Searching for Causes of Poor Classification Performance in a Brain-Computer Interface Speller

Michael Motro

Advisor: Leslie Collins

Department of Electrical and Computer Engineering

Duke University

April, 2014

Abstract

The BCI Speller is a machine that flashes different letters on a screen and collects EEG data from the viewer, then analyzes the data to determine what letter the viewer was concentrating on. This allows a user to spell words without any body movements, which is useful for cases of heavy paralysis where speech is inhibited. The neurological event that this machine searches for is the P300, a pulse of electrical activity throughout the brain about 300 milliseconds after an unlikely event. However, the P300 has been noted as a varying event, and likewise the effectiveness of the Speller is highly variable: some tests will result in perfectly accurate spelling, while another test with the same user and conditions can be so inaccurate as to guess no letters correctly.

The focus of this research was to search for a fundamental difference between high-scoring and poor-scoring Speller tests. This could point towards a physical cause of inaccuracy, or could guide the way for a classification system that performs better in the case of a poor session.

Two signal features were determined particularly likely to be the cause of classification error – variance in the latency time of the P300 and the occurrence of large-scale noise artifacts. It is difficult to mathematically define the quantity of either of these features; instead, classification methods that account for these features were designed and implemented on test sessions.

Background

The P300 event potential was discovered by Sutton *et al.* in 1965 [1]. It can be evoked using what is known as the oddball paradigm, where an unlikely target event is placed amidst likely events. The P300's reliable latency and low-frequency makeup makes it useful for measuring reaction to a stimulus. Farwell and Donchin [2] first developed a Spelling machine based around this paradigm. There are now many variations on the concept, but the basis is always a screen with characters that can be bright or dim, and an EEG cap to measure voltage across the brain.

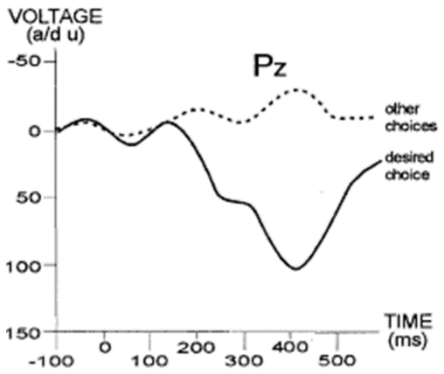


Figure 1: P300, from Wolpaw *et al.,* 2002 [3]

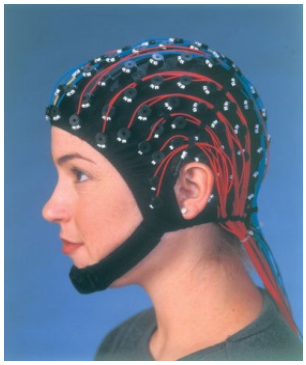


Figure 2: Example pictures of a row-column speller screen [4] and an EEG cap [5].

The user's task is to concentrate on the letter they wish to spell, and count the number of times it lights up briefly. Each flash of the correct letter should evoke a P300 response, but to the noise omnipresent in EEG measurements each letter is flashed many times. In addition, both the P300 and noise activity will vary across users and tests, so in order to determine the correct flashes a supervised system is needed. Usually several previously determined words will be spelled so that the algorithm learns how to differentiate between a correct letter's flash (target) and an incorrect letter's flash (non-target).

Data Used

The Speller machine data analyzed in this report was gathered during tests in summer 2010 by the SSPACISS group at Duke University. There were twenty one users in all, with most users performing multiple test sessions for a total of sixty test sessions.

Each session consisted of seven to nine words, always adding up to 35 letters spelled. For each letter to spell, every character on the speller grid was flashed 20 times. The speller grid was set up in a row-column paradigm so that every flash selected a row or a column, alternating. The number of possible rows and columns is 9, so that a single spelled letter requires 180 flashes. With a quarter of a second gap between flashes and pauses between letters, spelling proceeds at roughly a letter a minute.

The initial method of data processing and classification was based on the results from Krusienski *et al.*, 2007 [6]. Eight EEG electrodes were used, those determined to have robust accuracy. The voltage data from each electrode was filtered and downsampled to 20 Hz, which removes a significant amount of high-frequency noise and speeds up computation. For each speller flash, a feature set of 800 milliseconds was chosen. As the time between flashes is only a quarter of a second, each flash shares information with several before and after it. It is possible that adjacent target flashes would create P300's in close proximity to each other, and this issue is examined later.

In a machine learning context, each there were 180 observations per letter spelled, with 20 containing the target letter. For each 35-letter session, the number of observations totals 6300. 800ms of 20Hz voltage data from 8 channels was chosen as the feature set for each observation, totaling 128 features. The primary classification algorithm used was a stepwise Linear Discriminant Analysis, hitherto referred to as SWLDA. Despite its relative simplicity, this was determined by Krusienski *et al.* to be an effective classifier. Several other classifiers were examined in this session, but none significantly outperformed SWLDA, even in specific cases. Partial least squares discriminant analysis (PLSDA) was sometimes used as a verification of performance as it achieved roughly the same results as SWLDA.

There are two ways to turn the classification scores for each flash into an actual decision on which letter was spelled. The set of scores for each letter can be manipulated to create a decision, or the flashes for each letter can be averaged together and treated as a single observation. Merits and drawbacks exist for both points, but either way a layer of complexity must be added to the algorithm. To avoid this complexity, accuracy was usually measured in terms of the area under the Receiver Operating Characteristic curve of the per-flash classifier scores, rather than percent of letters guessed correctly.

