A Turing Test for B-Keeper: Evaluating an Interactive Real-Time Beat-Tracker

Andrew Robertson
Centre for Digital Music
Department of Electronic Engineering
Queen Mary, University of London
andrew.robertson@elec.qmul.ac.uk

Mark D. Plumbley
Centre for Digital Music
Department of Electronic Engineering
Queen Mary, University of London
mark.plumbley@elec.qmul.ac.uk

Nick Bryan-Kinns
Department of Computer Science
Queen Mary, University of London
nickbk@dcs.qmul.ac.uk

ABSTRACT

Off-line beat trackers are often compared to human tappers who provide a ground truth against which they can be judged. In order to evaluate a real-time beat tracker, we have taken the paradigm of the 'Turing Test' in which an interrogator is asked to distinguish between human and machine. A drummer plays in succession with an interactive accompaniment that has one of three possible tempo-controllers (the beat tracker, a human tapper and a steady-tempo metronome). The test is double-blind since the researchers do not know which controller is currently functioning. All participants are asked to rate the accompaniment and to judge which controller they believe was responsible.

This method for evaluation enables the controllers to be contrasted in a more quantifiable way than the subjective testimony we have used in the past to evaluate the system. The results of the experiment suggest that the beat tracker and a human tapper are both distinguishable from a steady-tempo accompaniment and they are preferable according to the ratings given by the participants. Also, the beat tracker and a human tapper are not sufficiently distinguishable by any of the participants in the experiment, which suggests that the system is comparable in performance to a human tapper.

Keywords

Automatic Accompaniment, Beat Tracking, Human-Computer Interaction, Musical Interface Evaluation

1. INTRODUCTION

Our research concerns the task of real-time beat tracking with a live drummer. In a paper at last year's NIME Conference [6], we introduced a software program, B-Keeper, and described the algorithm used. However, the evaluation of the algorithm was mainly qualitative, relying on testimonial from drummers who had tried using the software in performances and rehearsal.

In trying to find a scientific method for testing the program, we could not use previously established beat tracking tests, such as the MIREX Competition [4], since these did not involve the necessary component of interaction and our beat tracker was highly specialised for performance with input from drums. In MIREX, the beat trackers are compared to data collected from forty human tappers who collectively provide a ground truth annotation [5].

In order to test the real-time beat tracker, we wanted to make a comparison with a human tapper and to do so within a live performance environment, yet in a way that would be both scientifically valid and also provide quantitative as well as qualitative data for analysis.

In Alan Turing's 1950 paper, 'Computing Machinery and Intelligence' [9] he proposes replacing the question 'can a computer think?', by an *Imitation Game*, popularly known as the "Turing Test", in which it is required to imitate a human being[1] in an interrogation. If the computer is able to fool a human interrogator a substantial amount of the time, then the computer can be credited with 'intelligence'. Turing considered many objections to this philosophical position within the original paper and there has been considerable debate as to its legitimacy, particularly the position referred to as 'Strong A.I.'. Famously, John Searle [7] put forward the Chinese room argument which proposes a situation in which computer might be able to pass the test without ever *understanding* what it is doing.

The Imitation Game might prove to be an interesting model for constructing an experiment to evaluate an interactive musical system. Whilst we do not wish to claim the system posseses 'intelligence', its ability to behave *as if* it had some form of 'musical intelligence' is vital to its ability to function as an interactive beat tracker.

B-Keeper controls the tempo by processing onsets detected by a microphone placed in the kick drum with additional tempo information from a microphone on the snare drum. The beat tracker is event-based and uses a method related to the oscillator models used by Large[3] and Toiviainen[8]. Rather than processing a continuous audio signal, it processes events from an onset detector and modifies its tempo output accordingly. B-Keeper interprets the onsets with respect to bar position using an internal weighting mechanism and uses Gaussian windows around the expected beat locations to quantify the accuracy and relevance of the onset for beat tracking. A tempo tracking process to determine the best inter-onset interval operates in parallel with a synchronisation process which makes extra adjustments to remain in phase with the drums. The parameters defining

[1] As Turing formulates the problem, the computer imitates a man pretending to be a woman, so as to negate the element of bias due to the imitation process from the test

Figure 1: AR taps on the keyboard in time with drummer, Joe Caddy, during one of the tests

the algorithm's behaviour automatically adapt to suit the playing style of the drummer.

