

Statistical control of artifacts in dense array EEG/MEG studies

MARKUS JUNGHÖFER,^a THOMAS ELBERT,^a DON M. TUCKER,^{b,c} AND BRIGITTE ROCKSTROH^a

^aDepartment of Psychology, University of Konstanz, Germany

^bDepartment of Psychology, University of Oregon, Eugene, USA

^cElectrical Geodesics, Inc., Eugene, Oregon, USA

Abstract

With the advent of dense sensor arrays (64–256 channels) in electroencephalography and magnetoencephalography studies, the probability increases that some recording channels are contaminated by artifact. If all channels are required to be artifact free, the number of acceptable trials may be unacceptably low. Precise artifact screening is necessary for accurate spatial mapping, for current density measures, for source analysis, and for accurate temporal analysis based on single-trial methods. Precise screening presents a number of problems given the large datasets. We propose a procedure for statistical correction of artifacts in dense array studies (SCADS), which (1) detects individual channel artifacts using the recording reference, (2) detects global artifacts using the average reference, (3) replaces artifact-contaminated sensors with spherical interpolation statistically weighted on the basis of all sensors, and (4) computes the variance of the signal across trials to document the stability of the averaged waveform. Examples from 128-channel recordings and from numerical simulations illustrate the importance of careful artifact review in the avoidance of analysis errors.

Descriptors: Multichannel, EEG, MEG, Event-related potentials, Averaging, CSD

In recent years it has become evident that accurate recording of electrical or magnetic brain fields often requires adequate spatial sampling to avoid spatial aliasing (Tucker, Liotti, Potts, Russell, & Posner, 1994; Wikswo, Gevins, & Williamson, 1993). Dense sensor array electroencephalogram (EEG) systems (64–256 channels) are now used in many laboratories. Estimates of the spatial Nyquist frequency¹ of the human EEG and averaged event-related potential (ERP) suggest that an intersensor distance of 2–3 cm is required to achieve adequate spatial sampling (Spitzer, Cohen, Fabrikant, & Hallett, 1989; Srinivasan, Tucker, & Murias, 1998). With an even distribution of sensors across the head surface, a sampling density of less than 3 cm requires 128 sensors, and a density of less than 2 cm requires 256 sensors (Tucker, 1993). Similarly, magnetoencephalogram (MEG) systems have been scaled to whole-head coverage, and may now measure from 122 to 148 sensors at a time and twice as many in the near future. Both the

correctness of the scalp topography and the localization of neural generators depend on a sufficient spatial resolution (Junghöfer, Elbert, Leiderer, Berg, & Rockstroh, 1997; Tucker, 1993).

However, recording from dense arrays presents new problems for data acquisition. Although many creative theoretical approaches and some empirical studies have been advanced for the problem of electrical or magnetic source analysis, there has been little attention to the problems of statistical management of data quality in multichannel EEG and MEG systems. The results of topographical analysis as well as source analysis depend strongly on the quality of the data of each sensor that enters the analysis. The likelihood of errors due to noise or other artifacts increases with the number of sensors. If, in a given trial, artifacts are restricted to a few sensors, the trial still contains valuable information. However, simply removing the artifact-contaminated sensors from the average will introduce a specific class of errors. We propose a method for averaging multichannel event-related electromagnetic data that (1) optimizes data intake in high-resolution data acquisition, (2) minimizes errors of topography, current density analysis, or source localization due to missing sensors, and (3) provides statistical information about the data quality for each channel in the array.

ERP analysis typically begins with a three-dimensional matrix (trial \times sensor \times time) EEG_{*n,s,t*} with *n* denoting the number of trials or recording epochs, *s* the number of sensors, and *t* the number of time samples within a trial. Although we focus on electrical recordings in the present report, a similar structure is used for event-related MEG analysis. Data processing then comprises the following steps:

Research for this study was supported by the Deutsche Forschungsgemeinschaft.

We thank Ursula Lommen, Thomas Schäfer, and Stephan Schweinberger for providing the experimental data.

Address reprint requests to: Markus Junghöfer, Fachgruppe Psychologie, Universität Konstanz, P.O. Box D25, D-78457 Konstanz, Germany. E-mail: markus.junghoef@uni-konstanz.de.

¹The recorded potential across the scalp can be described by its spatial frequency, which reflects the variation in the signal across the head surface. As is true of sampling in time domain, the sampling rate has to be at least twice the highest-appearing frequency in the measurement. In the case of spatial frequency, use of a spatial sampling rate (sensor density) less than the Nyquist frequency leads to spatial aliasing.

- First, the influence of extraneous noise resulting from movement or sensor (electrode) artifacts is controlled by rejecting epochs with large amplitudes. A criterion is set such that within a given epoch n and for a given sensor s the range of data points $\text{EEG}_{n,s,t}$ for all time points t does not exceed a predefined absolute amplitude (for the EEG, for instance, a range of $100 \mu\text{V}$ is suggested; Elbert, Lutzenberger, Rockstroh, & Birbaumer, 1985). In case of violation of this requirement, the data recorded from a particular sensor will be declared as artifact contaminated for that particular trial.² If this problem recurs frequently in a given data set, the rejection strategy may be elaborated as follows: (a) If data of one or several identified sensors turn out to be of poor quality throughout a significant portion of the N trials, these sensors will be rejected completely from further analyses (from all trials). (b) Alternatively, an EEG_n epoch is rejected entirely if a significant portion of the S sensors turns out to be artifact contaminated.
- Second, artifacts from eye movements and blinks, as determined by periorbital electrooculogram (EOG) channels, are detected. Trials with artifacts may be rejected, or algorithms may be used to subtract the ocular artifact from the EEG channels (as described, for instance, by Berg & Scherg, 1991; Elbert et al., 1985).
- Third, the remaining trials are averaged for each sensor and the resulting averaged ERP is then analyzed further.

Although this procedure is commonly used, the selective elimination of artifactual trials or channels has significant drawbacks, particularly when applied to dense array data:

- First, if a sensor is noise contaminated in some but not all trials, the experimenter has to decide whether the rejection of that particular sensor, or the rejection of the noisy trials, will be appropriate. Often this decision is based on a rule of thumb that is not tailored to the specific data set: For example, if more than 20% of the sensors on a trial are noisy reject the trial, otherwise reject the data from individual sensors. Both trial and individual sensor data rejections result in a loss of signal information, and both actions may introduce a bias into the results.
- Second, according to the “all or none” excessive amplitude rule, that is, that a given amplitude range should not be exceeded at any sensor during a trial, all trials for which this criterion is not met will be rejected irrespective of how many sensors are problematic. Furthermore, because they have different positions in relation to skull conductivity and brain sources, different EEG sensors have different EEG signal amplitudes. This will result in different background EEG amplitudes, depending on their distance from the reference sensor, and this background EEG is considered the “noise” in ERP averaging. Artifactual amplitudes thus summate with different EEG amplitudes for different channels.

