

## A simple classification tool for single-trial analysis of ERP components

Christoph Bandt<sup>1</sup>, Mathias Weymar<sup>2</sup>, Daniel Samaga<sup>1</sup>, and Alfons O. Hamm<sup>2</sup>

1) Institute of Mathematics, University of Greifswald, Greifswald, Germany

2) Institute of Psychology, University of Greifswald, Greifswald, Germany

Running head: Single trial analysis of ERP components

Corresponding author:

Prof. Christoph Bandt

Institute of Mathematics and Computer Science

University of Greifswald

17487 Greifswald,

Germany

phone: +49 3834 864632

e-mail: [bandt@uni-greifswald.de](mailto:bandt@uni-greifswald.de)

## Abstract

Event-related potentials were recorded by measuring a dense sensor EEG from eight healthy volunteers in a visual oddball experiment. Single trials were analyzed with an extremely simple high-dimensional version of discriminant analysis. The question was how many of the target trials contribute to the average P3, and to test whether other components in the ERP are sensitive to discriminate between target and non-target trials. One common classification rule for all participants, expressing the P3 component, correctly classified 88% of the ERPs of all subjects in response to a target or non-target trial. For four of the eight participants there were strong differences in an early ERP component over the occipital recording sites. Their individual classification rules, obtained from the training data in the time interval up to 200 ms, correctly classified 85% of the trials of the test data.

The most common method to obtain different ERP components involves averaging samples of the EEG that are time-locked to repeated occurrences of a particular event, e.g., a sensory stimulus. Averaging suppresses the background EEG activity, changes in experimental conditions, and inferences which in single trials can all be much stronger than the brain response under study. On the other hand, averaging obscures intertrial variability (Spencer 2005), and it is not immediately clear to what extent properties of the average are inherent in the single trials.

To clarify this question, the simplest setting is to present in random order two different types of stimuli, and try to distinguish the corresponding brain responses trial by trial. Some of the first single-trial methods developed in this context. Stepwise discriminant analysis was used to classify auditory stimuli (Squires & Donchin, 1976) and to study the influence of the type of preceding stimuli (Squires, Wickens, Squires, & Donchin, 1976). While this work was based on the P300, stepwise discriminant analysis of small and early ERP components could detect ERPs elicited by patterns in the lower and upper visual half-field with an accuracy of 78 to 88 percent (Horst & Donchin, 1979). More recently, these classification methods found an important application in the development of *brain-computer interfaces* (BCI) which enable the direct communication between humans and computers by analyzing particular ERP components that reflect specific functions of the brain. In one line of research participants were instructed to imagine a movement with either their left or right hand (or finger) and the readiness potential was recorded and analyzed on the contra-lateral side of the brain (Blankertz et al., 2007; Wolpaw, Birbaumer, McFarland, Pfurtscheller, & Vaughan, 2002). More recently, slow cortical potentials were used in BCI paradigms to move cursors to specific targets helping locked-in patients to communicate with their environment (see Birbaumer, 2006 for review) and use these brain responses in biofeedback therapy.

Other BCI systems introduced by Farwell and Donchin are based on the single-trial detection of the P300 (Farwell & Donchin, 1988; Donchin, Spencer, & Wijesinghe, 2000; Jansen,

Allam, Kota, Lachance, Osho, & Sundaresan, 2004; Xu, Gao, Hong, Miao, Gao, & Yang, 2004) mostly using different variants of the visual oddball paradigm. In this paradigm one task-relevant visual stimulus is considered as target while the other visual stimulus can be ignored. The rate of correctly classified single trials in these BCI applications ranged from less than 50 % up to 98 %, depending on the task and on the person.

In the present study we use a very simple oddball design which requires less intentional effort than the BCI paradigms and thus seems more comparable to the typical experimental procedures used in psychology. Different checkerboard images (red/white vs. yellow/white) were used as target and non-target stimuli, and participants were instructed to count the targets, a task that has proven to elicit strong task-related mean P3 amplitude (Schupp, Junghöfer, Weike & Hamm, 2003). *The question was how many of the target trials contribute to the average P3, and to test whether other components in the ERP are sensitive to discriminate between target and non-target trials in the single trial analysis.*

Another aim of the experiment was to develop a simple screening methodology for single trials. Many powerful and sophisticated tools have been applied to single trial analysis (see for instance Poolman, Frank, Luu, Pedersen, & Tucker, 2008), ranging from independent component analysis (ICA) (Makeig, Debener, Onton, & Delorme, 2004; Xu et al., 2004; a corresponding open-source toolbox was presented by Delorme & Makeig, 2004) and wavelet representations (Effern et al., 2000; Quiroga & Garcia, 2003) to more specific methods like recurrence plots (Schinkel, Marwan, & Kurths, 2007). In the context of classification, discriminant analysis was replaced by its modern extension, the support vector machines, and other methods of machine learning (Blankertz et al., 2006; Kaper et al., 2004; Mensh, Werfel, & Seung, 2004; Wang et al., 2004). In a recent comparative study, however, Fisher's and stepwise discriminant analysis also performed well (Krusienski et al., 2006). If there are successful methods, why do they not become commonly applied in different areas of psychology? The reason might be a "tower-of-Babel effect": methods are too complicated to

be easily understood and adapted by a large group of researchers. The difficulties start with different approaches for pre-processing of the data, for instance low-pass and artefact filtering. Moreover, multivariate procedures often need a variety of parameters which have to be fine-tuned to the problem at hand, which can be done only by the expert. For the development of brain-computer interfaces this is no serious problem since the choice of parameters can be delegated to the computer and the best classification methods can be unified in an “expert system”, so that the machines will work. However, when single-trial analyses should be used for understanding the dynamics of human brain during cognitive processes, like attention and learning, or emotional modulations like fear acquisition and extinction, there is a need for simple and transparent methods that can be applied to various data sets without specific modifications.

*The extremely simple algorithm presented in this study comes from pattern recognition theory and combines averaging with single trial analysis. In contrast to other versions of discriminant analysis, this algorithm takes as input the original high-dimensional data with minimal pre-processing. Average responses to target and non-target stimuli are determined from the training data, and their difference is used as template for the classification of single trials. Each single trial in the test data is multiplied with the template function. For the target trials, the result tends to be positive, for the non-target trials negative.*

## Methods

### Participants

Eight healthy male volunteers, with normal or corrected-to-normal vision, participated in the present study. The age range was 19 to 26 years. All participants signed an informed consent prior to the study.

### Stimulus Materials, Procedure, and Recording

Task related stimuli were two distinct checkerboard images with yellow/white or red/white patterns. The target stimulus (either the red/white or the yellow/white checkerboard, balanced across participants) was presented 23 times while the non-target checkerboard image was presented 127 times. Images were presented on a 21 inch EIZO F77 computer screen located approximately 100 cm in front of the participant. Each trial started with the 0.5 s presentation of a fixation cross on the screen followed by the 0.75 s presentation of the checkerboard image. Inter-trial intervals varied between 1 and 1.5 s. The different checkerboards were presented in a pseudo random sequence allowing the presentation of 2 to 15 non-target stimuli between two successive target trials. Participants were instructed to count the number of target trials (either red/white or yellow/white checkerboards) and were offered a small monetary reward for counting correctly.

EEG was recorded with a 129 lead sensor net from Electrical Geodesics, Inc. (EGI), with a sampling rate of 500 Hz and on-line hardware band filter from 0.1-100 Hz. Data were continuously recorded with the vertex sensor as reference electrode, and the common median over all channels subtracted from each channel. (There was little difference between median and mean, except for time periods with eye artefacts, which changed the mean much more than the median.) Stimulus synchronized epochs lasting from 100 ms before until 900 ms after the onset of the checkerboard images were extracted. No other filters or pre-processing

procedures were applied. In particular, no trials were excluded, and the 50 Hz interference as well as possible artefacts were kept in the data.

### Data Analysis

All steps of data analysis were kept as simple and transparent as possible. Our study was restricted to 55 of the 128 sensors from the central, parietal and occipital regions of the brain where oddball effects are to be expected. A first screening of the data had shown that more temporal and frontal sensors did not show much differential activity during the task, and the frontal sensors also contained many eye movement artefacts. In Figure 1 the selected 55 sensors are shown, and sensors with high and low potential for classification are colored red and blue, respectively. Since the median of all 128 channels was subtracted from each channel, the reference sensor is given as the negative of this median.

