VALIDATING ATTACK PHASE DESCRIPTORS OBTAINED BY THE TIMBRE TOOLBOX AND MIRTOOLBOX

Kristian Nymoen
Department of Musicology
University of Oslo, Norway
kristian.nymoen@imv.uio.no

Anne Danielsen
Department of Musicology
University of Oslo, Norway
anne.danielsen@imv.uio.no

Justin London
Department of Music
Carleton College, USA
jlondon@carleton.edu

ABSTRACT

The attack phase of sound events plays an important role in how sounds and music are perceived. Several approaches have been suggested for locating salient time points and critical time spans within the attack portion of a sound, and some have been made widely accessible to the research community in toolboxes for Matlab. While some work exists where proposed audio descriptors are grounded in listening tests, the approaches used in two of the most popular toolboxes for musical analysis have not been thoroughly compared against perceptual results. This article evaluates the calculation of attack phase descriptors in the Timbre toolbox and the MIRtoolbox by comparing their predictions to empirical results from a listening test. The results show that the default parameters in both toolboxes give inaccurate predictions for the sound stimuli in our experiment. We apply a grid search algorithm to obtain alternative parameter settings for these toolboxes that align their estimations with our empirical results.

1. INTRODUCTION

Many music researchers rely on available toolboxes for quantitative descriptions of the characteristics of sound files. These descriptions are commonly referred to as *descriptors* or *features*, and they may be used as variables in experimental research or as input to a classification algorithm. In this paper we investigate the descriptors that concern the first part of sonic events, what we will refer to as *attack phase descriptors*.[1] Particularly, we are interested in the detection of salient time points, such as the point when the sound event is first perceived, the point when it reaches its peak amplitude, and the *perceptual attack* of the sound event, which will be introduced properly in Section 2. Robust detection of such points in time is essential to obtain accurate values for the attack phase descriptors commonly used in the music information retrieval community, such as Log-attack time and attack slope. We compare the results of two popular toolboxes for Matlab, the MIRtoolbox [1] (version 1.6.2) and the Timbre Toolbox [2] (version 1.4), in estimating these salient moments. Sound samples from 'real' music recordings are used to compare the toolbox analysis results to a listening test. As such, our research complements the work by Kazazis et al. [3], where the toolbox results are compared to strictly controlled synthesis parameters using additive synthesis.

In Section 2 we discuss previous work in auditory perception concerned with the attack portion of sonic events. In Section 3 we take a closer look at computational estimation of attack phase descriptors. Further, in Section 4, we show how the algorithms in the MIRtoolbox and Timbre toolbox compare to our own experimental results, and then we discuss these results and conclude in Section 5.

2. BACKGROUND

The attack of musical events has been studied from a range of perspectives. Pierre Schaeffer experimented with tape recordings of sounds in the 1960s [4]. By cutting off the beginning of sound events he showed the importance of the attack portion for the perception of sonic events.

A seminal paper in this field is John W. Gordon's study from 1987 of the perceptual attack time of musical tones [5]. Gordon manipulated synthesis parameters in a digital synthesizer and observed how the perceived moment of metrical alignment was affected. A range of parameters were found to be involved. Gordon introduced the term *perceptual attack time*, to describe the moment in time perceived as the rhythmic emphasis of a sound [5]. The term was further explained by Wright, saying that the perceptual attack time of a musical event is its "perceived moment of rhythmic placement" [6]. This definition emphasises that the perceptual attack time of a sound event acts as reference when aligning it with other events to create a perceptually isochronous sequence, as illustrated in Figure 1. Wright further argues that the perceptual attack of a sound event is best modelled not as a single point in time, but as a continuous probability density function indicating the likelihood as to where a listener would locate the maxima of rhythmic emphasis. These definitions are strongly linked to the concept of perceptual centres (P-centres) in speech timing, defining the moment in time when a brief event is perceived to occur [7]. In effect, when two sounds are perceived as synchronous, it is their P-centres that are aligned.

In addition to the perceptual attack time, Gordon [5] argues that it makes sense to talk about both *physical onset time* and *perceptual onset time* for a given sound event.

