Webappendix for Reanalysis of "Bedside detection of awareness in the vegetative state: a cohort study."

Introduction

Below we summarize the steps of Cruse et al.'s[1] experimental paradigm and analysis methodology, with specific attention devoted to the assumptions in their statistical model. We then provide the details of our tests of these assumptions that were described in the main text. Finally we describe analyses of the Cruse et al. dataset that don't require these assumptions, and comment on how these findings guide use of machine-learning approaches towards EEG analysis in brain-injured subjects.

Details of the experimental paradigm of Cruse et al.

Trials were grouped into "hand" and "toe" blocks, with no more than two blocks of each kind presented consecutively. 4 to 8 block-pairs were obtained for each patient, and 6 block-pairs for all normals. At the beginning of each block, subjects were instructed that when they heard a tone, they should imagine moving their right hand or their toes. Within each hand or toe block, there were 15 tones (each defined a trial) separated by 3 to 6·5 sec, on a random schedule, but the instructions were not repeated. Cruse et al. states that the interval was 4·5 to 9 seconds but this was a typographical error (personal communication, D. Cruse). This spacing may affect the interpretation of the results, because in some fraction of the trials the pre-tone "rest" data overlapped with the post-tone "task" data of other trials.

Data provided to us

We were graciously supplied the EEG data, metadata (block and task labels), and relevant analysis software by Cruse and colleagues. Subject data included all patients (numbered P1-P16 as in Cruse et al.) and a subset of the normals: five normals that covered the full range of performances (numbered N1-N5 in this manuscript). The SVM analysis of Cruse et al. found subjects P1, P12, P13, N1, N2 and N3 to be positive (Webappendix Table 1).

All EEG data provided had already been preprocessed including: segmentation of the continuous data into trial epochs spanning 1·5 seconds before the command tone to 4·0 seconds after the tone, filtering (1 to 40 Hz), application of a Laplacian montage, and manual removal of trials with significant artifact. We used the same 25 channels as in Cruse et al., and did not change the pre-processing.

Data visualization (Figure 1)

Figure 1A displays data in the time domain as provided to us, using the EEGLAB code 'eegplot'.[2] Power spectra of Figure 1B are calculated using the Chronux toolbox[3,4] code 'mtspectrumc'. As is standard for that analysis, three Slepian tapers were used on the 1·5 second data swatches, with a frequency resolution of 1·5 Hz. Each power spectral estimate (each curve in the figure) represents the average frequency content of the EEG from all the trials of one block of a subject. Note that the patient's power spectra flatten or have an upward slope above 10 Hz. This behavior is known to be associated with a preponderance of muscle artifact in the data[5].

Jackknife error bars (the I shape in lower left of figures) were calculated as the average of all jackknife error bars from each spectrum, as provided by the Chronux toolbox. They reveal that the spectra from the normals overlap (except for differences due to task), while the spectra of the patients do not. Moreover, although there is some systematic variation due to task in the data from the normal subject (i.e., the decrease in power in left and central channels in the hand task, chiefly in the 15-25 Hz range), no such behavior is apparent in the patient data.

The Cruse et al. SVM approach

All analyses ran in the Matlab (The Mathworks, Natick, MA) environment using built-in, Bioinformatics, and Signal Processing Toolbox functions. Cruse and colleagues also used g.BSAnalyze (g.Tec, Austria) code 'gBSdownupsampling' to downsample the EEG to 100 Hz; we instead used Matlab's 'resample'. With these codes we were able to replicate, precisely, their results on all subjects, other than as described below under "Cruse et al.'s analysis of data prior to tone"; we verified with the authors of Cruse et al. that this discrepancy was due to a typographical error in their report.

After downsampling, the EEG data are converted to the frequency domain by calculating power spectra in a one-second moving window, moving through the data in steps of 0·01 second (Matlab's 'spectrogram'). Cruse et al.'s primary SVM analysis used the spectra from windows with midpoints spanning 0·5 to 3·5 seconds post-tone. In each window the logarithm of the EEG power spectrum is determined, and averaged within four frequency ranges (7–13, 13–19, 19–25, and 25–30 Hz). The net result is that the SVM analysis used 30,000 features per trial: 300 (time points) x 4 (frequency ranges) x 25 (channels).

An SVM classifier is then determined from a "training" component of the dataset and its accuracy is determined by a "test" component. As in Cruse et al., the "training" component consisted of all of the data with one block of each type omitted, and the "test" component consisted of the two omitted blocks. To create the SVM classifier, training data are first normalized by subtracting off the mean and dividing by the standard deviation of the training features. A linear-kernel classifier is then created with Matlab's 'svmtrain' with all default settings, except that 'autoscale' was disabled as data had already been normalized. (For further details on the default settings, see http://www.mathworks.com/help/toolbox/bioinfo/ref/svmtrain.html).

The classifier is then tested with Matlab's 'svmclassify'. In Cruse et al., test components consisted of the trials of a single hand block and its paired toe block (e.g., hand block 1 and toe block 1), while the remainder of the trials from the other blocks formed the corresponding training component. We next describe this process in more detail.

Concretely, the classifier is a set of "voting weights" for the 30,000 features, determined from the differences between the values of these features in the hand and toe trials within the training set. The SVM then applies the classifier to the test component (which is ignored when the classifier is constructed), and determines the number of trials correctly classified. This process – cross-validation against out-of-sample data – is designed to protect against overfitting. The cross-validation is repeated with multiple partitions of the dataset into training and test components. Once the dataset has been repeatedly partitioned so that all trials have played the role of the test component, the overall performance of the SVM can be measured by the total fraction correct in out-of-sample testing.

