## Results: Experiment 2 – IRF

Dr. Raimondas Zemblys, Ph.D. kindly provided classification of our data by his algorithm (Zemblys et al., 2017), for which we are most grateful. As noted above, the algorithm is based on machine-learning but the paper claims that some logical post-processing of the machine learning results are also performed as part of the IRF algorithm. These post-processing steps are listed in Table 1.

| | **Table 1: Goals of the Post-Processing Steps in the IRF** |
|---|---|
| 1 | mark events that contain more than 75 ms of interpolated data as undefined. |
| 2 | merge fixations which are less than 75 ms and 0.5 deg apart. |
| 3 | make sure that all saccades have a duration of at least three samples, expand if required, which means that if we have a one sample saccade, we also label the preceding and following samples as saccade. |
| 4 | merge saccades that are closer together than 25 ms. |
| 5 | remove saccades that are too short (<6 ms) or too long (>150 ms). |
| 6 | remove PSOs that occur in other places than directly after a saccade and preceding a fixation. |
| 7 | remove fixations shorter than 50 ms. |
| 8 | remove saccades and following PSO events that surround episodes of missing data as these are likely blink events. |

A single rater (first author, LF) evaluated the same 20 recordings as in Experiment -1 as classified by the Zemblys et al (Zemblys et al., 2017) algorithm (the "IRF" algorithm). Since this evaluation occurred after the ONH-MNH comparison, the rater was not blind to the classification method. Reducing the evaluation to 1 vs 3 raters would lower the generalizability of the results to a potential population of raters. Also, with only 1 rater, the issue of inter-rater reliability is not relevant. All the single rater's evaluations of every type of error in the ONH and MNH algorithm were compared to the all of the rater's evaluations of the IRF algorithm.

Statistical significance of differences in error numbers and corresponding effect sizes were calculated as above for Experiment 1.

## Results: Experiment 2 – IRF

*General Impressions*

There are several general impressions that one has after reviewing the classification results from the IRF. First, on the positive side, this algorithm is amazingly accurate for saccade timing. It is a remarkable fact that we did not find a single saccade with an onset timing that was either too soon or too late. And the offset timing, while not as perfect as the onset timing of saccades, was extremely good also. For saccade timing, the IRF wins, decisively. This will clearly be born out when we review the evaluation results.

The second aspect that one notices is that this algorithm does a very poor job of rejecting unusual or artifactual events. As noted above, there are periods, typically during blinks, when the Eye-Link 1000 does not return a position value, but rather indicates missing data ("Not a Number" or NaN). The IRF interpolates across such blinks, treats the interpolated data as if it were good data, and attempts to classify such periods like all the good data in the recording. During post-processing, the IRF removes the longer blink periods from classification (Table 1, Step 2). For our evaluation of the IRF, we simply declared all these blink periods as noise/artifact. Both the ONH and the MNH exclude some data before and after each blink ("peri-blink" data). Although the IRF claims to do this during post-processing (Table 1, Step 8) our results indicate that no such post-processing step was actually conducted (see below). The IRF results are severely contaminated because the IRF attempts to classify these peri-blink recording periods. See for example, Figure 1. It shows what is obviously noise as misclassified by the IRF as fixation, saccade and PSO. If these types of artifactual events are not handled properly, we can have the situation illustrated in Figure 2, where noise is classified as a saccade of 1 msec duration with a peak velocity near 700 deg/sec. Again, this is not supposed to happen with the IRF (Table 1, Step 5). Another way to observe the effects of this poor artifact handling by the IRF is to view the main sequence relationship between saccade peak velocity and saccade length. This is illustrated in Figure 3. Finally, noise also affects the classification of fixation. With the IRF, we have found 6 fixations that are less than 10 msec, even though the post-processing steps (Table 1, Step 7) claim to remove fixations less than 50 msec. No such short fixations were noted for the ONH or the MNH. As noted in the caption to Figure 1, there are a number of fixation periods with extremely high velocity samples with the IRF.

