The Sound of One Eye Clapping: Tapping an Accurate Rhythm With Eye Movements

Anthony J. Hornof and Kyle E. V. Vessey
Computer and Information Science
University of Oregon
Eugene, OR 97403 USA
{hornof, kvessey}@cs.uoregon.edu

ABSTRACT
As eye-controlled interfaces becomes increasingly viable, there is a need to better understand fundamental human-computer interaction capabilities between a human and a computer via an eye tracking device. Prior research has explored the maximum rate of input from a human to a computer, such as key-entry rates in eye-typing tasks, but there has been little or no work to determine capabilities and limitations with regards to delivering gaze-mediated commands at precise moments in time. This paper evaluates four different methods for converting real-time eye movement data into control signals—two fixation-based methods and two saccade-based methods. An experiment compares musicians' ability to use each method to trigger the playing of sounds at precise times, and examined how quickly musicians are able to move their eyes to trigger correctly-timed, evenly-paced rhythms. The results indicate that fixation-based eye-control algorithms provide better timing control than saccade-based algorithms, and that people have a fundamental performance limitation for tapping out eye-controlled rhythms that lies somewhere between two and four beats per second.

Author Keywords
Eye tracking, gaze-mediated interaction, auditory, timing, anticipation, tapping, synchronization, rhythm.

ACM Classification Keywords
H.5.2. [Information Interfaces and Presentation]: User Interfaces – Input devices and strategies; H.5.2. [Information Interfaces and Presentation]: Sound and Music Computing – Methodologies and techniques; H.1.2 [Models and Principles]: User/Machine Systems – Human factors; Human information processing.

General Terms
Experimentation, Human Factors, Measurement.

University of Oregon
Computer and Information Science Department
Technical Report CIS-TR-2011-01

INTRODUCTION

As the science and practice of human-computer interaction embraces new and alternative modes of input (such as the current excitement around touch screens), it is important to understand the fundamental human-computer capabilities and limitations with each new mode. Though the widespread deployment of eye tracking remains just over the horizon, eye tracking is well-established as a means of interacting with a device [1].

Prior research has investigated the maximum rate of input from a human to a computer via an eye tracker and found maximum eye-typing rates of one character per 0.6 s [3], but there has been little or no work to determine how accurately a person can trigger eye-controlled events at precise moments in time. Such studies have been conducted for *finger* tapping and have found that people can accurately tap out rhythms with their fingers as fast as one tap every 100 ms, and that people tend to tap a few tens of milliseconds *before* the beat, but that this negative mean asynchrony decreases and disappears with musicians [4]. But few or no studies have yet been conducted to determine the fundamental characteristics, such as the fastest accurately-tappable rhythm, for *eye*-tapping.

Nearly all gaze-mediated computer interfaces trigger commands based on the location and duration of gaze *fixations*, which typically last from about 200 to 500 ms. But fixation-detection algorithms typically employ a minimum fixation duration of 100 ms which would impose an upper bound of ten eye-taps per second. Fixations typically alternate with saccades, which move the eyes quickly, on the order of 20 to 40 ms, to a new location. Though not typically used in eye-controlled interfaces, *saccade*-detection algorithms could also be used to trigger eye-controlled commands, and might be faster than *fixation*-detection algorithms and hence superior for precisely-timed eye-commands. Further, a saccade-based trigger might correspond more closely to the muscular control signals and proprioceptive feedback of an eye movement.

This paper describes an *eye-tapping* study that evaluates the best way to process eye tracking data to permit a user to trigger commands at precise moments in time with their eyes. The experimental paradigm is based on finger tapping studies, but conducted with real-time eye movement data. Four different methods are used to process the data to trigger sounds at precise time. Two are fixation-based and two are saccade-based. As with classic tapping

studies, the experiment also investigates the fastest rhythms that musicians are able to match with their eye movements.

METHOD

Participants moved their eyes back and forth between two small squares on a computer display to play handclap sounds (taps) to attempt to match a rhythm of woodblock sounds (beats). The two small squares were centered on the display and separated by 12° of horizontal visual angle; a vertical midline separated the two squares.

The experiment was a 4×3 within-subjects design. The two factors were *trigger method* and *tempo*.