The way in which Cruse et al. implemented cross-validation made an assumption concerning possible dependencies between blocks. Specifically, in Cruse et al., the subdivision into "training" and "test" always consisted of removing a single pair of adjacent hand and toe blocks. This cross-validation strategy assumes that there are no idiosyncratic statistical relationships between adjacent blocks (i.e., relationships between adjacent blocks that are not shared by more distant blocks); as mentioned in[6], the approach of using adjacent blocks can lead to erroneous conclusions when such relationships are present. We determine if this assumption is valid by extending the cross-validation procedure to include non-adjacent blocks as test sets (see below, 'Testing the relationship between blocks').

Once the average accuracy across all test blocks is determined, the Cruse et al. method determines statistical significance (i.e., p-value) of this accuracy, based on the total number of correctly classified trials out of the total number of trials in all blocks. The method assigns a p-value to the accuracy by assuming a binomial distribution for the count of correctly-classified test trials. In binomial statistics, each classification (of each test trial) is assumed to be an independent measure of the classifier's validity. We test this assumption two ways: by looking at the variability of the power spectral estimates within and across blocks; and by looking at the distribution of significance levels obtained from binomial statistics (see below, 'Testing independence of trials within blocks'). We also use a different approach to evaluate the significance level for each subject (a permutation test) that does not rely on this independence assumption; this yields very different results (see below, 'An SVM method that does not depend on assumptions of trial and block order independence').

Testing the relationship between blocks (Webappendix Figure 1)

We followed the recommendation of Lemm and colleagues[6] to calculate SVM accuracy using a cross-validation scheme in which all possible pairings of task blocks serve as test datasets. If there are no special relationships between blocks (these relationships could arise if the data are nonstationary, or as the result of dynamics with long time constants, such as reversible state changes), it should not matter how the pairs of blocks are chosen for validation, since the only differences between them should be from task performance

and random noise. Results are in Webappendix Figure 1. For two of the positive subjects (N1 and N2), no such dependence was found (i.e., accuracy did not change with block-pair spacing). But for N3, P1 and P12, classification accuracy decreased as the test-block-pairs were further apart, dropping into the range of chance performance (N3 and P1) or worse-than-chance performance (P12). This drop in accuracy implies that the idiosyncratic relationships between adjacent blocks contributed substantially to SVM performance in these subjects. In P13, the data show no evidence of this dependence on block-pair spacing, but the analysis is inconclusive since this dataset was limited to 4 block pairs.

Testing independence of trials within blocks (Figure 1B and Webappendix Figure 2)

Our first approached involved visualizing the data in the frequency domain. For the normals (typical example in Figure 1B, left), spectral estimates from different blocks of the same type (hand or toe) are consistent with each other. For the patients (typical example in Figure 1B, right), however, estimates from individual blocks are separated by more than their confidence limits would allow. This means for patients that individual trials are much more nearly matched within a block than across blocks (i.e., that the trials are not independent).

As a second approach to test the assumption of trial independence, we took an omnibus approach across all subjects. Here we used the same SVM method as Cruse et al., but applied to single placements of the spectral windows (1 sec. long) across the entire trial epoch (starting 1·5 sec pre-tone, spaced by 0·1 sec, ending at 4·0 sec post-tone). Thus, each analysis used only 100 features (25 channels x 4 frequency bands). We then calculated the significance of the classification at each time point using a one-sided binomial test, i.e., determining the probability that a chance classifier would yield as large (or larger) a fraction of correct classifications. If the classification was at random, this should produce a distribution of p-values evenly spread from 0 to 1. That is, a p-value of 0·05 or less should occur 5% of the time; a p-value of 0·3 or less should occur 30% of the time, etc. Note that the prediction of a flat distribution of p-values follows from very general principles: if the data are random, then the likelihood that a p-value will be less than some fraction f is, by definition, f. A classifier that was more often correct than expected from chance would show this behavior by having an excess of low p-values.

The results from this analysis are combined for all subjects of each group, rather than looking at each subject individually. This allows us to test the validity of the statistical model on the dataset as a whole, and improve its statistical power. To understand why it is valid to combine results from multiple subjects to test the model, consider the following analogy. Imagine that one is charged with doing quality control for a manufacturer of dies (6-sided playing pieces). The goal is to determine if the dies are fair, but the manufacturer only allows the tester to roll each die once (similar to present situation, in which there isn't enough data from any single patient to do a test of sufficient power). One can still assemble the probabilities of each of these roll-each-die-once trials, with the anticipation that if all the dies are fair, the distribution will be flat (i.e., the probability of each outcome will be the same). If the distribution is not flat, then the null hypothesis of fair dies is false. Even though one doesn't know which dies contributed to the failure of the null hypothesis, it can be rigorously ruled out for the population as a whole. (Note that pooling across dies results in a weak (conservative) test for fair dies, as the manufacturer could arrange to have each die unfair in a different way so that the ensemble distribution would be flat, though in fact the model of fair dies is still false.) But pooling will not lead to false-positives: a consistently non-flat distribution is inconsistent with fair dies, even if each is only rolled once.

For the normals (Webappendix Figure 2, left), we found the results expected from independent trials: the distribution of p-values was flat other than excess near 0, indicating that SVM classification was more accurate than chance, especially for analyses carried out post-tone. Some better-than-chance classification also occurred pre-tone, suggesting that subjects were still performing the task from the previous trial (also see Webappendix Figure 4A and 4C).

For the patients, the distribution of p-values (Webappendix Figure 2, right) is U-shaped: there is an excess of p-values not only near 0 but also near 1. P-values near 1 represent instances of apparently worse-than-chance classification (i.e., a trial from one task classified as belonging to the other), a phenomenon that is unlikely to be caused by a real difference between the two kinds of trials, or by spillover from the previous trial. The excess of this kind of outlier implies a violation of the key statistical assumption – independence -- that underlies the binomial test.