Also, the IRF has several problems classifying PSOs. For example, we have PSOs occurring after fixations, whereas, by definition PSOs occur after saccades only (Figure 4). The IRF is supposed to prevent this (Table 1, Step 8) but apparently this post-processing step also failed. Furthermore, there are some extremely short PSOs when classified by the IRF. Figure 5 illustrates a PSO that is only 1 msec duration. A couple of other issues of PSO scoring with the IRF are illustrated with the frequency histograms of PSO length for the three algorithms (Figure 6).

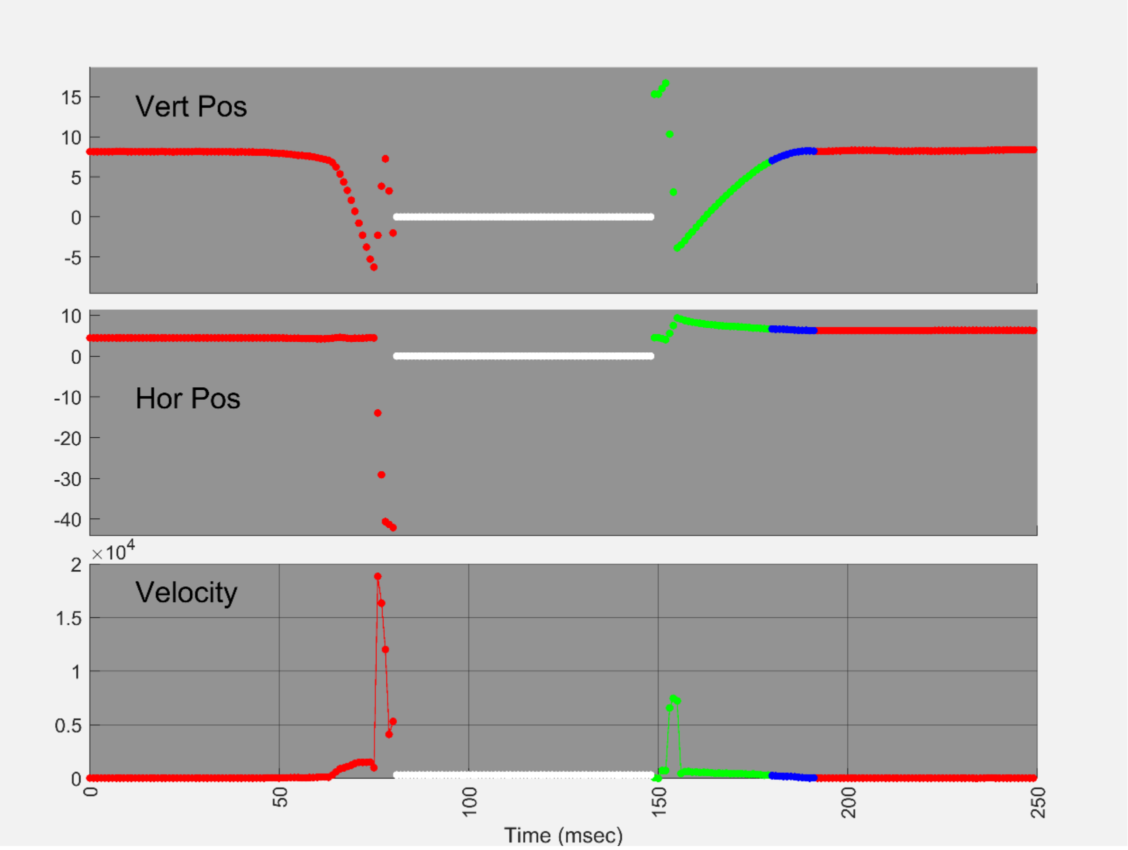


*Figure 1: Illustration of Peri-Blink scoring by the IRF. Fixations are red, saccades are green, PSOs are blue and noise/artifact is white. This is an illustration of the kinds of problems that can occur if one does not remove peri-blink eye-movement data from consideration. Prior to the blink, we have a classification of fixation with a peak velocity near 20,000 deg/sec. With the IRF, there were 128 "fixation" periods with peak velocity above 150 deg/sec and 45 "fixations" that have a peak velocity above 1000 deg/sec. The maximum peak velocity in a "fixation" with the IRF was above 49,000 deg/sec. For the ONH and the MNH the peak velocity during fixation was under 150 deg/sec. After the blink we have a classification of a saccade which is clearly incorrect. This "saccade" is followed by a "PSO" which is also just noise/artifact.*

Also, many of the PSOs as scored by the IRF would not meet our criteria for PSOs (see above). PSOs scored by the IRF often have no velocity peaks that are above that seen in the surrounding random fixation noise. Our criteria is similar to that used for the ONH, as indicated by the algorithm used to detect PSOs with the ONH and also with the example figures (Figures 1 and 9) illustrating PSOs in the original ONH paper (Nyström & Holmqvist, 2010). Figure 7 shows 4 events classified as PSOs by the IRF, that would not meet our criteria or the criteria for the ONH as PSOs[1].

[1] For a more full discussion of the PSO detection in the Zemblys et al. (Zemblys et al., 2017) paper, see the document labelled "Report on PSO Detection In the Zemblys et al (2017) Paper.docx" at: https://digital.library.txstate.edu/handle/10877/6874

