Cronbach et al., 1963; Brennan, 2003; Bodoff, 2008; Carterette et al., 2009; Salamon and Urbano, 2012; Bosch et al., 2016). Essentially, the GT framework decomposes the variability in the observed scores into variability due to actual differences between systems (σ_s^2), variability due to differences in track difficulty (σ_t^2), and residual variability (σ_e^2), which often refers to system-track interactions. The total variance of the observed scores is therefore modeled as:

$$\sigma^2=\sigma_s^2+\sigma_t^2+\sigma_e^2.\tag{4}$$

An evaluation with high σ_s^2 does not require large datasets, because the evaluated systems are very different to begin

Table 1: Popular public tempo datasets.

Dataset	Recordings	Tempo Ann.	Beat Ann.
ISMIR04 Songs (Gouyon et al., 2006)[1]	464	BPM	No
Ballroom (Gouyon et al., 2006; Krebs et al., 2013)[1]	698	BPM	Yes
RWC-C (Goto et al., 2002)[2]	50	BPM	Yes
RWC-G (Goto et al., 2003)[2]	100	BPM	Yes
RWC-J (Goto et al., 2002)[2]	50	BPM	Yes
RWC-P (Goto et al., 2002)[2]	100	BPM	Yes
RWC-R (Goto et al., 2002)[2]	15	BPM	Yes
GTzan (Tzanetakis and Cook, 2002; Marchand and Peeters, 2015)[1]	999	BPM	Yes
Hainsworth (Hainsworth, 2004)[1]	222	BPM	Yes
ACM Mirum (Peeters and Flocon-Cholet, 2012)[1]	1,410	BPM	No
SMC (Holzapfel et al., 2012)[1]	217	BPM	Yes
GiantSteps Tempo (Knees et al., 2015; Schreiber and Müller, 2018a)[3]	664	BPM/T1,T2,ST1	No
Extended Ballroom (Marchand and Peeters, 2016)[1]	4,180	BPM	No
LMD Tempo (Raffel, 2016; Schreiber and Müller, 2018b)[4]	3,611	BPM	No

[1] Excerpts available. [2] Requires application and purchase. [3] BeatPort previews, cached versions available from JKU. [4] 7Digital previews available.

with, but evaluations with high σ_t^2 or high σ_e^2 do require large datasets, because systems tend to perform similarly for the given tracks.

There are several coefficients in GT, but here we will report only the dependability index $\Phi \in [0, 1]$, which measures the ratio of system variance to itself plus error variance (Brennan, 2003):

$$\Phi = \frac{\sigma_s^2}{\sigma_s^2 + \frac{\sigma_t^2 + \sigma_e^2}{M}}, \qquad (5)$$

where M is the size of the dataset. A high Φ-value means that the dataset can reliably separate actual differences among systems from random variation due to sampling of tracks. Φ-values greater than 0.95 are generally considered high enough, but because this is rather arbitrary we focus more on qualitative comparisons among datasets and metrics.

We estimated Φ through an Analysis of Variance (ANOVA) for the datasets *ISMIR04 Songs*, *Hainsworth*, *GTzan*, *Ballroom*, *SMC*, *RWC* (here, the union of *RWC-C*, *RWC-G*, *RWC-J*, *RWC-P*, and *RWC-R*), and *GiantSteps Tempo* (**Figure 4, a–g**) using scores from five different systems (Davies et al., 2009; Percival and Tzanetakis, 2014; Böck et al., 2015; Schreiber and Müller, 2017, 2018b), closely following the approach described in Salamon and Urbano (2012). **Figure 4** shows $\hat{\Phi}$ as a function of the number of songs M, which lets us determine how many songs would be necessary for a reliable evaluation. The actual number of songs in the respective dataset is indicated by a vertical and the 0.95 reliability level by a horizontal dotted line. In other words, for a large enough dataset, $\hat{\Phi}$ should pass through the upper left quadrant (colored in pale orange). Using this criterion, only *ISMIR04 Songs*, *Ballroom*, and *GiantSteps Tempo*, are large enough to reliably differentiate system performance for the tested algorithms when using ACC_1 or ACC_2. In all cases but *GiantSteps Tempo*, both OE_1 and AOE_1 lead to similar or better $\hat{\Phi}$-values than ACC_1, i.e., we reach a greater reliability level for the given dataset. In fact, all seven tested datasets are large enough to reach the 0.95 threshold when using OE_1 as metric. For OE_2 and AOE_2 the picture is not quite as clear—for some datasets, like *GTzan* and *Ballroom*, they reach higher $\hat{\Phi}$-values than ACC_2, for others, like *SMC*, lower values.

In **Figure 4h**, we show an evaluation of the *MIREX* dataset (McKinney et al., 2007) based on the published MIREX 2018 results.[11] ‘One Correct’ reaches $\hat{\Phi} = 0.95$, P-Score reaches $\hat{\Phi} = 0.92$, but ‘Both Correct’ only $\hat{\Phi} = 0.67$, this means that the *MIREX* dataset is close to being large enough for P-Score but certainly not for ‘Both Correct.’

Note that all reported $\hat{\Phi}$-values depend on the tested systems. Removing older, worse performing systems from the evaluation may actually lower the $\hat{\Phi}$-value.

6.2 Annotation Quality

Serra (2014) states that among other aspects quality is an important criterion when creating research corpora. The audio has to be of high quality and annotations have to