[1] Note that *phase* in this context signifies that these descriptors concern a certain temporal segment of the sound. This differs from the technical meaning of *phase* in signal processing, such as in a *phase spectrum*.

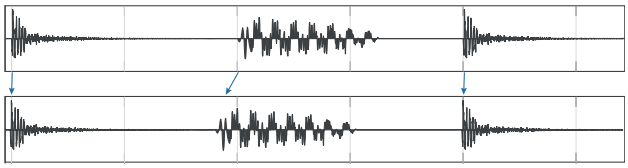

Figure 1. The top plot shows a sequence of sounds arranged isochronously by *physical onset*. The bottom plot shows the sounds arranged isochronously by *perceptual attack*. The latter will be perceptually isochronous.

The former refers to the moment in time when the sound event begins (before it is perceived), and the latter to the moment in time when the sound event is first perceived.

Attack phase descriptors may represent physical features of the sound signal, or be estimations of perceptual features. However, the distinction between the two types is rarely made clear. This may be due to the wide application of audio descriptors in various machine learning tasks, where physical features are sufficient to obtain the required result (e.g. in automatic classification of instrument type [8]). Another related machine learning task is the annual MIREX[2] competition, where the term *onset time* is used to denote the (approximate) time points in an audio file where new sound events occur, without addressing the distinction between physical and perceptual onset.[3] In studies of auditory perception, we argue that it is imperative to be aware of the distinction between the two and the inherent difficulty of estimating perceptual features [9].

2.1 Terminology

Table 1 and Figure 2 show a compilation of attack phase descriptors, most of them found in the Timbre Toolbox and MIRtoolbox. In our use of the term, attack phase descriptors include salient time points within or close to the attack portion of a sound event, in addition to relative measures, such as the time difference between two salient time points.

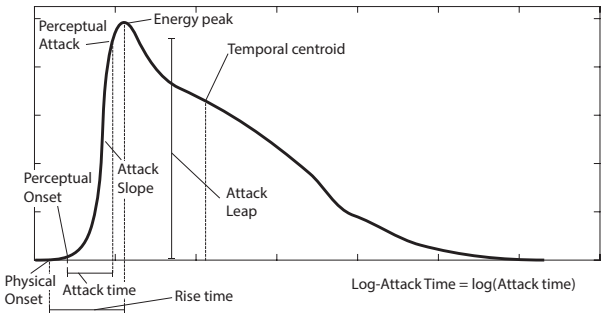

Figure 2. Illustration of various attack phase descriptors.

[2] http://www.music-ir.org/mirex

[3] The goal in the MIREX audio onset competitions is to detect discrete sound events in an audio file that has been annotated by a group of music experts. False positives (indicating a new sound event where an expert annotator did not) and false negatives (missing a sound event where an expert annotator indicated one) are penalised. In other words, high temporal precision for the extracted events is not a primary goal in the MIREX audio onset competition, which explains why the distinction between physical and perceptual onset is not addressed.

Name	Type	Description
Physical onset	phTP	Time point where the sound energy first rises from 0.
Perceptual onset	peTP	Time point when the sound event becomes audible.
Perceptual attack	peTP	Time point perceived as the rhythmic emphasis of the sound — what Gordon calls perceptual attack *time* [5].
Energy peak	phTP	Time point when the energy envelope reaches its maximum value.
Rise time	phTS	Time span between physical onset and energy peak.
Attack time	peTS	Time span between perceptual onset and perceptual attack.
Log-Attack Time	phTS	The base 10 logarithm of attack time.
Attack slope	peES	Weighted average of the energy envelope slope in the attack region [2].
Attack leap	peES	The difference between energy level at perceptual attack and perceptual onset [10].
Temporal centroid	phTP	The temporal barycentre of the sound event's energy envelope.

Table 1. Table of attack phase descriptors. ph: physical, pe: perceptual, T: time, E: energy, S: span, P: point.