B-Keeper is programmed as Java external within the Max/MSP environment. More details are given in our paper, 'B-Keeper: a real time beat tracker for live performance' [6], published at NIME2007.

2. EXPERIMENTAL DESIGN

The computer's role in controlling the tempo of an accompaniment might also be undertaken by a human controller. This, therefore, suggests that we can compare the two within the context of a "Turing Test" or Imitation Game. We also extend the test by including a control - a steady accompaniment which remains at a fixed tempo dictated by the drummer. For each test, the drummer gives four steady beats of the kick drum to start and this tempo is used as the fixed tempo.

The test involves a drummer playing along to the same accompaniment track three times. Each time, a human tapper (AR) taps the tempo on the keyboard, keeping time with the drummer, but only one of the three times will this be altering the tempo of the accompaniment. For these trials, controlled by the human tapper, we applied a Gaussian window to the intervals between taps in order to smooth the tempo fluctuation, so that it would still be musical in character. Of the other two, one will be an accompaniment controlled by the B-Keeper system and the other the same accompaniment but at a fixed tempo (see Figure 2). The sequence in which these three trials happen is randomly chosen by the computer and only revealed to the participants after the test so that the experiment accords with the principle of being *'double-blind'*: i.e. neither the researchers nor the drummer know which accompaniment is which. Hence, the quantitative results gained by asking for opinion measures and performance ratings should be free from any bias.

We are interested in the interaction between the drummer and the acommpaniment which takes place through the machine. In particular, we wish to know how this differs from the interaction that might take place with a person, or in this case, a human beat tracker. We might expect that, if our beat tracker is functioning well, the B-Keeper trials would be 'better' or 'reasonably like' those controlled by the human tapper. We would also expect them to be 'not like a metronome' and hence, distinguishable from the Steady Tempo trials. These expectations will form the basis of our hypotheses that are to be tested and we collected quantitative and qualitative data in order to do so.

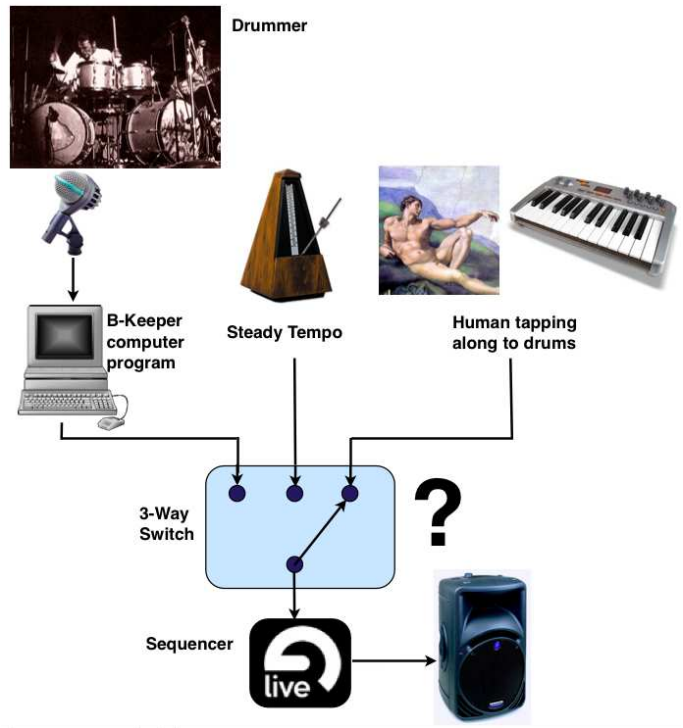

Figure 2: Design set-up for the experiment. Three possibilities: (a) Computer controls tempo from drum input; (b) Steady Tempo; (c) Human controls tempo by tapping beat on keyboard

