Towards user-informed beat tracking of musical audio

António Sá Pinto[1,2] and Matthew E. P. Davies[1] ⋆

[1] INESC TEC, Sound and Music Computing Group, Porto, Portugal
[2] Faculdade de Engenharia da Universidade do Porto, Porto, Portugal
{antonio.s.pinto,matthew.davies}@inesctec.pt

Abstract. We explore the task of computational beat tracking for musical audio signals from the perspective of putting an end-user directly in the processing loop. Unlike existing "semi-automatic" approaches for beat tracking, where users may select from among several possible outputs to determine the one that best suits their aims, in our approach we examine how high-level user input could guide the manner in which the analysis is performed. More specifically, we focus on the perceptual difficulty of tapping the beat, which has previously been associated with the musical properties of expressive timing and slow tempo. Since musical examples with these properties have been shown to be poorly addressed even by state of the art approaches to beat tracking, we re-parameterise an existing deep learning based approach to enable it to more reliably track highly expressive music. In a small-scale listening experiment we highlight two principal trends: i) that users are able to consistently disambiguate musical examples which are easy to tap to and those which are not; and in turn ii) that users preferred the beat tracking output of an expressive-parameterised system to the default parameterisation for highly expressive musical excerpts.

Keywords: Beat Tracking, Expressive Timing, User Input

1 Introduction and Motivation

While the task of computational beat tracking is relatively straightforward to define – its aim being to replicate the innate human ability to synchronise with a musical stimulus by tapping a foot along with the beat – it remains a complex and unsolved task within the music information retrieval (MIR) community. Scientific progress in MIR tasks is most often demonstrated through improved accuracy scores when compared with existing state of the art methods [18]. At the core of this comparison rest two fundamental tenets: the (annotated) data upon which the algorithms are evaluated, and the evaluation method(s) used to measure performance. In the case of beat tracking, both the tasks of annotating

⋆ António Sá Pinto and Matthew E. P. Davies are supported by Portuguese National Funds through the FCT-Foundation for Science and Technology, I.P., under the grant SFRH/BD/120383/2016 and the project IF/01566/2015.

datasets of musical material and measuring performance are non-trivial [6]. By its very nature, the concept of beat perception – how an individual perceives the beat in a piece of music – is highly subjective [15]. When tapping the beat, listeners may agree over the phase, but disagree over the tempo or preferred metrical level – with one tapping, *e.g.*, twice as fast as another, or alternatively, they may agree over the tempo, but tap in anti-phase. This inherent ambiguity led to the prevalence of multiple hypotheses of the beat, which can arise at the point of annotation, but more commonly appear during evaluation where different interpretations of ground truth annotations are obtained via interpolation or sub-sampling. In this way, a wide net can be cast in order not to punish beat tracking algorithms which fail to precisely match the annotated metrical level or phase of the beats; with this coming at the expense that some unlikely beat outputs may inadvertently be deemed accurate. Following this evaluation strategy, the performance of the state of the art is now in the order of 90% on existing datasets [3, 4] comprised primarily of pop, rock and electronic dance music. However, performance on more challenging material [10] is considerably lower, with factors such as expressive timing (*i.e.*, the timing variability that characterises a human performance, in opposition to a metronomic or perfectly timed rendition [7]), recording quality, slow tempo and metre changes among several identified challenging properties.

Although beat tracking has garnered much attention in the MIR community, it is often treated as an element in a more complex processing pipeline which provides access to "musical time", or simply evaluated based on how well it can predict ground truth annotations. Yet, within the emerging domain of creative-MIR [16, 11] the extraction of the beat can play a critical role in musically-responsive and interactive systems [13]. A fundamental difference of applying beat tracking in a creative application scenario is that there is a specific end-user who wishes to directly employ the music analysis and thus has very high expectations in terms of its performance [1]. To this end, obtaining high mean accuracy scores across some existing databases is of lower value than knowing "*Can the beats be accurately extracted (as I want them) for this specific piece of music?*". Furthermore, we must also be aware that accuracy scores themselves may not be informative about "true" underlying performance [17, 6].