The "Type"-column in Table 1 refers to particular characteristics of the descriptor. A descriptor marked 'ph' describes a physical aspect of the audio signal, and 'pe' describes perceptual aspects of the sound event. What we refer to as time point descriptors (TP) indicate a single point in time. Note that we choose not to use the word 'time' in TP descriptors, to emphasise that these are conceptually different from descriptors representing a time span. As noted by Wright [6] the reference of such time points may be a fixed point in the sound file (e.g., beginning), or an external time reference (e.g., a SMPTE clock). Further, the time points may be understood in reference to some other calculated time point, for instance the perceptual attack in relation to the physical onset. Gordon [5] used this measure, naming it *relative perceptual attack time*. It is worth mentioning that detection of physical onset may be unreliable. For natural sounds and recordings, the estimation of physical onset will depend on algorithmic variables; typically, a noise threshold to eliminate background noise, and some parameters as input to the envelope extraction algorithm. For digitally synthesized audio signals, however, the first nonzero signal value can be detected and the physical onset can confidently be used as a reference.

Time span descriptors (TS) describe the temporal relation between two time point descriptors. In effect, this category is not mutually exclusive to the TP category: a TP descriptor with a local time reference (e.g. the time of the physical onset of the sound event) can also be seen as a TS descriptor, such as Gordon's relative perceptual attack time.

We use the term energy span descriptors (ES) to denote descriptors that describe how the energy envelope develops between two time points. Finally, although not shown in Table 1, one could easily include EP descriptors to denote the energy level at salient time points.

3. COMPUTING ATTACK PHASE DESCRIPTORS

In this section we look more closely at how various attack phase descriptors are computed in the Timbre Toolbox and MIRtoolbox. With reference to the list of descriptors in Table 1, the computational estimation of *ph*-type descriptors is conceptually quite different from that of *pe*-type descriptors. While a physical feature of the signal can be calculated accurately, and only depends on algorithmic parameters like filter cutoff frequency or window length, any perceptual feature can only be an estimate of how the particular sound event would be perceived. Neither of the toolboxes state clearly whether the computed descriptors are estimates of perceptual or physical features. Both employ algorithms where the end of the attack range does not necessarily coincide with the peak of the energy envelope. This suggests that the toolboxes take into account the fact that the perceived attack of a sound event might end before the energy peak.

Below, we take a closer look at the attack phase descriptors as they are provided by the toolboxes, and in Section 4 we compare their output to perceptual results from our own listening experiment. Conceptually, the two toolboxes take similar approaches to computing attack phase descriptors. However, on an algorithmic level, the two are quite different. For now, we consider the default strategies and parameters in each toolbox and look at how the energy envelope and attack phase descriptors are calculated.

3.1 Energy envelope

The common basis of most algorithms for calculating attack phase descriptors is some function describing the *energy envelope* of the sound signal. Attack phase descriptors are calculated based on various thresholds applied to this envelope and its derivatives. The method and the parameters specified in the calculation of the energy envelope strongly influence the estimated attack phase descriptors.

The default strategy in the Timbre Toolbox is to apply a Hilbert transform to the audio signal, followed by a 3rd-order Butterworth lowpass filter with cutoff frequency at 5 Hz [2]. The Timbre Toolbox does not compensate for group delay in the filter when extracting the energy envelope. This is not crucial to the extraction of TS attack phase descriptors such as Log-Attack Time, but delays all TP descriptors by the group delay time of the filter.[4]

The MIRtoolbox implements a range of strategies, and the default strategy in onset detection with attack estimation[5] is to calculate the envelope as the sum of columns in a spectrogram using a hanning window of 100 ms with hop factor at 10%.

3.2 Extracting attack phase descriptors

The two toolboxes we explore in this paper use different terminology. In the MIRtoolbox the term 'onset' is used to describe the sound events extracted by peak detection of the energy envelope. This makes sense when seen in relation to the previously mentioned MIREX audio onset competition, but is distinctly different to the meaning of onset in our paper. Both toolboxes provide estimates of when the attack of a sound event begins and ends.[6] Both also provide several directly or indirectly related descriptors (such as attack slope and temporal centroid).