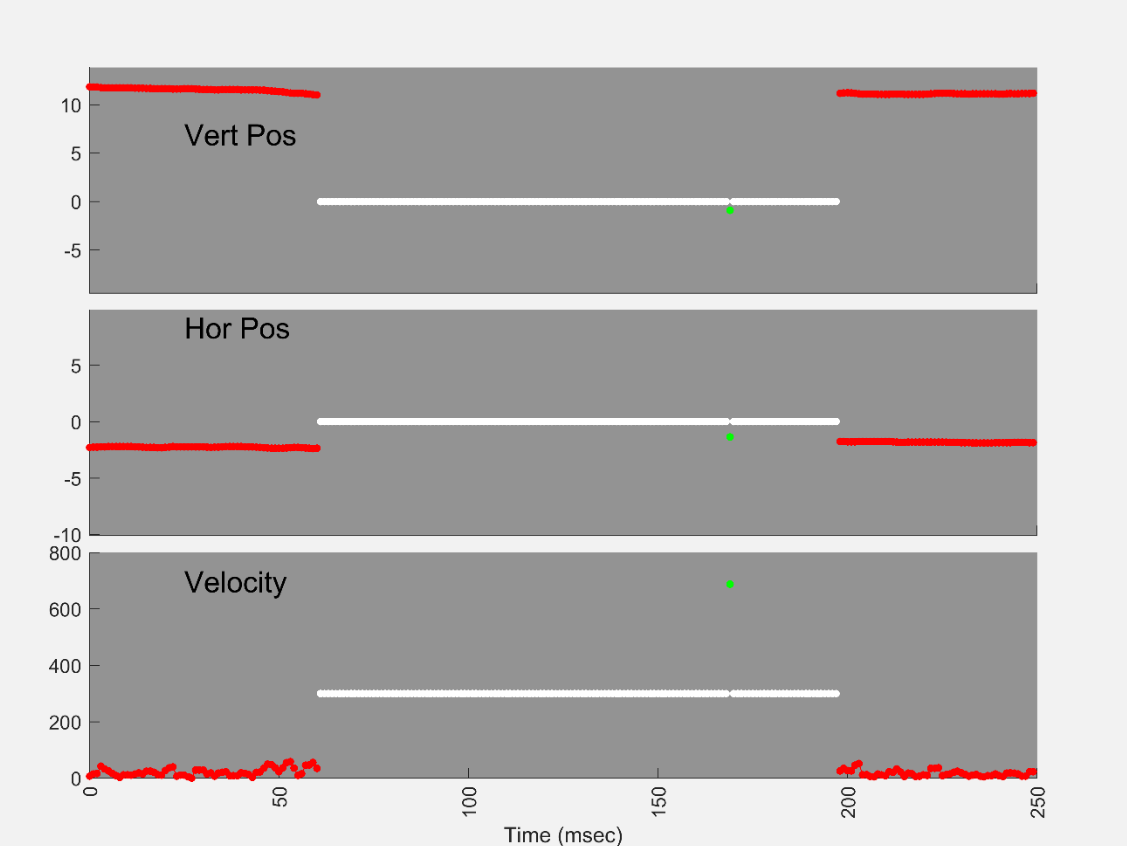


*Figure 2: See caption for figure 1. Here we have a noise period (blink), with a single sample declared by the Eye-Link 1000 as good data. The IRF scores this 1 msec sample as a saccade. Its duration is therefore 1 msec, and its peak velocity is about 700 deg/sec. Clearly, such events are going to distort any main sequence relationships, as shown in Figure 3 below.*

*Error Classification for the IRF versus the ONH and the MNH.*

Table 2 present the error numbers for errors of various types. Only comparisons with a total of more than 20 events between the two compared algorithms are included. These data are from the scoring by the first author only. Next to the error numbers are the p-values