To see how this U-shaped function can arise from non-independence, consider a simple scenario in which a coin is flipped repeatedly, and the number of heads is reported. If a fair penny is flipped a hundred times, it would be surprising to get fewer than 40% heads or more than 60%. Moreover, the binomial distribution based on 100 flips would accurately predict the probability of unusual outcomes (such as 70% heads). But if instead a 10 cent piece were flipped 10 times – with each coin flip falsely interpreted as 10 trials of flipping a penny - then outcomes with fewer than 40% heads or more than 60% heads would be reasonably common, and the outlier probabilities would be substantially underestimated by an "independent" model based on 100 pennies. That is, if dependencies between trials are ignored, datasets with large deviations from the average behavior will occur more often than expected. This finding on its own does not imply that all patients have correlation of trials within blocks, but only that there must be correlations in *some* subjects. It therefore shows that the overall statistical model is inappropriately used on this dataset, and a model that does not take into account this assumption of trial-independence should be used instead.

As is the case for relationships between blocks, a lack of trial-independence can arise from nonstationarities in the data, or from dynamics with long time constants.

Cruse et al.'s analysis of data prior to tone

We note that Cruse and colleagues also analyzed the performance of the SVM prior to the tone, but used only the spectral windows within the half-second prior to the tone. Their goal was not to test the statistical model (as in Webappendix Fig. 2), but to test whether there was evidence for task performance at all time periods in the patients. They reported that all positive patients had a $p>0{\cdot}05$, and took this as a control for the validity of their approach. We performed the identical analysis and found that both P13 and N1 had positive classifications in this time period ($p=0{\cdot}0103$ and 10^{-6} respectively; Webappendix Table 2), which are inconsistent with the claimed negative control. In a subsequent personal communication, D. Cruse confirmed that patient P13 had $p=0.0103$, and indicated that they had intended to state $p>0.01$; their stated $p>0.05$ was the result of a typographical error. Their manuscript did not comment on the p-values for this control in normal subjects, such as N1.

An SVM method that does not depend on assumptions of trial and block-order independence

We further re-analyzed the Cruse et al. datasets with a statistical approach using SVM that does not rely on the assumptions of trial and block independence (Webappendix Table 1). To take into account correlations between blocks, accuracy was defined as the average of the accuracies with all possible block-pairs used as the test datasets (i.e., fully cross-validated). This definition corresponds to averaging the multiple measures of accuracy for each subject shown in Webappendix Figure 1.

To take into account the lack of independence of trials within a block, we used a permutation test instead of the binomial test used in Cruse et al. Since the permutation test is based on relabeling the trials on a block-by-block basis, it avoids the need to assume statistical independence of individual trials. In more detail, our procedure for the permutation test was as follows. First, we calculated the fully cross-validated accuracy of the SVM classifier from each dataset (as described above, using all block pairs as test components). Then, we calculated the fully cross-validated accuracy of the SVM classifiers obtained from surrogate datasets. Surrogate datasets were created by randomly relabeling the blocks (i.e., hand blocks changed to toe blocks and vice versa), keeping half of the blocks as hand blocks and half as toe blocks. We then determined the significance of the SVM accuracy by comparing the accuracy in the actual dataset with the accuracies determined in the surrogates: if the SVM performance was better than chance, accuracy should be higher in the original dataset than in nearly all of the surrogates. Due to computational demands, for subjects whose datasets had 7 or more blocks, we considered a random set of 1000 random relabeling permutations; for subjects with 6 or fewer blocks, we considered all random relabelings. The permutation test p-value was then determined as the fraction of surrogate datasets that had higher classification accuracy than the original dataset. Note that the smallest possible p-value for each subject is limited by the number of ways to relabel the blocks. This limitation is non-negligible when there are only 4 block pairs for analysis (as in patient P13), as 4 block pairs only allow for 35 permutations, and a minimum p-value of $1/35=0{\cdot}0286$. (With 5 block pairs, the minimum p-value is $1/126=0{\cdot}0079$, and with 6 block pairs, as in subjects N1, N2, N3, and P12, it is $1/462=0{\cdot}0022$).

Results from this permutation test are in Webappendix Table 1 under the column "p-value with all block-pair by permutation test", and are discussed in the main manuscript.

Correction for multiple comparisons

As we mention in the main manuscript, correction for multiple comparisons is essential with a methodology with no gold standard for task performance. For the normals, it is reasonable to assume that all could do the task, so each p-value can be considered separately. But for the 16 patients, there was no reason to believe, *a priori*, that any could carry out the task. Thus, we needed to determine the likelihood of encountering the observed p-values, under the null hypothesis that the classifier was performing randomly in all 16 datasets. Put another way, without a correction for multiple comparisons, a classifier is expected to yield "positive" results in a fraction of patients just by chance, and we wanted to determine whether this phenomenon was a plausible interpretation of our findings. To determine the likelihood of encountering the observed p-values from the permutation approach, we chose the False-Discovery Rate (FDR) approach[7] as the primary measure, as it uses the overall distribution of p-values for determination of a multiple-comparisons-corrected significance level. Results are in Webappendix Table 1 ('**' next to subjects whose results remain significant). For the first patient (P13) to become significant, we would need to set an FDR-corrected p-value of 0·35. Note that results from the normal subjects N1 and N2 remain significant after FDR-correction with a threshold of 0·05. In Webappendix Table 1, normals and patients were analyzed as separate groups, but we also found that the results from FDR correction were the same when the groups were combined into a single calculation: results from N1 and N2 remained significant, and patient results did not become significant.