Poor Scorers

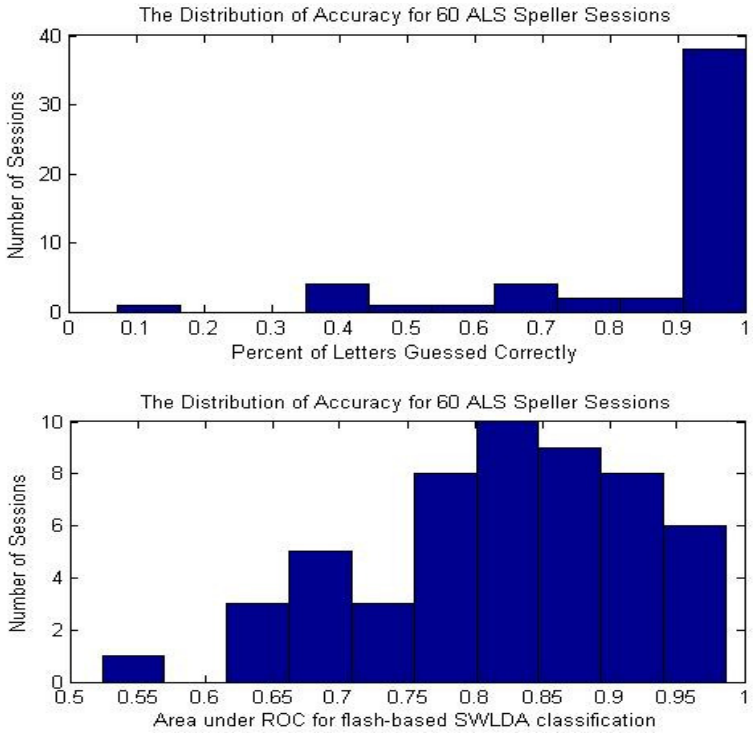


Figure 3: Histograms of the classifier performance for all test sessions.

Top- Percent of Letters Correct

Bottom- Area under ROC

The distribution of classifier performance for each of the sixty sessions is shown above. While the majority of sessions guess over 90% of letters correctly, fifteen out of the sixty sessions perform lower, with five failing to achieve 50% accuracy. These poor-scoring sessions seem to mostly score between .6 and .75 in terms of area-under-ROC.

Results

Noise Magnitude

The ratio of the target signal to the strength of noise is certainly a major factor in the accuracy of the classification – it is for this reason that many flashes across multiple EEG channels are necessary. It is clear from a visual examination of the signal that noise overpowers the P300.

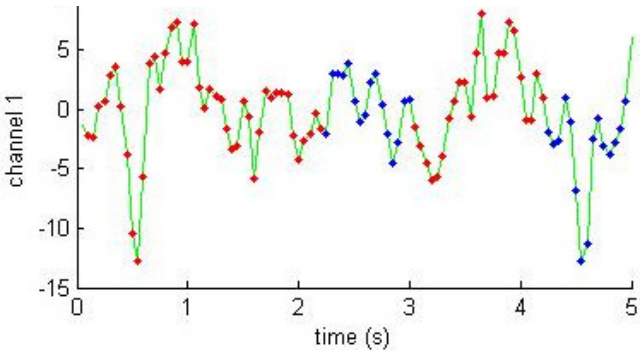


Figure 4: 5 seconds of data from an EEG channel, as an example. The 800ms following a target flash is colored blue.

As such, the signal-to-noise ratio of each session was plotted against accuracy to observe how closely the two were correlated. While there is a clear correlation, the relationship is not linear enough to suggest that this is the only determinant of accuracy.

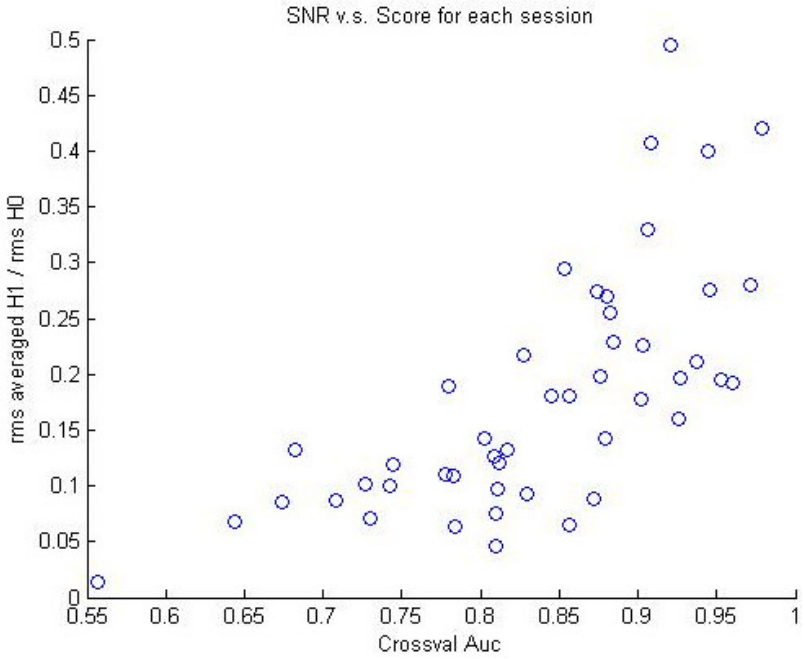


Figure 5: A scatter plot of each session's signal-to-noise ratio versus its area-under-ROC score.