After each trial, we asked each drummer to mark an 'X' on an equilateral triangle which would indicate the strength of their belief as to which of the three systems was responsible. The three corners corresponded to the three choices and the nearer to a particular corner they placed the 'X', the stronger their belief that that was the tempo-controller for that particular trial. Hence, if an 'X' was placed on a corner, it would indicate certainty that that was the scenario responsible. An 'X' on an edge would indicate confusion between the two nearest corners, whilst an 'X' in the middle indicates confusion between all three. This allowed us to quantify an opinion measure for identification over all the trials. The human tapper (AR) and an independent observer also marked their interpretation of the trial in the same manner.

In addition, each participant marked the trial on a scale of one to ten as an indication of how well they believed that test worked as 'an interactive system'. They were also asked to make comments and give reasons for their choice. A sample sheet from one of the drummers is shown in Figure 3.

We carried out the experiment with eleven professional and semi-professional drummers. All tests took place at the Listening Room of the Centre for Digital Music, Queen Mary, University of London, which is an acoustically isolated studio space. Each drummer took the test (consisting of the three randomly-selected trials) twice, playing to two different accompaniments. The first was based on a dance-rock piece first performed at Live Algorithms for Music Conference, 2006, which can be viewed on the internet [1]. The second piece was a simple chord progression on a software

version of a Fender Rhodes keyboard with some additional percussive sounds. The sequencer used was Ableton Live [2], chosen for its time-stretching capabilities.

We recorded all performances on video and audio and stored data from the B-Keeper algorithm. This allowed us to see how the algorithm processed the data and enabled us to look in detail at how the algorithm behaved and monitor how the tempo of the accompaniment was changed by the system.

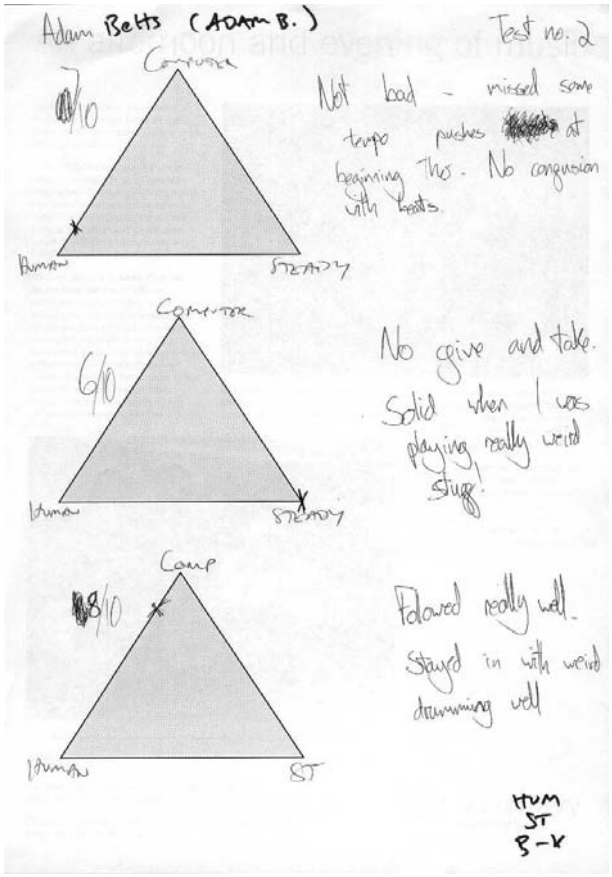

Figure 3: Sample sheet filled in by drummer Adam Betts.