After the screening each ERP was analyzed from 0 to 500 ms after stimulus onset, a time period which in general contains the P3 component as well as earlier components that might be sensitive to discriminate between target and non-target responses. The main research question was focussed on the correct categorization of single trials into target or non-target responses. To this end, the first 75 trials were taken to determine the classification rule. The rule was applied to the remaining 75 trials to test its reliability. This strategy was applied in different settings: (a) either using the data of single sensors, or groups of sensors, or all 55 sensors. The question is which channels contain most information about the response discrimination between target and non-target trials, and to what extent does the multivariate signal from several sensors improve the information from single channels. (b) either taking one common classification rule for all individuals, or taking individual classification rules for every participant. The common classification rule describes the general oddball effect of all subjects, perhaps even beyond the group studied here. The individual rules are adapted to specific reactions of each participant and thus could obtain a better classification.

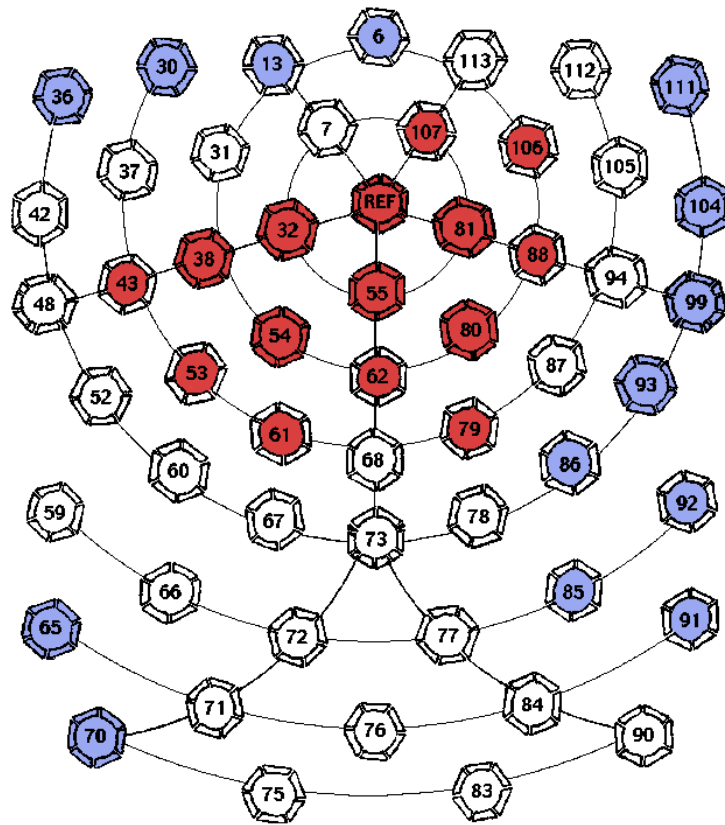


Figure 1. The 55 channels selected from the 129 lead EGI system. Shading indicates classification strength of the sensor: full blue  $AUC < 0.75$ , partly blue  $AUC < 0.78$ , full red  $AUC > 0.88$ , partly red  $AUC > 0.85$

Classification rules were determined directly from the training data, without using any knowledge of brain dynamics. In particular, no effort was made to adapt the method to changing trials, for instance latency jitter in the P3. Single trial analysis of the ERPs was conducted using a simplified version of Fisher's discriminant analysis in a high-dimensional version. This method has been used in pattern recognition, for instance for face recognition from visual images (McLachlan 1992). Since the application to EEG data seems to be new, the method is described in more detail below.

Simplified discriminant analysis

Let us first consider a single sensor. Every trial is a function  $x(t)$  of time,  $t=1, \dots, T$ . In our case we have  $T=250$  values in the 500 ms interval. All trials are normalized so that their mean over time equals zero:  $\sum x(t)=0$ . Moreover, we have an average function  $r(t)$  of all target trials, and another average function  $m(t)$  of all non-target trials (which then also have mean zero).

*Now we take the difference of group averages  $r(t)-m(t)$  as a template function. The score  $S(x)$  of the trial  $x$  is obtained by multiplying  $x$  with the template and taking the mean over time  $t$ :*

$$S(x) = (1/T) \sum (r(t)-m(t)) x(t).$$

If  $S(x)$  exceeds a certain constant  $C$  then  $x$  is classified as target, otherwise as non-target trial.

The underlying idea is that a target trial  $x$  should be similar to the average target  $r$ , so  $x(t)$  should be positive for those  $t$  for which  $r(t)>m(t)$ , and negative for  $r(t)<m(t)$ . Thus, all terms in  $S$  are positive. A non-target pattern will have negative values where  $r(t)<m(t)$ , and positive values where  $r(t)>m(t)$ , so that the terms in  $S$  will be negative. Of course, in reality this is only a tendency, as illustrated in Figure 2. The left column shows every second target trial (red) and every tenth non-target trial (blue) of the training set of channel Oz, for participant 1 in the time between 100 and 200 ms, and for participant 6 in the time between 200 and 300 ms. The score functions  $(r(t)-m(t)) x(t)$  shown in the middle are much better separated than the trial data themselves. Multiplication by the template amplifies the differences between the two groups. If we now take as score  $S(x)$  the mean of all values of the score function, the scores of the target trials should be larger than those of the non-target trials, as shown in the column on the right. This works well in the first row, even for the test data which have nothing to do with the template function. In the second row, the discrimination is still acceptable when it is combined with other data. The line represents our bound  $C=(S(m)+S(r))/2$ , halfway between  $S(m)$  and  $S(r)$ , indicated by the big symbols in the center. This bound is slightly better than

$C=0$  since  $m$  and  $r$  may have different magnitudes. In case of the P300, for instance,  $m$  may be constant zero, and  $r$  will form a peak.

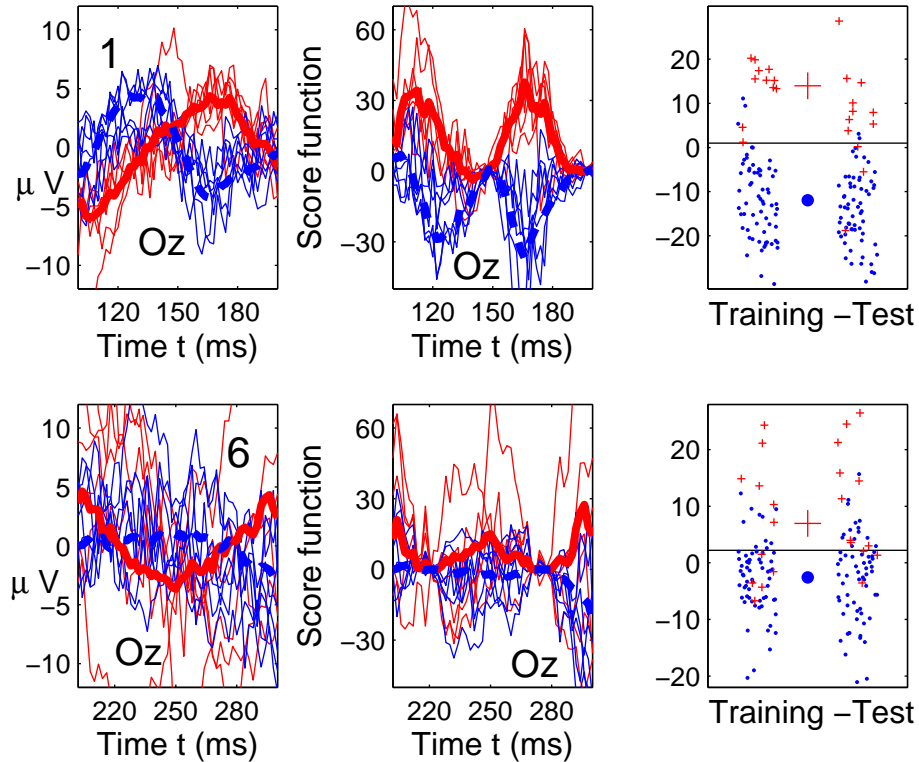


Figure 2. The template function is the difference between averages of target and non-target trials. Single trials of subjects 1 and 6 on the left, targets in red, non-targets blue, averages denoted by thick lines, non-target average dashed. Trials multiplied with template function are shown in the middle. Where the averages intersect, all score functions are zero. The right column shows the scores (means over time of score functions) of all training and test trials, with a slight random scatter in horizontal direction.