It is important to note that this is simply an estimate of SNR: the true SNR cannot be calculated because the exact nature of the target signal (the P300 for each flash) is not known. Instead, a theoretical target signal was constructed by averaging the signal from every target flash, and the RMS magnitude of this average was compared to the RMS magnitude of all non-target signals. Other factors than the pure magnitude of noise could easily affect this measurement – for example, variance in the location or shape of each target signal would lower the value of the average.

Cross-Correlation Results

As discussed previously, a signal-to-noise ratio was difficult to quantify when the target signal cannot be measured without noise. As a way of extracting target information without averaging flashes and potentially destroying or altering, all target flashes for a session were cross-correlated with each other. The goal was to find a difference in the magnitude of cross-correlation peaks between sessions, but that information was meaningless as the overall amplitude of each session varies drastically. However, the location of these cross-correlation peaks was fairly well correlated with the session's accuracy.

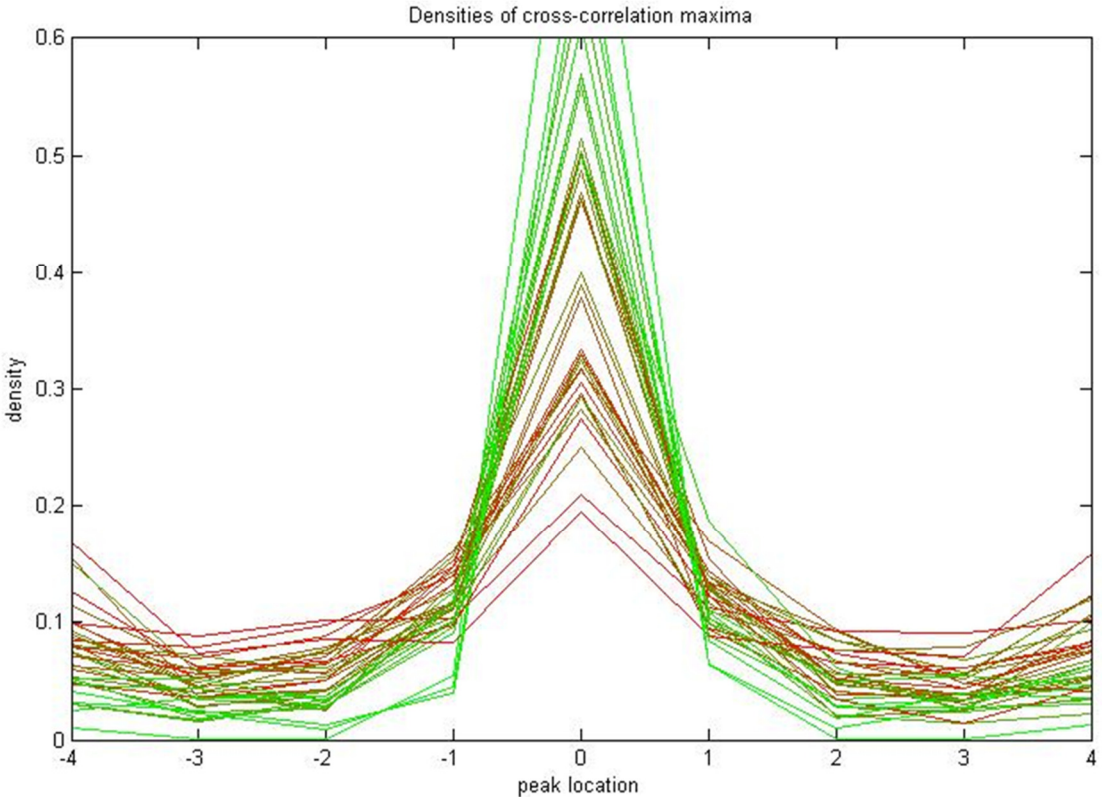


Figure 6: Histograms of the cross-correlation maxima between target signals for each session. The colors shift from green to red in descending order of session accuracy.

High-scoring sessions had most of their peaks in the dead center of the cross-correlation, as would be expected of similar or equal signals. Low-scoring sessions had more maxima located off center. This could simply be an effect of noise, but it could also indicate that some target flashes are shifted from other flashes. Comparing the close relationship of the cross-correlation peak latency and session accuracy to the less-strict relationship with RMS power, it seemed likely that variance in signal latency was causing problems.

The P300 response time is known to vary up to 200 milliseconds depending on the difficulty of the target/non-target distinction (Magliero *et al*., 1984 [7]). A shift of this magnitude would essentially force a classifier to look for a wider, more general, signal, thus exposing itself to noise. However, a measure of the variance in the target signal requires knowledge of the location of the target signal – a problem that would require classification of its own. Rather than attempt this initially, we resolved to test the current Speller classifier on synthetic data with known latency shifts, to see if latency shifts would have a significant effect on accuracy.