The *trigger method* included two fixation-based methods and two saccade-based methods. The two fixation-detection algorithms were the (a) dispersion-based and (b) velocity-based, both described in [5]. The dispersion-based imposed a threshold of 20 pixels (0.5° of visual angle) and the velocity-based a threshold of 20° per second. Both imposed a minimum duration of 100 ms and triggered taps with the first fixation to cross the midline. The two saccade-based methods were the (d) saccade start detection-method, in which the tap was triggered by the first gaze sample after maximum velocity, and (c) the midline condition, in which the tap was triggered by the sample across the midline. The two small squares on the display served as visual anchors but were not integral to any of the trigger methods.

Tempo refers to the speed of the beats. Beats were played every 0.25, 0.5, or 1.0 seconds. Figure 1 shows the exact timing of the beats within each tempo condition. As can be seen, the 0.25 s and 0.5 s tempos played in triplets whereas the 1.0 s tempo played at a constant rate.

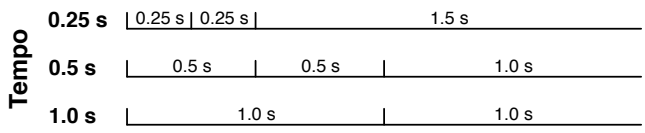

Figure 1. The spacing of the beats across a two-second span for each of the three tempos.

Twelve musicians (nine male and three female), each with an average of ten years of musical training or professional music experience, were recruited primarily from the School of Music and Dance at the University of Oregon. Each participated for about 1.5 hours and completed twenty-four 70-second sessions. Each session included one combination of factors (with the order counterbalanced across participants). The first twelve sessions were to practice all conditions, and the second twelve were to perform all conditions as accurately as possible. Participants earned US$20 plus a bonus of up to US$10 based on their speed and accuracy, which were determined based on the mean clap-to-beat asynchrony and the ratio of attempted handclaps to beats. An on-screen progress bar and textual feedback such as "Super!" and "Try Harder!" (inspired by the video game *Dance Dance Revolution*) provided real-time performance feedback.

Eye tracking data were collected by an LC Technologies monocular 60 Hz eye tracker and analyzed in real time using Cycling 74 Max/MSP 5, which in turn updated a 1280x1024 LCD visual display attached to a dual 2GHz PowerPC G5 running Mac OS X, as described in [2]. A chinrest maintained an eye-to-screen distance of 22 inches. Auditory stimuli and feedback were presented via a pair of Sennheiser HD 250 headphones connected to an M-Audio FireWire Solo interface.

The main performance measure in a tapping task is *asynchrony*, the time between the beat played by the system and the tap played by the participant. A perfect performance would produce asynchronies of zero. Early taps are reported as negative, and late taps are reported as positive. If the participant did not produce a tap for a beat, then no asynchrony was recorded for that beat.

Asynchronies were analyzed using a repeated measures ANOVA with the Greenhouse-Geisser correction. Five percent (1,127 beats) of all beats were excluded in the analysis because their taps were outliers that were more than two standard deviations from the grand mean.

RESULTS

Asynchrony as a Function of Beat Position

Figure 2 and Table 1 show asynchrony as a function of beat position for each trigger method and tempo. Statistical analyses were conducted with the 0.25 s and 0.5 s tempos (all of the 1.0 s tempo beats are essentially in position 1).

As can be seen in Figure 2, across all three tempos, the two fixation-based trigger methods consistently produce taps later than the two saccade-based methods ($F(1.75, 19.3) = 122$, $p < .001$). The 0.25 s and 0.5 s tempos result in overall different asynchronies ($F(1, 11) = 12.1$, $p = .005$), with the 0.25 s tempo producing asynchronies that are overall late and the 0.5 s that are overall early. The general trends *within* each tempo are relatively consistent across trigger method, with the 0.25 s tempo getting 36.9 ms later with each beat position, and the 0.5 s tempo getting 5.4 ms earlier with each beat position. Although the two tempos pull in different directions, the main effect of beat position was not canceled out—asynchrony was still significantly affected by beat position ($F(1.08, 11.9) = 6.95$, $p = .02$), suggesting that the increasing trend in 0.25 s tempo is dominating.