²For the sake of clarity, the term *sensor* will be used from here on, referring to SQUID (superconducting quantum interference device) sensors in MEG and electrodes in EEG recordings. For EEG recordings, it is important to remember that each channel is comprised of the scalp potential fields assessed by two sensors, one typically considered to be the reference, and that there are no privileged reference sites on the head where the potential remains constant across time. In this discussion, we assume the recording is with a common reference (rather than different bipolar electrodes for each channel).

- Third, once averaging has been accomplished, no statistical information about the noise level for particular sensors, or about the set of measurements as a whole, is available. As a consequence, data sets with different noise levels are compared within one statistical procedure. The lack of noise information limits the power of inverse source modeling methods such as the least squares (Press, Flannery, Teukolsky, & Vetterling, 1986), chi-square, maximum-likelihood (Sikihara, Ogura, & Hotta, 1992), or minimum-norm methods (Hämäläinen & Ilmoniemi, 1984). All these techniques can make good use of information on noise heterogeneity.

The crudeness of artifact screening and signal averaging contrasts with the effort that is invested in further data analysis, such as MRI-constrained source modeling with realistic head models. Empirical research (Braun, Kaiser, Kinces, & Elbert, 1997) has shown that the accuracy of current source modeling is highly dependent on the noise level of the data.

Overview

Therefore, we propose the following method for statistical control of artifacts in dense array studies (SCADS). The analysis requires two passes at the data, the first with the data kept in the recording reference, and the second with the data transformed to the average reference.

The first pass detects and rejects artifactual channels, in the recording reference (e.g., vertex referenced), to avoid contamination of all channels by the artifacts when transforming EEG data to the average reference. The average reference is computed by subtracting the potential average across all S sensors at one point of time from each single sensor potential at this point of time. Therefore artifacts of single sensors will contaminate the average reference (and thus all other sensors) by a factor of $1/S$.³ Once this pass is complete, the average reference may be computed to allow accurate topographic mapping and topographic waveform plots. An accurate average reference is unique to dense array studies. It requires a minimum of 64 channels, distributed to the inferior head surface as well as the top of the head (Junghöfer, Elbert, Tucker, & Braun, 1999; Tucker et al., 1994).

Some EEG analysis methods, such as source localization procedures, do not require transformation to the average reference (because the reference site may be modeled explicitly). In these cases, or in case of MEG data (which does not require a reference), the first stage can be omitted as it will be repeated in the second pass.

In the second pass, based on the average reference, global artifacts may be more clearly identified because the reference bias has been removed. Individual artifactual sensors that were identified in the first pass may be interpolated and replaced to complete the dataset and avoid the biases introduced by missing data.

³Sensors with artifactually attenuated signals, such as #114 in Figure 1, would be hard to detect after the transformation to averaged reference, as they then would include the signal at the reference location (appropriately estimated as zero minus the average reference) in addition to their own small signal. If such a sensor is distant from the reference, as in the present example, it is easily detected as artifact contaminated because its signal distribution differs from those of the neighboring sensors. If, however, a sensor is positioned adjacent to the reference, its signal deviation may go undetected, while still generating distortions of the spatial potential distribution.

Procedure

After outlining the steps of the analysis procedure, we will describe each in detail.

1. First Pass—Based on the Recording Reference:

- 1.1. Filter, thereby attenuating or removing artifacts in frequency bands that are not of interest for the analysis;
- 1.2. Construct editing data matrices;
- 1.3. Detect and reject consistently contaminated sensors (i.e., sensors exceeding a criterion of contamination throughout the experimental session);
- 1.4. Reject contaminated sensors in specific trials (to avoid the contamination of entire epochs when transforming to average reference);
- 1.5. Transform the edited data to average reference (to minimize the dependence of signal and noise amplitudes on the distance between the sensor and the chosen reference site).

2. Second Pass—Based on the Average Reference:

- 2.1. Construct editing data matrices (as step 1.2);
- 2.2. Determine and reject contaminated sensors in specific trials (based on the given editing matrices);
- 2.3. Reject contaminated trials;
- 2.4. Average the remaining epochs, using interpolated values for distinct contaminated sensors (to avoid a different number of averaged epochs for different sensors);
- 2.5. Compute the standard deviation across all trials included in the average.

1. First Pass Based on the Recording Reference

1.1. Filter

The decision to reject a given trial from the average should proceed after bandpass filtering within the frequency band of interest. For ERP studies, retention of near-DC variation is usually preferred because slow brain potential changes may be meaningful, and higher frequency information related to sensory potentials may also be important. It is therefore best to record with a broad bandpass (e.g., 0.01–100 Hz), then filter digitally, such as with a bandstop or notch to filter out the 50- or 60-Hz main power line.

The filter can be applied before segmentation of the ongoing stream of data into trials. If trials are filtered, any artifact produced by a fixed filter length must be minimized or subtracted from the beginning of the trial. Stimulus artifacts, such as from an electrical stimulus, must be removed before filtering. Otherwise, digital filtering will temporally smear the artifact, making its removal more difficult.

1.2. Construct Editing Matrices

Editing data matrices are constructed to remove or correct sensors that are artifact contaminated. For this matrix construction, the maximum absolute value over time, the standard deviation over time and the maximum of the gradient of values over time (first derivative) are determined for every epoch. These three parameters display different sensitivities for specific artifacts. For instance, a

sensor that is noisy throughout the entire epoch may produce normal amplitude values, whereas the noise contamination would be apparent in the standard deviation. Furthermore, transient artifacts may become obvious only from examining the first derivative.

