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

Anthony J. Hornof and Kyle E. V. Vessey
Department of Computer and Information Science
University of Oregon, Eugene, OR 97403 USA

As eye-controlled interfaces becomes increasingly viable, there is a need to better understand fundamental human-machine 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 examines 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.

As the science and practice of human-machine interaction embraces new and alternative modes of interaction, such as with touch screens, voice-interaction, and brain-computer interfaces, it is important to understand the fundamental human-computer capabilities and limitations with each new mode of interaction. Though the widespread deployment of eye tracking continues to appear top be just over the horizon (Jacob & Karn, 2003), eye tracking is well-established as a means of interacting with a device (Duchowski, 2007).

Prior research has investigated the maximum rate of input from a human to a computer via an eye tracker and found a maximum eye-typing rate of one character per 0.6 s (Majaranta, Ahola, & Špakov, 2009), but there has been little or no work to determine how accurately a person can trigger eye-controlled events at precise moments in time. The control of timing would be particularly important in some assistive technology applications. Previous research (Hornof & Sato, 2004; Hornof, 2009) identified task instances in which precise control of eye-controlled systems could create new opportunities to participate in musical activities, take part in rapid-fire conversations, or permit the careful timing of punchlines to jokes.

Previous research for finger tapping has shown 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 (Repp, 2005). But few or no similar studies have been conducted with eye movements to determine fundamental human capabilities and limitations for eye-tapping.

In everyday eye movements, people move their eyes with rapid *saccades* of roughly 20 to 40 ms, and then hold their eyes steady for *fixations* of roughly 200 to 500 ms (Rayner, 1998). Most eye-controlled computer interfaces (such as Hornof & Cavender, 2005; Majaranta et al., 2009) trigger gaze-based commands based on fixations rather than saccades in part because people can control the location of a fixation and thus issue commands based on looking at locations on a display. Fixations are typically identified by using a fixation-detection algorithm that requires 100 to 200 ms to determine that a fixation has started (Salvucci & Goldberg, 2000). Yet, if 100 to 200 ms are required to identify a fixation, this will introduce delays that a user must correctly anticipate if they are to issue a command at a precise moment in time, as needed for tasks such as playing music.

Saccade-detection algorithms could potentially be used in place of fixation-detection algorithms to trigger eye-controlled commands, such as by detecting that the gaze has crossed a line on the computer display during a saccade. This event could be reported more quickly by an eye tracker, requiring just one or two samples from the eye tracker to report the event. For a 60 Hz eye tracker, this would be just 16.7 ms. No additional 100 ms delay would need to be imposed. Saccade-based detection algorithms might be more responsive and provide superior control for precisely-timed eye-commands. Saccade-based triggers might also correspond more closely to the learned behavior of interacting with control devices through *movement* rather than by holding still.

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 (such as in Repp, 2005), but conducted with real-time eye movement data. Four different methods are used to process the data to trigger sounds at precise times. 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, the *taps*, to attempt to match a rhythm of woodblock sounds, the *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-based trig-

ger methods were (a) *dispersion-based*, in which a fixation is detected when six gazepoints (reported by the eye tracker sixty times per second, for a minimum fixation duration of 100 ms) are reported to be within 0.5° of each other, and (b) *velocity-based* in which a fixation is detected when the movement of the gazepoints across the display holds below 20° per second for 100 ms. In these two methods, the first fixation detected across the midline triggers the tap (the handclap). The two saccade-based methods were the (a) *maximum velocity* detection-method, in which the tap was triggered by the first gaze sample after maximum velocity of the saccade, and (b) the *midline crossed* condition, in which the tap was triggered by the first sample across the midline drawn on the display. 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. The three beats in each triplet are referred to as *beat positions* 1, 2, and 3.

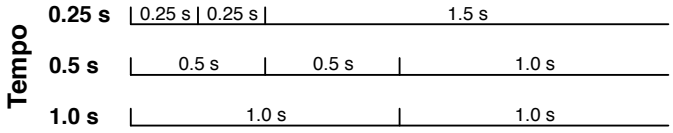

Figure 1. The spacing of the beats (the vertical tick marks) across each two-second span of the experiment, for each of the three tempos. Each timeline loops back to its start.

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 blocks. Each block included one combination of the two factors. The ordering of the blocks was randomized, and counterbalanced across participants. The first twelve blocks were to practice all conditions, and the second twelve were to perform all conditions as accurately as possible. Participants earned $10 plus a bonus of up to $10 based on their speed and accuracy, which were determined based on the time between the tap and the beat, and the ratio of attempted taps to beats. An on-screen progress bar and text such as “Super!” and “Try Harder!” provided real-time performance feedback.