Another standard approach to multiple-comparisons correction is the Bonferroni method; this is less conservative (i.e., more stringent) than the FDR approach. Here, the raw p-value required to achieve a desired significance level is set by dividing this desired significance level (typically 0·05) by the number of independent samples (i.e., subjects). No patients were significant after Bonferroni correction, since this would require a raw p-value of 0·05/16=0·003. By the Bonferroni method, we would need to set an overall significance level of 0·45 for the first patient to become significant, since the lowest raw p-value for patient subjects is 0·0286=0·45/16. Normal subjects N1 and N2 remained significant when analyzed as a separate subgroup or together with the patients.

As a final test of significance in the patient dataset as a whole, we asked whether the entire distribution of p-values across the 16 patients showed any deviation from chance. To do this, we used the Kolmogorov-Smirnov test (Matlab's 'kstest'), to check consistency with a flat distribution. The null hypothesis of a flat distribution could not be excluded (p=0·23).

Details of the univariate approach (Webappendix Figures 3 and 4)

In Webappendix Figures 3 and 4, we compare EEG spectra before and after the tone in a frequency-by-frequency, channel-by-channel fashion. We use the term "univariate" to denote this approach, and to contrast it with the "multivariate" SVM strategy of looking for patterns distributed across multiple frequencies, channels and time points. Specifically, we calculated EEG spectra from task (0·5 to 2 seconds post-tone) and rest (1·5 to 0 seconds pre-tone) periods separately for both right-hand and toe blocks. We then determined whether the spectra, across the 25 channels and frequencies (7-30 Hz) used in Cruse et al., were significantly different between the task and rest periods on average across all blocks.

Significance was determined by a z-statistic[8] at each frequency on each channel, for a total of 850 tests. We declared a subject positive on a task if any frequency / channel tests remained positive after an FDR-corrected threshold of 0·05. This is very similar to a previous approach used to detect evidence of motor imagery performance in patients with disorders of consciousness[9]. Note, though, that this analysis did not seek to determine whether the changes in the hand blocks differed from those in the toe blocks, as the method was not intended for a block-design setting.

For all 16 patients, we found no overall significant differences between rest and post-tone, (Webappendix Figure 3 – right). In contrast, in all 5 normals, we found that one or both tasks showed significant differences across a broad range of frequencies, with dependence on channels typical of motor imagery tasks (decreased power especially over left motor areas in the hand task and midline areas in the toes task; Webappendix Figure 3 – left)[10,11].

The time course of the task-versus-rest EEG changes in normals (Webappendix Figure 4A-C, left) showed the expected[10,11] task-related suppression of EEG power, with an orderly scalp topography. Normals also show evidence for task performance at the end of the trial, which explain the findings of Webappendix

Figure 2 left, discussed above. The patients (Webappendix Figure 4A-C, right) did not show changes with a clear spatiotemporal organization.

Two further points regarding the univariate analyses deserve mention. First, in normals, there was a general concordance between the univariate and SVM approaches, as subjects who had positive classification by SVM (N1, N2 and N3) had many more significantly different frequencies by univariate testing than those who did not (N4 and N5). This supports the findings above that the normals met the assumptions of the SVM model of Cruse et al. Second, the univariate approach found evidence for task performance in N4 and N5, where the SVM approach did not, suggesting that in this situation it is a more sensitive test. This does not imply that multivariate (e.g., SVM) approaches are intrinsically inferior to univariate approaches, as the multivariate approach may be more sensitive if a weak signal is spread across many EEG channels, frequency ranges, or time points. But because multivariate approaches carry a high risk of overfitting (i.e., finding chance associations), they must be carefully vetted by an appropriate statistical model. Multivariate approaches also require larger amounts of data, especially when the data are high dimensional. This problem is exacerbated when nonparametric statistics (e.g., permutation test) must be used to assess the SVM classifier, as is the case here, because of the correlation structure of the data.

Potential for variation in task performance

One potential concern for both the univariate and SVM approaches is that they use data from all blocks, assuming that the task was performed in the same way each time. If the changes in the EEG were not the same each time, then both techniques would have difficulty detecting them. This is not directly relevant to this manuscript since our goal was to test the Cruse et al. approach, which is run on multiple blocks at once. Nevertheless, to test this possibility, we ran the univariate analyses on P13 (highest-performing patient subject) separately for each block and still saw no evidence for task performance. When we ran the analyses for N2 (normal with similar classification rate to P13), we see evidence for task performance on each block, though with slightly different patterns. This is akin to what we found in our previous work[9] where evidence for task performance on individual blocks had slightly different patterns, but with sufficient commonality so that there was a stronger signal when all blocks were combined.

Interpretation of findings from P13

Finally, we wish to stress that our tests of the validity of the statistical model used by Cruse and colleagues do not determine with perfect confidence whether or not an individual patient demonstrated evidence of task performance. This is relevant for P13 where our permutation approach found that the SVM was able to classify better on the real data than on all possible permutations, but that the p-value was limited by the few number of blocks (only 4). The limited number of blocks is a crucial limitation in settings where there is a non-stationary signal due to fluctuating muscle artifact (Figure 1), as it may happen to vary with task by chance. Furthermore, our univariate findings from P13 (Webappendix Figure 3 and 4) indicate that this "positive finding" does not reflect the typical EEG sensori-motor rhythm modulation seen in normal subjects producing motor imagery, as suggested by Cruse et al.'s Figure 2.

REFERENCES (for Webappendix Text and Webappendix Figures):

1 Cruse D, Chennu S, Chatelle C, *et al.* Bedside detection of awareness in the vegetative state: a cohort study. *Lancet* 2011; **378**: 2088–94.