Of course, a user-specific beat annotation can be obtained without any beat tracking algorithm, by manually annotating the desired beat locations. However, manually annotating beat locations is a laborious procedure even for skilled annotators [10]. An alternative is to leverage multiple beat interpretations from a beat tracking algorithm, and then provide users with a range of solutions to choose from [8]. However, even with a large number of interpretations (which may be non-trivial and time-consuming to rank) there is no guarantee that the end-user's desired result will be present, especially if the alternative interpretations are generated in a deterministic manner from a single beat tracking output, *e.g.*, by interpolation or sub-sampling.

In this paper, we propose an alternative formulation which allows an end-user to drive how the beat tracking is undertaken. Our goal is to enable the

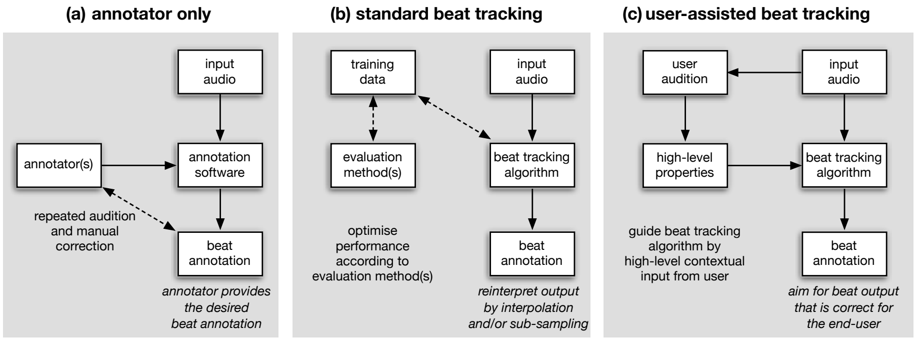
(a) annotator only
input audio
annotator(s)
annotation software
repeated audition and manual correction
beat annotation
annotator provides the desired beat annotation
(b) standard beat tracking
training data
input audio
evaluation method(s)
beat tracking algorithm
optimise performance according to evaluation method(s)
beat annotation
reinterpret output by interpolation and/or sub-sampling
(c) user-assisted beat tracking
user audition
input audio
high-level properties
beat tracking algorithm
guide beat tracking algorithm by high-level contextual input from user
beat annotation
aim for beat output that is correct for the end-user

Fig. 1. Overview of different approaches to obtaining a desired beat annotation. (a) The user annotates the beat positions. (b) A beat tracking algorithm is used – whose performance has been optimised on annotated datasets. (c) Our proposed approach, where user input guides the beat tracking.

user to rapidly arrive at the beat annotation suitable for their purposes with a minimal amount of interaction. Put another way, we envisage an approach to beat tracking where high-level contextual knowledge about a specific musical signal can be given by the user and reliably interpreted by the algorithm, without the need for extensive model training on annotated datasets, as shown in Fig. 1. In this sense, we put aside the concept of "universal" beat tracking models which target equal performance irrespective of the musical input signal, in favour of the more realistic goal of identifying different classes of the beat tracking problem, which require different beat tracking strategies. While the end goal of retrieving beat locations may be the same for fast-paced techno music and highly expressive classical guitar recordings, the assumptions about what constitutes the beat, and how this can be extracted from audio signals are not. Conversely, constraints should not be placed on what musical content can be creatively re-purposed based on the limitations of MIR algorithms.

The long term challenges of our approach are as follows: i) determining a low-dimensional parameterisation of the beat tracking space within which diverse, accurate solutions can be found in order to match different beat tracking conditions; ii) exposing these dimensions to end-users in a way that they can be easily understood; iii) providing an interpretable and understandable mapping between the user-input and the resulting beat annotation via the beat tracking algorithm; and finally iv) measuring the level of engagement among end-users who actively participate in the analysis of music signals.