Shift Detection: Synthetic Data

The goal in constructing a testing environment for signal shift was to keep all other elements as close to the true test data as possible. The synthetic target signal was obtained by averaging all target flashes for each session. The targets could be shifted up to 100 milliseconds in either direction – more of a shift is unlikely given current knowledge, and would in fact be impossible to classify because of the 250 ms space between flashes. A shift of +150 ms in one flash would also be read as a shift of -100ms in the following flash.

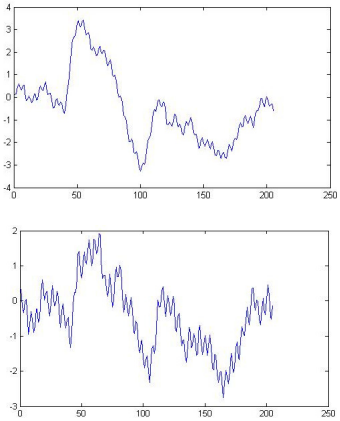
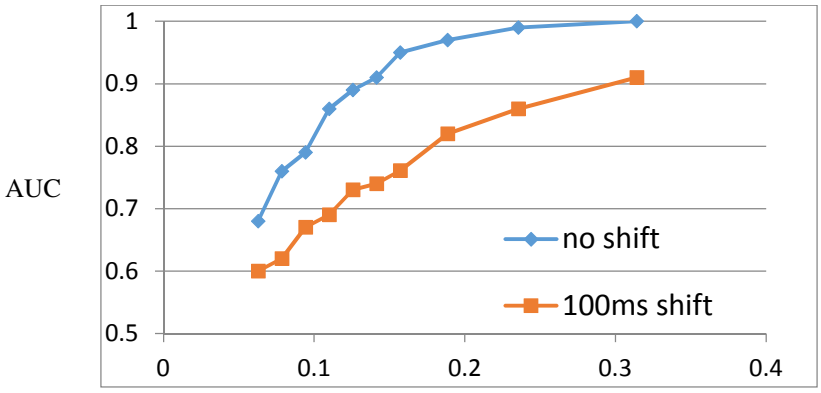


Figure 7: Average of all target flashes for a high-scoring (top) and low-scoring (bottom) session.

The synthetic non-target data was directly taken from real non-target session segments and added in a random order. In order to control the noise magnitude, the non-target data was multiplied by a gain so that its RMS power was at a given proportion to the target signal's power – effectively fixing the SNR as measured in the Noise Magnitude section to a certain value.

The performance of a PLSDA classifier at various SNR levels is shown below. Adding random shift to the target signal has a clear effect on the accuracy of the signal, though at sufficiently high or low levels of noise it has less of an effect. The area-under-ROC of the real test sessions fell almost entirely between 0.6 and 0.9, so the range where the shifts have a major effect is the range that corresponds to real data.

Figure 8: Classifier performance on synthetic data at varying SNR levels.



It is worth noting that the synthetic data assumes that the target signal has no mutations other than a potential shift in location, and that relying on an average-based SNR measurement may not be a perfect analysis of the system. As evidence of this, synthetic data was created for multiple sessions – some with high scores, and others with low – each fixed to the same SNR value. Were this value the only factor in accuracy these subjects would perform identically, so such a high variance shows that a significant factor or two is still missing from the false data.

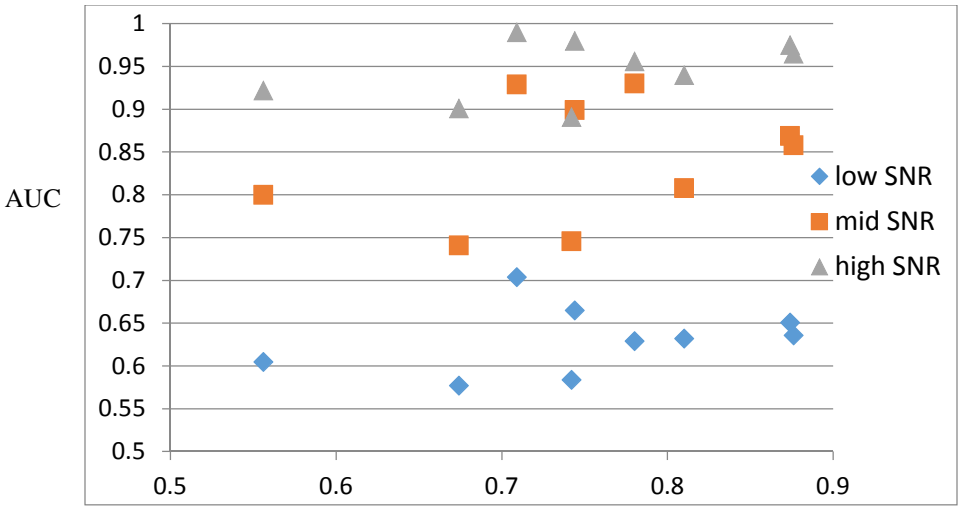


Figure 9: Comparison of classifier performance at various levels of SNR, with data extracted from different sessions.

Shift Detection: Modified Classifier

While it was proven through the synthetic data tests that target signal latency shifts could have an effect on data, there was still no way to quantify the amount of shift in any session. The approach left was to design a classifier that could detect and account for shifts in the target signal. If such a classifier could outperform standard classifiers, even only on a few sessions, this would be strong evidence that shifts were affecting accuracy in certain sessions.