3. RESULTS

We shall contrast the results between all three tests, particularly with regard to establishing the difference between the B-Keeper trials and the Human Tapper trials and comparing this to the difference between the Steady Tempo and Human Tapper trials. In Figure 4, we can see the opinion measures for all drummers placed together on a single triangle. The corners represent the three possible scenarios: B-Keeper, Human Tapper and Steady Tempo with their respective symbols. Each 'X' has been replaced with a symbol corresponding to the actual scenario in that trial. In the diagram we can clearly observe two things:

There is more visual separation between the Steady Tempo trials than the other two. With the exception of a relatively small number of outliers, many of the steady tempo trials were correctly placed near the appropriate corner. Hence, if the trial is actually steady then it will probably be identified as such.

The B-Keeper and Human Tapper trials tend to be spread over an area centered around the edge between their respective corners. At best, approximately half of these trials have been correctly identified. The distribution does not seem to have the kind of separation seen for the Steady Tempo trials, suggesting that they have difficulty telling the two controllers apart, but could tell that the tempo had varied.

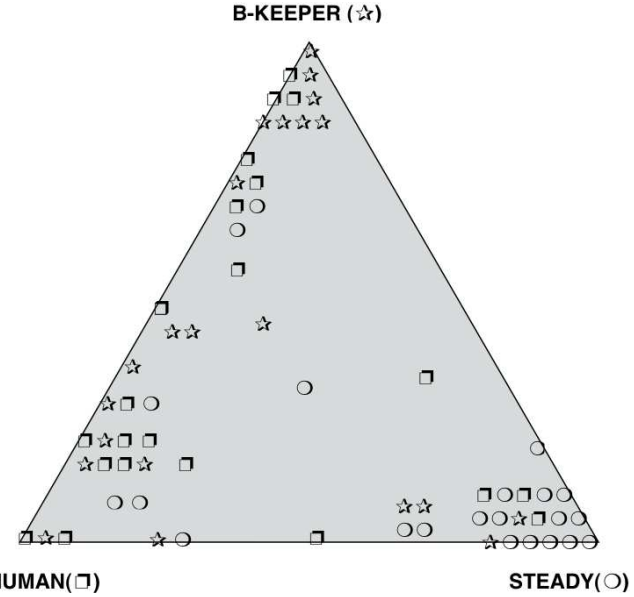

Figure 4: Results where the eleven different drummers judged the three different accompaniments (B-Keeper, Human Tapper and Steady Tempo) in the test. The symbol used indicates which accompaniment it actually was (see corners).

The deduction process used by participants generally worked by first trying to determine whether the tempo had been steady or not. In the majority of cases, this was successful, but some misidentifications were made, particularly if the drummer had *played to* the accompaniment and not made much attempt to influence the tempo. In these cases, the distinction between an interactive accompaniment, which will adapt to you, and one at a fixed tempo is harder to judge.

The second deduction to be made would be, in the case where the tempo varied or the music appeared responsive, to discern whether the controller had been B-Keeper or the Human Tapper. In order to do so, there needs to be some assumption as to the characteristics that might be expected of each. From interviews, we recognised that drummers expect the human to be more adaptable to changes in rhythm such as syncopation and they may also have felt that a human would respond better to changes within their playing. For instance, as drummer Tom Oldfield commented: "I felt that was the human, because it responded very quickly to me changing tempo."

3.1 Case Study: Joe Caddy

One dialogue exchange shows the kind of logical debate in action.[2]

JC: [talking about the trials]: "The first one I gave 8 and I put actually closer to human response. I played pretty simply and it followed it quite nicely. The second one had no response at all to tempo on the drums. The last one I gave 9 - great response to tempo change, I slowed it up, I slowed it down. It took a couple of beats to resolve, but I

[2]**JC** refers to Joe Caddy, session drummer and drummer for hip-hop band Captive State; **AR** refers to the first author, who acted as the Human Tapper in all experiments.

think I put it nearer the B-Keeper."

AR: "Is that because you have some experience of the system?"

JC: "If it was human, I would have expected it to catch up more quickly. I think because it took two or three beats to come in at the new tempo, it was the B-Keeper."