Peeters et al. [2] argue that while a common approach in estimating the beginning and end of an attack range is to apply fixed thresholds to the energy envelope, more robust results may be obtained by looking at the slope of the energy envelope before it reaches its peak value. Trumpet sounds are mentioned as a particular reason for this, as their energy envelopes may increase after the attack. Both of the toolboxes we discuss take such an approach, however with different strategies. The two strategies are illustrated in Figure 3, and explained further below.

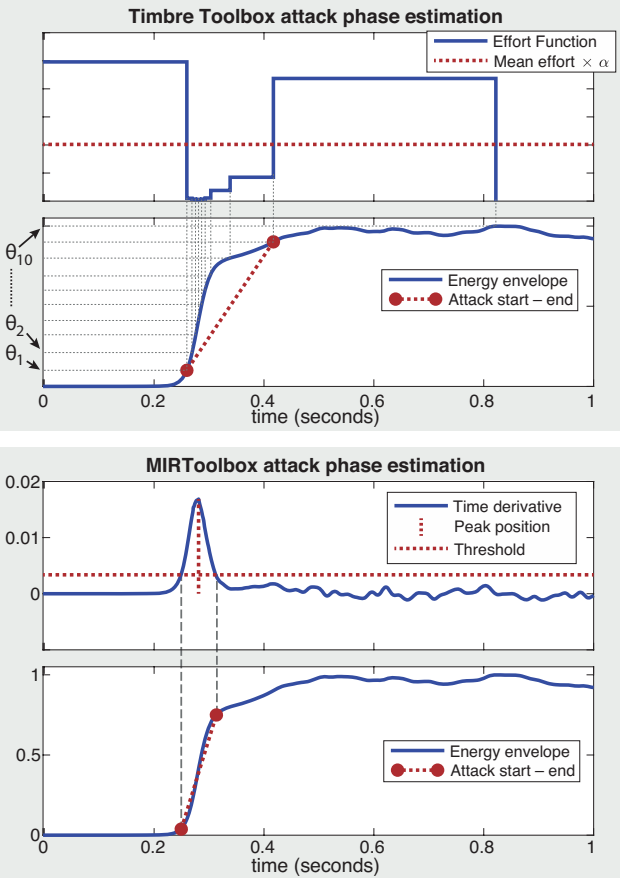

Figure 3. Attack estimation in the Timbre Toolbox and MIRtoolbox. The figure shows the process of estimating the attack start and attack end. The Timbre Toolbox uses an approach that examines the time interval between successive energy thresholds. The MIRtoolbox puts a threshold on the first derivative of the energy envelope. For comparability, this figure uses an identical energy envelope (from a clarinet sound) for both algorithms.

[4] Examples of delayed envelopes are shown Figure 4 in Section 4.2.

[5] This is the default strategy for MIRtoolbox *onset detection*, other functions in the toolbox rely on different calculations of the energy envelope.

[6] In the MIRtoolbox, the attack range is estimated by passing an optional 'Attacks'-argument to the `mironset()` function. As per version 1.6.2 of the toolbox, the `mirgetdata()` function will return the energy peak rather than the extracted attack region. For an object `A` containing the analysis of an audio file, the required code to obtain the attack start is `uncell(get(A,'AttackPosUnit'))` and to obtain the attack end, the code `uncell(get(A,'PeakPosUnit'))` may be used.

Peeters [11] suggested the *weakest effort method*, which is implemented in the Timbre Toolbox for estimating the beginning and end of the attack. In this method, a set of i equidistant thresholds θ_i is specified between zero and a peak value for the energy envelope. For each θ_i, a corresponding t_i denotes the first time when the energy envelope crosses the corresponding threshold. Then a set of 'effort' parameters $\omega_{i,i+1}$ is calculated as the time between each t. With the mean effort notated as $\overline{\omega}$, the beginning of the attack is calculated as the first $\omega_{i,i+1}$ that falls below $\overline{\omega} \times \alpha$, and the end as the first sucessive ω_i that exceeds $\overline{\omega} \times \alpha$. Both are also adjusted to the local maximum/minimum within the $\omega_{i,i+1}$ interval. The default value for α is 3.