It can be proved mathematically that the template method agrees with Fisher's discriminant analysis in the case that all possible target and non-target trials are modelled in the form  $\rho(t)+\epsilon(t)$  and  $\mu(t)+\epsilon(t)$ , respectively, where  $\rho(t)$  and  $\mu(t)$  are the ideal target and non-target functions, and  $\epsilon(t)$  denotes Gaussian white noise.  $r(t)$  and  $m(t)$  are then the estimates of the functions  $\rho(t)$  and  $\mu(t)$  from our data. The assumption of uncorrelated Gaussian noise is

certainly not very realistic, but reasonable – like Fisher’s assumption of Gaussian distribution with equal covariance structure in both groups in the low-dimensional setting. Fisher had proved that his weight function is the best possible linear discriminant function (McLachlan 1992, Huberty 1994), and this applies to our setting, too. Our focus is on simplicity more than optimality, however.

Here are the different settings for our classification. For the common classification rule we take  $r(t)$  and  $m(t)$  as grand averages of the target and non-target trials, respectively. For individual classification rules we have to determine  $r(t)$ ,  $m(t)$  for each participant separately. Next, consider the multivariate analysis of several, or all 55 sensors. The score of a trial was first determined separately for each sensor, and assigned a rank number within all scores of that sensor. The rank is between 1 and 75 for the 75 test trials, with 75 assigned to the largest value. The total score of a trial is then obtained as the average of the trial ranks in all sensors. This procedure is used in sailing contests, to balance between slow and fast wind conditions. In our case it guarantees that all sensors contribute to the total score in the same way, independently on the size of their amplitudes.

### Classification errors

The choice of the threshold  $C$  determines the classification errors and is not easy (cf. Huberty 1994). We must realize that like in statistical tests, there are *two errors*:  $\alpha$  – wrongly classifying a non-target trial as a target, and  $\beta$  - wrongly classifying a target trial as non-target. When  $C$  is increased,  $\alpha$  decreases and  $\beta$  increases. In a symmetric setting, like the left-right paradigm in motor imagery, one can take just the total number of misclassified cases as error. In an oddball experiment this does not make sense, however. Since target trials are rare (e.g., 10 %), classifying *all* trials as non-target trials will then result in an “overall error” of 10%, but this is not what we want. The  $\beta$  errors must be given greater weight, since there are fewer target cases, and since after all, their recognition is our goal. So it is custom to take

$(\alpha+\beta)/2$  as classification error, which gives equal weight to both groups independently of their size. Since both errors have different quality - less misclassified non-targets do not really compensate more misclassified targets – many authors recommend the *equal error rate*: choose  $C$  so that  $\alpha=\beta$ . Here we take  $(\alpha+\beta)/2$  as error and try to get equal error rates. However, we have just 12 oddball trials among the 75 test trials, so every misclassification counts for  $1/12=8.3\%$ , and  $\beta$  can assume just the values 0,  $1/12$ ,  $2/12$  etc. Thus in an experiment like ours, an equal error rate can only be realized as crude approximation.

To provide realistic error rates, we have to estimate the threshold  $C$  only from training data. This can increase the error values, even in the case of perfect separation of targets and non-targets, when we predict a wrong place for  $C$  (see Table 4). Since we have only 11 target trials for each individual in the training data, we use the simple estimator  $C=(S(m)+S(r))/2$ . Under the ideal conditions mentioned above, this estimator yields equal error rates (cf. McLachlan 1992, section 5.4).

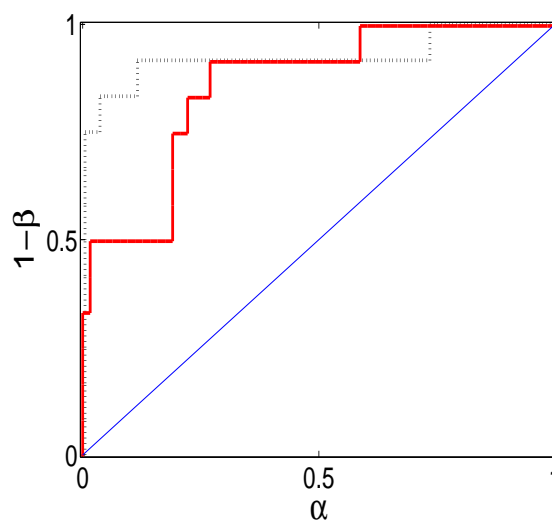


Figure 3. ROC curves for the two classifications of test data in Figure 2 with  $AUC=0.92$  and  $0.86$ . Classification by pure guess with  $AUC=0.5$  is indicated by the diagonal.

An objective estimate of classification errors which does not rely on a choice of  $C$  is the so-called *AUC value*. This parameter comes from signal analysis in engineering (ROC analysis,

‘‘receiver operation characteristic’’) and has become absolutely standard in clinical studies (Hanley 1989; Armitage & Berry, 1994), but apparently not in the analysis of EEG signals (see, however, Poolman, Frank, Luu, Pedersen, & Tucker, 2008). The ROC curve is drawn in a square; it represents the correct classification rate  $1-\beta$  of the target trials as a function of the classification error  $\alpha$  of the non-target trials, and AUC means the *area under this curve*. This is just an ‘‘average number of correctly classified cases’’ taken over all  $\alpha$ , in other words, all possible thresholds  $C$ . Statistics packages like SPSS contain the corresponding procedures.

Since we present AUC values below, let us explain how they are calculated. Suppose target trials have the larger classification scores  $y$ , and the classification scores of the non-target trials are  $z$ . We determine for each  $z$  the percentage of  $y$ -values (among all  $y$ -values) which are larger than  $z$ . We add all these percentages, and divide by the number of ordinary trials to obtain AUC. In the case of perfect separation – all  $y$  larger than any  $z$  – we get  $AUC=1$ . When the  $y$ - and  $z$ -values are randomly ordered, we obtain  $AUC=0.5$ . This corresponds to classification by pure guess, or coin-tossing (Figure 3).

## Results

Seven subjects counted the checkerboard targets entirely correct while one participant miscounted two of the targets. One would not expect the same accuracy when counting is based on the analysis of EEG data only. Our main result goes into this direction, however: *even with our rather crude method, about 90 % of the targets and non-targets can be recognized from the ERPs.* In other words, the oddball effect can be seen in almost every single trial. Let us start with the P3 component, the most prominent and most common oddball indicator.

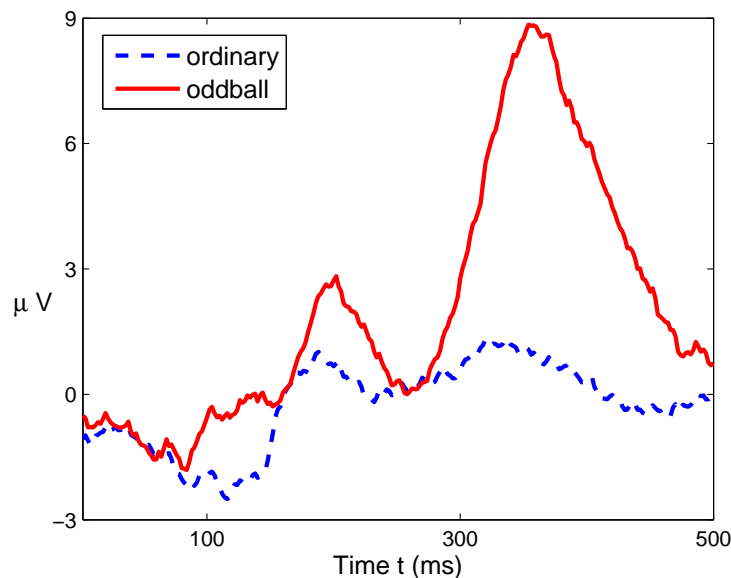


Figure 4. Grand average in channel Cz

### Averages of the P3.

As a typical example for a sensor in the centro-parietal region, Figure 4 shows the grand average of all ERP-responses over Cz (the reference of the EGI system) which was one of the best sensors for classification. P3 amplitudes elicited by the target stimuli were about 8  $\mu\text{V}$  larger than the responses to the non-target checkerboard. This is highly significant. More importantly, this effect can be seen in the averaged P3 of each participant although shapes, lengths and amplitudes of the P3 are quite different (Figure 5).

