COMPARING ONSET DETECTION & PERCEPTUAL ATTACK TIME

Dr Richard Polfreman
University of Southampton
r.polfreman@soton.ac.uk

ABSTRACT

Accurate performance timing is associated with the perceptual attack time (PAT) of notes, rather than their physical or perceptual onsets (PhOT, POT). Since manual annotation of PAT for analysis is both time-consuming and impractical for real-time applications, automatic transcription is desirable. However, computational methods for onset detection in audio signals are conventionally measured against PhOT or POT data. This paper describes a comparison between PAT and onset detection data to assess whether in some circumstances they are similar enough to be equivalent, or whether additional models for PAT-PhOT difference are always necessary. Eight published onset algorithms, and one commercial system, were tested with five onset types in short monophonic sequences. Ground truth was established by multiple human transcription of the audio for PATs using rhythm adjustment with synchronous presentation, and parameters for each detection algorithm manually adjusted to produce the maximum agreement with the ground truth. Results indicate that for percussive attacks, a number of algorithms produce data close to or within the limits of human agreement and therefore may be substituted for PATs, while for non-percussive sounds corrective measures are necessary to match detector outputs to human estimates.

1. INTRODUCTION AND MOTIVATION

This research forms part of a larger project involving evaluation of controller hardware and parameter mappings in the context of real-time physical modeling synthesis [10]. Thus a specific device (e.g. Microsoft Kinect) will have its control outputs (e.g. performer's 2D hand position) mapped onto synthesis model parameters (e.g. plectrum position in relation to a string). A number of techniques for controller evaluation have been proposed, e.g. [9], including qualitative and quantitative methods. One method of evaluation to be used will ask the performer to match as accurately as possible a given audio target phrase using a given combination of controller, mapping and synthesis configuration. The target and the attempt will then be compared to assess how well the task was completed, in addition to other qualitative assessments. Given that a number of participants, controllers and targets may be used, it would be helpful to complete the performance analysis computationally rather than rely on expert markup of the audio. While in some situations it would be possible to use the timing of control data such as MIDI NoteOn events directly, with perhaps a fixed latency, here the timing of a note or onset may vary significantly for a given control value dependant on other parameters. For example, the position of a plectrum along a string, pluck release threshold, current string displacement and velocity and tension (pitch) will all impact upon the distance from the string the plectrum will need to reach before releasing the string and generating the onset. This indirect control over event timing means that measuring the audio output is necessary. Previous work on onset detection generally does not consider timing accuracy in detail, justifiably prioritising detection rates (type 1 and type 2 errors) and using a temporal tolerance between ground truth and detections beyond which an onset is said to have been missed [3]. Here however, the detailed timing of the onsets is critical.

The measure of two performances being "in time" is a complex issue with a large number of contextual factors, but in this case the target and performance are short monophonic solo instrument phrases with a fixed tempo and it was felt that this case would be simple enough to be studied. More expressive timing feature are ignored and PAT synchronous events are considered the ideal.

2. ONSET TIME

2.1 When is a Note?

Three potential onset times are described in published work. *Physical onset time* (PhOT) is usually considered to be the audio signal first rising from zero, *perceptual onset time* (POT) the time at which a human listener can first detect this change and finally *perceptual attack time* (PAT) is the "perceived moment of rhythmic placement" [15], or rhythmic centre, and is similar to the p-centre concept in speech analysis [13]. A "correct" performance therefore places the events' PATs appropriately, rather than PhOTs or POTs.

While most studies have considered PAT to be a specific time, Wright proposes that PAT is distributed over a finite time period and should be considered as a probability density function describing the likelihood of a listener hearing the PAT at each time point (PAT-pdf) [15]. This could account for variation between listeners and by individuals in repeated trials and implies that there will be span of time over which an event can remain in musical time with another. This spread of time values is of interest here, since this governs how well-localized the PAT for a particular sound is and how accurately a detection algorithm must match the ground truth.