for the comparison between error rates. In the final column we present the estimated effect size (Cohen’s d) for each comparison (absolute value). Blue highlighting indicates that error numbers were significantly higher for the comparison algorithm and yellow highlighting indicates that the error numbers were higher for the IRF algorithm. For the ONH-IRF comparison, the ONH had higher numbers of errors than the IRF for 4 error types and the IRF had more errors than the ONH for 4 types of errors. For the MNH-IRF comparison, the MNH has significantly more errors than the IRF for one error type and the IRF had more errors than the IRF for 6 error types.

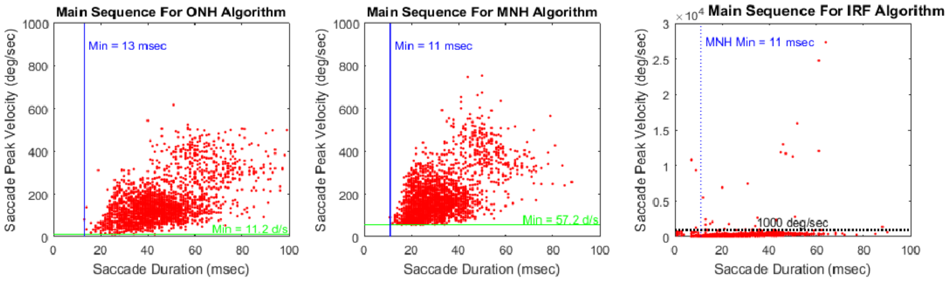
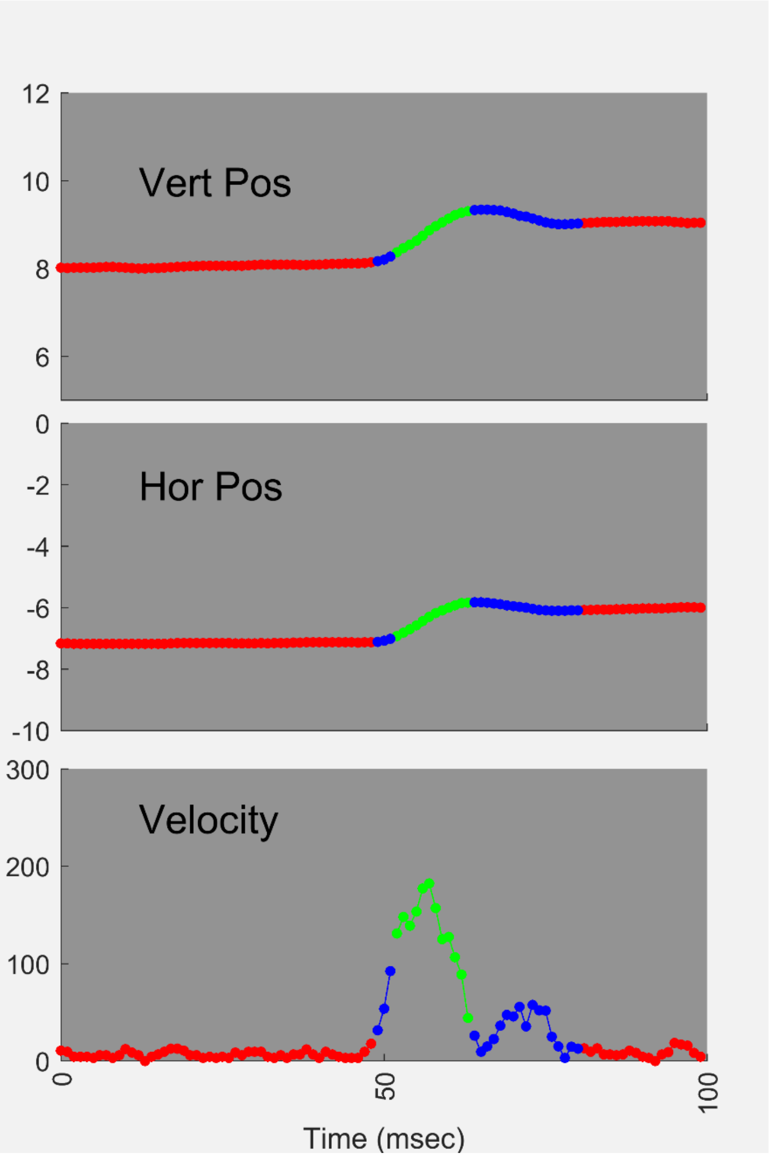