2 Delorme A, Makeig S. EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *J Neurosci Methods* 2004; **134**: 9–21.

3 Chronux.org: Home. http://chronux.org/ (accessed 5 Oct2010).

4 Mitra P, Bokil H. Observed Brain Dynamics, 1st ed. New York, NY, Oxford University Press, USA, 2007.

5 Whitham EM, Pope KJ, Fitzgibbon SP, *et al.* Scalp electrical recording during paralysis: Quantitative evidence that EEG frequencies above 20 Hz are contaminated by EMG. *Clin Neurophysiol* 2007; **118**: 1877–88.

6 Lemm S, Blankertz B, Dickhaus T, Müller K-R. Introduction to machine learning for brain imaging. *Neuroimage* 2011; **56**: 387–99.

7 Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J R Statist Soc B* 1995; **57**: 289–300.

8 Bokil H, Purpura K, Schoffelen J-M, Thomson D, Mitra P. Comparing spectra and coherences for groups of unequal size. *J Neurosci Methods* 2007; **159**: 337–45.

9 Goldfine AM, Victor JD, Conte MM, Bardin JC, Schiff ND. Determination of awareness in patients with severe brain injury using EEG power spectral analysis. *Clin Neurophysiol* 2011; **122**: 2157–68.

10 Pfurtscheller G, Lopes da Silva FH. Event-related EEG/MEG synchronization and desynchronization: basic principles. *Clin Neurophysiol* 1999; **110**: 1842–57.

11 Bai O, Lin P, Vorbach S, Li J, Furlani S, Hallett M. Exploration of computational methods for classification of movement intention during human voluntary movement from single trial EEG. *Clinical Neurophysiology* 2007; **118**: 2637–55.

Subject	# of blocks	# of trials used (after artifact rejection)	Accuracy based on matched block pairs*	Accuracy based on all block pairs	p-value based on matched block pairs, binomial test *	p-value based on all block pairs, permutation test
N1	6	166	0·9096	0·8534	$9x10^{-30}$	0·0022**
N2	6	164	0·7561	0·7459	$3x10^{-11}$	0·0022**
N3	6	170	0·6529	0·6157	$8x10^{-5}$	0·0498
N4	6	176	0·4716	0·4223	0·4976	0·5303
N5	6	178	0·4438	0·3895	0·1542	0·6126
P1	8	202	0·6139	0·5644	0·0015	0·0930
P2	4	114	0·6053	0·5373	0·0308	0·1429
P3	8	160	0·4750	0·4047	0·5801	0·6420
P4	5	69	0·4348	0·3623	0·3356	0·5238
P5	4	102	0·5196	0·5319	0·7666	0·2000
P6	9	132	0·5379	0·5076	0·4335	0·2460
P7	5	76	0·5658	0·6053	0·3019	0·0635
P8	4	86	0·4884	0·5436	0·9142	0·3143
P9	4	118	0·5847	0·5975	0·0798	0·2571
P10	4	114	0·3947	0·3882	0·0308	0·6857
P11	5	142	0·4859	0·4000	0·8013	0·4762
P12	6	146	0·7123	0·5947	$3x10^{-7}$	0·0649
P13	4	96	0·7812	0·8021	$3x10^{-8}$	0·0286
P14	6	150	0·4067	0·3433	0·0271	0·7879
P15	3	60	0·4167	0·4333	0·2451	0·6000
P16	4	98	0·4796	0·4031	0·7620	0·4857

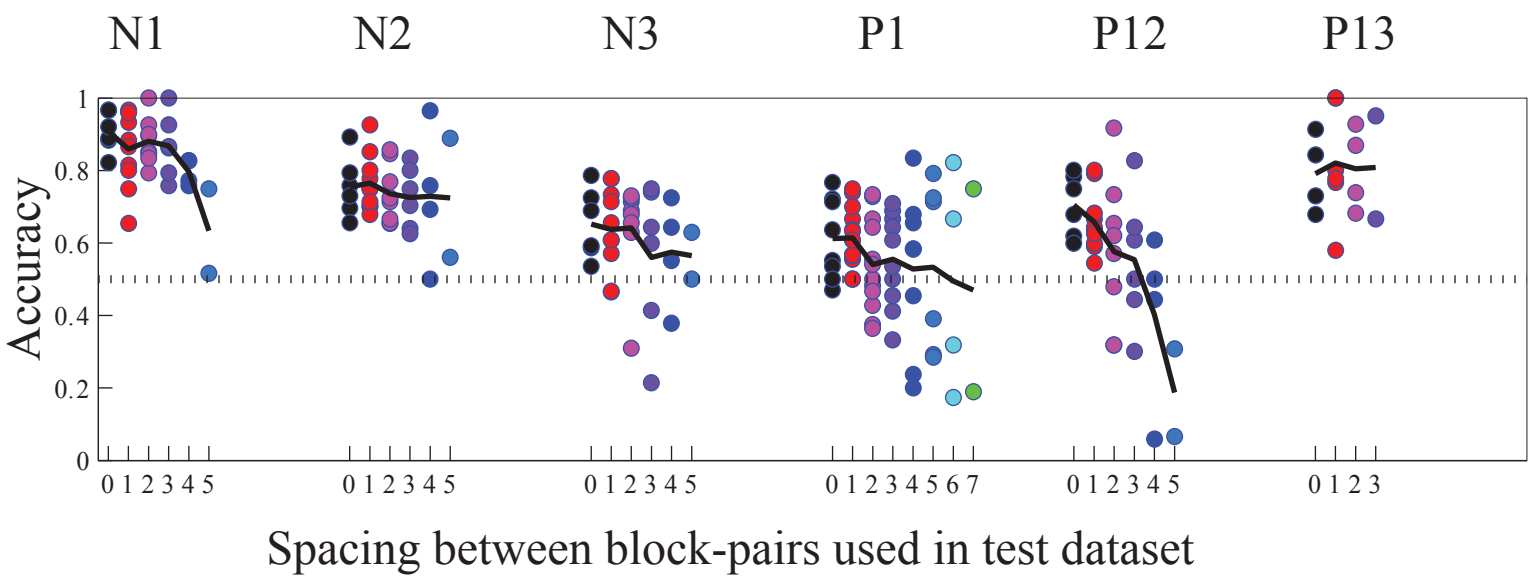
Webappendix Table 1 - Summary of results. Shaded rows are subjects with positive classification in Cruse et al.