AR: "Same. I think it's an 80 per cent chance that that was B-Keeper."

[Result is revealed: The first was B-Keeper; the last the Human Tapper , i.e. controlled by **AR** - the opposite to what both **JC** and **AR** have identified.]

AR: "I just didn't think it was that though. I guess it must have been."

JC: "The last test we did, I changed the tempo much more. Do they surprise you those results?"

AR: "The first I felt was me and I felt that the last wasn't me."

This exchange demonstrates how both a drummer and even the person controlling the tempo can both be fooled by the test. From the point of view of the key tapper, AR suggests that there is a *musical illusion* in which, by tapping along to the drummer playing, it can appear to be having an effect when in fact there is none. The illusion is strongest when the B-Keeper system was in operation as the music would respond to changes in tempo. This effect is reflected in the opinion measures reported by AR, which we initially expected to be higher for the Human Tapper trials than the others, but had a mean of only 45% (see Table 1).

3.2 Case Study: Adam Betts

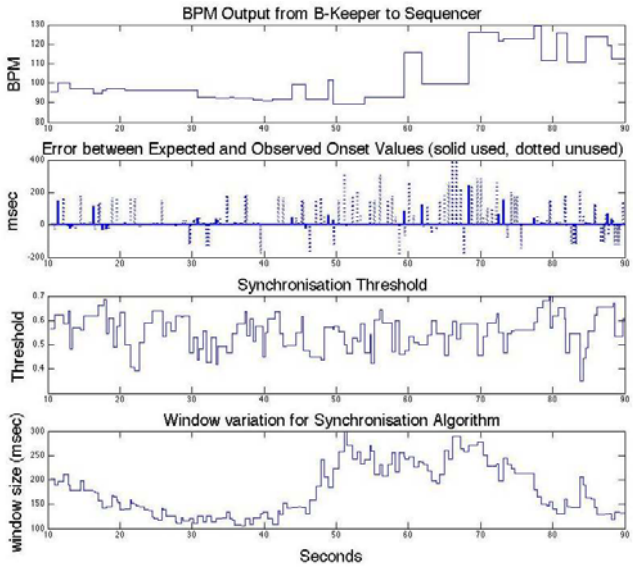

Figure 5: Data from the B-Keeper's interaction with drummer Adam Betts. The top graph shows the tempo variation. The second graph shows the errors recorded by B-Keeper between the expected and observed beats. The final two graphs show how the synchronisation threshold and window automatically adapt, becoming more generous when onsets fail to occur in expected locations.

The above study shows a scenario in which the B-Keeper fooled the drummer into guessing it was a human-controlled accompaniment. In one trial with James Taylor Quartet drummer, Adam Betts, the machine had been calibrated (to its usual setting) so as to be fairly responsive to tempo changes. However, when he played a succession of highly syncopated beats, the algorithm responded by making the synchronisation window so wide that the machine was thrown out of sync. In Figure 5, this can be seen happening after about fifty seconds, where the pattern has changed so the onsets are no longer used by the tracker to synchronise (dotted errors in second graph). When it eventually does so at sixty to seventy seconds, an erroneous adjustment easily occurs due to the size of the window and low threshold.

In this case, it was immediately apparent that it was B-Keeper since the tempo had varied and done so in a non-human manner. It had made an apparent mistake and all three involved in the experiment, the drummer, the human tapper and our independent observer, immediately concluded that this was B-Keeper. On the trial sheet, Adam commented:

> "Scary. Okay at beginning, but got confused and guessed tempo incorrectly with 16ths etc. When it worked, it felt good."

Such an event happened only one time out of the the twenty-two tests[3], but it is interesting since it suggests that the form of the experiment is viable for similar reasons to those suggested by Turing. In the scenario of the imitation game, if the machine did exhibit abnormal behaviour (for instance, as he suggests, the ability to perform very quick arithmetical calculations) or, as implied throughout Turing's paper, the inability to answer straight-forward questions such as the length of one's hair, then one could easily deduce it was the machine. In this case, the absence of human tolerance to extreme syncopation is the the kind of 'machine-like' characteristic that made it easily identifiable.