2.2 PAT Measurement

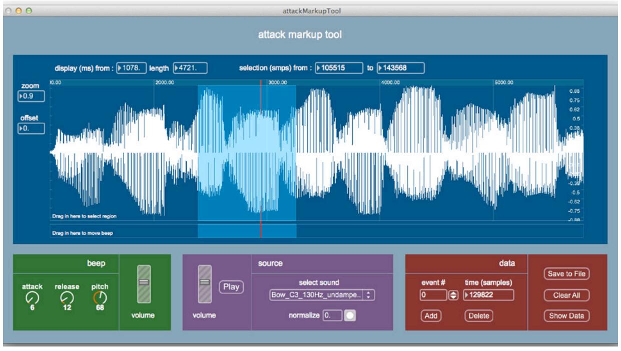
2.2.1 Measurement Methods

PAT can be measured in a number of ways as identified in [4, 7, 13, 14, 15]. The intrinsic PAT of a sound is typically not measured directly, but rather the delta-PAT (ΔPAT) [7], calculated via comparison against a reference tone. If the reference sound is very short in time, its PAT will be very close to its PhOT and so the target sound PAT can be estimated from the ΔPAT. PAT must be expressed relative to a zero point, usually either the sound's PhOT or the offset from the beginning of the audio file where a sequence is being considered [15].

The most common measurement is rhythm adjustment, where two sounds are aligned by the listener until they either appear synchronous (sound together) or isochronous (sound evenly spaced rhythmically, alternately presented) [14]. Both synchronous and isochronous methods have problems (such as event fusion in synchronous presentation) while isochronous cannot be used where the PATs for a musical sequence are to be measured, rather than isolated events. Likewise Villing's phase correction response (PCR) method [14] is unsuitable for sequences and so the synchronous method was used here.

A tool was created for participants to align a reference sound against a series of test sounds containing a number of onsets (Figure 1). While the reference sound should be short, Wright found that if it is too short there are problems for accurate alignment. He also found that a reference click based on matching the spectrum of the test sound aided PAT alignment [15]. In our experiment, the reference was a simple sine tone, which is the same as the target we will use for the performer to follow, which will include pitch changes at a later stage. Wright gave users control over amplitudes to help avoid fusion of the two events and this was included here. Our tool also allowed the user to change the pitch of the reference, again to help limit fusion ([4] suggests frequency independence of PAT). Gordon [7] indicated that subjects had difficulty matching sounds with very different attack times, and so a user variable attack time was included to ameliorate this, although clearly this has the potential to add uncertainty to the ground truth and so was limited to <127ms.

The participant can choose a sound, select any part of it to be looped and place a marker on the sound that triggers the reference tone. The marker can be dragged with the mouse and fine-tuned by changing the value in a number box, in samples at 44.1kHz sample rate. Thus the location of the reference can be adjusted by ~0.02ms. Participants were instructed to adjust this value until the test event and reference sounded musically synchronous. The visual display is to aid users in finding physical onsets quickly before searching those regions for perceptual alignment. For each event the tool recorded the PAT and the other user settings so that these could also be analysed if necessary. Participants were each given a training session (in addition to a written manual) and asked to complete the task using headphones.



display (ms) from : 1078. length 4721.
selection (smps) from : 105515 to 143568
zoom 0.9
offset 0.
attack 6 release 12 pitch 68
select sound Bow_C3_130Hz_undampe..
normalize 0.
event # 0 time (samples) 129822

Figure 1. Software tool for ground truth collection.

2.2.2 Test Sounds

Five test sounds were used, four were synthesized with the IRCAM's Modalys software [6] and the performances made deliberately imperfect, so that each event in the sequence would not be identical and the timing of events not strictly metrical. The sequences provide a set of variations in timbre and attacks as one might expect in an instrumental performance. The dynamics were generally stable but with occasional deviations. The models were: plucked string (un-damped), legato bowed string, struck plate (un-damped) and a single reed-tube. Only in the reed sound was complete silence reached between onsets and not for all of those. The final sound was a sine tone, which is used as the target for performance matching, in this case precisely metrical. These beeps were 95ms long (5ms attack, 90ms decay) with a 500ms inter-PhOT interval. Each sound was normalized, had a fundamental frequency of 130.81Hz (an octave below middle C) and contained 16 onsets, providing 80 events in total.