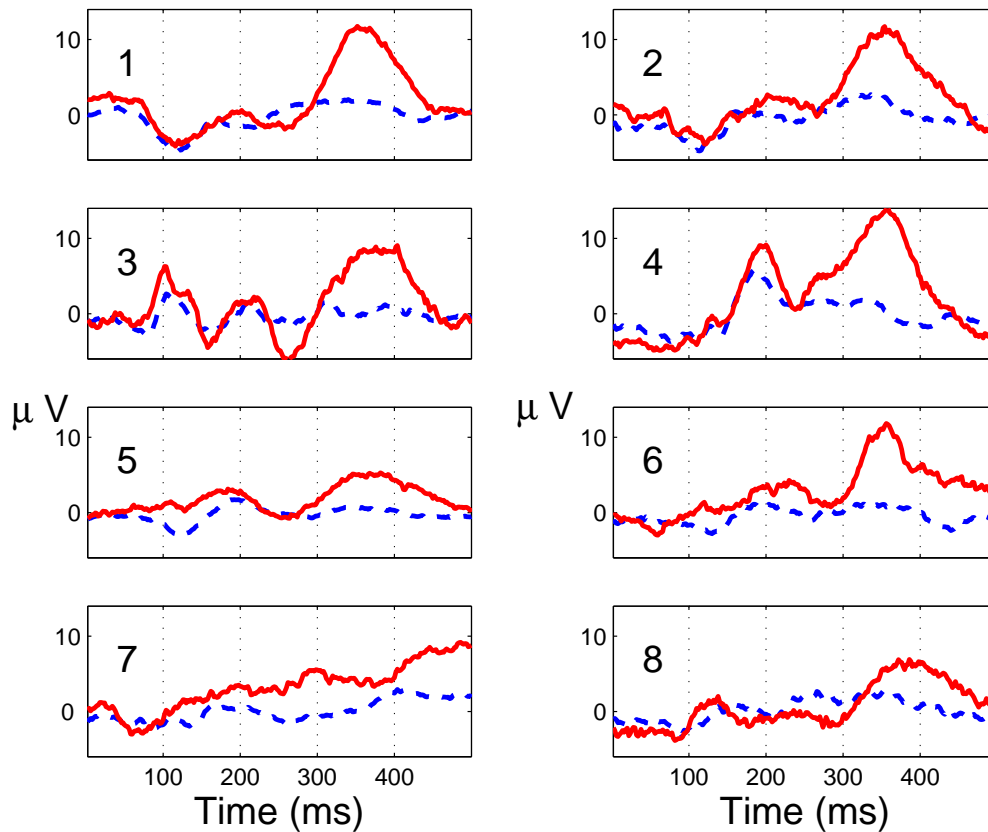


Figure 5. Average ERPs of non-target (dashed) and target (solid) trials in channel Cz for the eight individuals.

Thus the grand average of Cz expresses a common tendency of all individuals. We resisted the temptation to shift the curves in Figure 5 to the same origin, as well as to smooth the curves by moving average. The small waves in Figures 4 and 5 indicate the lack of low-pass filtering.

### Sliding box-plots

In this analysis the question is investigated whether the individual averages for the targets and non-targets express a common tendency of *all* respective trials. It would be confusing to see the curves of all the trials together (cf. Figure 2). So the 25% and 75% percentiles of all target trials, and all non-target trials, were determined for each time point  $t$  between 250 and 500 ms,

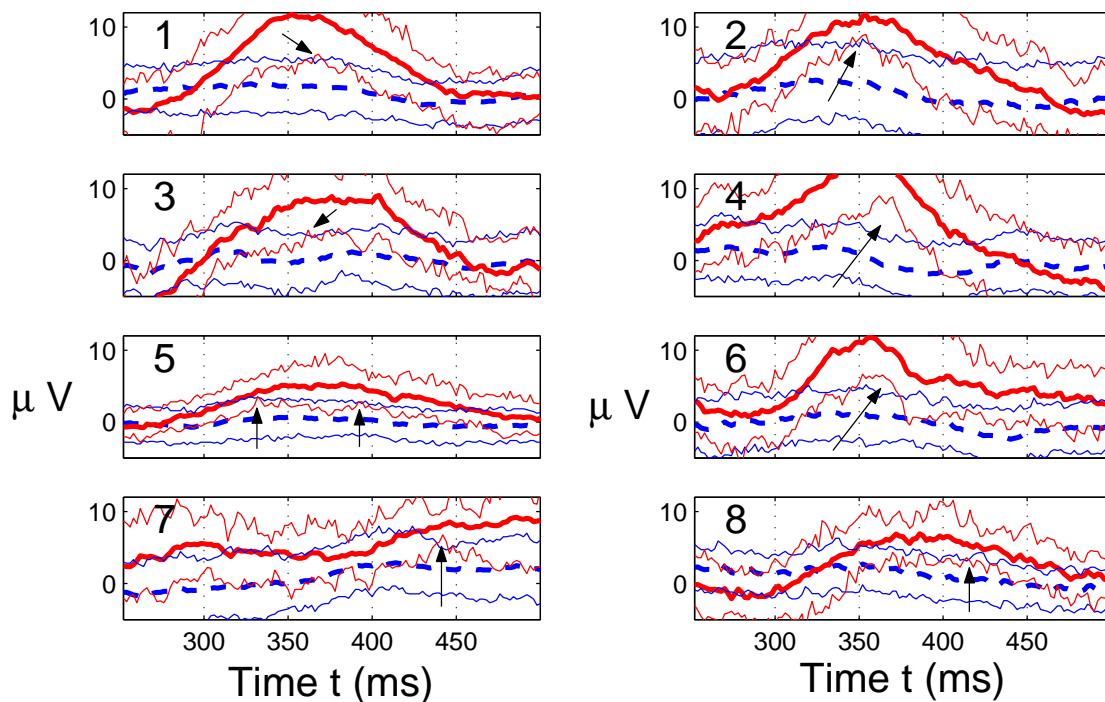


Figure 6. Sliding box-plots of channel Cz for all individuals. Arrows indicate time segments where the boxes do not intersect, so that the value of Cz at one time point classifies at least 75% of all non-target trials and all target trials.

and drawn in Figure 6 as curves accompanying the averages. There is a red trace for the target trials, and a blue trace for the non-targets. For every time point  $t$ , the middle half of the values  $x(t)$ , of the targets and of the non-targets, are contained in the respective trace – although only few trial curves will remain within the trace for all time. The two intervals given by a time point  $t$ , as vertical section with the traces, correspond to the boxes of box-plots of the values  $x(t)$  for targets and non-targets. Thus Figure 6 could be called a sliding box-plot. The arrows in Figure 6 indicate time points where the red and blue traces do not overlap. This means that we can put a threshold  $C$  between the two intervals, and can *separate the groups of target and non-target trials by one time point in one channel with an equal error rate of less than 25%*. In fact, this could only misclassify the smallest 25% of the larger group, and the largest 25% of the smaller group. For participant 4 and 6 we find a lot of such time points, indicating a

good separation between targets and non-targets. For the other subjects we also found such points (although less often) which still indicates that the P3 effect reproduces in single trials. A classification by time averages will be more useful than using special time points, however.

#### Single sensor classification with grand average

We determined the scores  $S(x)$  where the template function  $r(t)-m(t)$  was taken from the grand averages of target and non-target trials in a single sensor, as shown in Figure 4. The training data of the simplified discriminant analysis always contained 11 target and 64 non-target trials for every participant, the test data 12 target and 63 non-target trials. The corresponding average AUC values, taken over all trials of all individuals, were determined for each sensor. The best sensor was #38 with an average AUC of 0.896 over all individuals (Table 1). The maximum AUC was above 0.99 (participant 1), and the minimum was 0.72 (participant 7). The four sensors with highest AUC were all near to Cz: #38, 32, 81, and Cz itself, still with average AUC=0.881 (cf. Figure 1). With our simple estimate of C, the average equal error rate was almost 20 % since for the common rule, a common threshold had to be chosen for all individuals (Table 1 shows errors for sensor #38. For Cz, #32, and #81 they are almost the same). Thus, we can state that *with a single sensor in the central region and with one common classification rule based only on grand averages of the training data in a channel near to Cz, about 80 % of all trials of all 8 individuals were correctly classified*. It should be noted, however, that *rankings of the sensors* here and below are all based on the scores of the *test* data. It is not appropriate to evaluate the fit of the training data to the formula derived from the training data themselves. This was checked just for curiosity, and for the grand average classification occipital channels performed best.