Concerning the dimensions of beat tracking, it is well-understood that music of approximately constant (medium) tempo, with strong percussive content (*e.g.*, pop, rock music) is straightforward to track. Beat tracking difficulty (both for computational approaches and human tappers) can be due to musical reasons and signal-based properties [9, 10]. While it is somewhat nonsensical to consider a piece of music with "opposite" properties to the most straightforward case, it has been shown empirically that highly expressive music, without clear percussive content, is not well analysed even by the state of the art in beat tracking [10, 4]. Successful tracking of such pieces should, in principle, require input features which can be effective in the absence of percussion and a tracking model which can rapidly adapt to expressive tempo variation. While recent work [3] sought to develop multiple beat tracking models, these were separately trained at the level of different databases rather than according to musical beat tracking conditions.

In our approach, we reexamine the functionality of the current state of the art in beat tracking, *i.e.*, the recurrent neural network approach of Böck et al. [4]. In particular, we devise a means to re-parameterise it so that it is adapted for highly expressive music. Based on an analysis of existing annotated datasets, we identify a set of musical stimuli we consider typical of highly challenging conditions, together with a parallel set of "easier" examples. We then conduct a small-scale listening experiment where participants are first asked to rate the perceptual difficulty of tapping the beat, and subsequently to rate the subjective quality of beat annotations given by the expressive parameterisation vs the default version. Our results indicate that listeners are able to distinguish easier from more challenging cases, and furthermore that they preferred the beat tracking output of the expressive-parameterised system to the default parameterisation for the highly expressive musical excerpts. In this sense, we seek to use the assessment of perceptual difficulty of tapping as a means to drive the manner in which the beats can be extracted from audio signals towards the concept of user-informed beat tracking. To complement our analysis, we explore the objective evaluation of the beat tracking model with both parameterisations.

The remainder of this paper is structured as follows. In Section 2 we detail the adaption of the beat tracking followed by the design of a small-scale listening experiment in Section 3. This is followed by results and discussion in Section 4, and conclusions in Section 5.

2 Beat Tracking System Adaptation

Within this work our goal is to include user input to drive how music signal analysis is conducted. We hypothesise that high-level contextual information which may be straightforward for human listeners to determine can provide a means to guide how the music signal analysis is conducted. For beat tracking, we established in Section 1 that for straightforward musical cases, the current state of the art [4] is highly effective. Therefore, in order to provide an improvement over the state of the art, we must consider the conditions in which it is less effective, in particular those displaying expressive timing. To this end, we first summarise the main functionality of the beat tracking approach of Böck et al., after which we detail how we adapt it.

The approach of Böck et al. [4] uses deep learning and is freely available within the madmom library [2]. The core of the beat tracking model is a recurrent neural network (RNN) which has been trained on a wide range of annotated beat tracking datasets to predict a beat activation function which exhibits peaks at likely beat locations. To obtain an output beat sequence, the beat activation function given by the RNN is post-processed by a dynamic Bayesian network (DBN) which is approximated by a hidden Markov model [14].

While it would be possible to retain this model from scratch on challenging data, this has been partially addressed in the earlier multi-model approach of Böck et al. [3]. Instead, we reflect on the latter part of the beat tracking pipeline, namely how to obtain the beat annotation from the beat activation function. To this end, we address three DBN parameters: i) the minimum tempo in beats per minute (BPM); ii) the maximum tempo; and iii) the so-called "transition-λ" parameter which controls the flexibility of the DBN to deviate from a constant tempo[3]. Through iterative experimentation, including both objective evaluation on existing datasets and subjective assessment of the quality of the beat tracking output, we devised a new set of expressiveness-oriented parameters, which are shown, along with the default values in Table 1. More specifically, we first undertake a grid search across these three parameters on a subset of musical examples from existing annotated datasets for which the state of the art RNN is deemed to perform poorly, *i.e.*, by having an information gain lower than 1.5 bits [19]. An informal subjective assessment was then used to confirm that reliable beat annotations could be obtained from the expressive parameterisation.