2.2.3 Ground Truth Results

Nine participants completed the task for all 80 events and so each sound file had 144 marked-up onsets and there were 720 data points in total. All participants had some musical experience, typically in ensembles or bands and/or formal performance training. Where data seemed particularly erroneous, such as a missing or duplicated event, or in isolation extremely different to others , participants were asked to review and double-check their data to ensure they were content with the values originally supplied, and, only if not, amend them. As with other studies, participants reported that the task was challenging, particularly with the non-percussive sounds, while one reported that (in the reed case) there were a range of time values over which the reference and test sound were equally “in time”, and that they had simply tried to be consistent in where they placed the reference sound.

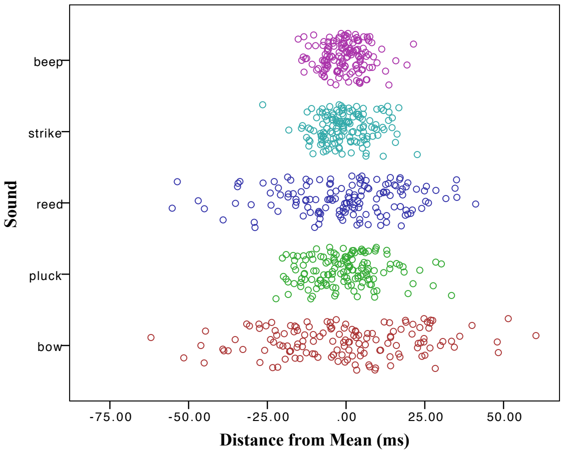


Figure 2. Scatterplot of ΔPAT ground truth data.

To group the results of ΔPAT values across different events within a particular sequence, the mean ΔPAT value was taken for each event and then each ΔPAT value replaced by its distance from that mean. Figure 2 shows scatter plots of these mean-shifted ΔPAT values (with vertical jitter to improve visibility). As expected, shorter attacks gave rise to more tightly clustered ΔPAT times, although outliers remain, while the longer attacks produce more widely spread results, as the location of the note is more ambiguous. We also expect smaller variation in the beep sounds since each event is almost identical, differing only in the phase of the sine in each. The plucks show greater spread than the other percussive attacks, again expected due to the more complex articulation: a double attack of the initial plectrum impact on the string followed rapidly by the release of the string creating the note (Figure 3). The time between impact and release was typically between 20ms and 40ms, averaging 23ms. The audio files were also annotated for PhOT for comparison with ΔPAT and onset detector times. For the percussive attacks this was straightforward as in each case there were discontinuities in the signal at the point where each new event began and which could be found through visual inspection. In the case of the pluck sounds, both the impact and string release times were noted. For the reed sound, onsets starting from silence were similarly clear, while others were estimated from the inflection point in amplitude between the decay of one note to the beginning of the next. The bow sound was particularly difficult and required inspection of the sonogram in addition to the time domain signal and PhOT was estimated from disruption to the harmonic structure as one event ends and the other begins. Table 1 shows the mean and standard deviation offsets from ΔPAT to PhOT for each sound, where the pluck sound is using the string release time. All apart from pluck are positive values as expected, where ΔPAT is later than PhOT. As can be seen from the table, ΔPAT appears very close to PhOT for the short attacks, although with some variation as reflected in Figures 2 and 4. Interestingly pluck is very close to the string release point, in fact slightly earlier, suggesting an effect of the preceding impact bringing the PAT forward. Given the close agreement between mean ΔPAT and PhOT for the short attacks, this is indicates that onset detectors which measure PhOT should provide timing data close to ΔPAT. For the non-percussive attacks ΔPAT is significantly later than PhOT, and so the utility of onset detectors will depend on whether they remain close to PhOT or are similarly delayed.

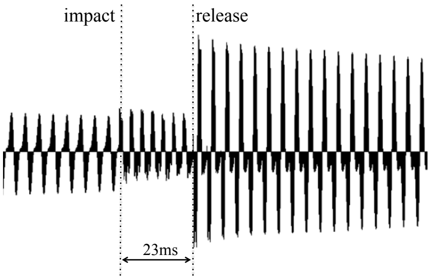


Figure 3. Section of pluck waveform showing decay of previous event followed by initial plectrum impact and release of string.

Sound	*Bow*	*Pluck*	*Reed*	*Strike*	*Beep*
Mean (ms)	23.52	-0.46	51.21	3.48	2.08
σ (ms)	22.83	10.64	18.46	7.42	5.91

Table 1. Mean and standard deviation for ΔPAT-PhOT distance.