Table 1: Mean Identification measure results for all judges involved in the experiment. Bold percentages correspond to the correct identification

		Judged as:		
Judge	**Accompn.t**	B-Keeper	Human	Steady
Drummer	B-Keeper	**44** %	37 %	18 %
	Human	38 %	**44** %	17 %
	Steady	12 %	23 %	**64** %
Human Tapper	B-Keeper	**59** %	31 %	13 %
	Human	36 %	**45** %	23 %
	Steady	15 %	17 %	**68** %
Observer	B-Keeper	**55** %	39 %	6 %
	Human	33 %	**42** %	24 %
	Steady	17 %	11 %	**73** %

3.3 Analysis and Interpretation

The mean scores recorded by the drummers are given at the top of Table 1. They show similar measures for correctly identifying the B-Keeper and Human Tapper trials, both have mean scores of 44%, with the confusion being predominantly between which of the two variable tempo controllers is operating. The Steady Tempo trials have a mean confidence score of 64% on the triangle.

Each participant in the experiment had a higher score for identifying the Steady Tempo trials than the other two. It appears that the Human Tapper trials are the least identifiable of the three and the confusion tends to be between the B-Keeper and the Human Tapper.

[3]This was due to incorrect parameter settings for the drumming style in question.

Table 2: Table showing the polarised decisions made by the drummer for the different trials.

	Judged as:		
Controller	B-Keeper	Human	Steady
B-Keeper	**9.5**	8.5	4
Human Tapper	8	**10**	4
Steady Tempo	2	4	**16**

Table 3: Table showing the polarised decisions made by the drummer over the Steady Tempo and Human Tapper trials.

	Judged as:	
Controller	Human Tapper	Steady Tempo
Human Tapper	12	4
Steady Tempo	5	14

Of the B-Keeper trials themselves, the drummers were least confident in identifying it as the controller. The researchers, who acted as independent observer and the tapper, were more confident. In an analogous result, we might expect the human tapper, the first author, to be able to distinguish the trials in which he controlled the tempo, however, this did not appear to be the case. He was more successful at discerning the other two trials.

We can polarise the decisions made by drummers by taking their highest score to be their decision for the that trial. In the case of a tie, we split the decision equally. The advantage of this method is that we can make pair-wise comparisons between any of the controllers, whilst also allowing the participants the flexibility to remain undecided between two possibilities. Table 2 shows the polarised decisions made by drummers over the trials. There is confusion between the B-Keeper and Human Tapper trials, whereas the Steady Tempo trials were identified over 70% of the time. The B-Keeper and Human Tapper trials were identified 43% and 45% respectively, little better than chance.

3.4 Comparative Tests

In order to test the distinguishablility of one controller from the other, we can use a Chi-Square Test, calculated over all trials with either of the two controllers. If there is a difference in scores so that one controller is preferred to the other (above a suitable low threshold), then that controller is considered to be chosen for that trial. Where no clear preference was clear, such as in the case of a tie or neither controller having a high score, we discard the trial for the purposes of the test.

Thus for any two controllers, we can construct a table for which decisions were correct. The table for comparisons between the Steady Tempo and the Human Tapper trials is shown in Table 3. We test the hypothesis that the distribution is the same for either controller, corresponding to the premise that the controllers are indistinguishable.

The Chi-Square Test statistic for this table is 8.24 which means that we reject the test hypothesis at the 5% significance level. This indicates a significant separation between the controllers. Partly this can be explained from the fact that drummers could vary the tempo with the Human Tapper controller but the Steady Tempo trials had the characersitic of being metronomic.