Table 1. Overview of default and expressive-adapted parameters.

Parameter	**Default**	**Expressive**
Minimum Tempo (BPM)	55	35
Maximum Tempo (BPM)	215	135
Transition-λ (unitless)	100	10

As shown in Table 1, the main changes for the expressive model are a shift towards a slower range of allowed tempi (following evidence about the greater difficulty of tapping to slower pieces of music [5]), together with a lower value for the transition-λ. While the global effect of the transition-λ was studied by Krebs et al. [14], their goal was to find an optimal value across a wide range of musical examples. Here, our focus is on highly expressive music and therefore we do not need a more general solution. Indeed, the role of the expressive model is to function in precisely the cases where the default approach can not.

3 Experimental Design

Within this paper, we posit that high-level user-input can lead to improved beat annotation over using existing state of the art beat tracking algorithms in a

[3] the probability of tempo changes varies exponentially with the negative of the "transition-λ", thus higher values of this parameter favour constant tempo from one beat to the next one [14].

“blind” manner. In order to test this in a rigorous way, we would need to build an interactive beat tracking system including a user interface, and conduct a user study in which users could select their own input material for evaluation. However, doing so would require understanding which high-level properties to expose and how to meaningfully interpret them within the beat tracking system. To the best of our knowledge, no such experiment has yet been conducted, thus in order to gain some initial insight into this problem, we conducted a small-scale online listening experiment, which is split into two parts: **Part A** to assess the perceptual difficulty of tapping the beat, and **Part B** to assess the subjective quality of beat annotations made using the default parameterisation of the state of the art beat tracking system versus our proposed expressive parameterisation.

We use **Part A** as a means to simulate one potential aspect of high-level context which an end-user could provide: in this case, a choice over whether the piece of music is easy or difficult to tap along to (where difficulty is largely driven by the presence of expressive timing). Given this choice, **Part B** is used as the means for the end-user to rate the quality of the beat annotation when the beat tracking system has been parameterised according to their choice. In this sense, if a user rates the piece as “easy”, we would provide the default output of the system, and if they rate it as “hard” we provide the annotation from the expressive parameterisation. However, for the purposes of our listening experiment, all experimental conditions are rated by all participants, thus the link between **Part A** and **Part B** is not explicit.

3.1 Part A

In the first part of our experiment, we used a set of 8 short music excerpts (each 15 s in duration) which were split equally among two categories: i) “easy” cases with near constant tempo in 4/4 time, with percussive content, and without highly syncopated rhythmic patterns; and ii) “hard” cases typified by the presence of high tempo variation and minimal use of percussion. The musical excerpts were drawn from existing public and private beat tracking datasets, and all were normalised to -3 dB.

We asked the participants to listen to the musical excerpts and to spontaneously tap along using the computer keyboard at what they considered the most salient beat. Due to the challenges of recording precise time stamps without dedicated signal acquisition hardware (*e.g.*, at the very least, a MIDI input device) the tap times of the participants were not recorded, however this was not disclosed. We then asked the participants to rate the difficulty they felt when trying to tap the beat, according to the following four options:

- Low - *I could easily tap the beat, almost without concentrating*
- Medium - *It wasn’t easy, but with some concentration, I could adequately tap the beat*
- High - *I had to concentrate very hard to try to tap the beat*
- Extremely high - *I was not able to tap the beat at all.*

Our hypothesis for **Part A** is that participants should consistently rate those drawn from the "easy" set as having Low or Medium difficulty, whereas those from the "hard" should be rated with High or Extremely High difficulty.