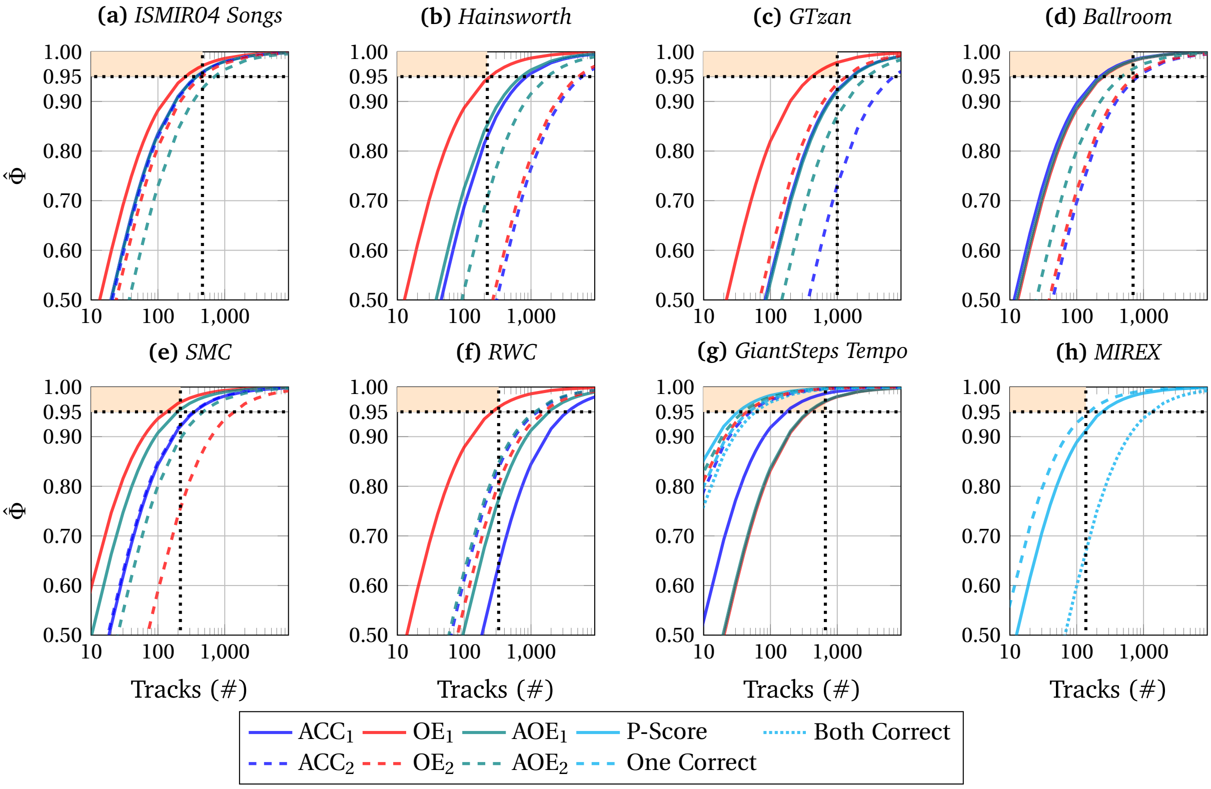


Figure 4: Dependability index $\hat{\Phi}$ as function of metric and track count. Vertical dotted line: actual number of tracks in dataset. Horizontal dotted line: $\hat{\Phi} = 0.95$. Desired quadrant shaded in pale orange. **(a–g)** $\hat{\Phi}$ based on estimates Davies et al. (2009); Percival and Tzanetakis (2014); Böck et al. (2015); Schreiber and Müller (2017, 2018b). **(h)** $\hat{\Phi}$ based on MIREX 2018 results.

be accurate. Ten years after *Ballroom* had been used for the first time, Percival and Tzanetakis (2014) investigated the accuracy of the annotations and corrected 32 (4.6%) of them. Corrections were also made to *ACM Mirum* (135, 9.6%) and *GTzan* (24, 2.4%). Interestingly, Percival and Tzanetakis emphasize the importance of using correct annotations, because testing systems on faulty data may lead researchers to optimize for these errors. This fear might be indicative for the state of MIR at the time. Machine learning was not ubiquitous yet and tuning hyperparameters using the test set was not perceived as quite the methodological faux-pas it is seen as now. But there are other good reasons to strive for quantifiable quality in test datasets: interpretability and comparability. If the quality of a test dataset is unknown, a metric like accuracy can at best be used to approximate the lower bound of a system's true performance. At worst it is simply useless. It is impossible to say whether any changes to the system can still increase performance. Additionally, it is impossible to compare results for different datasets in a meaningful way, if the dataset quality is unknown.

Schreiber and Müller (2018a), for example, noticed the fairly low ACC_2 performance of state-of-the-art tempo estimation systems on the original annotations of the *GiantSteps Tempo* dataset, and conducted a crowdsourced experiment to create a new ground truth. When comparing the performance of `böck` on the original annotations with the performance in the new annotations, ACC_1 jumps from 58.9% to 64.8% and ACC_2 from 86.4% to 94.0%.

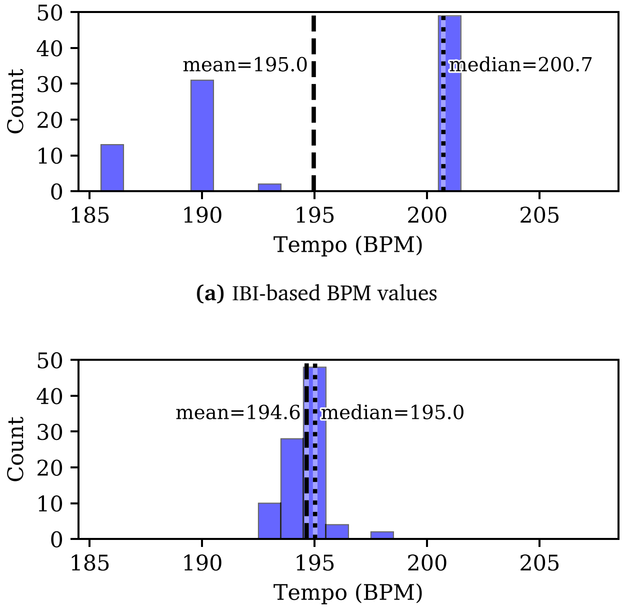
6.3 Modeling Global Tempo

It is well known that some of the tracks in popular datasets have varying tempi (Hainsworth, 2004; Peeters, 2007; Percival and Tzanetakis, 2014). To address this issue, Hainsworth defined the tempo for the tracks in his dataset as the *mean* of the Inter-Beat Intervals (IBI). Percival and Tzanetakis (2014) suggested using the *median* instead, to counter the influence of outliers—an idea already used by Peeters (2007) and Oliveira et al. (2010). Böck et al. (2015) followed this suggestion, but to the best of our knowledge did not publish their annotations. Subsequent publications still used the original *mean*-based annotations (Schreiber and Müller, 2017) or tempo values obtained in some other way. For example, Elowsson (2016) derived tempi from the peaks of smoothed IBI histograms.

In addition to changing tempi, some datasets (Hainsworth, 2004; Marchand and Peeters, 2015) contain recordings with microtiming variations. One may argue that for such recordings neither the mean nor the median IBI is an ideal solution, because the beats are not necessarily isochronous. As a result, one may see multiple peaks in an IBI histogram. For example, the IBI-based BPM histogram for the *GTzan* recording `jazz.00053` (**Figure 5a**) shows distinct peaks at 186, 190 and 201BPM even though the tempo of the track does not change over time. Choosing the median of the IBIs (200.7BPM) ignores the lower peaks at 186 and 190BPM. If we know a track's meter, we therefore may rather use the *median* of the intervals between *corresponding* beats, i.e., the intervals between beats that occur at the same position in subsequent measures divided by the number of beats per measure. Using this Inter-Corresponding-Beat Interval (ICBI) for tempo calculation, we can neutralize effects of variations in microtiming as well as outliers (**Figure 5b**).