The MIRtoolbox uses the first time derivative of the energy envelope (e') as basis for estimating attack phase descriptors. A peak in e' necessarily occurs before the energy peak itself, and the attack is predicted to begin when e' goes above 20% of its peak value, and end when it falls below 20% of its peak value.

4. ANALYSIS

After the previous look at the toolbox algorithms, we will now compare their calculations to results from a perceptual experiment. First, we describe the experiment, followed by an analysis of attack estimation by the two toolboxes using their default parameter settings. Subsequently, we perform a parameter optimisation on the toolbox algorithms to improve the alignment between the calculated estimations and our perceptual results.

4.1 Experiment

In our experiment 17 participants were asked to align a click track to a set of 9 sound stimuli. Each stimulus was presented repeatedly at an interval of 600 ms, along with a sequence of clicks every 600 ms. For each trial, the click track and stimuli started with a random offset. Participants adjusted the alignment of the clicks to the stimuli using a keyboard and/or a slider on the screen, until they perceived the two streams as synchronous. The task was repeated four times per stimulus. All were given a gift certificate worth NOK 200 (≈ 22 €) for their participation.

Eight isolated instrumental sounds were chosen as stimuli. These were selected with an aim to obtain stimuli with different perceptual characteristics and musical function. The click-sound from the click track was also included as a stimulus. The physical onset of each stimulus was annotated manually through inspection of the waveform and used as reference in the results presented below.

The results from the experiment show that our participants did not agree on an exact point in time where the stimulus and click tracks were aligned. Alignment of the click sound to itself was an exception, with a standard deviation of only 1 ms. The agreement on the location of the perceptual attack varied with the characteristics of the sound stimuli. Not counting the task of aligning the click track to itself, the standard deviations of the time delay between click track and stimulus track ranged between 7 ms (for bright sounds with fast attack and short duration) and 18 ms (dark sounds with slow attack and long duration). This verifies previous research, which suggests that perceived attack may best be modelled as a range (i.e. a 'beat bin' [12]) or probability distribution [6], rather than a single point in time. Consequently, we cannot evaluate the toolboxes based on a single value (e.g., their estimation of the beginning of the attack), but rather need to determine the amount of overlap between the calculated attack range and the distribution of perceptual responses.

The use of a simultaneous click track and stimulus track may provide fusion cues that are not inherent to the perceptual attack of the stimulus. Several scholars have argued that an alternating sequence between stimulus and click might be more reliable [9, 13]. We controlled for the effects of event fusion by a corresponding anti-phase alignment task, alternating click and stimulus. The results showed no significant difference between the two response modes [14], which is also in accordance with Villing's reported concistency across these measurement methods [7].

4.2 Energy envelope of the stimuli

In our analysis, we found that the toolboxes' default settings extract imprecise energy envelopes for certain types of sound. In particular, we observe that too long energy envelopes are calculated for short sounds, and that the energy envelopes fail to identify fast attacks in low-frequency sounds with brief high-frequency onsets (e.g. a bass drum sound). This confirms the finding of Kazazis et al. [3], that attack times for fast attack envelopes are largely overestimated in both toolboxes. As previously mentioned, the Timbre toolbox does not compensate for group delay in the lowpass filter. Consequently, the energy envelopes are not in sync with the waveform, and in the most extreme cases hardly overlap at all. The effect of the group delay can be seen in Figure 4, which shows estimated energy envelopes of a kick drum sound. The figure also shows the smearing that occurs with default parameters (D), and how a tighter fit to the waveform may be obtained with an alternate parameter setting (A) for each toolbox. The alternate settings in this figure correspond to a cutoff frequency of 40 Hz for the lowpass filter in the Timbre toolbox, and for the MIRtoolbox a frame size of 50 ms with 2% hop factor.

For the results presented in the following section, we used Matlab's `filtfilt` function to compensate for the group delay in the Timbre toolbox. These results were marginally better when compared to the version with group delay.