3.2 Part B

Having completed **Part A**, participants then proceeded to **Part B** in which they were asked to judge the subjective quality of beat annotations (rendered as short 1 kHz pulses) mixed with the musical excerpts. The same set of musical excerpts from **Part A** were used, but they were annotated in three different ways: i) using the *default* parameterisation of the Böck et al. RNN approach from the madmom library [2]; ii) using our proposed *expressive* parameterisation (as in Table 1); and iii) a control condition using a completely *deterministic* beat annotation, *i.e.*, beat times at precise 500 ms intervals without any attempt to track the beat of the music. In total, this created a set of $8 \times 3 = 24$ musical excerpts to be rated, for which participants were asked to: *Rate the overall quality of how well the beat sequence corresponds to the beat of the music.*

For this question, a 5-point Likert-type item was used with (1) on the left hand side corresponding to "Not at all" and (5) corresponding to "Entirely" on the right hand side. Our hypothesis for **Part B** was that for the "hard" excerpts, the annotations of the expressively-parameterised beat tracker would be preferred to those of the default approach, and for all musical excerpts that the deterministic condition would be rated the lowest in terms of subjective quality.

3.3 Implementation

The experiment was built using HTML5 and Node.js and run online within a web browser, where participants were recruited from the student body of the University of Porto and the research network of the Sound and Music Computing Group at INESC TEC. Within the experimental instructions, all participants were required to give their informed consent to participate, with the understanding that any data collected would be handled in an anonymous fashion and that they were free to withdraw at any time without penalty (and without their partial responses being recorded). Participants were asked to provide basic information for statistical purposes: sex, age, their level of expertise as a musician, and experience in music production.

All participants were encouraged to take the experiment in a quiet environment using high quality headphones or loudspeakers, and before starting, they were given the opportunity to set the playback volume to a comfortable level. Prior to the start of each main part of the experiment, the participants undertook a compulsory training phase in order to familiarise themselves with the questions. To prevent order effects, each participant was presented with the musical excerpts in a different random order. In total, the test took around 30 minutes to complete.

4 Results and Discussion

4.1 Listening Experiment

A total of 10 listeners (mean age: 31, age range: 23–43) participated in the listening test, 9 of whom self-reported amateur or professional musical proficiency.

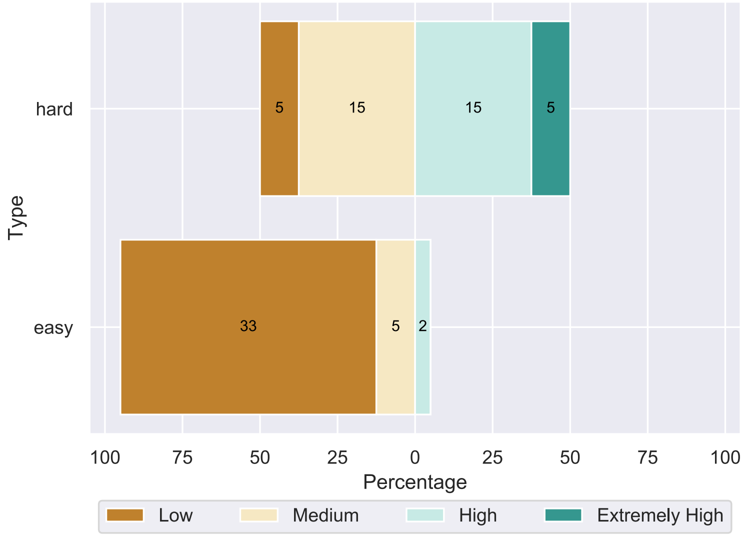

Fig. 2. Subjective ratings of the difficulty of beat tapping.

For **Part A**, we obtained 40 ratings for each stimuli group "easy" and "hard", according to the frequency distribution shown in Fig. 2. The most frequent rating for the first group was "low" (82.5%), followed by the "medium" rating (12.5%). For the "hard" group, a symmetrical rating was obtained: the adjacent ratings "medium" and "high" (37.5% each), complemented by the more extreme ratings "low" and "extremely high" (12.5% each). A Mann-Whitney test showed that there was a statistically significant difference between the ratings for both groups, with $p < 0.001$.