*Figure 3: Main sequence relationships (peak velocity of saccades plotted against saccade duration) for all three classification methods. On the left is data for the ONH, the MNH data is in the middle and the IRF data is on the right. For ONH, the minimum peak velocity for a saccade was 11.2 deg/sec. This is probably physiologically impossible, but the results stem from the problematic "adaptive threshold" feature of the ONH. For the MNH, the minimum peak velocity was 57.2 deg/sec, which seems reasonable. For the IRF, the minimum peak velocity was approximately 28 deg/sec, which also seems low. For the ONH and the MNH, all of the peak velocities of the saccades are less than 800 deg/sec. For the IRF, there are 34 saccades with peak velocity greater than 1000 deg/sec, and one with saccade with a maximum peak velocity above 27,000 deg/sec.*

Given the absence of methods to handle unusual or artifactual events in the IRF, there are many more noise periods that were classified as fixation with this method versus the ONH or the MNH methods. The IRF also had more noise periods classified as saccades than the MNH but not the ONH. There were many more fixation periods misclassified as PSOs for the IRF versus the ONH and more dramatically for the MNH. The IRF had fewer fixations and PSOs that were not detected than the ONH. The IRF had 0 saccades that start too early, whereas the MNH and especially the ONH had many more such events. Although the numbers are small, the IRF had more saccades that end too early than either the ONH or the MNH. The IRF has fewer "saccades that end too late" errors than both other algorithms, although the effect was statistically significant only versus the ONH. The IRF has significantly more PSOs that end too early than either other algorithm. The IRF also has more PSOs that end too late than both algorithms, but this increase was statistically significant versus the MNH only.



more end

*Figure 4: Illustration of a "PSO" occurring after fixation, as scored by the IRF. See caption for Figure 1. Although this only occurred 4 times in all of our recordings, the fact that it ever occurs is one indication that there is a problem with the classification of PSOs with the IRF.*