Three $N \times S$ data matrices M are produced with elements m_{ns} for the n th epoch or trial at the s th sensor. The elements m_{ns} of the first matrix contain the maximal absolute value of the sensor s and the trial n over time. The second matrix comprises the standard deviations across the trial time interval. The third matrix describes the maximal temporal gradient. If the time interval of interest does not correspond to the entire trial interval, for example, the analysis targets only the first part of a recorded epoch, the calculation of the elements m_{ns} should be based on the targeted time interval to avoid rejection of trials or sensors because of artifacts occurring in non-targeted time segments. Moreover, it might be necessary to exclude a distinct time segment with obvious stimulus-induced or other experiment-specific artifacts from this calculation to avoid rejection of trials or sensors just because of this specific artifact. The editing matrices thus allow a focused screening of the data for the unique conditions of the experiment.

Additional criteria and matrices can be created. Coherent biological noise, such as from alpha waves, often poses a problem for ERP source modeling. Therefore, alpha power might be specified in a further editing matrix in order to identify trials with large-amplitude alpha waves.

1.3. Detect and Reject Consistently Contaminated Sensors

The parameter (p) matrices computed in step 1.2 (absolute maximum, standard deviation, and first derivative) are used to determine contaminated sensors by creating a statistical measure of the degree and consistency of the artifactual values. A confidence coefficient is introduced to weight this measure. Medians are used to avoid the influence of extreme values.

We first calculate the corresponding boundary values for $Lim_{\pm}(p)$ for each parameter matrix p (Figure 1B):

$$Lim_{\pm}(p, \lambda) = med_S(med_N^s(p_{s,n}))$$

$$\pm \lambda_p \cdot \sqrt{\frac{\sum_{s=1}^S (med_N^s(p_{s,n}) - med_S(med_N^s(p_{s,n})))^2}{S}}$$

$$\Leftrightarrow Lim_{\pm}(p, \lambda) = \hat{a} \pm \lambda_p \cdot \sqrt{\frac{\sum_{s=1}^S (a_s - \hat{a})^2}{S}}$$

$p_{s,n}$: value of sensor s , trial n , and parameter p (absolute maximum across time; standard deviation across time; absolute maximum of first derivative across time)

med_N : median across all trials N ;

med_S : median across all sensors S ;

λ_p : confidence coefficient has to be chosen by the user.

Since $a = med_N^s(p_{s,n})$ is the median across all trials for a given sensor, then $\hat{a} = med_S(med_N^s(p_{s,n}))$ is the grand median across these S medians (a fixed value for each of the three parameter types p). Therefore, the root-part of the given equation

$\sqrt{\frac{\sum_{s=1}^S (a_s - \hat{a})^2}{S}}$ is similar to a standard deviation except that \hat{a} is

the median and not the mean across all S median values. The general form of the equation as a whole resembles a robust (because it is based on medians) confidence interval, with the confidence coefficient replaced by λ_p and the SE replaced by a median-based SE equivalent.

In computing the confidence intervals for each sensor, a physical property of the data recorded with respect to a common reference site must be considered. This property is illustrated in Figure 1A: the sensors close to the reference will measure only a

small potential difference from the reference site (because there is only a small amount of brain tissue generating a voltage difference between the index sensor site and the reference site). In the present example, the vertex sensor (Cz) served as reference. Figure 1A plots the amplitude (median of absolute maximum) at each sensor site as a function of its polar angle from the Cz reference, which was defined as zero (the pole). A larger polar angle means more brain tissue contributing to the potential difference, and thus a larger channel amplitude (sensor site minus reference site potential). This reference-dependence effect occurs with any (e.g., nose, mastoid, noncephalic) reference, and the effect can vary in a complex fashion between subjects and experimental conditions (Junghöfer et al., in press).