#### Single sensor classification with individual averages.

We determined scores where the template functions  $m(t)$  and  $r(t)$  were individual averages in a sensor, as shown in Figure 5. This was done for each person and each sensor.

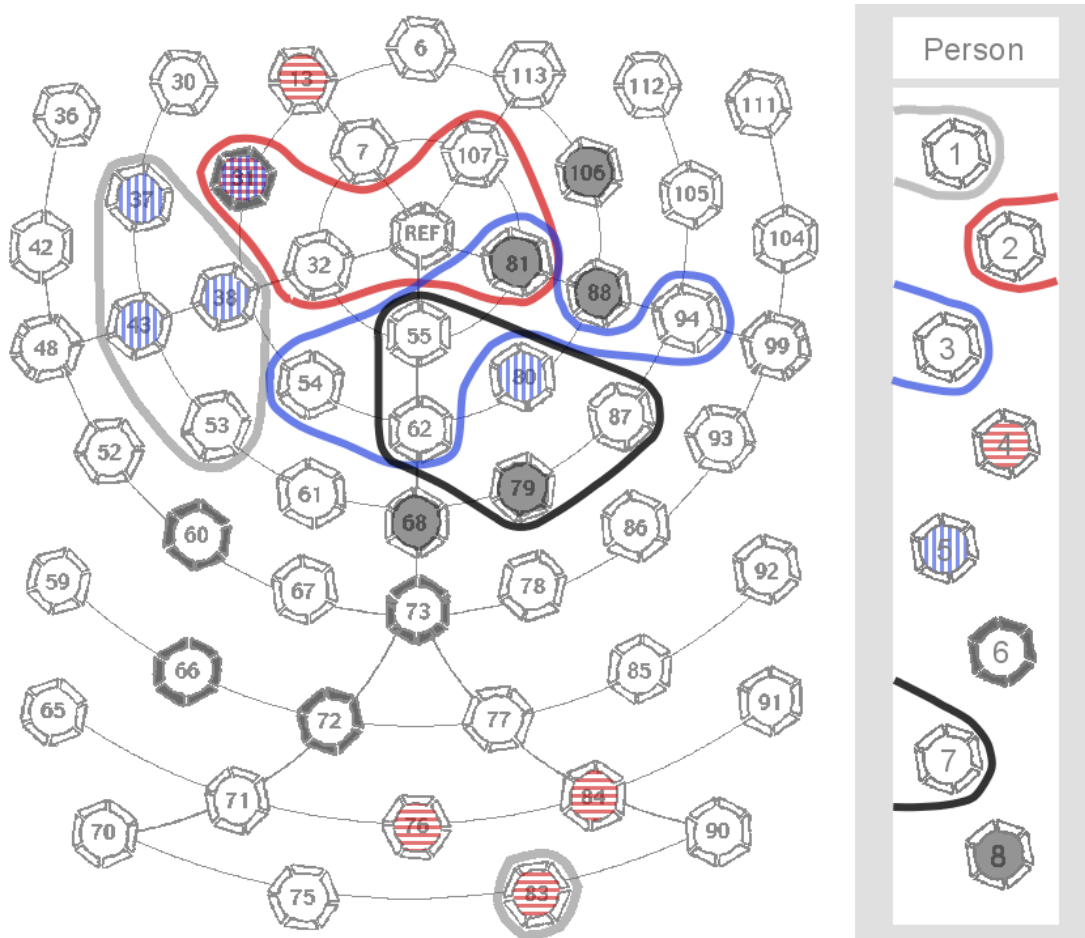


Figure 7. The five sensors with highest AUC for each individual. For participants 2,3, and 7 they form one cluster enclosed by a coloured curve. For the others, sensors with highest AUC are indicated by shadings as shown in the inset, and they lie in two separate regions. Sensor 31 has high AUC for participants 2,4,5, and 6.

When the same sensor is used for all individuals, the results become only slightly better than Table 1. In Figure 1, the sensors with  $AUC > 0.85$  were shaded red. They all belong to the centro-parietal region. This proves that a good “grand classification” – one formula which applies to all individuals – must be based on the P3 component. The sensors with  $AUC < 0.78$  were shaded blue. They all represent peripheral positions confirming our initial selection of 55 out of 128 sensors. No sensor was completely bad, the smallest AUC value was 0.64.

*The classification improves remarkably when optimal sensors are chosen individually.* Figure 7 shows the five sensors with largest AUC for each participant which are at quite different positions. At least one sensor for each subject lies in the centro-parietal region, but some occipital sensors also have high AUC.

Participant	1	2	3	4	5	6	7	8	Average
AUC	0.991	0.832	0.921	0.870	0.925	0.929	0.724	0.976	0.896
Error in %	11.1	26.2	19.4	24.0	13.3	17.7	36.7	4.8	19.1

Table 1. The best one-sensor grand average classification with sensor #38

Participant	1	2	3	4	5	6	7	8	Average
Best sensor	37	32	55	<b>76,83</b>	38	72	80	88	
AUC	0.921	0.925	0.934	<b>1.000</b>	0.931	<b>0.987</b>	0.913	<b>0.988</b>	0.950
Error in %	13.9	17.1	17.9	<b>1.2</b>	14.1	<b>4.8</b>	8.9	<b>5.0</b>	10.4

Table 2. Best one-sensor individual classification functions for every participant

Table 2 lists the sensor with highest AUC for each individual. It shows that *with one-sensor individual classification, an average AUC of 0.95 was obtained. Even with estimating the threshold  $C$  from the training data, the average error was only 10.4%.* The worst AUC was 0.913 (participant 7), the largest error 17.9% (participant 3). Participant 1 performed better with the grand average, due to a latency shift of the P3 between his training and test data. On the other hand, participant 4 had a 100 % classification of all test data with sensor #76 which corresponds to Oz as well as with the neighbouring sensor #83. This perfect one-sensor individual classification was not based on the P3 component! This leads us to early oddball indicators in the occipital region.

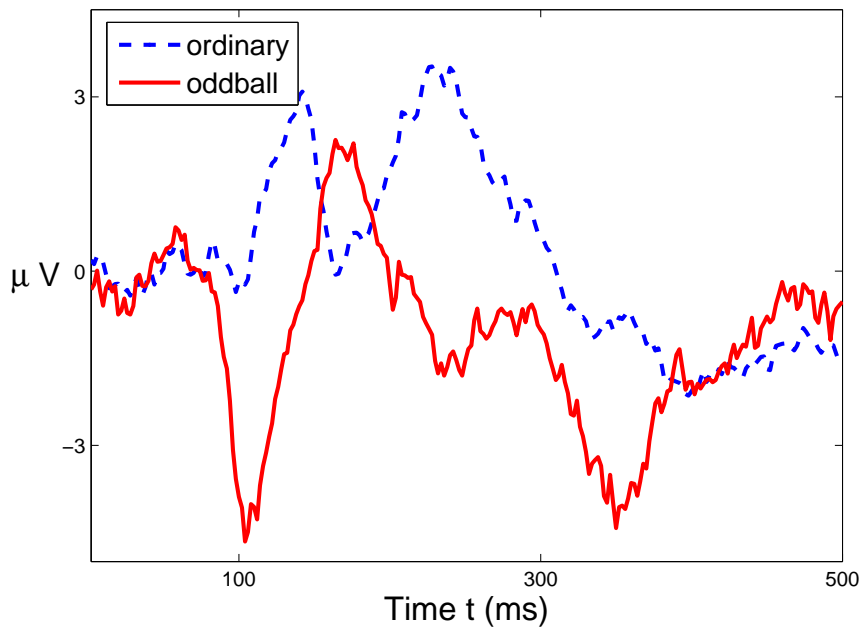


Figure 8. Grand average in channel Oz (sensor #76)

### Oddball effects beside P3

Figure 8 shows the grand average of sensor #76 which corresponds to Oz in the 10-20-system. Although this channel is not considered typical for oddball effects in the literature, there are three clear negative peaks at 100, 250, and 350 ms in the target average, and two positive peaks at 130 and 250 ms in the non-target average. Differences of  $3 \mu\text{V}$  between the grand average curves after 100, at 250 and 350 ms are clearly significant: since the standard deviation of single measurements is at most  $10 \mu\text{V}$ , a t-test gives z-values beyond 3. However, in contrast to Cz, the grand average of Oz does not express a common tendency of all individuals. The individual averages in Figure 9 show an N1 for targets in participants 3 up to 8, but the N2 and N3 only in participants 2, 4, 6, and 8. The P1 for non-targets can be seen in participants 1, 2, 4, and 5. Thus, some of the participants do significantly influence the grand averages. In the same way, individual averages could be determined only by part of the trials. To study this problem, sliding box-plots were drawn. It was found that at least in participants

1, 4, and 8, the vast majority of the trials follow the pattern of the individual average. Figure 8 shows several time points before 200 ms where the red and blue traces are separated.