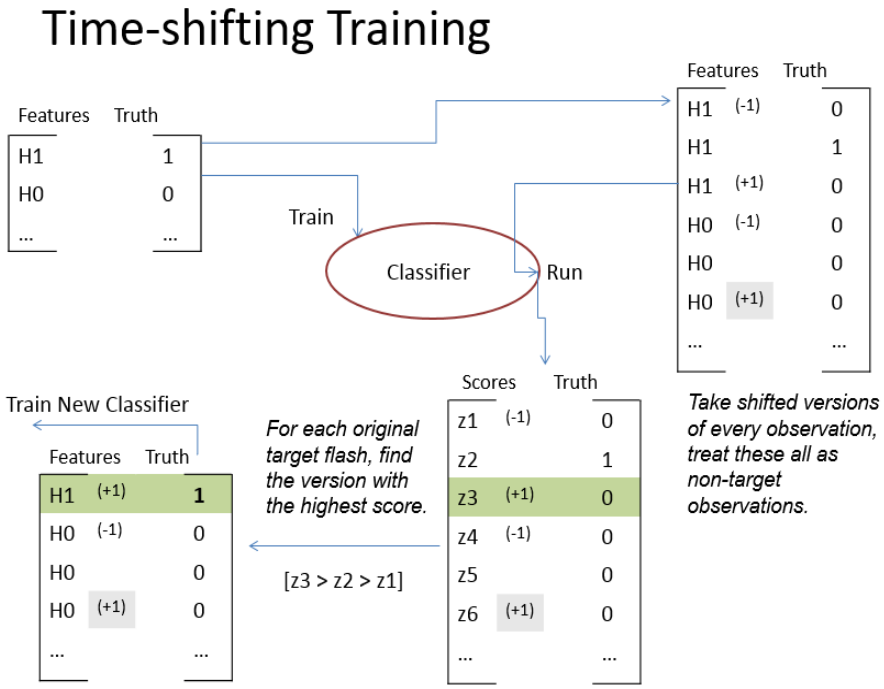


Figure 10: A diagram of the training method designed to detect shifts. The fundamental concept is train normally, then test the training data with this classifier, along with shifted versions of the target flashes. The shifted version of the target signal that scores the highest is considered the correct target signal, and a new set of training data is created to classify with.

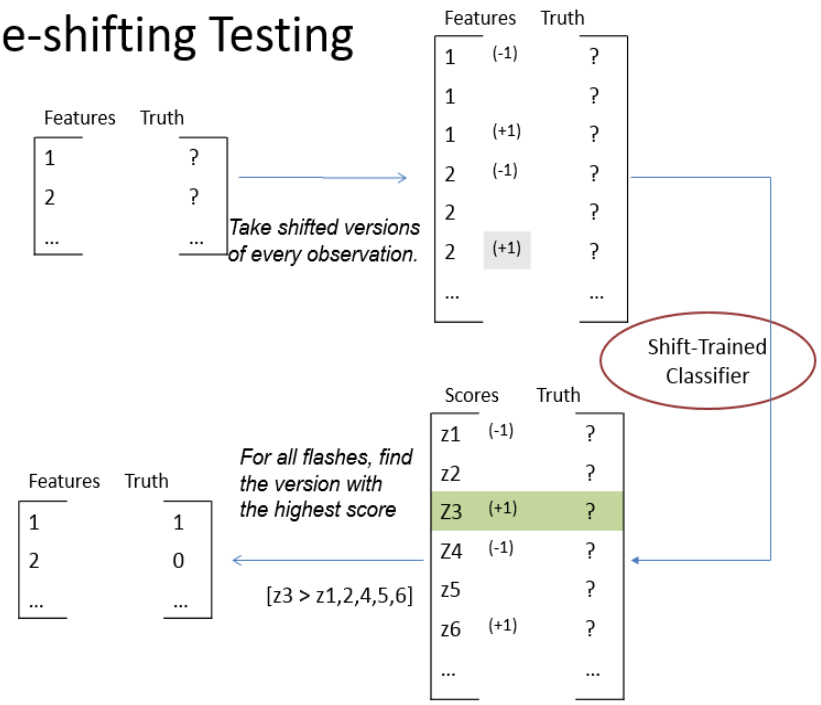


Figure 11: The shift-detecting algorithm for test data. A shifted version of each flash is run through the classifier, and any shift of any flash with the highest score is considered indicative of the correct flash.

A classifier with shift-detecting properties had been designed by Thompson *et al.*, 2013 [8]. This classifier trains normally, but for testing it takes shifted versions of every target flash and tests them as well, picking the single observation (across every shift of every flash) with the highest score. However, its performance was poor and its use, according to Thompson, was as an indicator of session accuracy. We adapted this model to also function on training data, on the grounds that shifts in the training data would cause as many problems as those in the test data.

The figure below compares the performance of this model on shifted synthetic data to the normal classifier on shifted and unshifted data. It consistently outperforms the normal classifier, though it is worth pointing out that this synthetic data has a shift range of +/- 100 milliseconds, as well as a uniformly random distribution of shifts. When the target signals are left unshifted this shift-detecting classifier performs worse than the normal one, probably because when multiple shifted versions of a noisy non-target signal are being examined, it is all the more likely that one of them will be mistaken for a target signal.