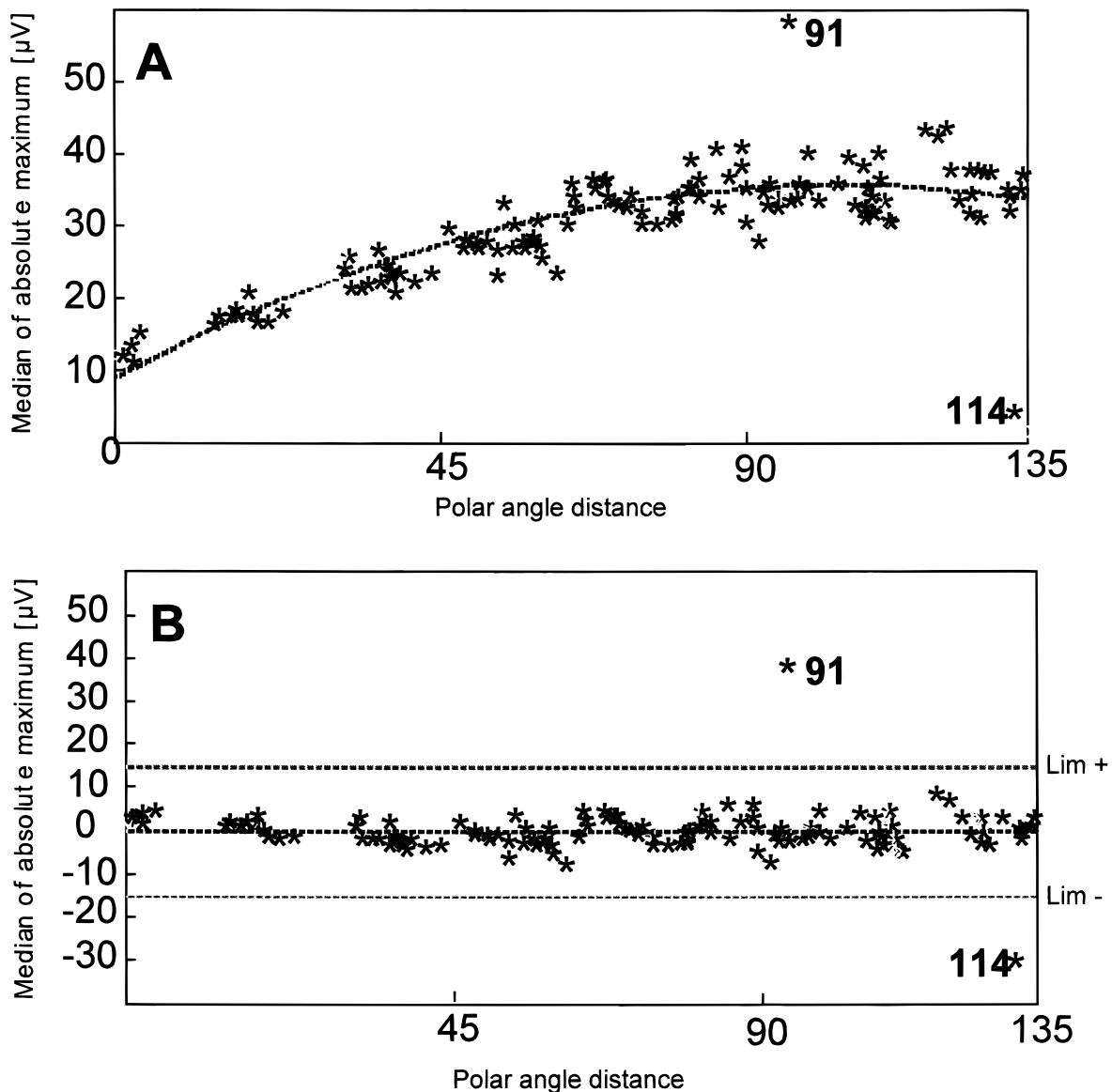


Figure 1. Removal of artifact contaminated channels based on SCADS step 1.3. (A) Each asterisk corresponds to an electrode, sorted according to their polar (theta) angle on the abscissa. The z -axis points from the center of the head through the reference electrode (in this case the vertex). The ordinate indicates the noise level computed as the median across trials of the maximum of absolute values within a sensor epoch. The dashed line shows the least square regression of second degree. (B) After subtraction of the regression function, data from sensors that lie above or below the calculated confidence interval with boundary values $Lim_{\pm}(p)$ (upper and lower dashed line) can easily be identified and removed.

To remove this effect, and thus equate the confidence intervals across the recording channels, the sensors are arranged according to their polar angular distance from Cz and the resulting function $a(s)$ is interpolated using a second degree polynomial least square regression. The original data $a(s)$ are corrected to $c(s)$ by the resulting spatial dependency $b(s)$, such that $c(s) = a(s) - b(s)$ (see Figure 1B). Now the final confidence interval for each parameter p can be calculated:

$$Lim_{\pm}(p, \lambda) = \hat{c} \pm \lambda_p \cdot \sqrt{\frac{\sum_{s=1}^S (c_s - \hat{c})^2}{S}}$$

If the spatial corrected median across all trials for a given sensor $c(s)$ exceeds the confidence intervals for any of the three parameters, it would be rejected from the analysis. Based on the analysis of a large assortment of data sets a p independent value of $\lambda = 2$ seems to be a good choice to reject consistently noisy sensors while keeping data from sensors that show large signals that may be just large amplitude ERPs.

Figure 1B illustrates the confidence interval for an analysis based on the absolute amplitude maximum (first parameter). Sensors 91 and 114 can be seen to be contaminated by artifacts throughout the measurement interval, and are thus candidates for rejection and replacement by interpolation.

Figure 2A illustrates a data set from which only sensor #82 (bottom of sensor array) was rejected completely.

1.4. Detect and Reject Contaminated Sensor Epochs

In the next step, the individual sensor epochs (i.e., a single sensor channel s for a single trial n) are removed if the value of one of the three parameters for the sensor epoch $p_{s,n}$ exceeds the following confidence interval (calculated across all trials for that sensor channel):

$$Lim_{\pm}(p_s) = med_N^s(p_{s,n})$$

$$\begin{aligned} & \pm \mu_p \cdot \sqrt{\frac{\sum_{n=1}^N (p_{s,n} - med_N(p_{s,n}))^2}{N}} \\ & = \hat{d} \pm \mu_p \cdot \sqrt{\frac{\sum_{n=1}^N (d - \hat{d})^2}{N}} \end{aligned}$$

med_N^s : median across all trials N of sensor s ;

μ_p : confidence coefficient has to be chosen by the user;

Again a p independent value of $\mu = 2$ is a good choice to select sensor epochs with excessive noise content.

1.5. Transform the Edited Data to Average Reference

Trials for which the resulting number of sensors with adequate data is less than a specified threshold are removed from further analysis. This threshold varies from experiment to experiment. For a clinical study of dementia it may be necessary to accept data with

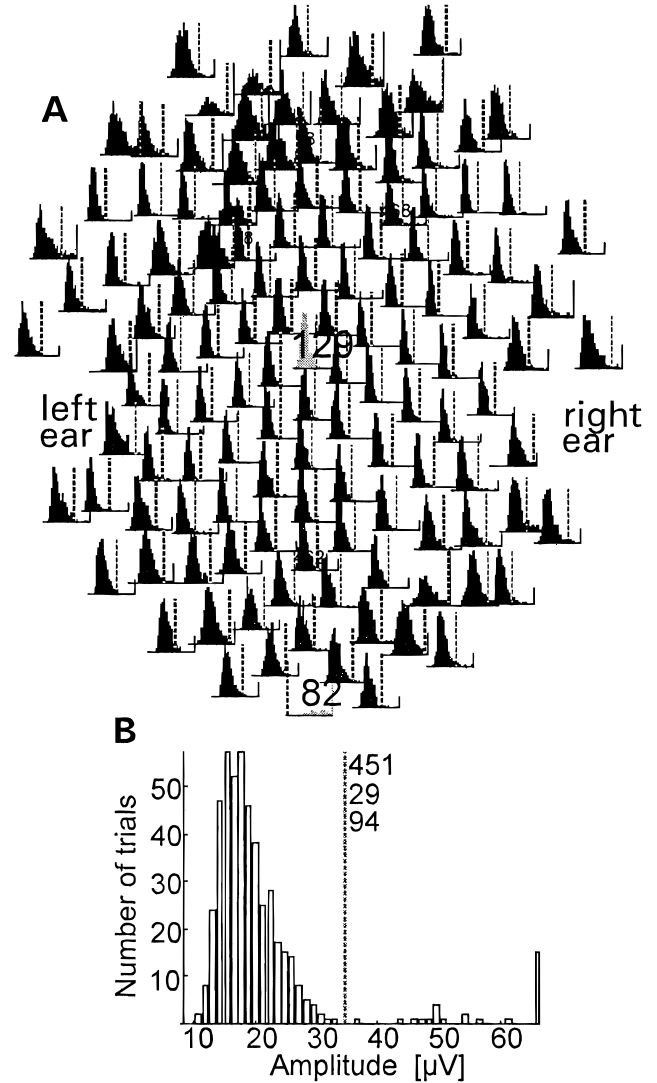


Figure 2. (A) Histogram of the maximum absolute values of single sensor epochs calculated for each sensor. The user must define interactively the upper boundary of the amplitude range that is accepted. A total of 480 trials was distributed among 12 conditions, 40 trials each. The upper limit was determined individually per channel (the distribution marked in gray indicates data from the acquisition reference Cz [129]). Note that data from sensor 82 was removed in step 1.3 of the data processing. The histograms for the standard deviation and for the largest gradient are typically similar to those of the largest amplitudes. (B) Frequency distribution magnified for a single channel. Of the 480 original trials, 29 are rejected and 451 (94%) are accepted for further analysis.