6.4 Dataset Suitability

While improving and versioning annotations is commendable, it does not ensure that the dataset fits the use case. Obviously, if the use case focuses on Ballroom, using a Reggae dataset for testing is the wrong approach. Similarly, if a metric is chosen that was designed for a certain use case, which may imply a certain kind of music, one must ensure that it is suitable for the actually used kind of music (**Figure 6**). As pointed out above, a precondition for using ACC_1 and ACC_2 with 4% tolerance is a stable tempo in each test track. We can visualize whether this precondition is met for a dataset by converting IBIs to normalized tempi and plotting their distribution. Concretely, given a track's IBIs $b = \{b_0, b_1, \ldots, b_{N-1}\}$ in seconds with $b_n \in \mathbb{R}_{>0}$ and the



(a) IBI-based BPM values

(b) ICBI-based BPM values

Figure 5: Histograms of BPM values for *GTzan* `jazz.00053` based on **(a)** IBIs and **(b)** ICBIs.

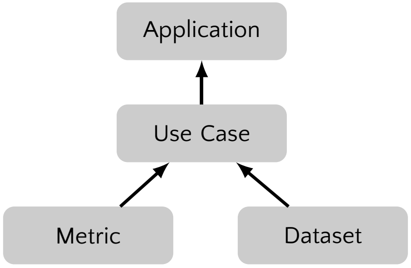


Figure 6: Dependencies between application, use case, metric, and dataset (an arrow from *A* to *B* denotes that *A* depends on *B*).

number of IBIs $N \in \mathbb{N}_{>0}$, we define its (local) tempo values $t = \{t_0, t_1, ..., t_{N-1}\}$ in BPM as

$$t_n = \frac{60}{b_n}. \tag{6}$$

The normalized tempi $t^{\text{norm}} = \{t_0^{\text{norm}}, t_1^{\text{norm}}, \ldots, t_{N-1}^{\text{norm}}\}$ for *each track* are then defined as:

$$t_n^{\text{norm}} = \frac{t_n}{\frac{1}{N}\sum_{i=0}^{N-1} t_i} \tag{7}$$

Figure 7 depicts distributions of t^{norm}. For *SMC*, only half the normalized local tempi fall into the ±4% interval [0.96,1.04] (shown in gray). For *Hainsworth*, it is 75.6%. *GTzan* and *Ballroom* have values of 91% or more and are thus much better suited for ACC$_1$ and ACC$_2$.

To get an impression of how many tracks in a dataset have large tempo variability, we can use the standard deviation σ of the normalized tempi t^{norm}—also known as the coefficient of variation c_{var}:

$$c_{\text{var}}(t) = \frac{\sigma(t)}{\mu(t)} = \sigma\left(t^{\text{norm}}\right). \tag{8}$$

Figure 8 shows the percentage of tracks for which $c_{\text{var}}(t) < \tau$. with $\tau \in [0,0.5]$. Among the shown datasets, *SMC* contains the highest percentage of tracks with large tempo variability. For only 61.3% of the tracks is $c_{\text{var}}(t) < \tau$. In contrast, $c_{\text{var}}(t)$ is less than 0.1 for 99.4% of all *Ballroom* tracks. This affects accuracy. To demonstrate, we measure ACC$_2$ using `böck`, `schr`, and `perc` for subsets of the datasets containing only tracks with $c_{\text{var}}(t) < \tau$, $\tau \in [0,0.5]$.[12] The used tempo annotations are based on *median* IBI-values. For *SMC* (**Figure 9a**), all three systems reach higher scores at $\tau = 0.1$ than for greater τ. Comparing ACC$_2$ for $\tau = \infty$ to $\tau = 0.1$, accuracy increases for `böck` by 18.4 pp, for `schr` by 11.5 pp, and for `perc` by 10.0 pp. For *Hainsworth* (**Figure 9b**) the systems also achieve higher scores at $\tau = 0.1$, but not as much in absolute numbers. For *GTzan* (**Figure 9c**) the increase is still a little smaller, and for *Ballroom* (**Figure 9d**) there is none, because almost all tracks have small $c_{\text{var}}(t)$. This relationship between τ and ACC$_2$ reveals that of the four datasets only *Ballroom* is suitable for

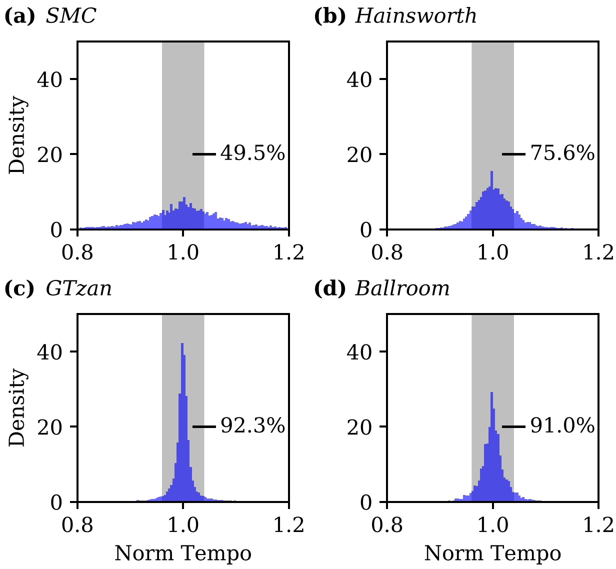


Figure 7: Distributions of normalized tempi. The gray area marks the interval [0.96,1.04]. The shown percentage is the fraction of normalized tempi within the interval.

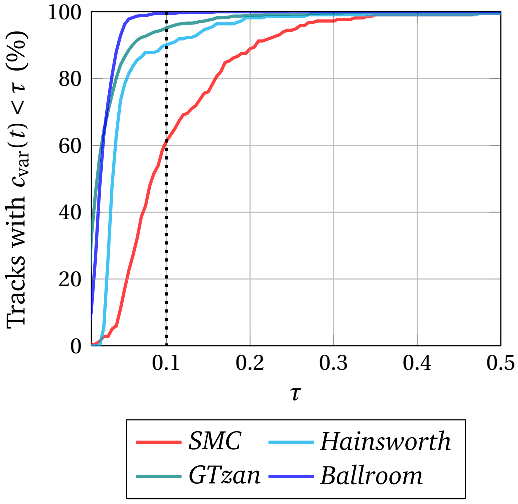


Figure 8: Percentage of tracks with $c_{\text{var}}(t) < \tau$.