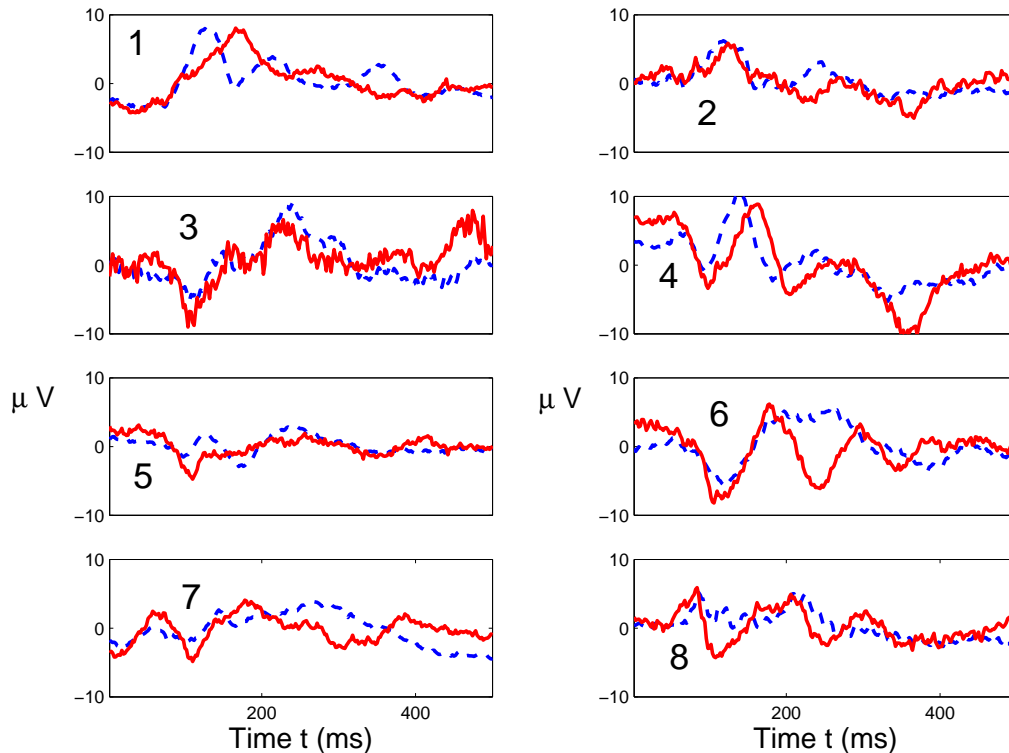


Figure 9. Average ERPs of non-target (thick) and target (thin) trials in channel Oz for all individuals.

The curves for individuals 1, 4, and 8 in Figure 10 look rather similar. Since targets have smaller values around 120 ms and higher values around 170 ms, which can also be seen for single trials of participant 1 in Figure 2, template functions should apply well to this situation. A discriminant analysis was applied to the individual averages in the time interval [0, 200 ms], instead of [0, 500 ms], and to all sensors. The results for participants 2, 3, 6, and 7 were bad: AUC was at most 0.75, and the “best” sensors were not in the occipital region. However, the results for participants 1, 4, 5 and 8, presented in Table 3, seem quite remarkable. The best sensors were Oz and its neighbours, and AUC was far beyond 0.9. *In four of eight individuals, an occipital oddball effect before 200 ms could be verified in 85 % of all single trials.*

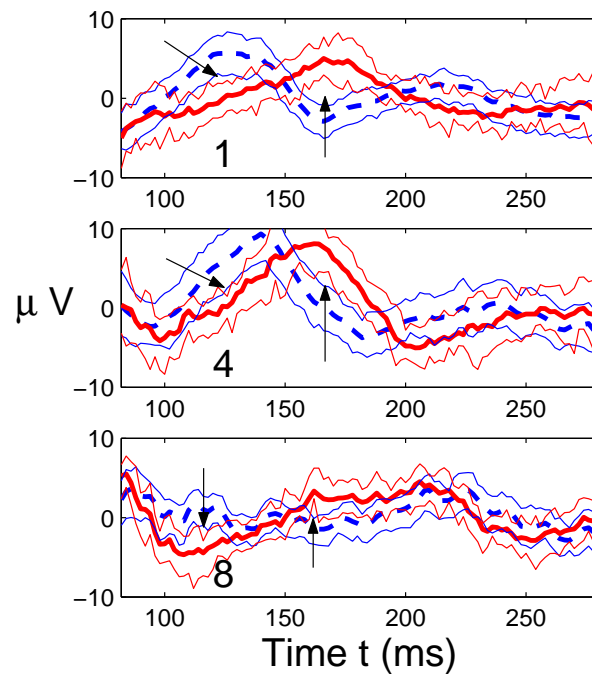


Figure 10. Sliding box-plots of channel Oz for participants 1, 4, and 8

Participant	1	4	5	8	Average
Best sensor	75	83	83	76	
AUC	0.935	<b>0.993</b>	0.954	0.938	0.955
Error in %	11.5	<b>5.0</b>	24.0	18.8	14.8

Table 3. Best one-sensor classifications between 0 and 200 ms, before the onset of P3

### Classification with several sensors

The grand classification was done for each of the 1485 possible pairs of sensors. Compared with a single sensor (Table 1), the improvement was not impressive. Similarly, all pairs of sensors were tested for each individual. The best pairs are given in Table 4. The average AUC was above 0.97, and the minimum AUC in participant 7 still above 0.92. For participant 4, there were 47 most different pairs of sensors which gave perfect classification.

Subject 8 also achieved perfect classification, and participant 6 almost did. It seems interesting to note that the pairs with maximal AUC were never neighbour sensors. Also clusters of few neighbouring sensors did not yield better AUC with our method than the best single sensor of the cluster. However, the grand classification with all 55 sensors, documented in Table 5, gave much better AUC. The error rates in Table 5 are based on a common rule for all individuals: 19 of the 75 test trials were assumed to be targets. On the other hand, individual classifications with all 55 sensors were much worse than those in Table 2 with a single sensor.

Participant	1	2	3	4	5	6	7	8	Average
Sensor pair	43,77	13,32	52,94	many	31,79	30,73	80,107	84,88	
AUC	0.962	0.950	0.967	1.000	0.975	0.996	0.921	1.000	0.971

Table 4. Best individual classifications with two sensors

Participant	1	2	3	4	5	6	7	8	Average
AUC	0.971	0.906	0.951	0.935	0.927	0.963	0.825	0.989	0.934
Error in %	10.5	15.5	10.5	10.5	10.5	10.5	20.4	5.5	11.7

Table 5. Grand average classification with 55 sensors

## Discussion

An extremely simple version of discriminant analysis was used for classification of single trials. Essentially, the method involves only a single multiplication step. Fisher's linear discriminant analysis requires feature extraction, estimation of a covariance matrix and solution of a system of linear equations for the weights of the features. Stepwise linear discriminant analysis, introduced to ERP research by Donchin and co-authors in the seventies (see the references in Horst & Donchin, 1979) selects optimal features and optimal weights in a more complex way. The forward method starts with the single feature which discriminates best, and adds step by step those features which provide most additional information for the separation of the groups. The backward method begins with all features, and step by step discards features with little discriminative value. Both methods can be combined.