Eye tracking data were collected by an LC Technologies monocular 60 Hz eye tracker and processed in real time using Cycling 74 Max/MSP 5 (similar to Repp, London, & Keller, 2005), which in turn updated a 1280x1024 LCD visual display attached to a dual 2GHz PowerPC G5 running Mac OS X, as described in Hornof and Sato (2004). A chinrest maintained an eye-to-screen distance of 22 inches. Auditory stimuli 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. Consistent with Repp (2005), 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.

As can be seen in Figure 2, across all three tempos, the two fixation-based trigger methods consistently produce taps

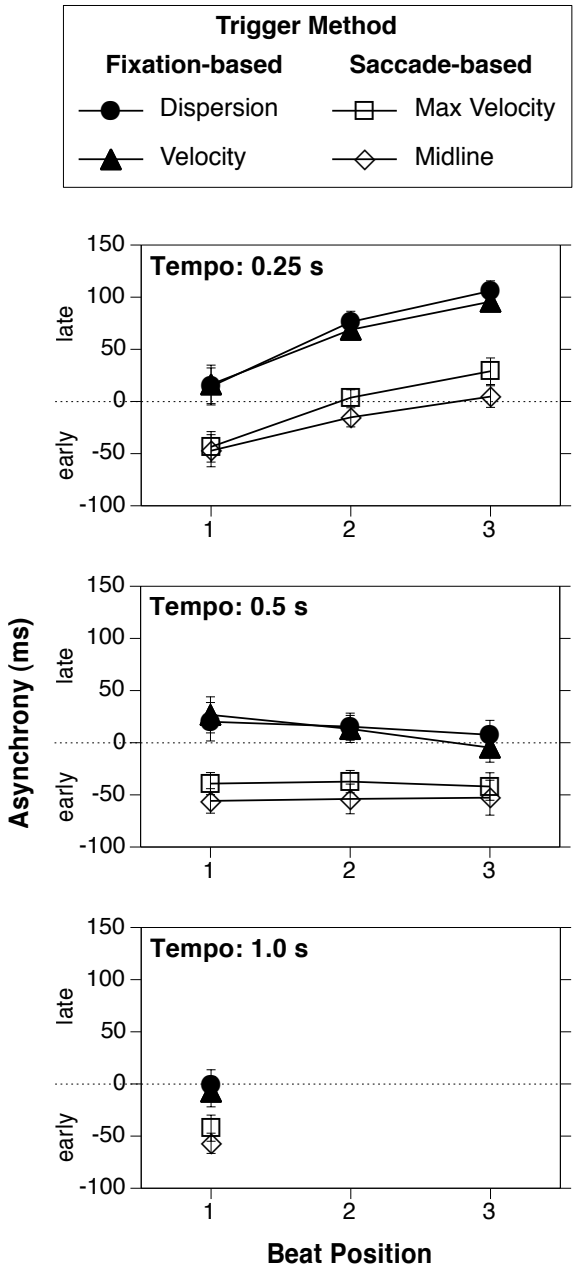

Figure 2. Mean asynchrony, in milliseconds, as a function of beat position, separated by trigger method and tempo. Also, the standard error of the 12 participant means.

			Beat Position		
Trigger Method		Tempo	1	2	3
Fixation-Based	Dispersion-Based	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-Based	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)		
Saccade-Based	Maximum Velocity	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)		
	Midline Crossed	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)		

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

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. For the 0.25 s and 0.5 s tempos, the general trend across beat position is relatively consistent across all four trigger methods. Figure 3 shows these trends.

Figure 3 shows how asynchrony increases across the three beats for the 0.25 s tempo but not for the 0.5 s tempo. The 0.25 s tempo is 36.9 ms later with each beat position, and the 0.5 s tempo is 5.4 ms earlier with each beat position. (The 1.0 s tempo cannot be compared because all of its beats are essentially in Position 1.) Figure 3 illustrates the only significant two-way interaction that was found, between beat position and tempo ($F(1.18, 13.0) = 41.5$, $p < .001$). The figure also shows how the overall accuracy is better for the 0.5 s tempo than for the 0.25 s tempo. 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 dominates.

First-Beat Asynchrony

Beat Position 1 across all three graphs in Figure 2 shows that participants could accurately tap on a beat after an interval of 1.0 s or 1.5 s. The general patterns of asynchrony are quite consistent across the three tempos, with the fixation-based methods more on the beat (closer to 0) and the saccade-based methods consistently about 50 ms early. 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 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.