From these results we interpret that there was greater consistency in classifying the "easy" excerpts as having low difficulty, with only two excerpts rated above "medium", than for the "hard" excerpts which covered the entire rating scale from low to extremely difficult, albeit with the majority of ratings being for medium or high difficulty. We interpret this greater variability in the rating of difficulty of tapping to be the product of two properties of the participants: their expertise in musical performance and/or their familiarity with the pieces. Moreover, we can observe a minor separation between the understanding of the perceptual difficulty in tapping on the part of the participant and the presence of expressive timing in the musical excerpts; that experienced listeners may not have difficulty in tapping along with a piece of expressive music for which they knew well. Thus, for expert listeners it may be more reasonable to ask a direct question related to the presence of expressive timing, while the question of difficulty may be more appropriate for non-expert listeners who might lack familiarity with the necessary musical terminology.

For **Part B**, we again make the distinction between the ratings of the "easy" and the "hard" excerpts. A Kruskal-Wallis H test showed that there was a statistically significant difference between the 3 models (*expressive*, *default* and *deterministic*): $\chi^2(2) = 87.96$, $p < 0.001$ for "easy" excerpts, $\chi^2(2) = 70.71$, $p < 0.001$ for "hard" excerpts. A post-hoc analysis performed with the Dunn test with Bonferroni correction showed that all the differences were statistically significant with $p < 0.001/3$ (except for the pair *default–expressive* under the "easy" stimuli, for which identical ratings were obtained). A descriptive summary of the ratings (boxplot with scores overlaid) for each type of stimuli, and under the three beat annotation conditions are shown in Fig. 3.

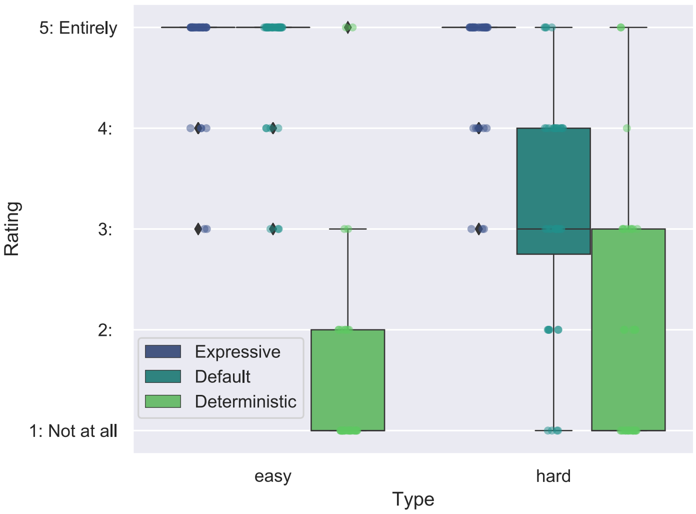

Fig. 3. Subjective ratings of the quality of the beat annotations.

The main results from Part B are as **follows. For** the "easy" excerpts there is no difference in performance for the *default* and *expressive* parameterisations of the beat tracking model, both of which are rated with high scores indicating high quality beat annotations from both systems. We contrast this with the ratings of the *deterministic* output (which should bear no meaningful relationship to the music) and which are rated toward the lower end of the scale. From these results we can infer that the participants were easily able to distinguish accurate beat annotations and entirely inaccurate annotations, which is consistent with the Beat Alignment Test [12]. Concerning the ability of the expressively parameterised model to achieve such high ratings, we believe that this was due to very clear information concerning the beat in the beat activation functions from the RNN.