Here we use the original data instead of extracting features, and determine weights just as differences of the group averages at each time point, without further calculation. The drawback of such a simple method is that we have no degrees of freedom, no parameters to adapt to changing situations. This method can classify only events which repeat very regularly.

### The oddball effect is reproduced

Our main question was to investigate whether the ERPs to target and non-target trials can reliably be discriminated on the basis of single trial analysis. The experiment has clearly shown that the oddball effect is regularly reproduced. About 90% of all target and non-target trials of the test data could be correctly classified with formulas derived from the training data. The main indicator of the oddball effect was the P3 component which could be verified in all eight participants of the experiment in central and centro-parietal channels. With one common formula for all individuals, based on the grand average in an arbitrary channel around Cz, about 80% of all trials were correctly classified. Further improvements of the

classification were obtained either by using individual averages and optimal choices of sensors, or by using the grand averages of all 55 sensors.

Similar results were obtained in the classical papers by Squires & Donchin (1976) on auditory stimuli, Horst & Donchin (1979) on patterns in the upper and lower visual half-field, as well as in a number of more related recent studies. Precise classification rates can hardly be compared, due to the variety of experimental conditions and classification schemes. Bayliss & Ballard (2000) report 80 % and 85 % recognition rates for a computer game where virtual drivers had to stop at red traffic lights. Steuer, Grieszbach, Krause, & Schack (2002) correctly classified 85-94 % pattern comparisons with a combination of EEG and MEG. Similar recognition rates were obtained in BCI motor imagery experiments (Mensch, Werfel, & Seung, 2004; Blankertz et al., 2006, data set I; Blankertz et al., 2007). The P300 speller of Farwell & Donchin (1988) is a more complex BCI classification task where a given letter has to be recognized from a 6x6 letter array when rows and columns are repeatedly highlighted. In a corresponding task of the BCI competition 2003, a recognition rate of 100% was achieved with an extensive training data set (Kaper et al., 2004; Xu et al., 2004). A recent study of the P300 speller with eight subjects (Krusienski et al., 2006) compared different classification techniques. Six subjects obtained recognition rates above 90%, the other two below 80%. Fisher's linear discriminant analysis and stepwise discriminant analysis performed better than the more complicated support vector machines. The choice of eight electrodes as well as the methods of pre-processing were optimized before comparing methods, however (Krusienski, Sellers, McFarland, Vaughan, & Wolpaw, 2008). As in other studies for detecting the P3, down-sampling of the data to 20 Hz was a first step, so the method presented here does not fit into this scheme. Moreover, in all BCI studies some intentional effort was required from the participants.

### Grand average and individual classification

Classification by grand average templates is motivated by the desire to understand general rules of brain response which do not depend on the person. There is also a statistical advantage: a classification based on the large training set of all participants, with 88 target and 512 non-target trials, should be more stable. The P3 component was present in each individual, with differing shape, amplitude and latency (Figure 5), and with the maximum strength in different channels. Subject 7 was a borderline case with weak and very late P3. Nevertheless, the grand average in one central channel was sufficiently similar to the single trials of all individuals to give a classification with less than 20% error (Table 1). There were 6 centro-parietal sensors where the AUC of the classification was larger than 0.88, and 8 other channels with  $AUC > 0.85$  (Figure 1). When all 55 sensors were taken together, the grand average classification with an overall ranking over the results of the channels gave a classification error of only 11.7% (Table 5). This seems surprising since individual differences were completely neglected in this classification. However, taking all sensors together may be considered as a spatial filter which for every subject includes the regions of strongest P3.

Classification by individual averages is appropriate for BCI applications. It should be more accurate because it is adapted to the person. On the other hand, each of the 8 participants saw 23 target patterns and 127 non-targets, and only half of them could be used as training data. 11 target trials in the training set seem not sufficient to estimate reliably a lot of individual features. One untypical target trial, or a few non-target trials which in the beginning of the experiment may also show an oddball effect, could spoil the whole classification. In fact, individual classification by one central channel, or by all 55 channels, was not better than the classification by grand average. The reason is that the sensors with largest differences between targets and non-targets for different individuals are on quite different positions (Figure 7). So each particular sensor will be bad for at least some persons,

and taking all 55 sensors will include some weak channels where the individual average is more or less random while the grand average still makes sense.

Thus individual classification depends on individual choice of the sensors. With one optimal sensor, the classification error was only 10.4% (Table 2). With two sensors, the classification was still better (Table 4), but it seems questionable to optimize two sensor positions with such a small database. Nevertheless, it was interesting to see that optimal sensor pairs were never at neighbouring positions. Responses at neighbouring channels were too correlated to provide meaningful additional information for discrimination.

### Early oddball effects

One goal of our study was to see whether other ERP components beside P3 contribute to the discrimination of target and non-target trials. In the individual one-sensor classification, participant 4 achieved a 100% separation of all trials, for two choices of the sensor in the occipital region (Table 2). This led us to a study of the channel Oz (sensor #76) and its neighbours (Figures 8-10). For three participants, there were very clear differences of targets and non-targets in the time between 100 and 200 ms after stimulus onset. A one-sensor classification of single trials in occipital channels in the time period between 0 and 200 ms gave excellent results for four participants, but no discrimination for the other four individuals. For participants 1, 4, 5, and 8 it was possible to classify targets and non-targets with less than 15% error before 200 ms, without any use of the P3 component (Table 3). This could be an early visual mismatch negativity as described in Horvath, Czigler, Jacobsen, Maess, Schröger, & Winkler (2008). It should be noted that already Horst & Donchin (1979) in their study of patterns from the lower and upper visual half-field determined discriminating time points which were mostly before 200 ms. Recent studies that show the capacity of the brain for extremely fast image processing (Thorpe, Fize, & Marlot, 1996; Poolman, Frank, Luu, Pedersen, & Tucker, 2008) indicate that early occipital responses to visual stimuli

deserve more careful single-trial investigation. We note that also in the N2 domain, between 200 and 300 ms, we found individual oddball effects, as demonstrated for participant 6 in Figure 2. These observations have to be checked with a larger data set.

### Methodological questions

One can ask why we performed no pre-processing. Actually, we first worked with gentle filtering and got almost the same results as reported here. Then we were eager to see whether such kind of pre-processing is really necessary. It turned out that for classification, filters were not necessary, and we can live even with the 50 Hz interference. It seems that simplified discriminant analysis involves some kind of low pass characteristics when it uses slowly varying averages as templates. As far as the P3 is concerned, our study confirms that the common low-pass filters are appropriate. The early oddball effects might have become obscured by such filtering, however. The data indicated even more individual oddball responses within time windows of length smaller 50 ms, but the size of our database and the sample frequency of 500 Hz were not sufficient to draw conclusions. Experiences without pre-processing could be helpful when it comes to the study of finer details in single trials where filtering really matters.

The advantage of our method is its transparency and simplicity. It is clear why it works, and when it does not work. It can be useful as a screening tool, or for finding an initial configuration in BCI training. Moreover, instead of the original data, the method can use other templates, like differences of channels, spectrograms or correlograms. In online applications, template functions can be updated with almost no effort so that the method may be included as a small procedure in more complex software packages.

However, the method has severe limitations since there are no parameters to adapt to changing situations. Our participant 1 had a P3 latency which almost continuously increased with each target trial. The result was that the template of the training data did not at all fit the

test data. In such cases, low-pass filtering can be an improvement since it produces a more slowly varying template function. In contrast, the Oz response between 100 and 200 ms of participant 1 was detected very well (Figure 2).

In our experiment, healthy young men were kept in a quiet environment and were shown a slow sequence of only two different patterns. It is not clear whether target and non-target trials remain recognizable when further patterns come in, when subjects lack concentration or have to respond under the influence of stress or disturbance. Perhaps it will be possible to measure confusion, fatigue, and learning effects on a trial-by-trial basis. An important question is whether spatio-temporal patterns of response will remain stable over time, so that repetitions of the experiment lead to similar results. Another problem is to find the patterns in the ongoing EEG, without knowing at which time stimuli were presented. The ultimate goal would be to study brain activity directly under everyday conditions, but there is still much work to be done before this goal will be a reality.