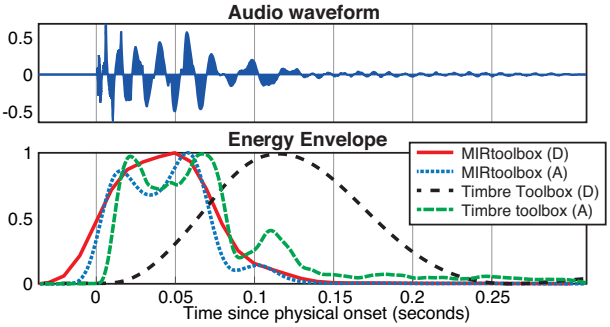

Figure 4. Normalised energy envelopes extracted by the toolboxes. D = default settings, A = alternate settings.

4.3 Attack detection

The toolboxes estimate the attack range using the algorithms presented in Section 3.2. In addition to envelope extraction parameters, each toolbox relies on a predefined threshold, the default values being α = 3 for the Timbre Toolbox and 0.2 times e'_{peak} for MIRtoolbox.

We have compared the attack range estimated by the two toolboxes to the time span covered by mean±SD in our perceptual experiment. We acknowledge the problem of comparing two such different measures (one denoting a time range, and the other a distribution of single time points). However, we argue that a good computational estimate of a perceptual attack range should overlap with the majority of responses in the perceptual data.

The results of attack detection in both toolboxes are illustrated in Figure 5. The figure shows results from default parameters and an optimised set of parameters as will be explained in Section 4.4. For each stimulus, the four vertical lines indicate the four different approaches. The boxes and black horizontal lines indicate the results (mean±SD) from our perceptual test. In summary, the default settings for both toolboxes result in quite long onset periods, compared to the range indicated by the results from the perceptual experiment. The mean estimated attack time (duration of attack) for all the sounds were 48 ms (MIRtoolbox) and 96 ms (Timbre toolbox), while the mean interval corresponding to two standard deviations from our perceptual results was 22 ms.

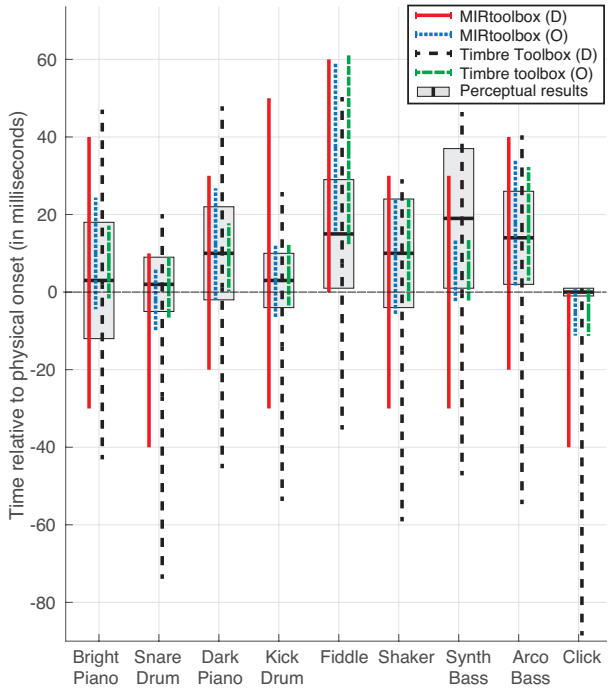

Figure 5. Default (D) and optimised (O) estimates of attack ranges of the 9 stimuli. Boxes show perceptual results for each sound file (mean±SD). Physical onset (time 0) was manually determined by inspecting the waveform.

4.4 Optimisation

Although several of the algorithmic parameters are hard-coded (i.e. not intended for user adjustment), one may modify the code in order to run an optimisation algorithm on the parameters. We have used a two-dimensional grid search optimisation on two parameters for each toolbox: one envelope parameter and one threshold parameter, to minimise the difference between the toolbox results and our perceptual data. For each parameter setting, we computed the Jaccard Index [15] between the estimated attack range and the time span covered by our experimental results (mean±SD). This is a measure of the amount of overlap between the two, and takes a value of 1 if the two ranges are identical, and 0 if there is no overlap. For the default parameter settings, the mean Jaccard Index across all sounds were 0.41 (MIRtoolbox) and 0.25 (Timbre toolbox).