Conversely, the ratings of the "hard" excerpts show a different picture. Here, the ratings of the expressively parameterised model are similar to the "easy" excerpts, but the ratings of the *default* model [2] are noticeably lower. This suggests that the participants, in spite of their reported higher perceptual difficulty in tapping the beat, were able to reliably identify the accurate beat predictions of the *expressive* model over those of the *default* model. It is noteworthy that the ratings of the *deterministic* approach are moderately higher for the "hard" excerpts compared to the "easy" excerpts. Given the small number of samples and participants for this experiment, it is hard to draw strong conclusions about this difference, but for highly expressive pieces, the *deterministic* beats may have inadvertently aligned with the music in brief periods compared to the "easy" excerpts, which may have been unambiguously unrelated.

4.2 Beat Tracking Accuracy

In addition to reporting on the listening experiment whose focus is on subjective ratings of beat tracking, we also examine the difference in objective performance of using the *default* and *expressive* parameterisations of the beat tracking model. Given the focus on challenging excerpts for beat tracking, we focus on the SMC dataset [10]. It contains 217 excerpts, each of 40 s in duration. Following the evaluation methods described in [6] we select a common subset: F-measure, CMLc, CMLt, AMLc, AMLt, and the Information Gain (D) to assess performance. In Table 2, we show the recorded accuracy on this dataset for both the default and expressive parameterisations. Note, for the default model we use the version in the madmom library [2] which has been exposed to this material during training, hence the accuracy scores are slightly higher than those in [4] where cross fold validation was used. In addition to showing the performance of each parameterisation we also show the theoretical upper limit achievable by making a perfect choice (by a hypothetical end-user) among the two parameterisations.

Table 2. Overview of beat tracking performance on the SMC dataset [10] comparing the default and expressive parameters together with upper limit on performance.

	F-measure	CMLc	CMLt	AMLc	AMLt	D
Default[2]	0.563	0.350	0.472	0.459	0.629	1.586
Expressive	0.540	0.306	0.410	0.427	0.565	1.653
Optimal Choice	0.624	0.456	0.611	0.545	0.703	1.830

From Table 2, we see that for all the evaluation methods, with the exception of the Information Gain (D), the default parameterisation outperforms the expressive one. This is an expected result since the dataset is not entirely comprised of highly expressive musical material. We consider the more important result to be the potential for our *expressive* parameterisation to track those excerpts for which the *default* approach fails. To this end, the increase of approximately 10% points across each of the evaluation methods demonstrates how these two different parameterisations can provide greater coverage of the dataset. It also implies that training a binary classifier to choose between expressive and non-expressive pieces would be a promising area for future work.

5 Conclusions

In this paper we have sought to open the discussion about the potential for user-input to drive how MIR analysis is performed. Within the context of beat tracking, we have demonstrated that it is possible to reparameterise an existing state-of-the-art approach to provide better beat annotations for highly expressive music, and furthermore, that the ability to choose between the default and expressive parameterisation can provide significant improvements on very challenging beat tracking material. We emphasise that the benefit of the expressive model was achieved without the need for any retraining of the RNN architecture, but that the improvement was obtained by reparameterisation of the DBN tracking model.

To obtain some insight into how user input could be used for beat tracking, we simulated a scenario where user decisions about perceptual difficulty of tapping could be translated into the use of a parameterisation for expressive musical excerpts. We speculate that listener expertise as well as familiarity may play a role in lowering the perceived difficulty of otherwise challenging expressive pieces. Our intention is to further investigate the parameters which can be exposed to end-users, and whether different properties may exist for expert compared to non-expert users. Despite the statistical significance of our results, we recognise the small-scale nature of the listening experiment, and we intend to expand both the number of musical excerpts uses as well as targeting a larger group of participants to gain deeper insight into the types of user groups which may emerge. Towards our long-term goal, we will undertake an user study not only to understand the role of beat tracking for creative MIR, but also to assess the level of engagement when end-users are active participants who guide the analysis.

References

1. K. Andersen and P. Knees. Conversations with Expert Users in Music Retrieval and Research Challenges for Creative MIR. In *Proc. of the 17th Intl. Society for Music Information Retrieval Conf.*, pages 122–128, 2016.
2. S. Böck, F. Korzeniowski, J. Schlüter, F. Krebs, and G. Widmer. Madmom: A new python audio and music signal processing library. In *Proc. of the 2016 ACM Multimedia Conf.*, pages 1174–1178, 2016.