Figure 3 shows how asynchrony increases across the three beats with the 0.25 s tempo but not the 0.5 s tempo; this figure illustrates the only significant two-way interaction, between beat position and tempo ($F(1.18, 13.0) = 41.5$, $p < .001$). Figure 3 also shows how the overall accuracy is better for the 0.5 s tempo than for the 0.25 s tempo.

First-Beat Asynchrony

Beat position 1 in Figure 2 and Table 1 shows how accurately participants could tap on a beat after an interval of 1.0 s (for the 0.5 s and 1.0 s tempos) or an interval of 1.5 s (for the 0.25 s tempo). The general patterns of asynchrony are quite consistent across the three tempos, with the fixation-based methods tapping after the saccade-

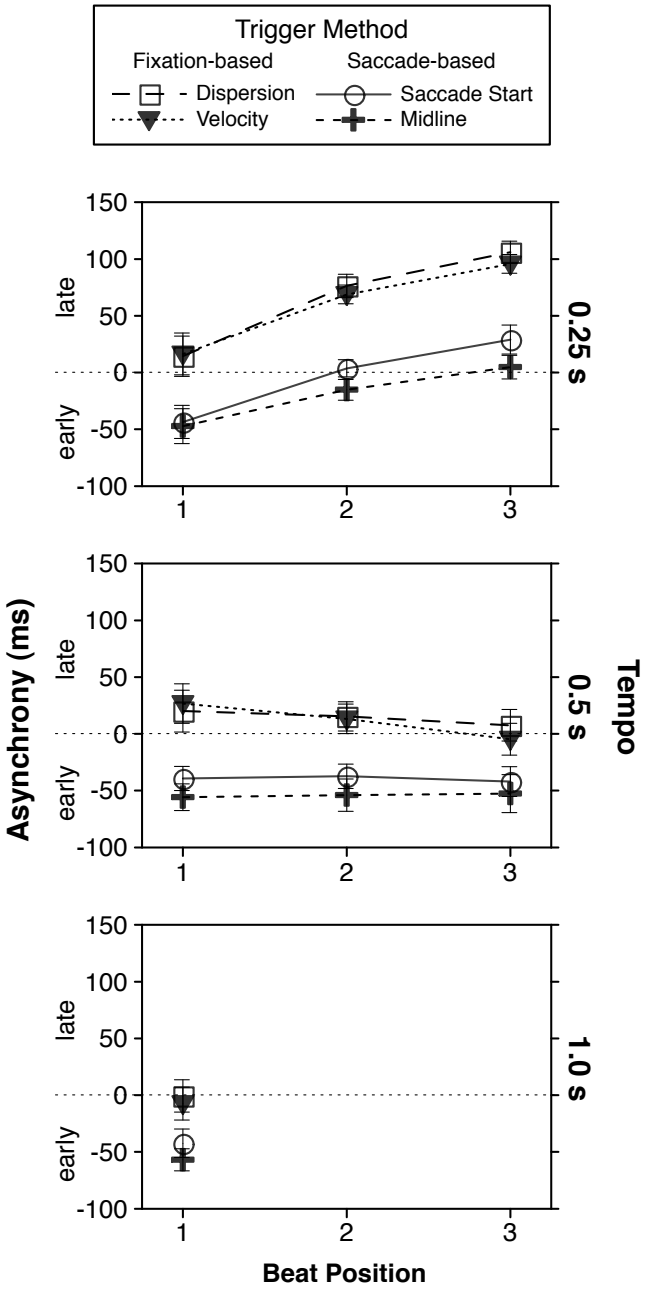

Figure 2. Mean asynchrony, in milliseconds, as a function of beat position, separated by trigger method and tempo, and the standard error of the 12 participant means.

based, but with the fixation-based closer to the beat. Analyses confirm that this first-beat asynchrony is affected by trigger method ($F(1.88, 20.6) = 64.7$, $p < .001$) but not by tempo ($F(1.29, 14.2) = 0.836$, $p = .400$). The consistency of this trend across tempos is supported by the lack of an interaction between trigger method and tempo ($F(4.29, 47.2) = 1.94$, $p = 0.115$).