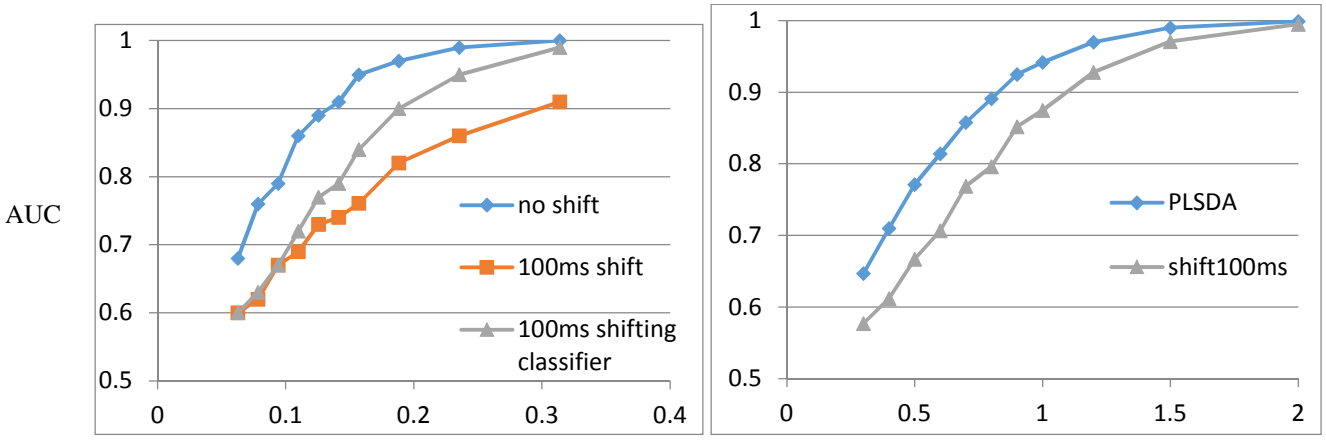


Figure 12: Classifier performance v.s. SNR with 100ms shift (left) and no shifts (right)

Shift Detection: Modified Classifier on Real Data

After fair success on synthetic data, this classifier was used to score the real test sessions. It underperformed consistently across accuracy ranges, despite detecting more shifts in the lower sessions. This still cannot conclusively prove that target signal shifts are not affecting the classifier accuracy, as this classifier performs best at high SNR and heavy shifting is present. However, are shifts present they will clearly require a more sophisticated method to detect.

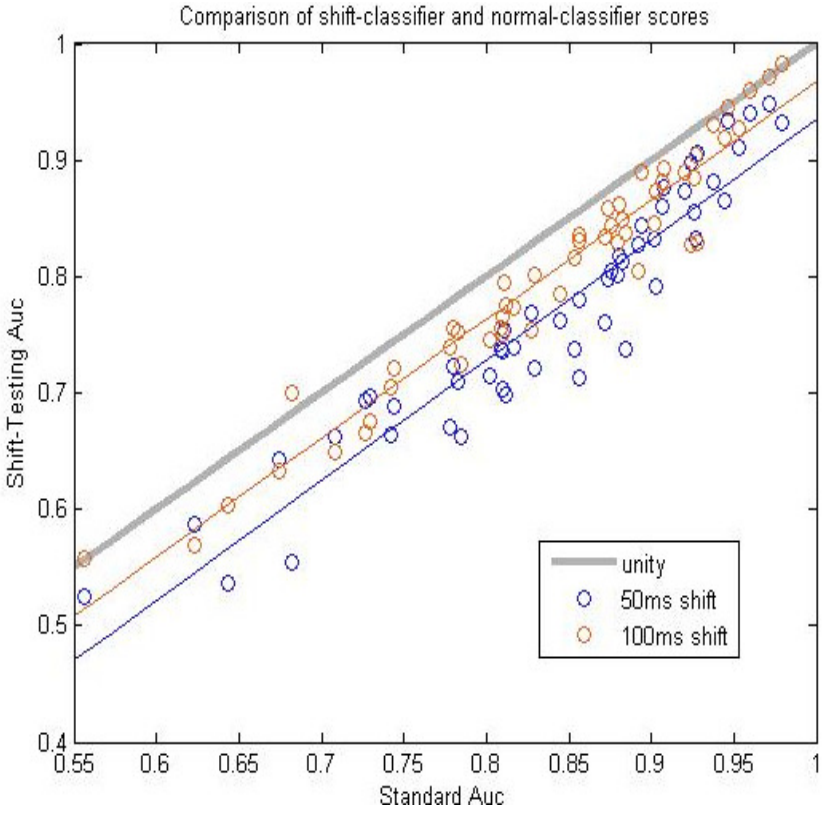


Figure 13: Classifier performance on real test sessions

Non-target Artifacts

Artifact removal is a common technique in EEG signal processing, as human brains are constantly functioning in more ways than we currently understand. However, there is little record of artifact removal techniques for the BCI Speller system. The heavy low-pass filtering of the BCI non-target data might remove some artifacts, but the possibility of other is undeniable.

As the EEG noise is not visually comprehensible in the time domain, frequency domain plots were made to search for particularly strong frequency bands or patterns. While no particularly frequencies stood out as strong, there were occasional spikes of amplitude across all frequencies, lasting for several seconds. To see these on a larger scale, the test sessions were split into four second blocks and the root-mean-squared power of each block was measured.