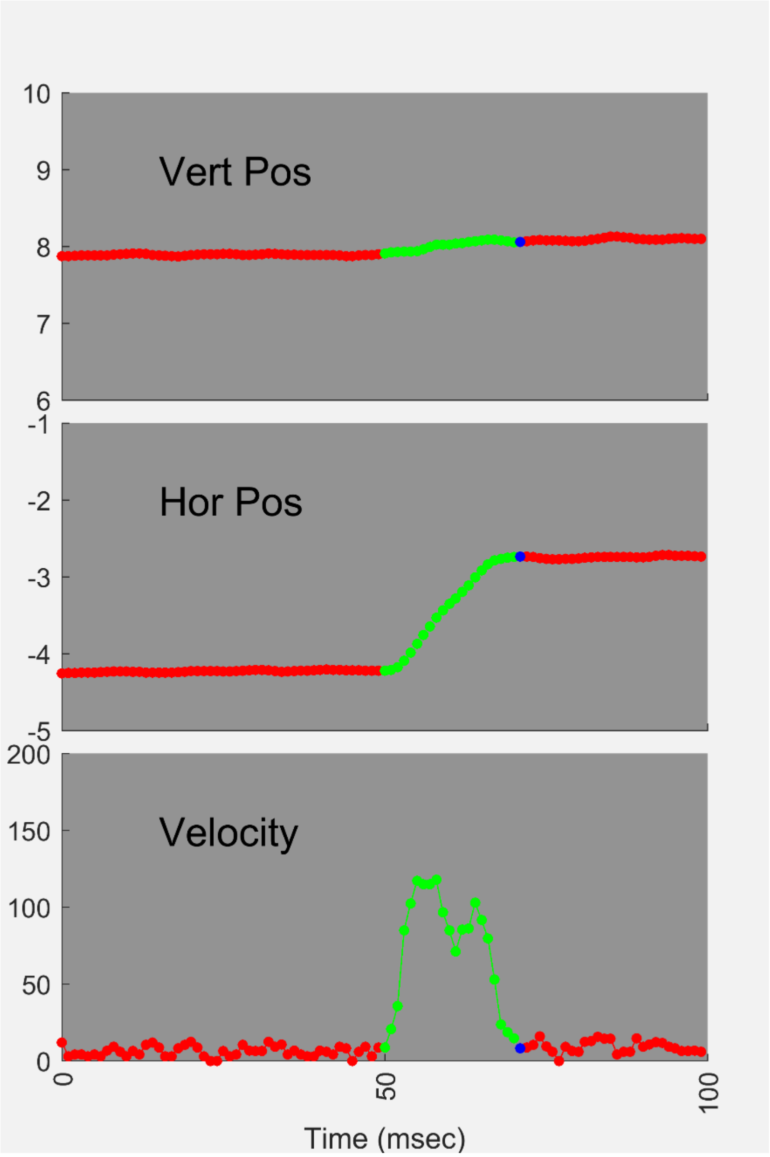


*Figure 5: Illustration of a 1 msec PSO scored by the IRF.  There we 23 PSOs that were 1 msec in duration with the IRF.*