80 of 128 sensor channels with acceptable data; for a normal study with well-trained volunteers it may be possible to require 128 of 128 sensor channels. At this point, the accepted data are transformed to the average reference.

2. Second Pass Based on Average Reference

2.1 Construct the Average-Referenced Editing Matrices

If step 1.2 is repeated using the average reference data, artifacts produced by the reference sensor can be taken into account. Fur-

thermore, the dependency of signal and noise power on the distance from the reference is no longer relevant. All calculations up to this step are accomplished by an automatic computer algorithm without necessity of interactive control.

2.2 Detect and Reject Artifact-Contaminated Sensor Epochs

Identifying distinct sensors from particular trials should be based on visual inspection—in contrast to the automated detection of the first stage—for the following reasons: Because the noise may not be normally distributed and signal amplitude may not be constant across the trial, the contour of the frequency distribution of a sensor does not clearly indicate its noise. However, trials with abnormal amplitudes can be identified—as illustrated in Figure 2B for one sensor selected from the whole-head array in Figure 2A. Outliers in the distribution typically result from artifacts such as eye blinks, eye movements, or body movement. The probability of such artifacts, as well as those resulting from marked drifts, increases with the number of channels and with the duration of the data acquisition period.

Figure 2B illustrates a frequency distribution of amplitude maxima (amplitude histogram) for a given sensor. These are created by selecting the data point in each trial with the largest absolute deviation from zero. The critical upper limit for the amplitude maxima was chosen based on inspection of the histogram. All trials with values above this limit (indicated by the dashed line in Figure 2B) are removed—in the given example this number amounts to 29, so that 451 (94%) trials remain for further analysis.

The removal of individual sensor epochs is illustrated in Figure 2A. The maximum absolute amplitudes across trials are log normally distributed, that is, they are skewed toward lower amplitudes. In an interactive manner, the experimenter must determine the upper and lower boundaries between which the data are acceptable. If a distribution is extremely broad or exhibits multiple peaks, the sensor may be removed completely (although such sensors should have been already removed when applying step 1.2 of the suggested first pass data processing). Under certain circumstances, bimodal distributions for some sensors may be observed. In general, this bimodality could indicate that a sensor—for instance due to movement—has lost its skin contact in the course of the acquisition period. This disruption will produce an abrupt increase in noise amplitude. In such cases a good strategy is to define the boundary amplitudes just above the distribution with the lower signal amplitudes.⁴

Boundaries have to be determined for each of the three parameters (absolute maximum, largest standard deviation, largest gradient). We have observed that the standard deviation and temporal gradient distributions resemble the distribution of the maximum amplitudes illustrated in Figure 2. Sensor epochs for which one or more of the three measures is outside of the respective predefined boundaries are removed from further analysis, whereas all others are included in the average.

A computerized automatic processing of this outlined step would be desirable, particularly to standardize the described criteria for rejection. Automatic processing would also be complex, because

⁴Inspection of Figure 2A points to yet another noteworthy effect: sensors in the center of the sensor configuration show narrower distributions and smaller amplitudes than sensors in more inferior regions. This difference is a consequence of the inadequate sampling from the inferior head surfaces (face and neck) that does not meet the criteria for determining the average reference. The effects of this “polar average reference effect” are discussed in Junghöfer et al. (1998).

the distributions of the three editing matrices M depend on the unique characteristics of the signals.⁵

2.3 Reject Contaminated Trials

To identify trials contaminated by artifacts at many or all sensors, a histogram of the distribution of artifact-contaminated trials per sensor is constructed (Figure 3). On the basis of the histogram, the experimenter determines a lower boundary for trial rejection. This procedure improves the accuracy of the subsequent sensor epoch interpolation.⁶

With 118 selected as the minimum number of good sensors per trial, of all trials showing more than 10 bad sensors within one trial are rejected from further data analysis. In the present example, 90 (23%) of the 480 trials are rejected due to artifacts that affect at least 11 sensors. Even if the boundary for trial rejection is lowered—for instance to 110 good sensors per trial—only a few more trials would have been saved for further analysis, all of them with in-

⁵Though the authors recommend visual inspection to guide data screening, the following calculation provides reasonable default values for the upper threshold for each sensor s and parameter p : $Lim_+(p_s, v) =$

$$\max \left(\frac{\sqrt{N}}{\left(\frac{\left(\frac{Lim_+(p_s, v) - (p_s - \hat{p}_s)^2}{Lim_-(p_s)} \right)^\mu}{\sqrt{Lim_+(p_s, v) - Lim_-(p_s)}} \right)} \right) \quad (\text{In most cases } Lim_-(p_s) \text{ is the}$$

simply the minimum of p_s .) The root expression in the denominator describes the standard deviation of the given parameter p for one sensor s within the interval from the lower to the upper threshold. Because the signal-to-noise ratio (S/N) of the final average increases with the square root of the contributing number of trials N , this ratio increases with expansion of the threshold interval. However, the extension of the upper threshold also leads to a reduction in S/N, because dissimilarity of trials will increase. The inverse standard deviation function of the parameter distribution depending on the threshold interval can be used to quantify this S/N decrease. The aim of the overall function is, therefore, to determine the thresholds that maximize S/N depending on the number of trials with consideration of trial likelihood. The coefficient exponent μ weights the importance of trial likelihood in comparison to the number of trials and has to be chosen by the user. We found 0.5 to be a reasonable value to account for an appropriate trial variance.

In order to take into account the spatial scalp distribution of the final upper thresholds (i.e., neighboring sensors should show similar thresholds), we calculate a “threshold deviation value” for each sensor. This value is the sum of the “nonlinearly weighted” differences between each sensor threshold and all other thresholds. The weighting corresponds roughly to the inverse of the spatial distance (the spherical angle distance). However, using spherical spline interpolation this weighting is not linear.

Finally, the program suggests rejecting all sensors for which threshold deviations fall beyond two standard deviations of the median over all threshold sensor deviations. If based on a large number of trials (i.e., > 100), this procedure does not require additional manual editing and could be used to standardize sensor and trial rejection criteria.