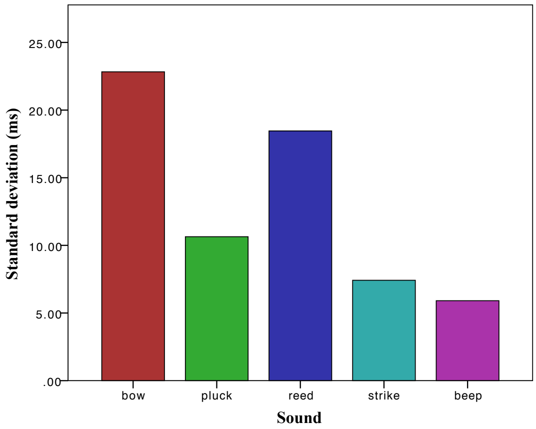


Figure 4. Mean ΔPAT standard deviations.

Figure 4 shows the σ across events for ΔPAT, indicating how consistently human listeners can determine ΔPAT for each sound (against the reference tone). Thus for bowed sounds, ±σ gives a spread of ~42ms and for the sine beep ~11.0ms. While only bow and pluck passed Shapiro Wilks normality tests, over 70% of data for each sound were within ±σ of the mean. The limit of discrimination of temporal events is typically considered to be ~10ms [4]. Wright logically proposed a system for automatic mark-up of audio using onset detection followed by a PAT model to correct for the difference between PhOT and PAT [15]. However, if the time differences between the ground truth and onset times reported by onset detectors are within similar limits to human listeners it indicates that these may be used directly to provide PAT data without adding a specific PAT-PhOT model.

3. ONSET DETECTION

3.1 Onset Detection Algorithms

Onset detection algorithms are typically based on PhOT or POT, with a time tolerance to decide successful detections. The task usually comprises three main steps: (optional) pre-processing; generation of an onset detection function (ODF) that indicates the probability of an onset at each moment in time; and peak selection across the ODF. While some methods are psychoacoustically motivated, differences between PhOT, POT and PAT are usually ignored. Here those differences are important if an onset detector is to provide ΔPAT estimates.

Several comparative studies of the performance of onset detection algorithms have been published, while the MIREX event compares a number of new algorithms annually. Studies, including [1, 3, 5], compare the rates of false positives and false negatives against a selection of test sounds. Collins [3] compared 16 onset detection algorithms with NPP (non-pitched percussive) and PNP (pitched non-percussive) monophonic sounds, finding that for the NPP case, a spectral difference function based on work by Klapuri [8] was most effective, while for the PNP case all algorithms performed less well, with a phase deviation method being the most successful [1]. While comparing algorithms against PhOT rather than PAT, Collins used detection tolerances of 50ms for PNP sounds and 25ms for NPP, which compare well with the figures shown in Figure 4 [3].

3.2 Onset Measurement

3.2.1 Implementation

A Max patch was developed to run a number of onset detection algorithms against the test audio. This displays the ODF for each detector as well as the detection hits. Initially 8 algorithms were tested, including two widely known Max objects *bonk~* and *sigmund~*, both later rejected as unable to provide sufficiently accurate results. To compare results with a commercial onset detection system, the audio software Melodyne 3.2[1] was also included in the experiment. For each sound Melodyne's percussive mode detection was used, as this outperformed the other options, even on non-percussive sounds, and it should be noted that no detection parameters were user adjusted in this case.

A modified[2] version of the *aubioonset~* MSP object by Andrew Robertson [11], itself a port of algorithms implemented by Paul Brossier [2] was used for high frequency content (HFC), energy based, modified Kullback-Leibler (MKL), complex, spectral difference (SD) and phase deviation (PD) functions, the equations for which can be found in the literature [2]. In each case the FFT size was 2048 with a hop of 128 samples. While there are more recent algorithms, these were chosen as being widely available and frequently referred to in the literature as the basis for other algorithms or tests. Due to difficulties with non-percussive attacks, two adaptations were implemented as Max patches – weighted phase deviation (WPD) and spectral flux (SF) following Dixon [5], the latter rectifying the difference between frames in SD, important in distinguishing between onsets and offsets. Peak-picking for WPD and SF involved taking the difference between the outputs of two moving average filters (using *average~*) and passing the result to a Schmitt trigger (*thresh~*). One filter was coarse providing an adaptive threshold (averaging over ~130ms), the other fine to smooth the ODF (typically ~20ms).