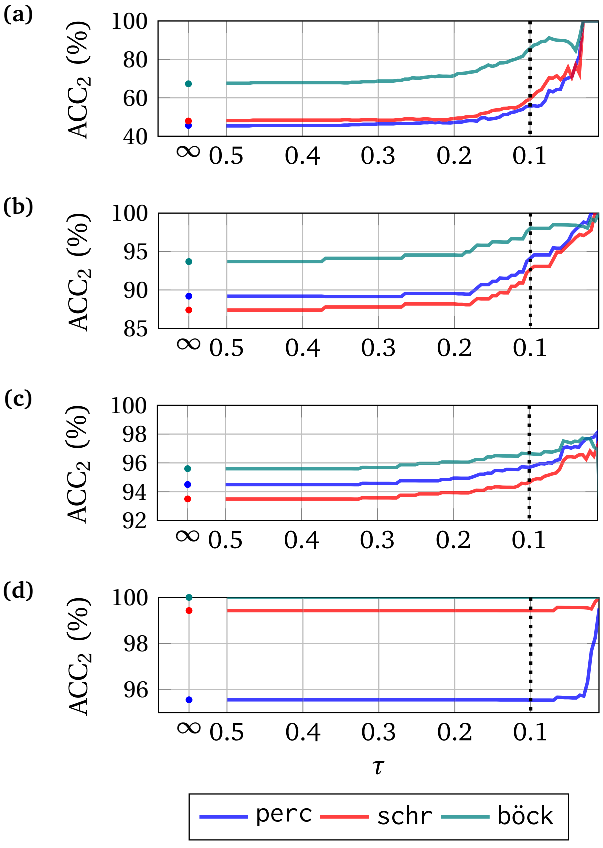


Figure 9: ACC$_2$ for tracks with $c_{\text{var}}(t) < \tau$. Lower τ coincides with higher accuracy. Datasets: **(a)** *SMC* **(b)** *Hainsworth* **(c)** *GTzan* **(d)** *Ballroom*. Different y-scales used for clarity.

ACC_2 (and thus ACC_1) without reservations, because it meets the required degree of stability.

7. Public Repository

To help overcome issues like opaque one-figure evaluations with binary metrics, differently derived annotations, closed source evaluation code, and the inability to evaluate the evaluation, we have created a public GitHub repository called `tempo_eval` (https://tempoeval.github.io/tempo_eval/) that hosts different versions of corpus annotations (Section 7.1), estimates for these corpora (Section 7.2), and evaluation code that goes beyond single figure binary metrics. It provides a basis for the needed collaborative improvement of data and metrics. Section 7.3 demonstrates how the repository can be used for evaluation.

7.1 Reference Annotations

The `tempo_eval` repository allows the continuous improvement of reference annotations without shadowing past versions. This makes it possible to evaluate against all reference versions, improving comparability to older published results and thus transparency as well as interpretability. To provide easy access to reference data we converted published annotations to JAMS (Humphrey et al., 2014) for which tools already exist (Raffel et al., 2014).

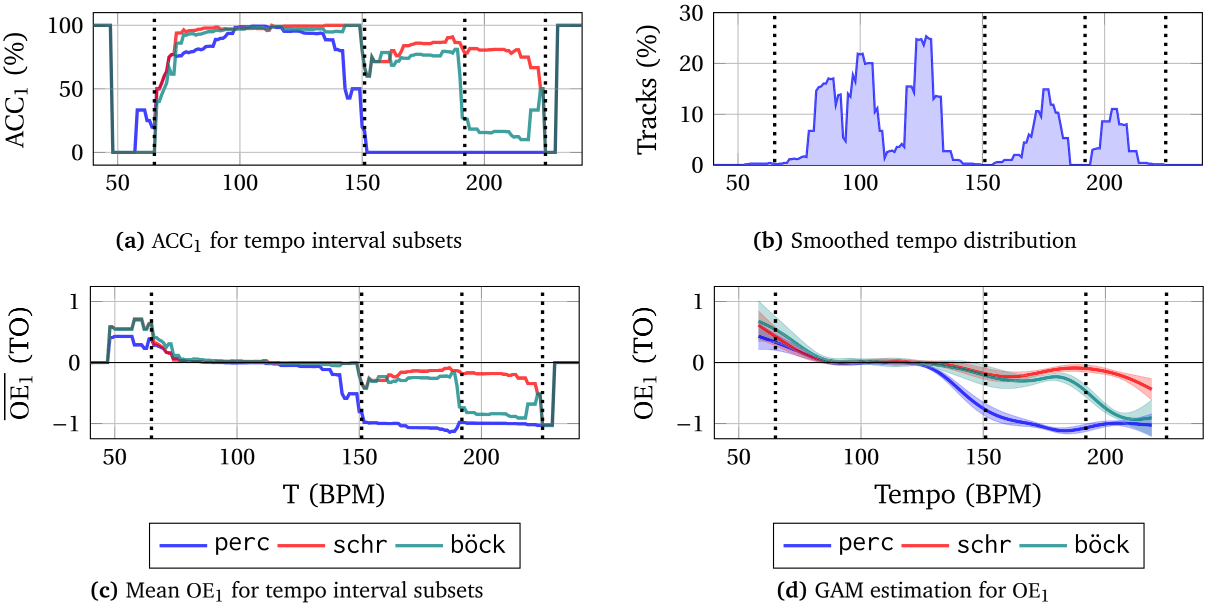
7.2 Estimated Annotations

Rather than just serving as a static source of reference data, the `tempo_eval` repository offers a place for researchers to publish and archive their algorithms' estimates instead of just mentioning single value metrics in their publications. This allows re-evaluation with new and old reference annotations and proper development of new metrics, which may ultimately lead to a better understanding of tempo estimation systems and the tempo estimation task. For example, **Figure 3** shows values for a proposed metric (Section 5) for historic estimates measured against a ground truth, which has been newly derived from *median* ICBI-values. Because the repository is open and public, contributing is easy, e.g., via pull requests. As a starting point, we have added estimates by many recent and classic systems for commonly used datasets.

7.3 Evaluation Code

The `tempo_eval` repository also contains evaluation code. Implemented are ACC_1, ACC_2, and P-Score, along with McNemar's test for significant differences for ACC_1 and ACC_2, OE_1, OE_2, their corresponding absolute incarnations, and t-tests for estimates from algorithm pairs. Results can be rendered in a publishable report (Markdown/HTML), and figures and data are exportable in several formats. As argued above, reporting single value metrics is not sufficient for an in-depth evaluation. We have therefore implemented visualizations for system performance depending on tolerances (**Figure 2**), tempo stability (**Figure 9**), tempo range (**Figure 10**), and—if available—genre- or free-form-tags (**Figure 11**). As an example, we will discuss tempo- and genre-dependent evaluation using the *Ballroom* dataset with annotations from Percival and Tzanetakis (2014).