⁶The simple choice of a minimum number of intact channels per trial (also used in SCADS step 1.5) does not take into account the spatial arrangement of the rejected sensors. If they are all in the same region, then interpolation will be inadequate for that region. To solve this problem we would like to have a parameter that describes the overall accuracy of interpolation. Using nearest neighbor interpolation, this parameter could be the averaged spherical angle distance from the index sensor (the one to be interpolated) to the neighboring sensors used for interpolation (see, e.g., Nunez, 1981). This is actually the average across all interpolation weights (decreasing linearly with increasing angular distance in the case of the nearest neighbor method). However, using spherical spline interpolation, the sensors are not linearly weighted. In this case an elegant solution is the interpolation of given spatial test functions (actually the spherical spline functions themselves based on the given configuration of intact and rejected channels using the same spherical spline functions chosen for the following average procedure). The accuracy of this interpolation depends not only on the number of rejected channels but also on their spatial relations.

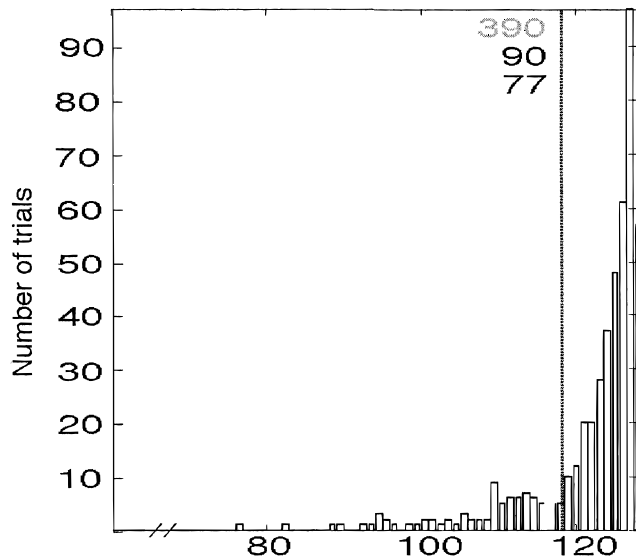


Figure 3. Histogram of a typical distribution of channels that result from SCADS step 2.3. With the given minimum number of good sensors per trial of 118, all trials—90 of the total 480 (23%)—showing 11 or more bad sensors within one trial are rejected from further data analysis.

complete sensor sets. In experiments with many trials per condition, it is desirable to choose a high boundary, because the loss of trials will then have a tolerable effect on the signal-to-noise ratio. If a distribution with a clear peak, like the one in Figure 3, is not obtained, the choice of the lower boundary may be based on the number of remaining trials.

2.4. Average the Artifact-Free Trials with Interpolated Values for the Artificial Sensor Sites

Before averaging the remaining trials (i.e., all trials with a sufficient number of intact sensors remaining at this step), all sensor epochs that have been identified above as artifact contaminated are replaced by interpolation. The interpolation is achieved with weighted spherical splines fit to the intact sensors. In contrast to nearest neighbor interpolation, spherical splines are fit to all remaining sensors, such that the influence of each sensor is weighted nonlinearly by the inverse of its distance to the missing sensor and by the specific noise level (i.e., more distant and/or noisier sensors are weighted less than closer and/or cleaner sensors). This estimation and reconstruction of rejected sensor epochs is of particular importance to maintaining the accuracy of the dense array representation. The complete dataset, including the estimated interpolated data, may be computed and stored if single-trial analyses (such as spectral, latency-adjusted averaging, or nonlinear analyses) are desirable. Otherwise, the estimation and reconstruction can proceed with the averaged ERP epochs.

Accurate surface potential distributions are particularly important for estimating radial current source density (CSD) from the two-dimensional (2D) Laplacian of the potential. Assuming no sources in the scalp, the 2D (surface) Laplacian is proportional to the radial current flow in or out of the skull at the computed point. The estimation of CSD based on spherical spline interpolation of EEG scalp potentials was first developed by Perrin, Pernier, Bertrand, and Echallier (1989). The calculation of the weighted spher-

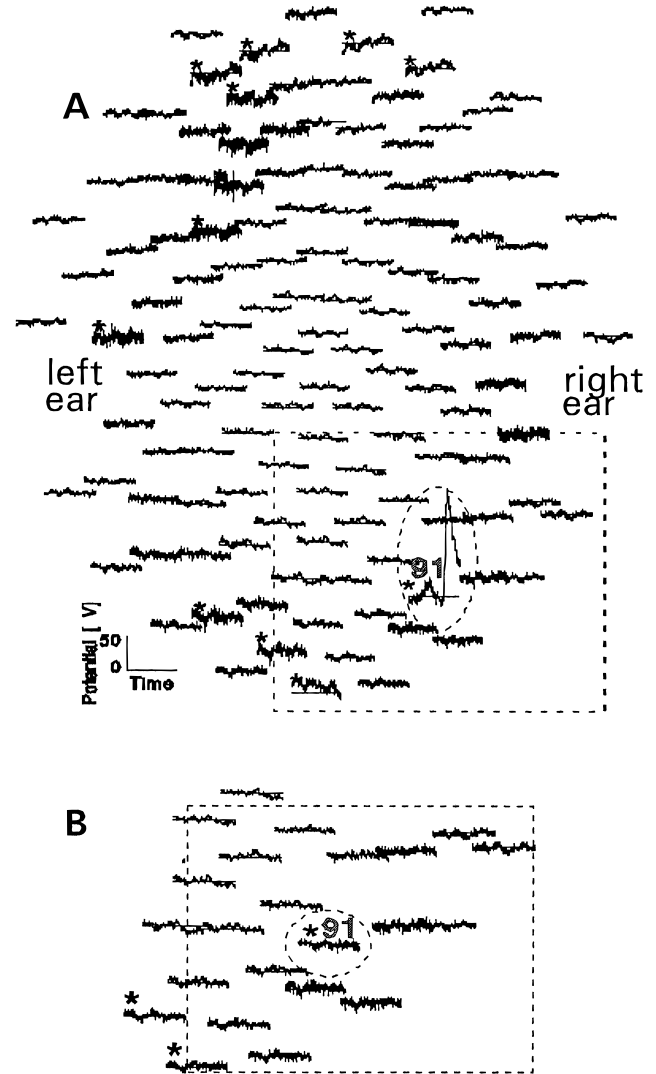


Figure 4. Results for a typical noise-weighted interpolation of single sensor epochs. The rejected sensors are marked with an asterisk. (A) The original data set and (B) the marked detail view of (A) after single trial correction. (Channel 91 exhibits the consequences of a reset, occurring during the trial, in which a sensor with a large electrochemical offset exceeds the input range of the amplifier, and the input capacitor of the AC coupling is reset to zero.)