* - Method used in Cruse et al·

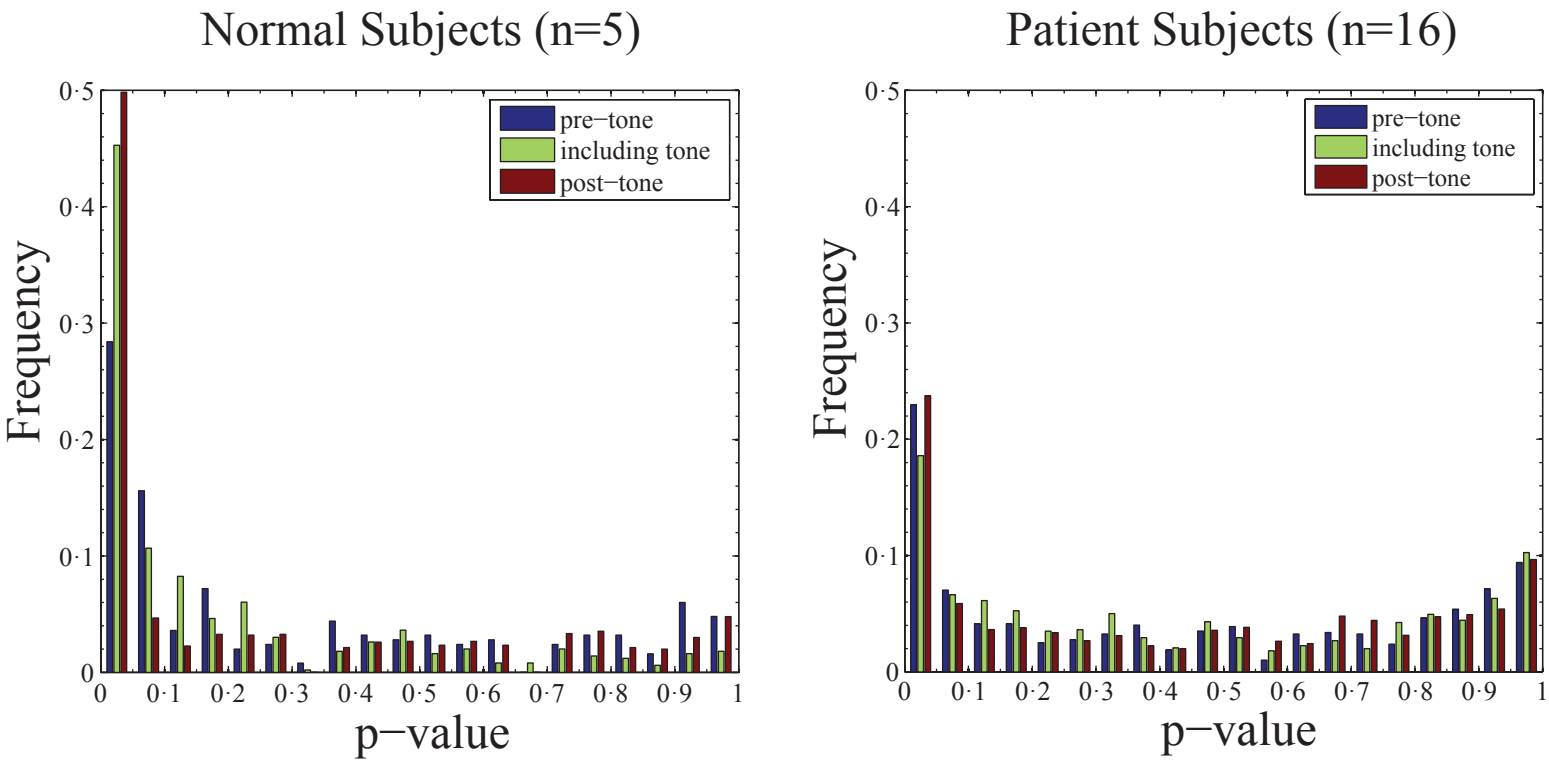
** - Permutation p-values significant after FDR correction for multiple comparisons, corrected alpha=0.05.

Subject ID	Accuracy Pre-tone (-0·5 to 0)	P-value by binomial test
N1	0·6867	$2x10^{-6}$
N2	0·5854	0·0347
N3	0·4941	0·9389
P1	0·5198	0·6225
P12	0·5411	0·3627
P13	0·6354	0·0103