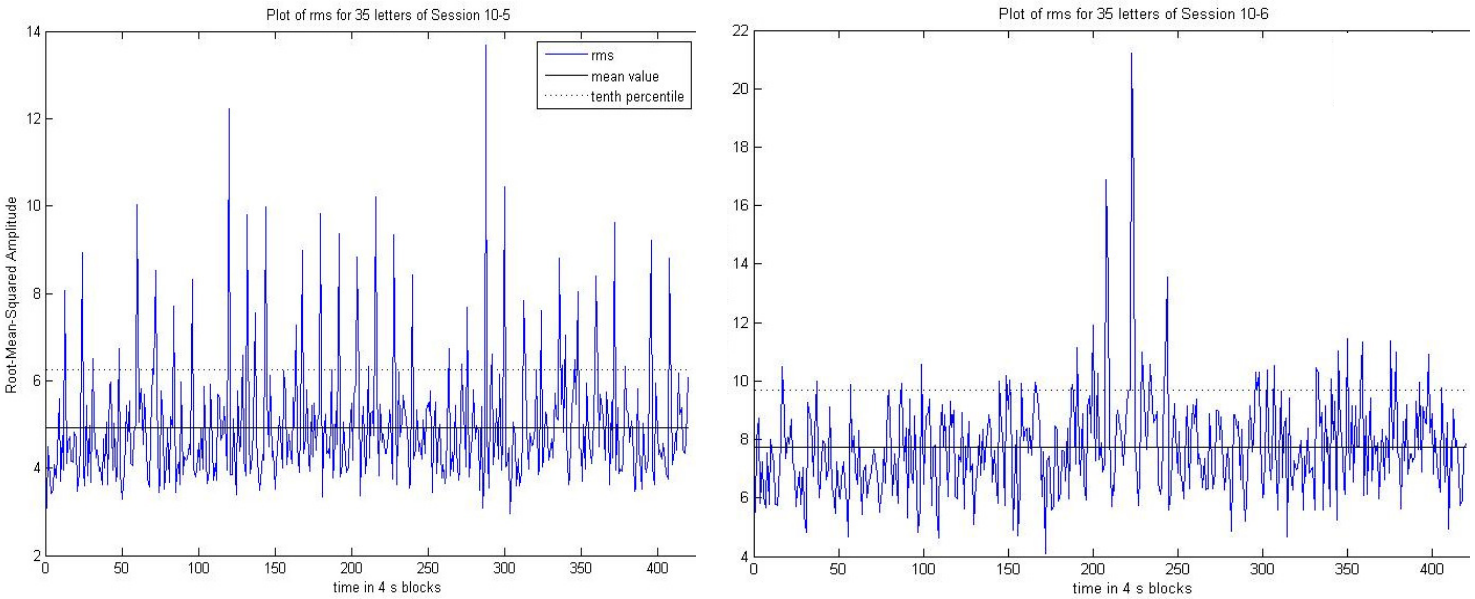


Figure 14: RMS power calculated in 4 second blocks, for a high scoring (left) and a poor scoring (right) session

For all sessions examined, there were significant amplitude spikes in these several-second periods. However, the rms power of the spikes in the poor scoring sessions were generally higher than those in high scorers, at least relative to the mean power of each session. Currently no numeric quantification of the extremity of the high-amplitude sections has been decided, so a correlation with classifier performance is not yet possible. However, as a test of potential relation to accuracy, a classifier was run on sessions with all signals past the tenth percentile of magnitude (past the dotted line on the plot) cut from the observations.