3. S. Böck, F. Krebs, and G. Widmer. A multi-model approach to beat tracking considering heterogeneous music styles. In *Proc. of the 15th Intl. Society for Music Information Retrieval Conf.*, pages 603–608, 2014.
4. S. Böck, F. Krebs, and G. Widmer. Joint beat and downbeat tracking with recurrent neural networks. In *Proc. of the 17th Intl. Society for Music Information Retrieval Conf.*, pages 255–261, 2016.
5. R. Bååth and G. Madison. The subjective difficulty of tapping to a slow beat. In *Proc. of the 12th Intl. Conf. on Music Perception and Cognition*, pages 82–55, 2012.
6. M. E. P. Davies and S. Böck. Evaluating the evaluation measures for beat tracking. In *Proc. of the 15th Intl. Society for Music Information Retrieval Conf.*, pages 637–642, 2014.
7. P. Desain and H. Honing. Does expressive timing in music performance scale proportionally with tempo? *Psychological Research*, 56(4):285–292, 1994.
8. M. Goto, K. Yoshii, H. Fujihara, M. Mauch, and T. Nakano. Songle: A Web Service for Active Music Listening Improved by User Contributions. In *Proc. of the 12th Intl. Society for Music Information Retrieval Conf.*, pages 311–316, 2011.
9. P. Grosche, M. Müller, and C. Sapp. What makes beat tracking difficult? a case study on chopin mazurkas. In *Proc. of the 11th Intl. Society for Music Information Retrieval Conf.*, pages 649–654, 2010.
10. A. Holzapfel, M. E. P. Davies, J. R. Zapata, J. Oliveira, and F. Gouyon. Selective sampling for beat tracking evaluation. *IEEE Transactions on Audio, Speech and Language Processing*, 20(9):2539–2460, 2012.
11. E. J. Humphrey, D. Turnbull, and T. Collins. A brief review of creative MIR. In *Late-breaking demo session of the 14th Intl. Society for Music Information Retrieval Conf.*, 2013.
12. J. R. Iversen and A. D. Patel. The Beat Alignment Test (BAT): Surveying beat processing abilities in the general population. In *Proc. of the 10th Intl. Conf. on Music Perception and Cognition*, pages 465–468, 2010.
13. C. T. Jin, M. E. P. Davies, and P. Campisi. Embedded Systems Feel the Beat in New Orleans: Highlights from the IEEE Signal Processing Cup 2017 Student Competition [SP Competitions]. *IEEE Signal Processing Magazine*, 34(4):143–170, 2017.
14. F. Krebs, S. Böck, and G. Widmer. An efficient state space model for joint tempo and meter tracking. In *Proc. of the 16th Intl. Society for Music Information Retrieval Conf.*, pages 72–78, 2015.
15. D. Moelants and M. McKinney. Tempo perception and musical content: what makes a piece fast, slow or temporally ambiguous? In *Proc. of the 8th Intl. Conf. on Music Perception and Cognition*, pages 558–562, 2004.
16. X. Serra et al. Roadmap for music information research, 2013. Creative Commons BY-NC-ND 3.0 license, ISBN: 978-2-9540351-1-6.
17. B. L. Sturm. Classification accuracy is not enough. *Journal of Intelligent Information Systems*, 41(3):371–406, 2013.
18. J. Urbano, M. Schedl, and X. Serra. Evaluation in Music Information Retrieval. *Journal of Intelligent Information Systems*, 41(3):345–369, 2013.
19. J. R. Zapata, A. Holzapfel, M. E. P. Davies, J. L. Oliveira, and F. Gouyon. Assigning a confidence threshold on automatic beat annotation in large datasets. In *Proc. of the 13th Intl. Society for Music Information Retrieval Conf.*, pages 157–162, 2012.