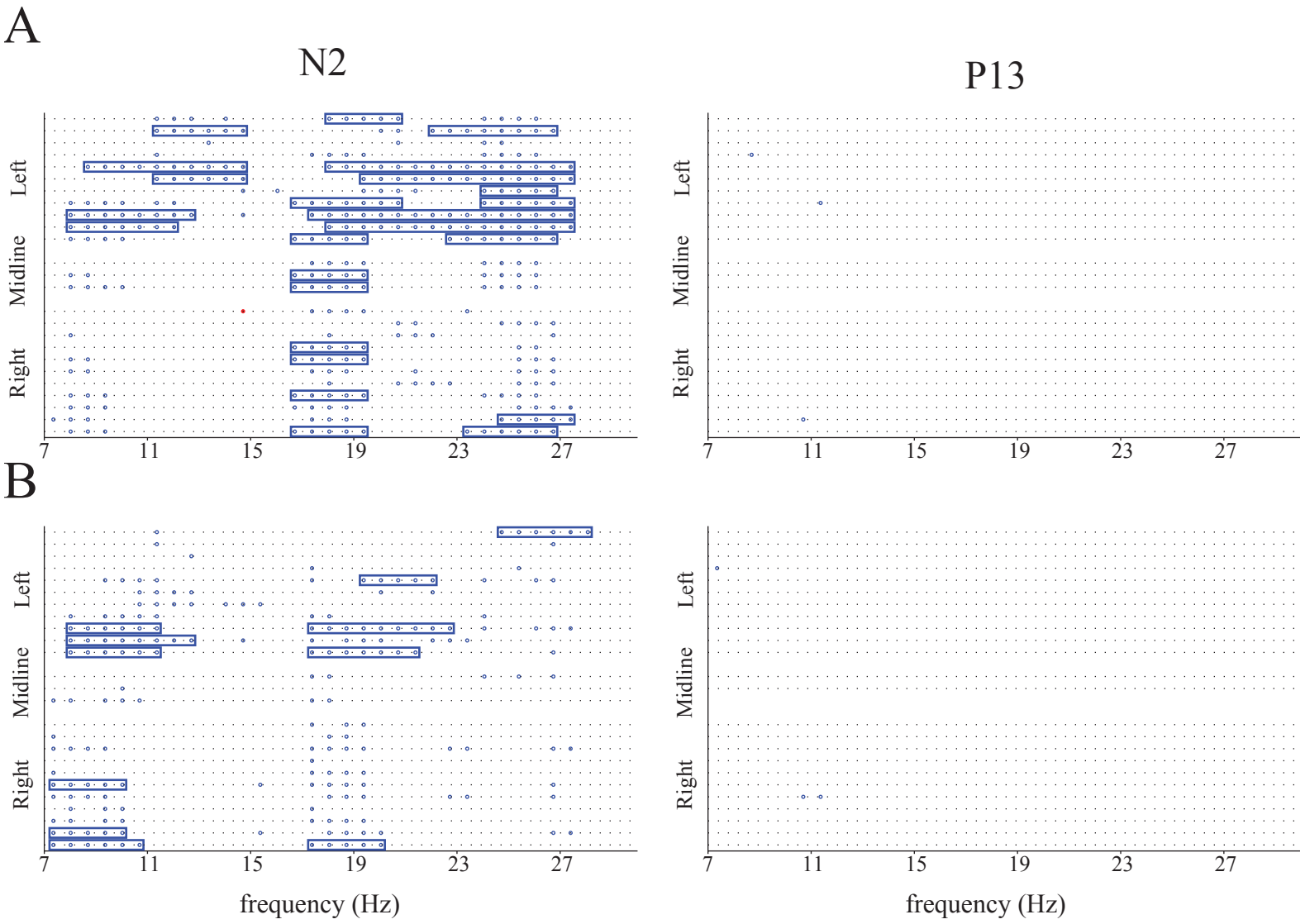
Webappendix Table 2: A re-examination of the negative control used by Cruse et al, consisting of an SVM analysis of the spectral windows within the half-second prior to the tone for the three positive patients and three of the positive normals reported as positive in Cruse et al. Since this period was prior to the tone, failure to detect a response during this period was taken by Cruse et al. as evidence for the method's validity. The above re-examination shows that two subjects, N1 and P13, have p-values substantially less than 0·05, and, by this measure, are inconsistent with the claimed negative control. Calculation is by same methods as in Cruse et al. (i.e., matched-block-pairs for calculation of accuracy and binomial test for calculation of significance). In Cruse et al., results of applying this control to normals were not given, and it was stated that no patients were significant at $p>0·05$; in later personal correspondence, the authors confirmed the above-listed value of $p=0·0103$ for P13, and indicated that they had intended to state $p> 0·01$ in their manuscript.