Comparing the B-Keeper trials and the Human Tapper trials, we get the results shown in table 4. The Chi-Square test statistic is 0.03 which is extremely low, suggesting no significant difference in the drummers' identification of the controller for either trial. Whilst B-Keeper shares the characteristic of having variable tempo and thus is not identifiable simply by trying to detect a tempo change, we would expect that if there was a *machine-like* characteristic to the B-Keeper's response, such as an unnatural response or unreliability in following tempo fluctuation, syncopation and drum fills, then the drummer would be able to identify the machine. It appeared that, generally, there was no such characteristic and drummers had difficulty deciding between the two controllers. It may appear that having the Human Tapper visible to them would give them an advantage, however, this did not prove to be the case as the similarity between the computer's response and a human tapping along was close enough that often the observer and the human tapper were also unsure of the controller.

Table 4: Table contrasting decisions made by the drummer over the B-Keeper and Human Tapper trials.

	Judged as:	
Controller	Human Tapper	B-Keeper
Human Tapper	9	8
B-Keeper	8	8

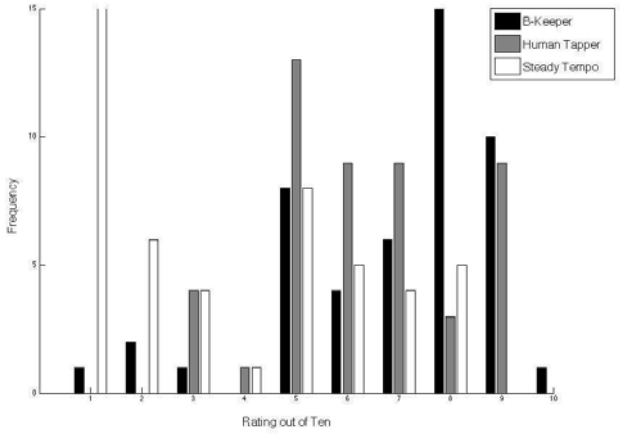

Figure 6: Bar Graph indicating the different frequency of cumulative ratings for the three scenarios - B-Keeper (black), Human Tapper (grey) and Steady Tempo (white).

The difficulty of distinguishing between controllers was a common feature of many tests and whilst the test had been designed expecting that this might be the case, the results were often surprising when revealed. In addition, we did not expect drummers to believe steady accompaniments had sped up or slowed down with them or the human tapper that he had controlled the tempo when he had not. This indicates a subjectivity to the perception of time. It seems that some drummers had an enhanced ability to spot a fixed tempo without even varying much, perhaps gained through extensive experience. Matt Ingram, session drummer, who professed to have been "playing to click for the last ten days, all day every day", remarked of the Steady Tempo trial: "It felt like playing to a metronome, cause it was just there. Either that or your time's great, cause I was trying to push it and it wasn't letting me."

3.5 Ratings

In addition to the identification of the controller for each trial, we also also asked each participant to rate each trial with respect to how well it had worked as an interactive

Table 5: Median ratings given by all participants for the different scenarios. The combined total median is given in bold.

	Median Rating		
Judge	B-Keeper	Human Tapper	Steady Tempo
Drummer	7.5	5.5	5
Human	8	6.5	4
Observer	8	7	5
Combined	**8**	**6**	**5**

accompaniment to the drums. Our reasoning in obtaining ratings for the accompaniments is that in addition to trying to establish whether the beat tracker is distinguishable human tapper controller, it is also desirable to compare the controllers through a rating system. Partly we are interested in how the drums and accompaniment sounded together, but also we are interested in its response to the drums.

The cumulative frequency for these ratings over all participants (drummers, human tapper and independent observer) is shown in Figure 6. The Steady Tempo accompaniment was consistently rated worse than the other two. The median values for each accompaniment are shown in Table 5. The B-Keeper system has consistently been rated higher than both the Steady Tempo and the Human Tapper accompaniment.