Figure 6 shows the output of our grid search algorithm for the MIRtoolbox. The mean Jaccard Index across the nine sounds was used as fitness measure. 30 settings were tested per parameter, in total 900 parameter settings per toolbox. The results from the parameter optimisation process are shown in Table 2, and the corresponding optimised estimates of attack ranges are shown in Figure 5. The Jaccard Index scores for the optimised parameters were 0.57 for the MIRtoolbox and 0.61 for the Timbre toolbox. The main improvement is the reduced spread of the attack times compared with the default attack estimates, as a result of more detailed energy envelopes.

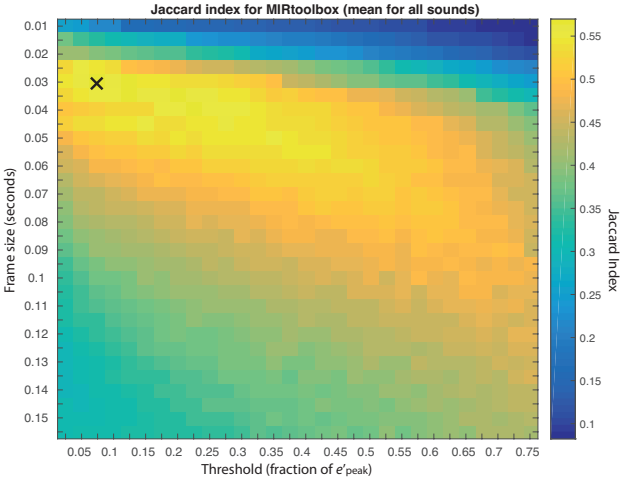

Figure 6. Output of grid search algorithm optimising frame size and threshold parameters in the MIRtoolbox.

Toolbox	Envelope parameter	Threshold parameter
Timbre toolbox	LPfilter cutoff frequency Default: 5 Hz Optimised: 37 Hz	α Default: 3 Optimised: 3.75
MIR-toolbox	Frame size Default: 0.1 s Optimised: 0.03 s	fraction of e'_{peak} Default: 0.2 Optimised: 0.075

Table 2. Parameters for optimisation, default values and optimised results

5. DISCUSSION

Estimating perceptual parameters computationally is difficult. Our results show that attack phase descriptors are no exception. The common first stage in attack detection is extraction of an energy envelope. Both of the toolboxes we have investigated render slowly changing energy envelopes by default. In the MIRtoolbox this comes as result of a large frame size, and in the Timbre toolbox a lowpass filter cutoff frequency of 5 Hz. The default parameters have the advantage of limiting the impact on the energy envelope of rapidly oscillating amplitude modulations in the audible range, in effect reducing the likelihood of false positives in MIREX-style onset detection tasks. However, the low cutoff frequency also hinders accurate representation of fast, non-oscillating amplitude changes (e.g. sounds with a short rise time). Interestingly, the authors of the Timbre toolbox acknowledge the need for a higher cutoff frequency and also address the filter group delay problem in a discussion of sound sample duration [2], but no solution is provided in the toolbox. Furthermore, only one of the four parameters subjected to optimisation in our experiment is easily accessible to the user. The MIRtoolbox allows the user to specify frame size as an input parameter to the `mironsets()` function. The Timbre toolbox cutoff frequency, and the threshold parameters for each toolbox (α and fraction of e'_{peak}) are hard-coded, basically leaving them inaccessible to less experienced users.

As noted by a number of scholars, the perceptual attack of a sound event cannot be measured directly [5–7, 9]. The common approach is to estimate the perceptual attack of a sound by comparing its alignment to some other sound with short duration. This in itself induces some uncertainty into perceptual attack experiments, as the perceptual attack of the reference sound is also unknown. The latter is, however, not a problem in our experiment, given the very precise perceptual results for the test sound (the perceived P-center of the click is 0 and the SD is less than 1 ms). Another complicating factor is that the perceptual attack of most sounds might best be modelled as a range, or a probability distribution, rather than as a single point in time [6]. Consequently, the collection of measured single-point indications of perceptual attack must be considered to represent a time range. We have chosen to represent this range by the by the mean$\pm$SD of our perceptual results. Thus, the widths of all ranges depend on this purely statistical measure. Ranges of different widths, with different corresponding sets of parameter optima, would have been obtained if a dispersion measure other than the standard deviation had been chosen. In future research we aim to investigate if Wright's distribution models for attack time could provide better estimates for the perceived width of the attack range [6].