ical spline interpolation and the algorithms for calculating both CSD and intracranial potential maps were described by Junghöfer et al. (1997). Using spherical splines, data for a rejected sensor may be interpolated accepting all valid sensors, with the contribution of each weighted according to its noise level. This interpolation allows estimates for sensor sites for which one or several neighbors are missing. In addition, the global noise level of the remaining sensor epochs is used to calculate the regularization or “smoothing” factor. As described by Whaba (1981) or Freedman (1981), larger values of the regularization or smoothing factor indicate a smaller contribution of a single sensor’s data relative to the other remaining sensors. A sufficient number of remaining sensors for accurate is guaranteed by the minimum threshold (as described in steps 1.5, and 2.3); otherwise the trial would have been rejected before this point. Each sensor is weighted according

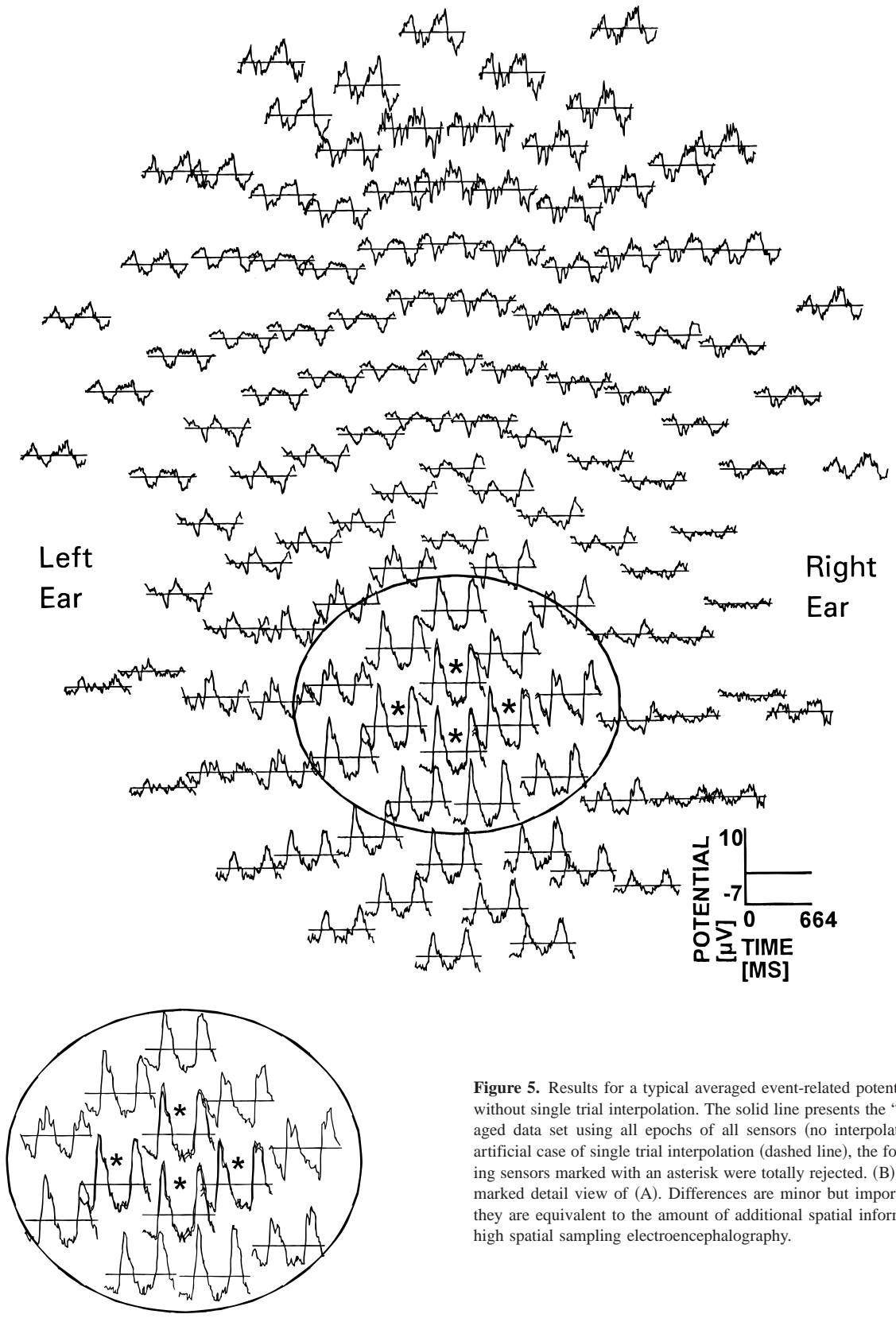


Figure 5. Results for a typical averaged event-related potential with and without single trial interpolation. The solid line presents the “clean” averaged data set using all epochs of all sensors (no interpolation). In the artificial case of single trial interpolation (dashed line), the four neighboring sensors marked with an asterisk were totally rejected. (B) presents the marked detail view of (A). Differences are minor but important because they are equivalent to the amount of additional spatial information using high spatial sampling electroencephalography.

to its signal-to-noise ratio, which is deduced from the histogram distributions in the editing matrices in step 2.2.

Figure 4 illustrates a typical interpolation for missing sensor epochs. The noise-contaminated sensor epochs (marked with an asterisk) were interpolated trial-by-trial. A trial before (A) and after (B, detail view) single trial interpolation was chosen to illustrate the effect of channel resetting (channel 91 in this example). This example illustrates how an artifact-contaminated sensor could influence all other sensors in the average reference transformation.

2.5. Calculate the Standard Deviation Across All Included Trials

Finally, the standard deviation is computed (for each time point) across all of the trials included in the average. In this matrix each element roughly describes the quality of each time sample in the average waveform. The standard deviation allows a comparison of signal quality between differing time points or differing sensors. This comparison could provide a cautionary note for further analysis.⁷ This caution is of crucial significance whenever the topographical distribution of averaged data is mapped using weighted spline interpolations. It also allows comparison between different data sets. Finally, this information on noise levels may help improve the accuracy of source modeling.