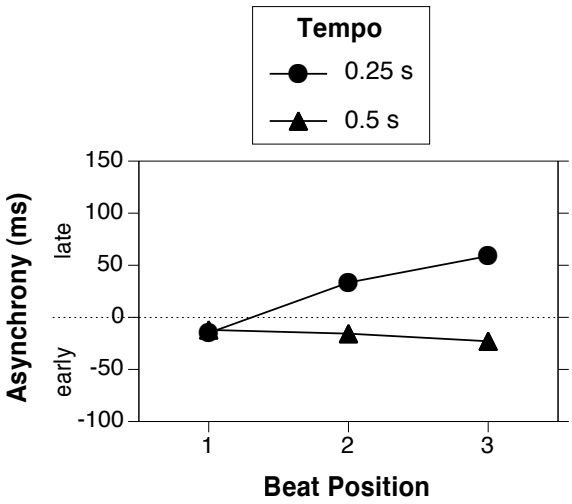

Figure 3. A two-way interaction between beat position and tempo for the 0.25 s and 0.5 s tempos.

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.

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, in which participants just cannot keep up.

It appears as if, with practice and optimal trigger techniques, musicians can eliminate negative mean asynchrony in eye-tapping as they have been observed to do in finger-tapping (Repp, 2005). In the best fixation-based conditions, the negative mean asynchronies were –4.0 ms (for the 1.0 s tempo) and 1.4 ms (for Beat Position 3 of the 0.5 s tempo).

Figure 4 shows a hypothesized timeline of the system and human information processing involved in the task. Each tap likely proceeds as such: The system plays the beat and tap, with the beat played precisely on the rhythm and the tap played based on the person's last eye movement. The person perceives the beat and tap as a single sound event, analyzes it to determine whether the tap was early or late, and adjusts the timing of the next planned eye movement to move the tap

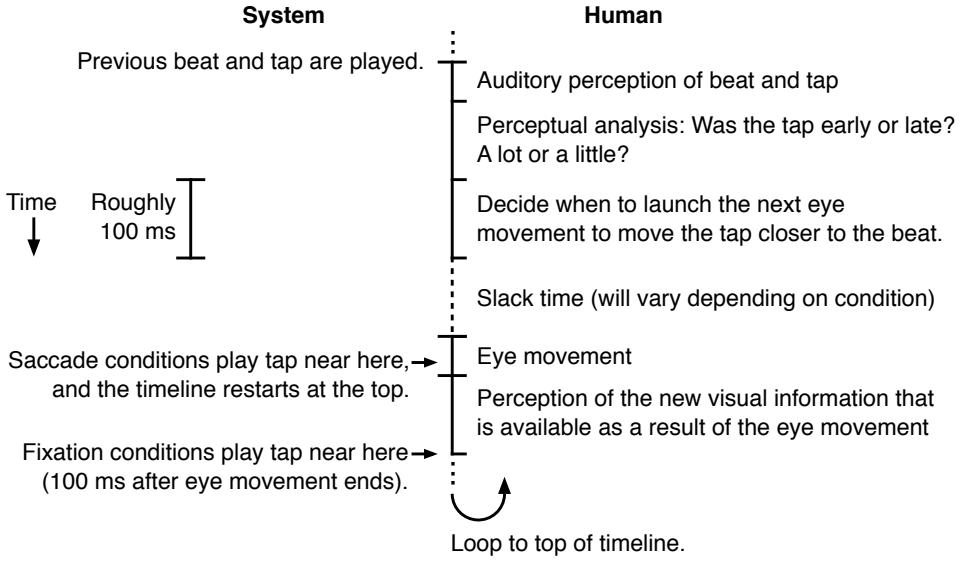

Figure 4. A hypothesized, approximate timeline of the system events and the human information processing involved in tapping on each beat during the eye-tapping task. The target beat at the end of the timeline should occur close to the saccade- or fixation-triggered tap.

closer to the beat. Some slack time elapses during which the person waits for the right time to initiate the movement, after which they launch the eye movement. The system plays the tap, near the end of the eye movement in the saccade-based methods, or roughly 100 ms *after* the end of the eye movement in the fixation-based methods.

The timeline illustrates the unusual nature of the task, which is to initiate a carefully-timed eye movement based on the perception and analysis of a sound event that was triggered by a previous carefully-timed eye movement. The eye-tapping task requires a person to quickly analyze the *auditory* results of a previous eye movement to decide when to make the next eye movement. The processing time required for this analysis and planning appears to be too great to support eye-tapping of four beats per second, even though four eye movements per second are commonly observed in tasks in which the movements are driven entirely by visual processing requirements, such as reading or visual search (Rayner, 1998). There is enough time between beats for all the processing that is necessary to keep up with the 0.5 s tempo, but not the 0.25 s tempo.