## References

- Armitage, P., & Berry, G. (1994). *Statistical methods in medical research*. Blackwell Scientific Publications.
- Azizian, A., Freitas, A.L., Parvaz, M.A. & Squires, N.K. (2006). Beware misleading cues: perceptual similarity modulates the N2/P3 complex. *Psychophysiology*, 43, 253-260.
- Bayliss, J.D., & Ballard, D.H. (2000). Single trial P3 recognition in a virtual environment. *IEEE Transactions on Rehabilitation Engineering*, 8, 188-190.

- Birbaumer, N. (2006). Brain-computer interfaces in research: coming of age. *Clin. Neurophysiol.* 117, 479-483.
- Blankertz, B., Müller, K.-R., Krusienski, D., Schalk, G., Wolpaw, J.R., Schlögl, A., et al. (2006). The BCI competition III. *IEEE Transactions on Neural System and Rehabilitation Engineering*, 14, no. 2, 153-159.
- Blankertz, B., Dornhege, G., Krauledat, M., Müller, K.-R., & Curio, G. (2007). The non-invasive Berlin brain-computer interface: fast acquisition of effective performance in untrained subjects. *Neuroimage*, 37, 539-550.
- BramHayston, N.M., El-Deredy, W., & McGlone, F.P. (2005). Changes in neural complexity of the EEG during a visual oddball task. *Clinical Neurophysiology*, 116, 151-159.
- Debener, S., Ullsperger, M., Siegel, M., & Engel, A.K. (2007). Towards single-trial analysis in cognitive brain research. *Trends in Cognitive Science*, in press.
- Delorme, A., & Makeig, S. (2004). EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *J. Neuroscience Methods*, 134, 9-21.
- Donchin, E., Spencer, K.M., & Wijesinghe, R. (2000). The mental prosthesis: Assessing the speed of a P300-based brain-computer interface. *IEEE Trans. Rehab. Eng.* 8, 174-179.
- Dornhege, G., Blankertz, B., Curio, G., & Müller, K.-R. (2004). Boosting bit rates in noninvasive EEG single-trial classifications by feature combination and multiclass paradigms. *IEEE Transactions on Biomedical Engineering*, 51, 993-1002.
- Effern, A., Lehnertz, K., Grunwald, T., Fernández, G., David, P., & Elger, C.E. (2000). Time adaptive denoising of single trial event-related potentials in the wavelet domain. *Psychophysiology*, 37, 859-865.
- Farwell, L.A. & Donchin, E. (1988). Talking off the top of your head: Toward a mental prosthesis utilizing event-related brain potentials. *Electroenceph. Clin. Neurophysiol.* 70, 510-523.

- Fox, M.D., Snyder, A.Z., Vincent, J.L., & Raichle, M.E. (2007). Intrinsic fluctuations within cortical systems account for intertribal variability in human behaviour. *Neuron* 56, 171-184.
- Hanley, J.A. (1989). Receiver operating characteristic (ROC) methodology. The state of the art. *Critical Reviews in Diagnostic Imaging* 29 (3); 307-335.
- Horst, R.L. & Donchin, E. (1980). Beyond averaging II. Single-trial classification of exogenous event-related potentials using stepwise discriminant analysis. *Electroencephalography and Clinical Neurophysiology* 48, 113-126.
- Horvath, J., Czigler, I., Jacobsen, T., Maess, B., Schröger, E., & Winkler, I. (2008). MMN or no MMN: No magnitude of deviance effect on the MMN amplitude. *Psychophysiology* 45, 60-69.
- Huberty, C.J., (1994). *Applied discriminant analysis*. Wiley and Sons, New York.
- Jansen, B.H., Allam, A., Kota, P., Lachance, K., Osho, A., & Sundaresan, K. (2004). An exploratory study of factors affecting single trial P300 detection. *IEEE Transactions on Biomedical Engineering*, 51, 973-978.
- Jaśkowski, P. & Verleger, R. (2000). An evaluation of methods for single-trial estimation of P3 latency. *Psychophysiology*, 37, 153-162.
- Kaper, M., Meinicke, P., Grossekhoefer, U., Lingner, T., & Ritter, H. (2004). BCI competition 2003 – data set IIb: Support vector machines for the P300 speller paradigm. *IEEE Transactions on Biomedical Engineering*, 51, 1073-1076.
- Krusienski, D.J., Sellers, E.W., Cabestaing, F., Bayouth, S., McFarland, D.J., Vaughan T.M., & Wolpaw, J.R. (2006). A comparison of classification techniques for the P300 speller. *J. Neural Eng.* 3, 299-305.
- Krusienski, D.J., Sellers, E.W., McFarland, D.J., Vaughan T.M., & Wolpaw, J.R. (2008). Toward enhanced P300 speller performance. *J. Neuroscience Methods* 167, 15-21.

- Li, Y., Cao, X., & Gao, S. (2004). Classification of single-trial electroencephalogram during finger movement. *IEEE Transactions on Biomedical Engineering*, 51, 1019-1025.
- Makeig, S., Debener, S., Onton, J., & Delorme, A. (2004). Mining event-related brain dynamics. *Trends in Cognitive Science*, 8, no. 5, 204-210.
- McLachlan, G.J. (1992). *Discriminant analysis and statistical pattern recognition*. Wiley and Sons, New York.
- Mensh, B.D., Werfel, J., & Seung, H.S. (2004). BCI competition 2003 – data set 1a: combining gamma-band power with slow cortical potentials to improve single-trial classification of electroencephalographic signals. *IEEE Transactions on Biomedical Engineering*, 51, 1052-1056.
- Parra, L.C., Spence, C.D., Gerson, A.D., & Sajda P. (2005). Recipes for the linear analysis of EEG. *Neuroimage*, 28, 326-341.
- Poolman, P., Frank, R.M., Luu, P., Pederson, S.M., & Tucker, D.M. (2008). A single-trial analytic framework for EEG analysis and its application to target detection and classification, *Neuroimage*, in press.
- Quiroga, R.Q. & Garcia, H. (2003). Single-trial event-related potentials with wavelet denoising. *Clinical Neurophysiology*, 114, 376-390.
- Schinkel, S., Marwan, N., & Kurths, J. (2007). Order pattern recurrence plots in the analysis of ERP data. *Cognitive Neurodynamics*, in press.
- Schupp, H.T., Junghöfer, M., Weike, A., & Hamm, A. (2003). Attention and emotion: An ERP analysis of facilitated emotional stimulus processing. *NeuroReport*, 14, 1107-1110.
- Spencer, K.M. (2005). Averaging, detection, and classification of single-trial ERPs. In Handy T.C. (Ed.), *Event-related potentials* (pp. 209-227), Massachusetts Institute of Technology.

- Squires, K.C. & Donchin, E. (1976). Beyond averaging: the use of discriminant functions to recognize event-related potentials elicited by single auditory stimuli. *Electroencephalography and Clinical Neurophysiology* 41, 449-459.
- Squires, K.C., Wickens, C., Squires, N.K., & Donchin, E. (1976). The effect of stimulus sequence on the waveform of the cortical event-related potential. *Science* 193, 1142-1146.
- Steuer, D., Grieszbach, G., Krause, W., & Schack, B. (2002). Single-trial classification of elementary comparison processes on the basis of instantaneous EEG and MEG coherences. *Brain Topography*, 15 (2), 125-137.
- Stefanics, G., Jakob, A., Bernàth, L., Kellényi, L., & Hernádi, I. (2004). EEG early evoked gamma-band synchronization reflects object recognition in visual oddball tasks. *Brain Topography*, 16 (4), 261-264.
- Thorpe, S., Fize, D., & Marlot, C. (1996). Speed of processing in the human visual system. *Nature* 381, 520-522.
- Wang, Y., Zhang, Z., Li, Y., Gao, X., Gao, S., & Yang, F. (2004). BCI competition 2003 – data set IV: an algorithm based on CSSD and FDA for classifying single-trial EEG. *IEEE Transactions on Biomedical Engineering*, 51, 1081-1086.
- 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.
- Xu, N., Gao, X., Hong, Bo, Miao, X., Gao, S., & Yang, F. (2004). BCI competition 2003 – data set IIb: enhancing P300 wave detection using ICA-based subspace projections for BCI applications. *IEEE Transactions on Biomedical Engineering*, 51, 1067-1072.