3.2.2 Comparison with Ground Truth

Detection function parameters were adjusted to achieve as close as possible to 100% success rate, i.e. 0 false positives (FP) or false negatives (FN). This was achieved for all the percussive attacks with all detectors, but the reed and bow sounds proved more problematic. Figure 5 shows the mean distance of each algorithm from the ground truth for the percussive attacks. The error bars indicate one standard deviation above and below the mean, the first set being those of the ground truth.

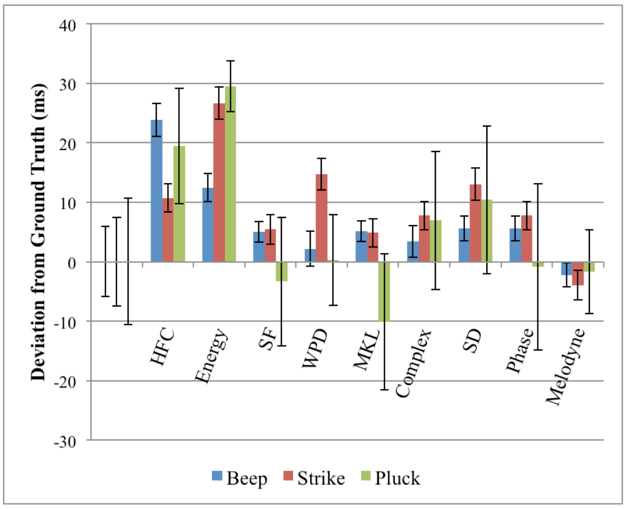


Figure 5. Mean distance from ground truth for each algorithm, percussive attacks (positive values are later).

As can be seen in the figure, HFC and Energy are typically late detectors, and fall outside the standard deviation (σ) ranges for the ground truth, while Melodyne performed very well across all three percussive attacks, preempting the ΔPAT values (PhOT earlier than PAT) as expected. In fact comparison with the PhOT data shows that Melodyne very accurately tracked PhOT (e.g. -0.1ms average distance for the beep), performing slightly worse

[1] http://www.celemony.com/

[2] Modified to include audio rate output of the onset detections as 1-sample delta functions, rather than Max *bangs*, to improve timing accuracy.

with pluck as it marked the impact time rather than string release for two events. Strike and beep across all the detectors show relatively consistent offsets from the ground truth, albeit varying by sound and detector, the σ values are all < 3ms. For pluck, HFC, Energy, WPD and Melodyne achieved a σ less than the ground truth.

To decide whether an onset detector can be used as a ΔPAT measurement tool we must define how closely the outputs of the detector must correspond to the ground truth. Table 2 shows a summary of each detector against each percussive attack for three simple tests. The first test (a) is simply whether the standard deviation of the detector output is less than that of the ground truth – not in itself sufficient, but indicative of relative stability. The second (b) and third (c) state whether combinations of the detector mean and standard deviation lie within limits of the 1 or 2 standard deviations of the ground truth:

$$(|\mu_D| + \sigma_D) < \sigma_{GT} \tag{1}$$

$$(|\mu_D| + (2 \times \sigma_D)) < (2 \times \sigma_{GT}) \tag{2}$$

where μ_D is the detector mean deviation from ground truth, σ_D and σ_{GT} the detector and ground truth standard deviations respectively. Equation 1 implies (for normal distributions) that we expect ~64% of detector values to lie within one standard deviation of the ground truth mean, changing to ~95% of detector outputs within two standard deviations of ground truth mean in equation 2.

Sound	*Beep*			*Strike*			*Pluck*		
Detector	*a*	*b*	*c*	*a*	*b*	*c*	*a*	*b*	*c*
HFC	Y	-	-	Y	-	-	Y	-	-
Energy	Y	-	-	Y	-	-	Y	-	-
SF	Y	-	Y	Y	-	Y	-	-	-
WPD	Y	Y	Y	Y	-	-	Y	Y	Y
MKL	Y	-	Y	Y	Y	Y	-	-	-
Complex	Y	-	Y	Y	-	Y	-	-	-
SD	Y	-	Y	Y	-	-	-	-	-
Phase	Y	-	Y	Y	-	Y	-	-	-
Melodyne	Y	Y	Y	Y	Y	Y	Y	Y	Y

Table 2. Onset detector summary for percussive attacks.