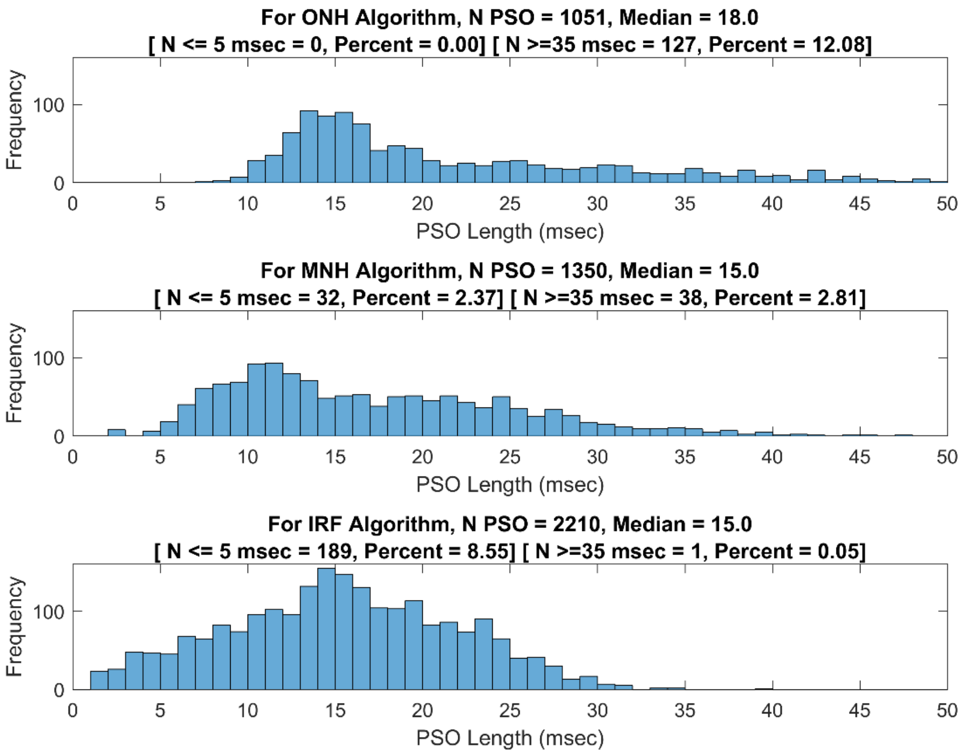


*Figure 6: Frequency histograms of PSO length for the three algorithms. The first thing to notice is that the IRF scores more than twice as many PSOs than the ONH (2,210 vs 1,051) -- an increase of 110%. Similarly, the IRF scores 860 more PSOs than the MNH – an increase of 64%. Also, some of the PSOs scored by the IRF are very short. There were 189 PSOs scored by the IRF with a duration less than or equal to 5 msec, whereas there were no such short PSOs for the ONH and only 32 for the MNH. As noted above, there were 23 PSOs scored by the IRF that were 1 msec in duration.*