In the most accurate conditions observed in the data, the saccade-based methods produced taps roughly 50 ms earlier than the fixation-based methods; this suggests that the participants adjusted to the timing characteristics of the different trigger methods rather than relying on one strategy for all trigger methods. The saccade-based trigger methods did not produce taps earlier than the fixation-based methods simply because the saccade-based methods capture a moment earlier in the timeline. If this were the case, the saccade-based methods would cause the tap to be played more than 100 ms earlier than the fixation-based methods, because this is the time required for a fixation-detection algorithm to detect a fixation. Participants did not simply allocate a fixed amount of slack time across all trigger conditions, but instead clearly adjusted their eye movement times based on perceived differences in the the fixation- versus saccade-based methods.

Understanding how people plan an eye movement to trigger a sound using each type of method might be better understood by exploring more tempos between 0.25 s and 0.5 s. Since the saccade-based methods respond to an eye movement 100 ms earlier than the fixation-based methods, if the two types of methods require the same time to plan the next movement, the saccade-based methods would seem to have a 100 ms advantage over the fixation-based methods. If there is no tempo with which, across the three beat positions, people can maintain a constant asynchrony using the saccade-based but not the fixation-based methods, this would suggest that the saccade-based methods require more analysis and planning time. The question of how people do the task could similarly be explored by reducing the 100 ms required to detect a fixation down to just two eye tracker samples, assuming that as soon the eyes arrive at a trigger location (the small squares on the display), the fixation has started.

That people can accurately hit the beat with fixation-based methods but are consistently early with saccade-based methods might relate to a person's ability to imagine and predict the temporal relationship between (a) perceptual and motor events that a person can easily and correctly anticipate and (b) the sound of the tap being played. It may be that the fixation-based methods, which play the tap roughly 100 ms after the eye movement, play the tap at a moment that corresponds to perceptually salient events such as visual features becoming available after the eye movement, or to a point in time that can be imagined in relationship to salient events such as the world having just changed its position after an eye movement. And perhaps the saccade-based methods play taps at moments that are more difficult to connect to salient perceptual or motor events. In fact, the visual system even suppresses sensory information during a saccade (Rayner, 1998), creating a sort of *non*-event, which might be particularly difficult to anticipate. Perhaps proprioception of the launch or the middle of a saccade is simply not readily available to a person

such that they can imagine, anticipate, and plan around either event. Repp (2005) discusses previous models of sensorimotor synchronization that incorporate the imagining of beats and taps for finger-tapping. Perhaps eye-tapping can be explained by similar models.

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 eye movements 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 with severe motor impairments to use his or her eyes to trigger musical events at precise times. This experiment looked at perhaps the simplest possible eye-tapping task, moving the eyes between two small squares. 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 many large buttons on a display.

This study advances the field of human factors by establishing new knowledge regarding fundamental human capabilities and limitations in an emerging mode of human-machine control.

ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under Grant No. IIS-0713688. The authors thank Yunfeng Zhang and the reviewers of this paper for their excellent suggestions for improving this work.

REFERENCES

Duchowski, A. T. (2007). *Eye Tracking Methodology: Theory and Practice.* London: Springer-Verlag.

Hornof, A., & Sato, L. (2004). EyeMusic: Making music with the eyes. In *Proceedings of the 2004 Conference on New Interfaces for Musical Expression (NIME04)*. Shizuoka, Japan, June 3-5.

Hornof, A. J. (2009). Designing with children with severe motor impairments. In *Proceedings of ACM CHI 2009: Conference on Human Factors in Computing Systems.* New York: ACM.

Hornof, A. J., & Cavender, A. (2005). EyeDraw: Enabling children with severe motor impairments to draw with their eyes. In *Proceedings of ACM CHI 2005: Conference on Human Factors in Computing Systems.* New York: ACM.

Jacob, R. J. K., & Karn, K. S. (2003). Eye tracking in human-computer interaction and usability research: Ready to deliver the promises (Section commentary). In J. Hyona, R. Radach, & H. Deubel (Eds.), *The Mind's Eyes: Cognitive and Applied Aspects of Eye Movements.* (pp. 573-605). Oxford: Elsevier Science.

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

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, *124*(3), 372-422.

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

Repp, B. H., London, J., & Keller, P. E. (2005). Production and synchronization of uneven rhythms at fast tempi. *Music Perception*, *23*(1), 61-78.

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