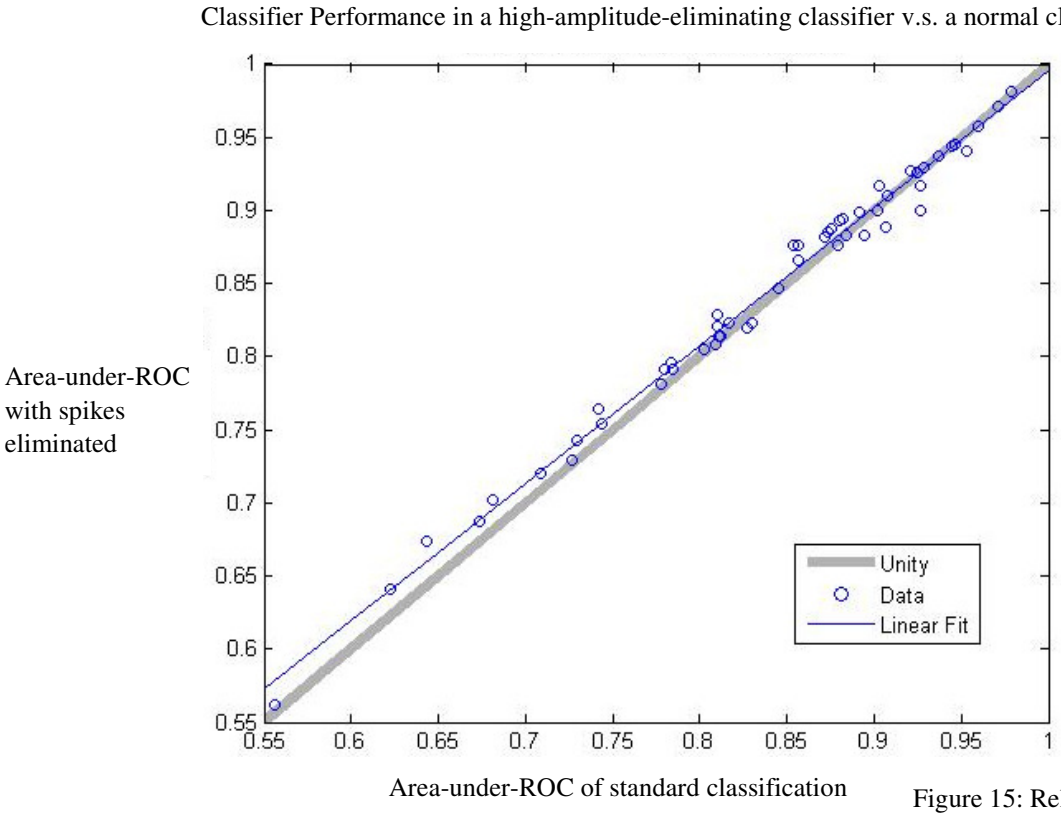


Figure 15: Relative performance of a prototype artifact-removing classifier.

The improvement in scores is minor but fairly consistent. Moreover, the sessions that were improved the most were the poor-scoring ones, as would be expected if these high-amplitude sections were a significant source of inaccuracy. This is not yet a viable classification method: the high-amplitude time segments were removed before classification so in a way this method selectively picked data to test. However, these initial results definitely encourage a deeper examination of high-amplitude artifacts in the EEG data.

Future Directions

There are several other ways one could examine shifts or transformations in the target signal. For one, higher-quality synthetic data could be constructed so that target shape transformations and other signal effects could be accounted for. Also, a classifier-based algorithm would not be the only way to detect shifts. Unsupervised methods like clustering or factor analysis would probably prove more powerful in this situation, since if classification were reliable these experiments would not be necessary in the first place.

Of the potential accuracy-affecting features not included in this report, the variance in signal across different EEG channels stands out as a well-discussed factor. There is a fair bit of literature on reliable channel sets as well as algorithms that choose channels to classify on for a particular session. This field was not covered because it was deemed more worthwhile to examine less-researched signal properties first, and to either prove or disprove their effectiveness. However, it is not entirely impossible that signal shift and the channels chosen are linked. The P300 reaches different parts of the brain at different times (Halder et al., 2013 [9]), so it is possible that a shift in the target signal would have a different effect on different channels.

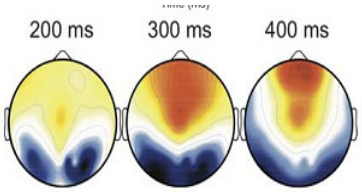


Figure 16: P300 voltage across the brain at various times [9]

It may be possible to quantify the amount of high-amplitude artifacts by calculating the variance of the amplitude measurements, or a similar value. However, this would not necessarily account for single sections of unusual activity, such as the short time span with a great deal of activity in figure 14. Instead of simply quantifying or removing these amplitude spikes, it may be worthwhile to examine them in greater detail and search for patterns or causes. Either way this is a relatively simple feature to measure and manipulate, and it has already demonstrated at least some relation to classifier accuracy. I would advise that research towards enhanced Speller accuracy be focused on these artifacts.

Acknowledgements

Thanks to Leslie Collins for introducing me to this project, and to the field of machine learning in general.

Also thanks to the other members of SSPACISS involved with the BCI Speller: Sandy Throckmorton, Kenneth Morton, Ken Colwell, and Boyla Mainsah.

Thanks to Martha Absher and Ellen Currin for making student research incredibly accessible and smooth-running.

And thanks to all the volunteers who tested out the Speller. Especially the ones who didn't do too well.

References

1. Sutton S., Braren M., Zubin J., John E.R., 1965. Evoked-potential correlates of stimulus uncertainty. *Science*, 150, 1187-1188.

2. Farwell, L. A. and Donchin, E. 1988. Talking off the top of your head: toward a mental prosthesis utilizing event-related brain potentials. *Electroenceph. and Cl. Neurophys.*, 70, 510-523.

3. Wolpaw J.R., Birbaumer N., McFarland D.J., Pfurtscheller G., Vaughan T.M., 2002. Brain-computer interfaces for communication and control. *Clinical Neurophysiology*, 113, 767-791.

4. Schalk Lab. User Tutorial: Introduction to the P300 Response. bci2000.org, 2013.

5. Jennifer Sisto, 2009. An introduction to multimodal imaging. Electrical and Systems Engineering, Washington University.

6. Kruienski D.J., Sellers E.W., McFarland D.J., Vaughan T.M., Wolpaw J.R., 2007. Toward enhanced P300 speller performance. *Journal of Neuroscience Methods*, 167, 15-21.

7. Magliero A., Bashore T.R., Coles M.G.H., Donchin E., 1984. On the Dependence of P300 Latency on Stimulus Evaluation Processes. *Psychophysiology*, 21-2, 171-186.

8. Thompson D.E., Warschausky S., Huggins J.E., 2013. Classifier-based latency estimation: a novel way to estimate and predit BCI accuracy. *J. Neural Eng*, 10, 016006.

9. Halder S., Hammer E.M., Kleih S.C., Bogdan M., Rosentiel W., Birbaumer N., Kubler A., 2013. Prediction of Auditory and Visual P300 Brain-Computer Interface Aptitude. PLos ONE 8(2): e53513.