The overall median ratings, calculated over all participants, were: B-Keeper: 8, Human Tapper: 6 and Steady Tempo: 5. It is important that not only was the the beat tracker not significantly distinguishable from the human tapper, but it performed as well when judged by both the drummer and an independent observer. The fact that the median rating is towards the top end of the scale suggests that musically the beat tracker is performing its task well. As the experiment was double-blind, there was no bias within the scaling of the different controllers.

If we look at pair-wise rankings, we can assess the the significance of this difference between ratings. Firstly, we convert the rating out of ten into a strict ordinal rating (allowing equality where necessary). The Wilcoxon signed-rank test is a non-parametric statistical test that can apply to test the hypothesis that the controllers' rankings have the same distribution. For more than twenty trials, the distribution for this test statistic is approximately normal.

When contrasting the rankings given by drummers to B-Keeper with the Steady Tempo and Human Tapper trials, the approximate Z ratios[4] are 2.97 and 2.32 respectively. Thus, we would reject the hypothesis that the controllers are equally preferable at the 5% significance level in both cases. The fact that the ratings are significantly higher for B-Keeper is highly important as the primary aim is to create a musically successful beat tracker for live drums.

4. CONCLUSIONS AND FUTURE WORK

In this experiment, we contrast a computer-based beat tracker with a human tapper and metronome for the purposes of providing interactive accompaniment to drums. The Turing Test has proved an interesting scenario for a scientific evaluation of the beat tracker. By contrasting it with a human tapper in an experiment analogous to that described by Turing for language imitation, we were able to assess its performance against human abilities which are the standard against which beat trackers are best judged. This provides a more informative comparison for evaluation than subjective interviews.

The beat tracker has proved to be comparable in performance to the human tapper and is not distinguishable in any statistically significant way. The Steady Tempo accompaniment was perceived as a less successful accompanist and was considerably more distinguishable from the variable tempo accompaniments. In addition, the resulting accompaniment was judged as being aesthetically comparable with that resulting from using a human tapper.

We are currently working on incorporating the beat tracker into a live rock music band. By interfacing with Ableton Live, the beat tracker provides a framework for the triggering of loops, samples and electronic parts within a rock performance without recourse to click tracks or other compromises. We aim to evaluate its efficiency by case studies with users of the system. We are also concentrating on improving the ability of the system to correctly interpret extended syncopation and expressive timing within drum patterns within its analysis of onsets.

5. ACKNOWLEDGMENTS

The authors would like to thank Adam Stark, Enrique Perez Gonzales and Robert Macrae for acting as independent observers during the tests. We would also like to thank all drummers who kindly participated in the experiment: Joe Yoshida, Rod Webb, Joe Caddy, Matt Ingram, Jem Doulton, Greg Hadley, Adam Betts, Tom Oldfield, David Nock, Hugo Wilkinson and Mark Heaney.

AR is supported by a studentship from the EPSRC.

6. REFERENCES

[1] `http://www.elec.qmul.ac.uk/digitalmusic/b-keeper`.

[2] `http://www.ableton.com`, viewed on 4th April, 2008.

[3] E. W. Large. Beat Tracking with a Nonlinear Oscillator. In *Working Notes of the IJCAI-95 Workshop on Artificial Intelligence and Music, Montreal*, pages 24–31, 1995.

[4] M. F. McKinney. Audio beat tracking contest description, 2006. `http://www.music-ir.org/mirex2006/index.php/Audio_Beat_Tracking` as viewed on 4th april 2008.

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

[6] A. Robertson and M. Plumbley. B-keeper: A beat-tracker for live performance. In *Proc. International Conference on New Interfaces for Musical Expression (NIME), New York, USA*, 2007.

[7] J. Searle. Minds, Brains and Programs. *Behavioural and Brain Sciences*, 3:417–457, 1980.

[8] P. Toiviainen. An interactive midi accompanist. *Computer Music Journal*, 22(4):63–75, 1998.

[9] A. Turing. Computing Machinery and Intelligence. *Mind*, 59:433–460, 1950.

[4] normal with zero mean and unit variance