Figure 10a shows ACC_1 values for subsets defined by tempo ranges [T – 10, T + 10] BPM. Clearly visible, `perc's` ACC_1 drops to zero for $T > 150$BPM, and `böck's` ACC_1 sharply decreases to 27.3% or less for $T > 190$BPM. Both systems seem to exhibit some form of octave bias (Schreiber and Müller, 2017), i.e., the ability to estimate the tempo appears tied to certain tempo ranges. **Figure 10c** depicts mean OE_1 values for the same scenario and shows what kind of errors lead to the observed low accuracy. Apparently, `perc` suffers from octave errors of –1 TO for $T > 150$BPM. The same is true for `böck` and



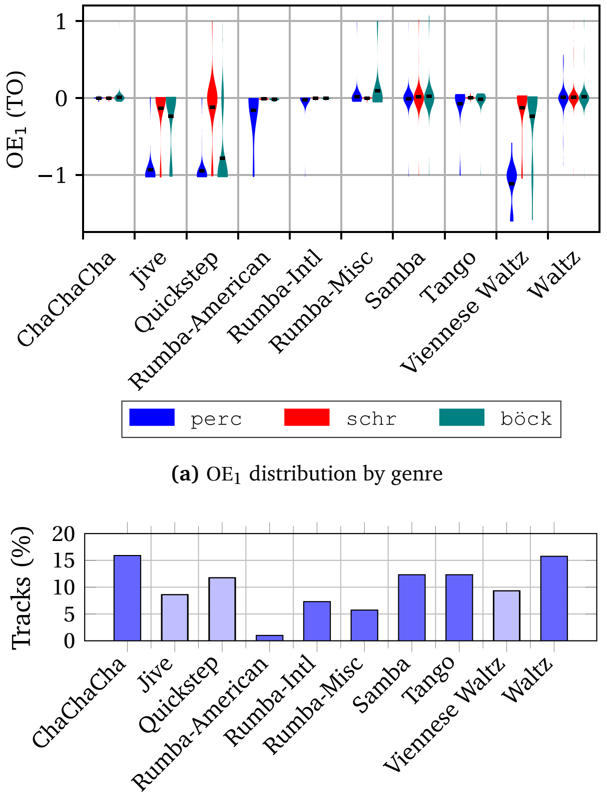
(a) ACC_1 for tempo interval subsets

(b) Smoothed tempo distribution

(c) Mean OE_1 for tempo interval subsets

(d) GAM estimation for OE_1

Figure 10: (a), **(c)** ACC_1 and mean OE_1 for T ± 10BPM intervals. **(b)** Smoothed tempo distribution of tracks in *Ballroom* according to the ground truth from Percival and Tzanetakis (2014). **(d)** OE_1 predictions of generalized additive models (GAM). Shaded areas correspond to 95% confidence intervals.



(a) OE_1 distribution by genre

(b) Genre distribution

Figure 11: (a) Per genre OE_1 distributions based on kernel density estimation (KDE) for tracks from *Ballroom* using the ground truth from Percival and Tzanetakis (2014). Mean OE_1 values are marked in black. **(b)** Genre distribution in *Ballroom*.

T > 190BPM. None of the systems seem to do well for tracks with T < 66 BPM or T > 225BPM, but as we can see in **Figure 10b**, the dataset contains only very few songs in these tempo ranges. **Figure 10d** combines error magnitude, error direction, and significance in a single graph. It shows the predictions and their 95% confidence interval of generalized additive models (GAMs) fitted on the respective systems' OE_1 results. A large confidence interval indicates tempo regions with few samples or large variability in performance. In **Figure 10d** this can be seen for less than 75BPM (few tracks), around 150BPM (performance starts to shift), and for more than 210BPM (few tracks, low performance).

Because JAMS supports additional annotations like genre, tags, and beat positions, these can be incorporated into the evaluation. For example, **Figure 11a** shows OE_1 distributions by genre. Mostly due to −1TO octave errors, `perc` does poorly on Jive, Quickstep, and Viennese Waltz—the three genres with the highest average tempo. `böck` faces the same issue with Quickstep. This is noteworthy, because Jive, Quickstep, and Viennese Waltz combined make up almost 30% of the *Ballroom* dataset, as shown in **Figure 11b** (light-blue bars).

Note that evaluation by ballroom genre is just an example. The code picks up on any JAMS annotation declared in the `tag_open` namespace.

8. Conclusions

In this article we asked the question whether the task of global tempo estimation is solved yet. To find out, we investigated what applications global tempo estimation is used for, discussed currently used metrics, analyzed popular datasets with emphasis on tempo stability and size, and presented the results of a survey among domain experts. We found that applications and use cases for global tempo estimation are somewhat ill-defined, the binary nature of ACC_1 and ACC_2 is problematic and the metrics are not suitable for some use cases, the construct-validity of ACC_2 is questionable, the industry has not adopted P-Score, and that some currently used datasets are too small or do not have a tempo that is stable enough for ACC_1 and ACC_2. Because of these issues, our answer to the opening question, whether the task of global tempo estimation is solved yet, is *no*. Not because estimation systems are not good enough—we do not really know whether that is the case or not, but because it is impossible to solve a task for which neither use cases with success criteria have been well motivated and properly defined, nor the suitability of metrics or datasets has been shown.

Going forward, we need to recognize that global tempo estimation is a task serving different possible applications, each with its own accuracy requirements. Performance synchronization may need as accurate a tempo estimate as possible, while a general musicological interest, playlist building, or some other downstream algorithm may only require a rough estimate or tempo markings like *andante* and *allegro*. Actually achievable accuracy depends on tempo stability, on how tempo is modeled, and annotations are derived. ACC_1 and ACC_2 with their fixed 4% tolerance do neither different accuracy requirements nor tempo stability levels justice. We therefore recommend using the complementary OE metrics, which do not suffer from this limitation and deliver meaningful results for music with different degrees of tempo stability. If reporting ACC_1 and ACC_2 is a necessity, one might also want to plot results for tolerance ranges (**Figure 2**). In accordance with the industry and despite its popularity among scholars, we see no practical use for P-Score, until larger datasets with the required annotations become available.