Finally, artifacts from subtle yet systematic eye movements may survive the artifact rejection process and thus contaminate the averaged ERP. Therefore, specific artifact correction procedures

⁷The calculation of the global power is such an example of the utility of the standard deviation across trials. The global power g could be weighted using the inverse standard deviation s (using the values of the matrix of calculated standard deviations across all trials described above):

$$g(t) = \frac{\sum_{j=1}^S x_j(t)^2 \cdot s_j(t)^{-1}}{S \cdot \sum_{j=1}^S s_j(t)^{-1}};$$

$x_j(t)$: Potential of the sensor j at time t .

S : Number of sensors

Compared to the common formula

$$g(t) = \frac{\sum_{j=1}^S x_j(t)^2}{S}$$

this calculation results in less noise, because sensors with low reliability are weighted less than sensors with high reliability.

should also be implemented, such as subtractive correction procedures (e.g., Elbert et al., 1985) or modeling approaches (Berg & Scherg, 1991).

To test the accuracy of interpolation against measured data, we selected four adjacent sensors with good data (no rejected trials) in a 200-trial visual ERP experiment. We then treated them as if they were artifactual, interpolated them according to the SCADS methodology described above, and compared the interpolations to the actual data. The overplot of interpolated vs. actual data are shown for the four sensors (marked with asterisks) in Figure 5A, and in detail in Figure 5B. Although the interpolation was not perfect, Figure 5B shows that the waveform was fairly well reconstructed even for this large missing region. The fact that the interpolation is only approximate indicates that sampling with lower sensor density (e.g., 64 or 32 channels) would not accurately reflect the spatial frequency of the scalp fields (Srinivasan et al., 1998).

The major advantage of the interpolation method of SCADS can be emphasized at this point: Averaging of trials without substituting the rejected sensor epochs by interpolated values will result in different numbers of trials per sensor site in the averages. This method would produce a temporal and spatial correlation of signal and noise, which would not be equally distributed across trials.

Conclusion

By interpolating the artifactual sensors in individual (raw EEG and MEG) trials of the ERP, the SCADS methodology maximizes the data yield in dense array ERP and MEG studies. Furthermore, SCADS avoids analysis artifacts caused by correlated signal and noise. However, this methodology requires both extensive computing and the attention of the experimenter, requiring on the order of 5–10 min per condition per recording session. This interactive processing might be automated if a large amount of data of the same kind are to be analyzed. However, the SCADS methodology clearly requires more experimenter time and computing resources than the conventional averaging method. This methodology may not be necessary for experiments such as from university subjects, most of whom can provide data with minimal artifacts. However, for experiments that are valuable and difficult to collect without artifacts, such as from children or clinical populations, the additional investment may be justified.

Another benefit of SCADS is the statistical information about data quality, which provides objective criteria for rejection or inclusion of the data from a subject. Finally, in the subsequent steps of surface field mapping and electrical and magnetic source analysis, the SCADS methodology may provide substantial information on the noise and the variance of the average as well as the average signal represented by the ERP.

REFERENCES

- Berg, P., & Scherg, M. (1991). Dipole models of eye movements and blinks. *Electroencephalography and Clinical Neurophysiology*, 79, 36–44.
- Braun, C., Kaiser, S., Kincses, W., & Elbert, T. (1997). Confidence interval of single dipole locations based on EEG data. *Brain Topography*, 10, 31–39.
- Elbert, T., Lutzenberger, W., Rockstroh, B., & Birbaumer, N. (1985). Removal of ocular artifacts from the EEG: A biophysical approach to the EOG. *Electroencephalography and Clinical Neurophysiology*, 60, 455–463.
- Freedman, W. (1981). On spherical spline interpolation and approximation. *Mathematical Methods in the Applied Sciences*, 3, 551–575.
- Hämäläinen, M., & Ilmoniemi, R. (1984). *Interpreting measured magnetic fields of the brain; estimates of current distribution*. Report TTK-F-A 559. Espoo, The Netherlands: Helsinki University of Technology.
- Junghöfer, M., Elbert, T., Leiderer, P., Berg, P., & Rockstroh, B. (1997). Mapping EEG-potentials on the surface of the brain: A strategy for uncovering cortical sources. *Brain Topography*, 9, 203–217.
- Junghöfer, M., Elbert, T., Tucker, D., & Braun, C. (1999). The polar effect of the average reference: A bias in estimating the head surface integral in EEG recording. *Electroencephalography and Clinical Neurophysiology* 110, 1149–1155.
- Nunez, P. (1981). *Electric fields of the brain: The neurophysics of EEG*. New York: Oxford University Press.

- Perrin, F., Pernier, J., Bertrand, O., & Echallier, J. (1989). Spherical splines for potential and current density mapping. *Electroencephalography and Clinical Neurophysiology*, 72, 184–187.
- Press, W., Flannery, B., Teukolsky, S., & Vetterling, W. (1986). *Numerical recipes: The art of scientific computing*. Cambridge, UK: Cambridge University Press.
- Sikihara, K., Ogura, Y., & Hotta, M. (1992). Maximum likelihood estimation of current dipole parameters for data obtained using multichannel magnetometer. *IEEE Transactions in Biomedical Engineering*, 39, 558–562.
- Spitzer, A., Cohen, L., Fabrikant, J., & Hallett, M. (1989). A method for determining optimal interelectrode spacing for cerebral topographic mapping. *Electroencephalography and Clinical Neurophysiology*, 72, 355–361.
- Srinivasan, R., Tucker, D., & Murias, M. (1998). Estimating the spatial Nyquist of the human EEG. *Behavioral Research Methods, Instruments, and Computers*, 30, 8–19.
- Tucker, D. (1993). Spatial sampling of head electrical fields: The geodesic sensor net. *Electroencephalography and Clinical Neurophysiology*, 87, 154–163.
- Tucker, D., Liotti, M., Potts, G., Russell, G., & Posner, M. (1994). Spatiotemporal analysis of brain electrical fields. *Human Brain Mapping*, 1, 134–152.
- Wahba, G. (1981). Spline interpolation and smoothing on the sphere. *SIAM, Journal of Scientific and Statistical Computing*, 2, 5–16.
- Wikswow, J., Gevins, A., & Williamson, S. (1993). The future of MEG and EEG. *Electroencephalography and Clinical Neurophysiology*, 87, 1–9.

(RECEIVED April 21, 1998; ACCEPTED August 20, 1999)