As can be seen in Table 2, Melodyne passed each test for each percussive attack, indicating that it is likely to provide a useful equivalent to ΔPAT data, while WPD is effective for beep and pluck, and MKL for strike.

With the reed sounds several algorithms conflated onset and offsets, with Complex, SD, MKL and PD often showing stronger peaks, sometimes double peaks, on offsets in the detection function (see Figure 6), making their FN rate unacceptably large[1]. Melodyne also suffered from offset conflation with the reed sound, although the detection function could not be examined. As expected, the half-wave rectification introduced in the SF algorithm eliminated this problem, with offset peaks significantly lowered, as did WPD by shaping the response by amplitude. The remaining detectors were able to provide zero FN and FP rates, and for HFC, SF and WPD with σ less than the ground truth. All means were delayed with respect to the ΔPAT ground truth and none were contained within ±σ of the ground truth (see Figure 7).

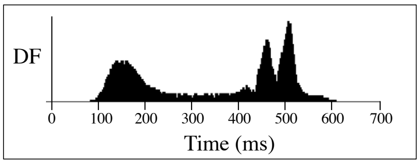


Figure 6. *Complex* onset detection function over the duration of a single reed event, showing large offset peaks.

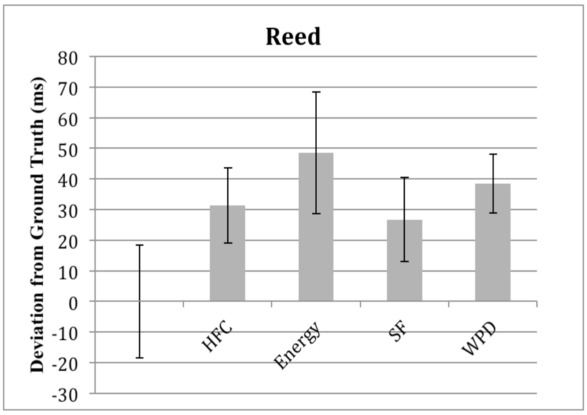


Figure 7. Mean distance from ground truth for selected algorithms, reed sound.

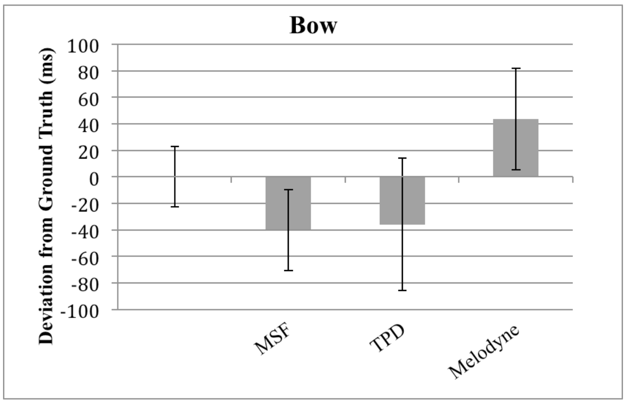


Figure 8. Mean distance from ground truth for selected algorithms, bow sound.

The bowed sound was most problematic (Figure 8), with only Melodyne achieving 0 FN and FP rates, while others (e.g. HFC) resulted in an FP rate of ~33% if adjusted to zero FN rates. SF had one FP with rectification off - i.e. as SD but with the dual-filter peak picker (labeled MSF). MKL only identified a single onset, while WPD achieved zero FN and FP, when rather than weighting each phase contribution by the magnitude from the FFT frequency

[1] Testing with a single clarinet sample indicated that these algorithms suffer offset conflation there also, rather than this being a product of the physical model synthesis.

bins, a threshold was used (TPD). None of the three best detectors were within the ground truth range or had σ lower than the ground truth.

4. DISCUSSION AND FUTURE WORK

The aim of this work was to assess whether automatic onset detection methods might be used to provide metrics for measuring performance accuracy, where the phrases to be assessed would be monophonic but the sounds potentially complex. This required testing since performance timing is considered to be PAT based, while onset detection PhOT or POT based. Further, the reported performance of onset detectors is often reduced to type I and II errors, rather than distances from ground truth.