Almost regardless of metric or use case, we recommend not to use *Hainsworth* or the combined *RWC* datasets. Even though technically an evaluation with OE_1 is possible, they are too small for metrics that allow easy summarization like ACC_1 or AOE_1. Because of its borderline size, we also do not recommend *GTzan*. Due to its large tempo instabilities and small size, the *SMC* dataset should probably only be used to evaluate for low accuracy use cases using OE_1 or AOE_1, if at all. Of the tested datasets, we endorse using *ISMIR04 Songs*, *Ballroom*, and *GiantSteps Tempo*, if appropriate for the use case. To ensure comparable evaluations, we suggest using open source code like `mir_eval` or `tempo_eval`. All estimates and used annotations should be published, to improve reproducibility of the evaluation. The `tempo_eval` repository is meant as a home for this. Since annotations often exist in different versions, we explicitly warn against comparisons with accuracy figures reported by others.

Finally, to actually solve the task, we must clearly define use cases and success criteria (e.g., ballroom dance tournament, 99.9% accuracy, integer precision), choose the appropriate metrics for the use case, and—if not available—curate suitable datasets. Only then an estimation system may succeed at solving whatever the resulting task may be.

Notes

[1] Estimates produced using *madmom TempoDetector 2016 version 0.17.dev0*.

[2] MIREX 2006 to 2014.

[3] Because some links on the MIREX website are broken, we were unable to check all 74 distinct submissions. Some teams submitted multiple algorithms in a given year. Re-submissions were ignored.

[4] http://www.music-ir.org/mirex/wiki/2005:Audio_Tempo_Extraction.

[5] http://www.music-ir.org/mirex/wiki/2006:Audio_Tempo_Extraction.

[6] Private correspondence.

[7] Confusingly, the MIREX tempo estimation task requires estimation systems to estimate the salience, but has not used it in any evaluation since 2005.

[8] In 2018, an additional evaluation with 4% was conducted.

[9] We were unable to obtain either implementations of the approaches taken by Elowsson (2016) and Foroughmand and Peeters (2019) or estimates of their systems for the *Ballroom* dataset. We also do not include Böck et al. (2019) here, because their *Ballroom* results were apparently achieved using cross-validation during training (genre bias).

[10] The criticism voiced about ACC_2 in Section 3.1 obviously applies to OE_2 as well.

[11] Data from https://nema.lis.illinois.edu/nema_out/mirex2018/results/ate/mck/files.html. Based on 137 tracks, since some estimates for three tracks are missing.

[12] ACC_2 is appropriate, because the question of suitability does not hinge on octave errors.

Additional Files

The additional files for this article can be found as follows:

- **Appendix A.** Illustrated Survey Results. DOI: https://doi.org/10.5334/tismir.43.s1
- **Raw Data.** The raw survey results. DOI: https://doi.org/10.5334/tismir.43.s2

Acknowledgements

This work was supported by the German Research Foundation (DFG MU 2686/10-1) and European Union's H2020 programme (770376-2–TROMPA). The International Audio Laboratories Erlangen are a joint institution of the Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU) and Fraunhofer Institute for Integrated Circuits (IIS).

We would like to thank participants of our survey, G. Percival for kindly allowing us to use his results (Percival and Tzanetakis, 2014) to help start the estimate repository, M. Goto for permitting us to redistribute AIST Annotation for the *RWC* Music Database (Goto et al., 2002, 2003), and A. Klapuri and T. Virtanen for releasing beat annotations for the *Klapuri* dataset (Klapuri et al., 2006).

Competing Interests

The authors have no competing interests to declare.

References

Allen, P. E., & **Dannenberg, R. B.** (1990). Tracking musical beats in real time. In *Proceedings of the International Computer Music Conference*, pages 140–143. International Computer Music Association.

Alonso, M., David, B., & **Richard, G.** (2003). A study of tempo tracking algorithms from polyphonic music signals. In *4th COST 276 Workshop*.

Bock, S., Davies, M. E., & **Knees, P.** (2019). Multitask learning of tempo and beat: Learning one to improve the other. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 486–493. Delft, The Netherlands.

Bock, S., Krebs, F., & **Widmer, G.** (2015). 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. Malaga, Spain.

Bodoff, D. (2008). Test theory for evaluating reliability of IR test collections. *Information Processing & Management*, *44*(3), 1117–1145. DOI: https://doi.org/10.1016/j.ipm.2007.11.006

Bosch, J. J., Marxer, R., & **Gomez, E.** (2016). Evaluation and combination of pitch estimation methods for melody extraction in symphonic classical music. *Journal of New Music Research*, *45*(2), 101–117. DOI: https://doi.org/10.1080/09298215.2016.1182191

Brennan, R. L. (2003). Generalizability theory. *Journal of Educational Measurement*, *40*(1), 105–107. DOI: https://doi.org/10.1111/j.1745-3984.2003.tb01098.x

Carterette, B., Pavlu, V., Kanoulas, E., Aslam, J. A., & **Allan, J.** (2009). If I had a million queries. In *European Conference on Information Retrieval*, pages 288–300. DOI: https://doi.org/10.1007/978-3-642-00958-7_27

Cornelis, O., Six, J., Holzapfel, A., & **Leman, M.** (2013). Evaluation and recommendation of pulse and tempo annotation in ethnic music. *Journal of New Music Research*, *42*(2), 131–149. DOI: https://doi.org/10.1080/09298215.2013.812123

Cronbach, L. J., Rajaratnam, N., & **Gleser, G. C.** (1963). Theory of generalizability: A liberalization of reliability theory. *British Journal of Statistical Psychology*, *16*(2), 137–163. DOI: https://doi.org/10.1111/j.2044-8317.1963.tb00206.x

Davies, M. E., Plumbley, M. D., & **Eck, D.** (2009). Towards a musical beat emphasis function. In *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WAS-PAA)*, pages 61–64. New Paltz, NY, USA. IEEE. DOI: https://doi.org/10.1109/ASPAA.2009.5346462

Dixon, S. (2001). Automatic extraction of tempo and beat from expressive performances. *Journal of New Music Research*, *30*(1), 39–58. DOI: https://doi.org/10.1076/jnmr.30.1.39.7119

Downie, J. S. (2008). The music information retrieval evaluation exchange (2005–2007): A window into music information retrieval research. *Acoustical Science and Technology*, *29*(4), 247–255. DOI: https://doi.org/10.1250/ast.29.247

Ellis, D. P., & **Poliner, G. E.** (2007). Identifying 'cover songs' with chroma features and dynamic pro-gramming beat tracking. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* volume 4, Honolulu, Hawaii, USA. DOI: https://doi.org/10.1109/ICASSP.2007.367348

Elowsson, A. (2016). Beat tracking with a cepstroid invariant neural network. In *Proceedings of the 17th International Society for Music Information Retrieval Conference (ISMIR)*, pages 351–357. New York, NY, USA.