Saccade-Based vs. Fixation-Based Trigger Methods

Given the similar performance between the two fixation-based methods and the similar performance between the two saccade-based methods, and given that there was no

Trigger Method	Tempo	Beat Position 1	2	3
Dispersion	0.25 s	14.4 (61.5)	76.3 (35.3)	106.0 (33.5)
	0.5 s	20.1 (63.8)	15.4 (44.9)	7.6 (47.9)
	1.0 s	–0.6 (49.2)		
Velocity	0.25 s	16.4 (63.9)	68.8 (28.6)	95.7 (28.4)
	0.5 s	26.7 (59.9)	13.2 (44.9)	–4.7 (48.4)
	1.0 s	–7.4 (50.4)		
Midline	0.25 s	–47.2 (53.3)	–15.2 (31.7)	4.7 (35.9)
	0.5 s	–55.8 (40.7)	–53.9 (49.3)	–52.7 (58.0)
	1.0 s	–56.9 (33.6)		
Saccade Start	0.25 s	–43.5 (50.6)	3.7 (26.1)	29.0 (44.1)
	0.5 s	–39.2 (36.7)	–37.3 (37.1)	–42.0 (45.7)
	1.0 s	–42.4 (43.4)		

Table 1. Mean asynchrony, in milliseconds, and standard deviations (of the 12 participant means).

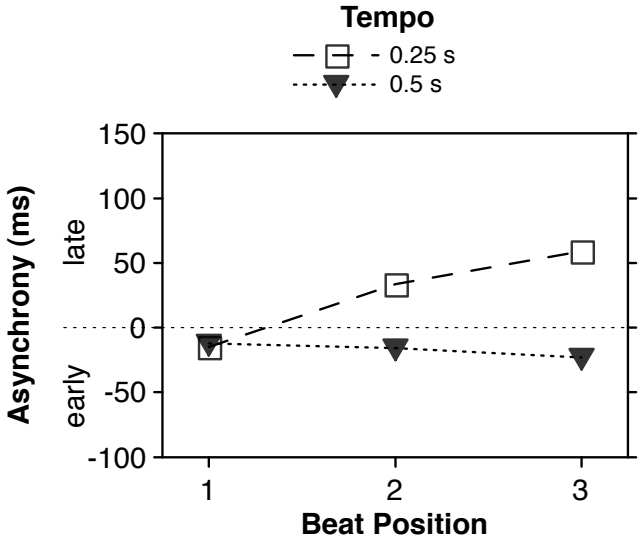

Figure 3. A two-way interaction between beat position and tempo.

significant difference when comparing each pair separately, the same analyses as above were conducted again after collapsing the data by saccade-based method and by fixation-based method. All of the same significant differences appear as when the four trigger methods were analyzed separately. This demonstrates that the differences that were reported above that relate to the trigger-method result from the *type* of trigger method—fixation-based versus saccade-based—that was used, and not the specifics within the two types of methods.

DISCUSSION

The data suggest that, if the goal is to to use an eye tracker to trigger commands at precise moments in time, it is best to process the eye tracking data using a fixation-detection algorithm rather than a saccade-detection algorithm. This is most clearly illustrated in that, across all three tempos, the first-beat asynchrony is consistently more accurate for the

fixation-based methods than for the saccade-based methods. Presumably, the 1.0 s or 1.5 s interval before these first-beats gives the participant the time needed to carefully anticipate, prepare, and execute an optimally-timed eye movement, and that this planning is best-executed in tandem with a fixation-based trigger method.

It might be that saccade-based taps are consistently triggered (an average of 67.1 ms) earlier than fixation-based taps in part because saccade-based methods capture a point in time that occurs *earlier* in the process of moving the eyes, and fixation-based methods capture a later point in time. Given this, it might be the case that neither technique has an overall advantage in terms of triggering a command at a precise point in time, but that together they just capture the middle of a saccade and the start of the subsequent fixation roughly 100 ms later. In other words, it might be that participants simply moved their eyes to the rhythm with little regard for the feedback that was provided by the playing of the handclap sound.