Ground truth data captured via rhythm adjustment with synchronous presentation indicated levels of agreement of approximately 12-20ms for percussive attacks and 42ms for non-percussive (within a single standard deviation). Given the likelihood that there is indeed a span of time offset over which two sounds may be said to remain in time rhythmically, it would seem useful to develop a new method of ΔPAT measurement that does not force the participant to select a single time value, but rather supports identification of a range. Such a method could make the task easier for participants, speeding up the annotation process and increasing accuracy.

All of the onset detectors managed zero type I and type II errors with the percussive attacks, but only some produced results close enough to the ground truth to be regarded as PAT equivalent data. For the non-percussive sounds, achieving 100% detection even in these short sequences proved challenging, and the timing did not match the ground truth closely enough, requiring some form of PAT model to correct for this. Future work should investigate existing models, such as those tested in [4].

The algorithms used are well known and therefore results may usefully be compared with other studies, but it would be helpful to test more recent algorithms for performance improvements. Recent work has explored the influence of peak-picking algorithms on the performance of onset detection and it would be useful to test alternative methods in this context, particularly as the temporal location of the peak is so critical here [12]. Similarly, preprocessing could be explored.

It would be useful if the MIREX onset detection test data were additionally annotated for PAT so that algorithms could be assessed against PAT as well as PhoT/POT data and against a large data set.

5. REFERENCES

[1] J. Bello, L. Daudet, S. Abdallah, C. Duxbury, M. Davies, and M. Sandler: "A Tutorial on Onset Detection in Music Signals", *IEEE Trans. On Speech And Audio Proc.,* Vol 13, No 5, pp.1035-1047, 2005.

[2] P. Brossier: *Automatic Annotation of Musical Audio for Interactive Applications*, Ph.D. Thesis, Queen Mary University of London, UK, 2006.

[3] N. Collins: "A Comparison of Sound Onset Detection Algorithms with Emphasis on Psychoacoustically Motivated Detection Functions", *Proceedings of AES118 Convention*, 2005.

[4] N. Collins: "Investigating computational models of perceptual attack time", *Proceedings of the 9th International Conference on Music Perception & Cognition (ICMPC9)*, pp. 923-929, 2006.

[5] S. Dixon: "Onset Detection Revisited", *Proceedings of the 9th Int. Conference on Digital Audio Effects (DAFx'06)*, pp. 133-137, 2006.

[6] N. Ellis, J. Bensoam and R. Caussé: "Modalys Demonstration", *Proceedings of the International Computer Music Conference (ICMC),* 2005.

[7] J. W. Gordon: "The perceptual attack time of musical tones" *J. Acoust. Soc. Am.,* Vol. 82, No. 1, pp. 88–105, 1987.

[8] A. Klapuri: "Sound onset detection by applying psychoacoustic knowledge", *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Proc. (ICASSP),* pp. 3089–92, 1999.

[9] S. O'Modhrain: "A Framework for the Evaluation of Digital Musical Instruments", *Computer Music Journal*, Vol. 35, No. 1, pp. 28–42, 2011.

[10] R. Polfreman: "Multi-Modal Instrument: Towards a Platform for Comparative Controller Evaluation", *Proceedings of the International Computer Music Conference (ICMC),* pp. 147-150, 2011.

[11] A. Robinson: "Queen Mary University of London: Andrew Robertson – Software", http://www.eecs.qmul.ac.uk/~andrewr/software.htm, accessed July 2013.

[12] C. Rosão, R. Ribeiro, and D. Martin de Matos: "Influence Of Peak Selection Methods On Onset Detection", *Proceedings of the 13th International Society for Music Information Retrieval Conference* (ISMIR 2012), pp. 517-12, 2012.

[13] S. Scott: "The point of p-centres", *Psychological Research*, Vol. 61, pp. 4–11, 1998.

[14] R. Villing: *Hearing the Moment: Measures and Models of the Perceptual Centre*, Ph.D. Thesis, National University of Ireland Maynooth, 2010.

[15] M. Wright: *The Shape Of An Instant: Measuring And Modeling Perceptual Attack Time With Probability Density Functions,* Ph.D. Thesis, Stanford University, Stanford, CA, 2008.