Elowsson, A., & **Friberg, A.** (2015). Modeling the perception of tempo. *Journal of the Acoustical Society of America*, *137*(6), 3163–3177. DOI: https://doi.org/10.1121/1.4919306

Font, F., & **Serra, X.** (2016). Tempo estimation for music loops and a simple confidence measure. In *Proceedings of the 17th International Society for Music Information Retrieval Conference (ISMIR)*, pages 269–275. New York, NY, USA.

Foroughmand, H., & **Peeters, G.** (2019). Deep-rhythm for tempo estimation and rhythm pattern recognition. In *Proceedings of the International Society for Music Information Retrieval Conference (IS-MIR)*, pages 636–643. Delft, The Netherlands.

Friberg, A., & **Sundberg, J.** (1995). Time discrimination in a monotonic, isochronous sequence. *Journal of the Acoustical Society of America*, *98*(5), 2524–2531. DOI: https://doi.org/10.1121/1.413218

Gartner, D. (2013). Tempo detection of urban music using tatum grid non negative matrix factorization. In *Proceedings of the 14th International Society for Music Information Retrieval Conference (ISMIR)*, pages 311–316. Curitiba, Brazil.

Gkiokas, A., Katsouros, V., Carayannis, G., & **Stafylakis, T.** (2012). Music tempo estimation and beat tracking by applying source separation and metrical relations. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Kyoto, Japan. DOI: https://doi.org/10.1109/ICASSP.2012.6287906

Goto, M., Hashiguchi, H., Nishimura, T., & **Oka, R.** (2002). RWC music database: Popular, classical and jazz music databases. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR).* Paris, France.

Goto, M., Hashiguchi, H., Nishimura, T., & **Oka, R.** (2003). RWC music database: Music genre database and musical instrument sound database. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, pages 229–230. Baltimore, MD, USA.

Goto, M., & **Muraoka, Y.** (1994). A beat tracking system for acoustic signals of music. In *Proceedings of the Second ACM International Conference on Multimedia*, pages 365–372. San Francisco, CA, USA. DOI: https://doi.org/10.1145/192593.192700

Gouyon, F., Klapuri, A. P., Dixon, S., Alonso, M., Tzane- takis, G., Uhle, C., & **Cano, P.** (2006). An experimental comparison of audio tempo induction algorithms. *IEEE Transactions on Audio, Speech, and Language Processing*, *14*(5), 1832–1844. DOI: https://doi.org/10.1109/TSA.2005.858509

Hainsworth, S. W. (2004). *Techniques for the Automated Analysis of Musical Audio.* PhD thesis, University of Cambridge, UK.

Holzapfel, A., Davies, M. E., Zapata, J. R., Oliveira, J. L., & **Gouyon, F.** (2012). Selective sampling for beat tracking evaluation. *IEEE Transactions on Audio, Speech, and Language Processing*, *20*(9), 2539–2548. DOI: https://doi.org/10.1109/TASL.2012.2205244

Humphrey, E. J., Salamon, J., Nieto, O., Forsyth, J., Bittner, R. M., & **Bello, J. P.** (2014). JAMS: A JSON annotated music specification for reproducible MIR research. In *Proceedings of the 15th International Society for Music Information Retrieval Conference (ISMIR)*, pages 591–596. Taipei, Taiwan.

Klapuri, A. P., Eronen, A. J., & **Astola, J.** (2006). Analysis of the meter of acoustic musical signals. *IEEE Transactions on Audio, Speech and Language Processing*, *14*(1), 342–355. DOI: https://doi.org/10.1109/TSA.2005.854090

Knees, P., Faraldo, A., Herrera, P., Vogl, R., Bock, S., Horschlager, F., & **Le Goff, M.** (2015). 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. Malaga, Spain.

Knees, P., & **Schedl, M.** (2016). *Music Similarity and Retrieval: An Introduction to Audio- and Web-based Strategies.* The Information Retrieval Series. Springer Berlin Heidelberg. DOI: https://doi.org/10.1007/978-3-662-49722-7_1

Krebs, F., Bock, S., & **Widmer, G.** (2013). Rhythmic pattern modeling for beat and downbeat tracking in musical audio. In *Proceedings of the 14th International Society for Music Information Retrieval Conference (ISMIR)*, pages 227–232. Curitiba, Brazil.

Lamere, P. (2009). In search of the click track. Blog post https://musicmachinery.com/2009/03/02/in-search-of-the-click-track/, last accessed 12/9/2018.

Marchand, U., & **Peeters, G.** (2015). Swing Ratio Estimation. In *Proceedings of the International Conference on Digital Audio Effects (DAFx).* Trondheim, Norway.

Marchand, U., & **Peeters, G.** (2016). The extended ballroom dataset. In *Late Breaking Demo Session of the 17th International Society for Music Information Retrieval Conference (ISMIR).* New York, NY, USA.

McKinney, M. F., Moelants, D., Davies, M. E., & **Klapuri, A. P.** (2007). Evaluation of audio beat tracking and music tempo extraction algorithms. *Journal of New Music Research*, *36*(1), 1–16. DOI: https://doi.org/10.1080/09298210701653252

Moelants, D., & **McKinney, M. F.** (2004). Tempo perception and musical content: What makes a piece fast, slow or temporally ambiguous. In *Proceedings of the 8th International Conference on Music Perception and Cognition*, pages 558–562.

Oliveira, J. L., Gouyon, F., Martins, L. G., & **Reis, L. P.** (2010). IBT: A real-time tempo and beat tracking system. In *Proceedings of the 11th International Society for Music Information Retrieval Conference (ISMIR)*, pages 291–296.

Peeters, G. (2005). Time variable tempo detection and beat marking. In *Proceedings of the International Computer Music Conference (ICMC)*. Barcelona, Spain.

Peeters, G. (2007). Template-based estimation of time-varying tempo. *EURASIP Journal on Advances in Signal Processing*, *2007*(1), 158–171. DOI: https://doi.org/10.1155/2007/67215

Peeters, G., & **Flocon-Cholet, J.** (2012). 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. ACM. DOI: https://doi.org/10.1145/2390848.2390861

Percival, G., & **Tzanetakis, G.** (2014). Streamlined tempo estimation based on autocorrelation and cross-correlation with pulses. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, *22*(12), 1765–1776. DOI: https://doi.org/10.1109/TASLP.2014.2348916

Raffel, C. (2016). *Learning-Based Methods for Comparing Sequences, with Applications to Audio-to-MIDI Alignment and Matching*. PhD thesis, Columbia University, USA. DOI: https://doi.org/10.1109/ICASSP.2016.7471641

Raffel, C., McFee, B., Humphrey, E. J., Salamon, J., Nieto, O., Liang, D., & **Ellis, D. P. W.** (2014). mir_eval: A transparent implementation of common MIR metrics. In *Proceedings of the 15th International Society for Music Information Retrieval Conference (ISMIR)*. Taipei, Taiwan.