But the evidence suggests that the fixation-based methods are simply superior. The only condition in which the saccade-based method is closer to zero than the fixation-based method is in beat positions 2 and 3 for the 0.25 s tempo. It could be that the saccade-based methods provide better control at this fast tempo, and participants are adjusting across the beats to get closer to zero. It seems more likely, though, that for both classes of trigger methods, participants simply could not keep up with the 0.25 s tempo. A followup experiment might explore what happens when there is a similar beat position 4, 5, and 6.

The data also suggest that the shortest interval that can be achieved between successive eye-taps is somewhere between 0.25 s and 0.5 s, which is faster than the optimal eye-typing rate of one key every 0.6 s, but slower than the optimal finger-tapping rate of one tap every 0.1 s. This maximum eye-tapping rate between 0.25 s and 0.5 s is evidenced by taps getting 37 ms later with each beat in the 0.25 s tempo, but 5.4 ms earlier with each beat the 0.5 s tempo. With the 0.25 s tempo, it appears as if participants just cannot keep up, and that lateness accrues at a rate of 37 ms per beat (which suggests that a 0.3 s tempo might be eye-tappable).

With the 0.5 s tempo, the saccade-based methods consistently tap roughly 45 ms early across all three beat positions, whereas the fixation-based methods tap just 23 ms late on beat 1, and bring the taps to *exactly* on the beat (just 1.4 ms late) by beat 3. It is possible that the consistent (45 ms) early performance is akin to the negative mean asynchrony of a few tens of milliseconds that is routinely-observed in finger tapping experiments [4]. It is also possible that the musicians were able to eliminate the negative mean asynchrony, as has also been observed in finger tapping studies (ibid.). Either way, it is evident that eye-tapping two taps per second is readily attainable.

CONCLUSION

An experiment was conducted to investigate the best way to process gaze samples from an eye tracker to provide the optimal control over the timing of commands issued via an eye tracker. The specific task was to follow three different rhythms with the eyes. Four different methods were used to monitor and capture eye movement data to trigger handclaps. The outcome indicates that fixation-based eye-control algorithms provide more accurate rhythmic and timing control than saccade-based eye-control algorithms, and that people have a fundamental performance limitation for tapping out an eye-controlled rhythm somewhere between two and four beats per second. People can "clap along" with an eye movement two times a second, but not four times a second.

The research presented here is of immediate use in the design of eye-controlled interfaces that require the issuing of commands at precise times, or in rapid sequence, such as for a musician to use his or her eyes to trigger an event at a precise time, or in the design of an eye-controlled interactive experience. This experiment looked at perhaps the simplest possible eye-tapping task, moving the eyes between two dots. Future research will examine how quickly, and with what precision, control decisions can be triggered when there are numerous possible command options, such as with four large buttons on a display.

This study advances the field of human-computer interaction by establishing new knowledge regarding fundamental human capabilities and limitations in an emerging alternative mode of interaction.

ACKNOWLEDGMENTS

This research was funded in part by the National Science Foundation under Grant No. IIS-0713688. Yunfeng Zhang assisted in the experimental design and data analysis.

REFERENCES

[1] Duchowski, Andrew T., *Eye Tracking Methodology: Theory & Practice,* 2nd ed., Springer-Verlag, London, UK, 2007.

[2] Hornof, A., & Sato, L. (2004). EyeMusic: Making Music with the Eyes. *Proceedings of the 2004 Conference on New Interfaces for Musical Expression (NIME04)*. Shizuoka, Japan, June 3-5, 185-188.

[3] Majaranta, P., Ahola, U.-K., & Špakov, O. (2009). Fast gaze typing with an adjustable dwell time. *Proceedings of ACM CHI 2009: Conference on Human Factors in Computing Systems.* New York: ACM, 357-360.

[4] Repp, B. H.. (2005). Sensorimotor synchronization: A review of the tapping literature. *Psychonomic Bulletin & Review, 12* (6), 969-992.

[5] Salvucci, D. D., & Goldberg, J. H. (2000). Identifying fixations and saccades in eye-tracking protocol. *Proceedings of the Eye Tracking Research and Applications Symposium.* New York: ACM Press, 71-78.