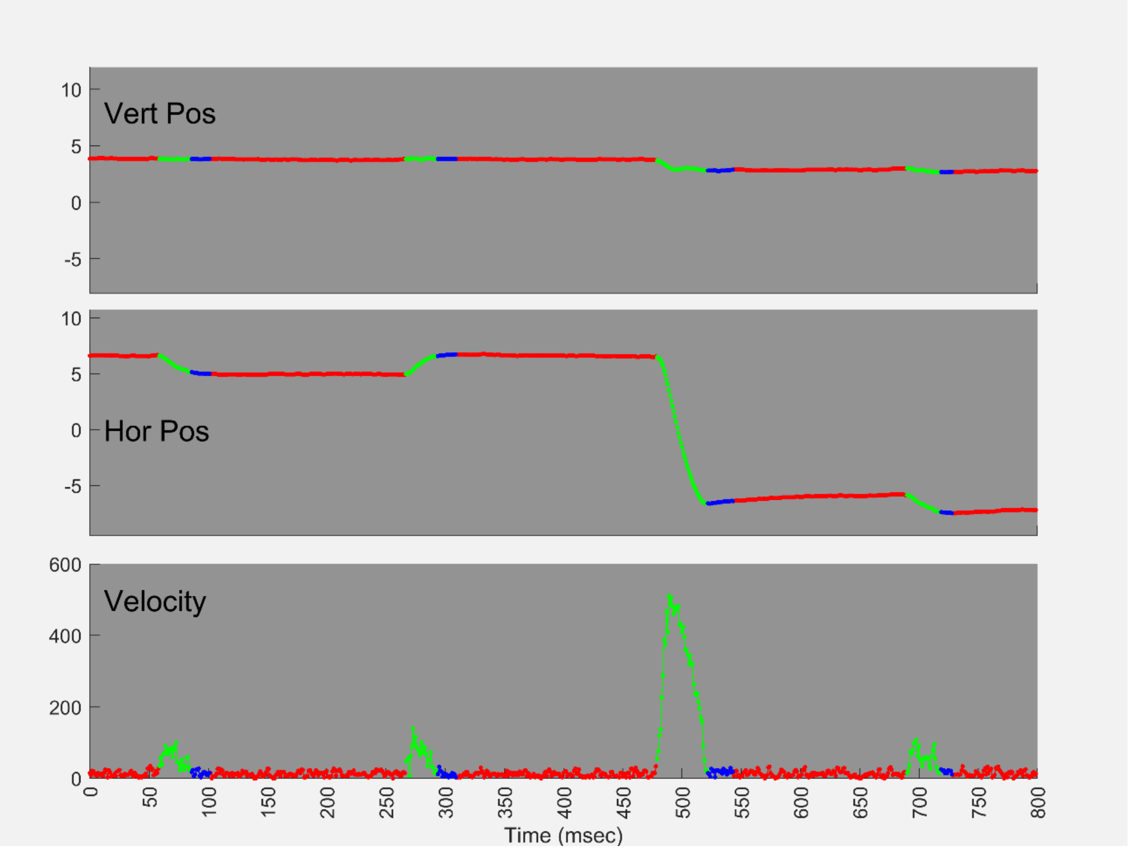


*Figure 7: Illustration of 4 PSOs classified by the IRF that do not meet typical criteria for PSOs. See caption for Figure 25. The PSOs identified here have no velocity peaks that are distinctly above random noise peaks during fixation. There were many hundreds of such PSOs as classified by the IRF.*

**Table 2: Comparing Error Numbers between the IRF and the ONH (Top) and the MNH (Bottom) Algorithms**

| | | ONH vs IRF | | | |
|---|---|---|---|---|---|
| Error Number | Error Name | Total Errors: ONH | Total Errors: IRF | p-value | Abs(d) |
| 1 | Number of Noise Events Misclassified as Fixation | 0 | 94 | 0.0000 | 11.93 |
| 2 | Number of Noise Events Misclassified as Saccade | 15 | 29 | ns | |
| 7 | Number of Fixations Misclassified as PSO | 93 | 524 | 0.0127 | 1.39 |
| 8 | Number of Fixations Not Detected | 231 | 1 | 0.0000 | 37.93 |
| 12 | Number of Saccades Not Detected | 33 | 0 | 0.0001 | 5.77 |
| 25 | Number of Saccades That Start Too Early | 1515 | 0 | 0.0000 | 37.93 |
| 27 | Number of Saccades That End Too Early | 3 | 33 | 0.0063 | 1.70 |
| 28 | Number of Saccades That End Too Late | 824 | 76 | 0.0000 | 37.93 |
| 31 | Number of PSOs That End Too Early | 13 | 50 | 0.0129 | 1.38 |
| 32 | Number of PSOs That End Too Late | 33 | 72 | ns | |

| | | MNH vs IRF | | | |
|---|---|---|---|---|---|
| Error Number | Error Name | Total Errors: MNH | Total Errors: IRF | p-value | Abs(d) |
| 1 | Number of Noise Events Misclassified as Fixation | 0 | 94 | 0.0000 | 11.93 |
| 2 | Number of Noise Events Misclassified as Saccade | 1 | 29 | 0.0005 | 3.54 |
| 7 | Number of Fixations Misclassified as PSO | 2 | 524 | 0.0000 | 127.78 |
| 25 | Number of Saccades That Start Too Early | 263 | 0 | 0.0000 | 37.93 |
| 27 | Number of Saccades That End Too Early | 6 | 33 | 0.0063 | 1.70 |
| 28 | Number of Saccades That End Too Late | 104 | 76 | ns | |
| 31 | Number of PSOs That End Too Early | 20 | 50 | 0.0063 | 1.70 |
| 32 | Number of PSOs That End Too Late | 2 | 72 | 0.0000 | 22.51 |