Salamon, J. (2019). What's broken in music informatics research? Three uncomfortable statements. In *36th International Conference on Machine Learning (ICML), Workshop on Machine Learning for Music Discovery*. Long Beach, CA, USA.

Salamon, J., Gomez, E., Ellis, D. P., & **Richard, G.** (2014). Melody extraction from polyphonic music signals: Approaches, applications, and challenges. *IEEE Signal Processing Magazine*, *31*(2), 118–134. DOI: https://doi.org/10.1109/MSP.2013.2271648

Salamon, J., & **Urbano, J.** (2012). Current challenges in the evaluation of predominant melody extraction algorithms. In *Proceedings of the 13th International Society for Music Information Retrieval Conference (ISMIR)*, pages 289–294. Porto, Portugal.

Schedl, M., Flexer, A., & **Urbano, J.** (2013). The neglected user in music information retrieval research. *Journal of Intelligent Information Systems*, *41*(3), 523–539. DOI: https://doi.org/10.1007/s10844-013-0247-6

Schedl, M., Zamani, H., Chen, C.-W., Deldjoo, Y., & **Elahi, M.** (2018). Current challenges and visions in music recommender systems research. *International Journal of Multimedia Information Retrieval*, *7*(2), 95–116. DOI: https://doi.org/10.1007/s13735-018-0154-2

Scheirer, E. D. (1998). Tempo and beat analysis of acoustical musical signals. *Journal of the Acoustical Society of America*, *103*(1), 588–601. DOI: https://doi.org/10.1121/1.421129

Schreiber, H., & **Muller, M.** (2014). Exploiting global features for tempo octave correction. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 639–643. Florence, Italy. DOI: https://doi.org/10.1109/ICASSP.2014.6853674

Schreiber, H., & **Muller, M.** (2017). A post-processing procedure for improving music tempo estimates using supervised learning. In *Proceedings of the 18th International Society for Music Information Retrieval Conference (ISMIR)*, pages 235–242. Suzhou, China.

Schreiber, H., & **Muller, M.** (2018a). A crowd-sourced experiment for tempo estimation of electronic dance music. In *Proceedings of the 19th International Society for Music Information Retrieval Conference (ISMIR)*. Paris, France.

Schreiber, H., & **Muller, M.** (2018b). A single-step approach to musical tempo estimation using a convolutional neural network. In *Proceedings of the 19th International Society for Music Information Retrieval Conference (ISMIR)*. Paris, France.

Serra, X. (2014). Creating research corpora for the computational study of music: The case of the CompMusic project. In *Proceedings of the AES International Conference on Semantic Audio*. London, UK. Audio Engineering Society.

Serra, X., Magas, M., Benetos, E., Chudy, M., Dixon, S., Flexer, A., Gomez, E., Gouyon, F., Herrera, P., Jorda, S., Paytuvi, O., Peeters, G., Schluter, J., Vinet, H., & **Widmer, G.** (2013). Roadmap for Music Information ReSearch. http://mires.eecs.qmul.ac.uk/files/MIRES_Roadmap_ver_1.0.0.pdf

Slaney, M. (2011). Web-scale multimedia analysis: Does content matter? *IEEEMultiMedia*, *18*(2), 12–15. DOI: https://doi.org/10.1109/MMUL.2011.34

Sturm, B. L. (2013a). Classification accuracy is not enough. *Journal of Intelligent Information Systems*, *41*(3), 371–406. DOI: https://doi.org/10.1007/s10844-013-0250-y

Sturm, B. L. (2013b). The GTZAN dataset: Its contents, its faults, their effects on evaluation, and its future use. *CoRR*, abs/1306.1461.

Sturm, B. L. (2014). Faults in the ballroom dataset. Blog post http://media.aau.dk/null_space_ pursuits/2014/01/ballroom-dataset.html, last accessed 4/29/2020.

Sturm, B. L. (2016). Revisiting priorities: Improving MIR evaluation practices. In *Proceedings of the 17th International Society for Music Information Retrieval Conference (ISMIR)*. New York, NY, USA.

Sturm, B. L., Bardeli, R., Langlois, T., & **Emiya, V.** (2014). Formalizing the problem of music description. In *Proceedings of the 15th International Society for Music Information Retrieval Conference (IS-MIR)*, pages 89–94. Taipei, Taiwan.

Tzanetakis, G., & **Cook, P.** (2002). Musical genre classification of audio signals. *IEEE Transactions on Speech and Audio Processing*, *10*(5), 293–302. DOI: https://doi.org/10.1109/TSA.2002.800560

Tzanetakis, G., & **Percival, G.** (2013). An effective, simple tempo estimation method based on selfsimilarity and regularity. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. Vancouver, Canada. DOI: https://doi.org/10.1109/ICASSP.2013.6637645

Urbano, J., Schedl, M., & **Serra, X.** (2013). Evaluation in music information retrieval. *Journal of Intelligent Information Systems*, *41*(3), 345–369. DOI: https://doi.org/10.1007/s10844-013-0249-4

Vignoli, F., & **Pauws, S.** (2005). A music retrieval system based on user driven similarity and its evaluation. In *Proceedings of the 6th International Conference on Music Information Retrieval (ISMIR)*, pages 272–279. London, UK.

Waterhouse, J., Hudson, P., & **Edwards, B.** (2010). Effects of music tempo upon submaximal cycling performance. *Scandinavian Journal of Medicine & Science in Sports*, *20*(4), 662–669. DOI: https://doi.org/10.1111/j.1600-0838.2009.00948.x

Zapata, J. R., & **Gomez, E.** (2011). Comparative evaluation and combination of audio tempo estimation approaches. In *42nd AES Conference on Semantic Audio*. Ilmenau, Germany.

How to cite this article: Schreiber, H., Urbano, J., & Müller, M. (2020). Music Tempo Estimation: Are We Done Yet? *Transactions of the International Society for Music Information Retrieval,* 3(1), pp. 111–125. DOI: https://doi.org/10.5334/tismir.43

Submitted: 16 October 2019 **Accepted:** 07 July 2020 **Published:** 24 August 2020

]u[*Transactions of the International Society for Music Information Retrieval* is a peer-reviewed open access journal published by Ubiquity Press.

OPEN ACCESS