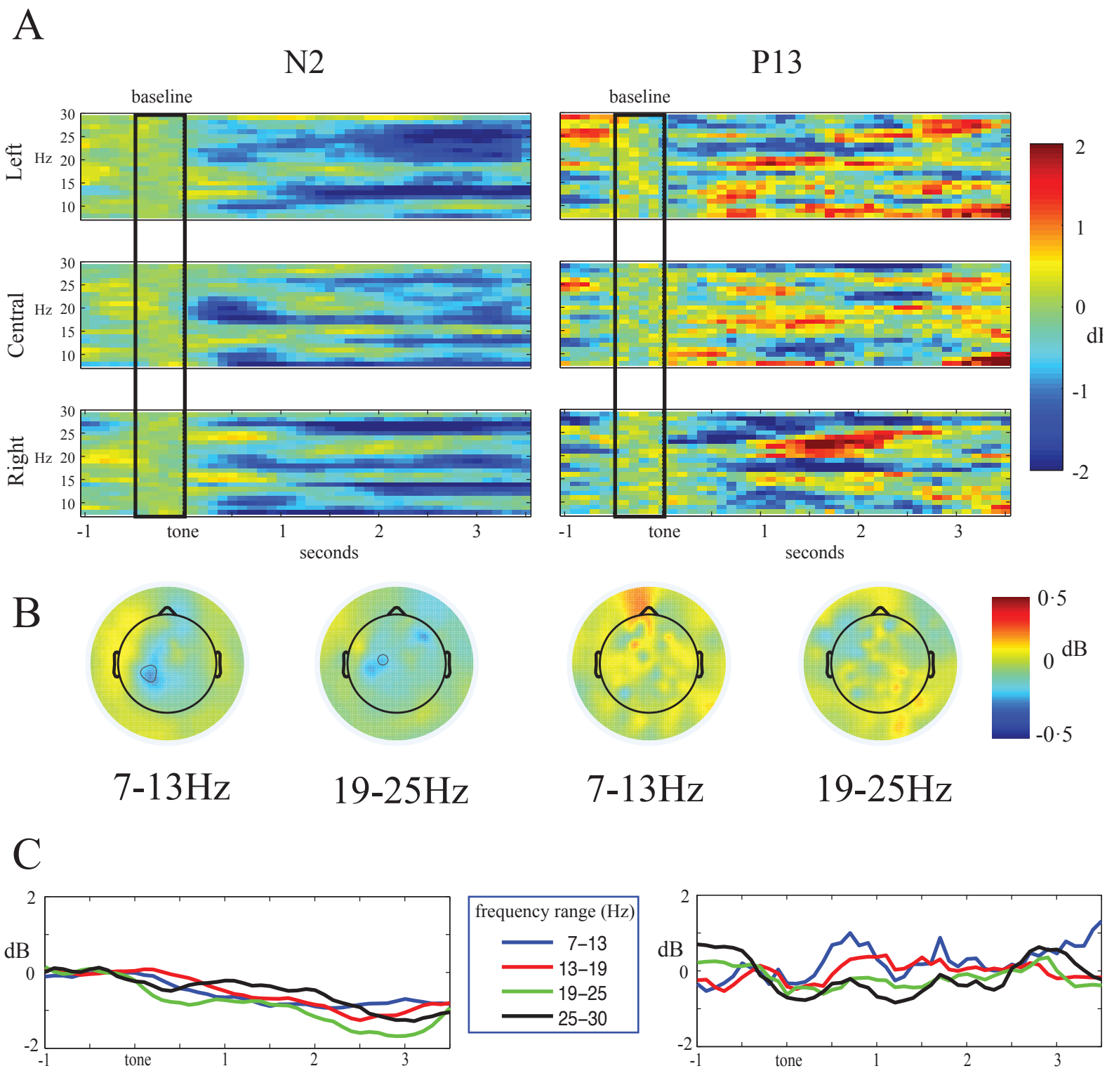
Webappendix Figure 1: Dependence of SVM classification accuracy on the temporal relationship between the test-set blocks. Each filled-in circle represents the accuracy obtained from SVM using a different pair of blocks (one hand and one toes) as the test dataset. Colors correspond to the different spacings displayed on the horizontal axis. Results from three positive normals and the three positive patients are shown. The black solid curves indicate mean values for each separation. When the test-set blocks are adjacent (points plotted over a separation value of 0), accuracy is high; this was the measure used by Cruse et al. In several of the subjects, classification accuracy declines to chance or below-chance levels as the separation increases.



Webappendix Figure 2: Histograms of one-sided p-values of SVM analysis as assessed by binomial statistics, for classifications centered at the full range of time points. Left: results from five normals; Right: results from 16 patients. Results are separated by whether the time point corresponded to a window that was entirely preceding the tone (i.e., a window that began from 1·5 to 1·0 sec pre-tone), included the tone (i.e., a window that began from 1·0 sec pre-tone to simultaneously with the tone), or was entirely post-tone. The vertical axis indicates the frequency that each p-value is obtained within the indicated time range. The null hypothesis (i.e., with a classifier performing at chance) is a flat distribution. An excess of p-values less than 0·05 indicates that the SVM apparently performed substantially better than chance (i.e., evidence for task performance); an excess of p-values greater than 0·95 indicates that the SVM apparently performed substantially worse than chance (i.e., EEG from one task was typically classified with the training-set EEGs of the other task).



Webappendix Figure 3: Summary of univariate tests calculated from all blocks of A hand and B toe tasks, for a typical normal and a typical patient subject. These subjects had nearly equal classification accuracy (N2: 0.75; P13: 0.78) in Cruse et al. In each figure, blue and red circles represent frequencies and channels at which the average EEG power spectra were significantly ($p \leq 0{\cdot}05$) different between task (0·5-2 sec. post-tone) and rest (1·5 to 0 sec. pre-tone) periods. Significant values with more power in rest condition appear in blue; those with more power in task condition appear in red (only one isolated point in A, left). Rectangles around contiguous circles signify adjacent significant values wider than the frequency resolution of the power spectra[8]. N2 shows expected pattern of task-related power decreases[10,11] while P13 shows no statistically significant changes. By the methods of [9] the changes seen for N2 are statistically significant, but those seen for P13 are not. Note that data used for A are the same as for Webappendix Figure 4.



Webappendix Figure 4: N2 reveals expected EEG power changes with hand motor imagery; P13 has a random pattern of change. A. The time-frequency log spectrogram of the EEG, averaged across all trials of all hand blocks, in the same normal (left) and patient (right) as Figure 1 using the same channels as Figure 1B (same data as in Webappendix Figure 3A). The calculation is performed with a window of 1 second and a shift of 0·1 second. To highlight task-related changes, spectrograms are normalized to the baseline (average power from 0·5 sec pre-tone to 0 is subtracted). **B.** Topographic maps of the log EEG power during the task period (1 to 1.5 seconds post tone) normalized to the baseline (0.5 to 0 pre-tone) in two frequency ranges. Contour lines on topographic maps encircle channels with statistically significant power differences between task and baseline at $p \leq 0{\cdot}05$ via a z-test[8]. Note that significant power differences are only present in N2. **C.** Baseline-subtracted spectrograms from A displayed as traces with values averaged across the frequency ranges used in Cruse et al.