The limited size of our experiment, with only 9 sound stimuli and 17 participants, engenders a chance of overfitting the parameters to our data. Before the results of parameter optimisation can be generalised, a larger corpus of sound files and more perceptual data must be investigated.

Our results show that the attack estimates provided by the two Matlab toolboxes are largely dependent on the input parameters used. Both toolboxes seem by default not to be oriented towards sounds with a fast rise time. With appropriate parameters, however, both toolboxes may provide estimates closer to perceptual results for a wider range of sounds. The toolboxes are in general excellent tools for sonic research, and may also be used where accurate timing of events is essential. However, we advise to take caution against using the default parameters as perceptual estimates and note that one must carefully select the parameters used in estimation of attack phase descriptors.

6. REFERENCES

[1] O. Lartillot and P. Toiviainen, "Mir in matlab (ii): A toolbox for musical feature extraction from audio," in *Proceedings of the Int. Conf. on Music Information Retrieval*, Vienna, 2007.

[2] G. Peeters, B. Giordano, P. Susini, N. Misdariis, and S. McAdams, "The timbre toolbox: Extracting audio descriptors from musical signals," *The Journal of the Acoustical Society of America*, vol. 130, no. 5, pp. 2902–2916, 2011.

[3] S. Kazazis, E. Nicholas, P. Depalle, and S. McAdams, "A performance evaluation of the timbre toolbox and the mirtoolbox on calibrated test sounds," in *Proc. of the Int. Symposium on Musical Acoustics (ISMA)*, Montreal, 2017 (forthcoming).

[4] P. Schaeffer and G. Reibel, *Solfège de l'objet sonore*, INA-GRM 1998 ed. Paris, France: ORTF, 1967.

[5] J. W. Gordon, "The perceptual attack time of musical tones," *J. Acoust. Soc. Am.*, vol. 82, no. 1, pp. 88–105, 1987.

[6] M. Wright, "The shape of an instant: Measuring and modeling perceptual attack time with probability density functions," Ph.D. dissertation, Stanford University, 2008.

[7] R. Villing, "Hearing the moment: Measures and models of the perceptual centre," Ph.D. dissertation, National University of Ireland, 2010.

[8] J. D. Deng, C. Simmermacher, and S. Cranefield, "A study on feature analysis for musical instrument classification," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 38, no. 2, pp. 429–438, 2008.

[9] N. Collins, "Investigating computational models of perceptual attack time," in *Proceedings of the 9th Int. Conference on Music Perception and Cognition*, 2006, pp. 923–929.

[10] O. Lartillot, "Mirtoolbox 1.6.1 user's manual," Aalborg University, Denmark, Tech. Rep., 2014.

[11] G. Peeters, "A large set of audio features for sound description (similarity and classification) in the CUIDADO project," IRCAM, Tech. Rep., 2004.

[12] A. Danielsen, M. R. Haugen, and A. R. Jensenius, "Moving to the beat: Studying entrainment to micro-rhythmic changes in pulse by motion capture," *Timing & Time Perception*, vol. 3, no. 1-2, pp. 133–154, 2015.

[13] R. Polfreman, "Comparing onset detection and perceptual attack time," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, 2013.

[14] J. London, K. Nymoen, M. Thompson, D. L. Code, and A. Danielsen, "Where is the beat in that note? comparing methods for identifying the p-center of musical sounds," To appear at the *16th Rhythm Production and Perception Workshop*, Birmingham, UK, 2017.

[15] M. Levandowsky and D. Winther, "Distance between sets," *Nature*, vol. 234, no. 5323, pp. 34–35, 11 1971.