ORIGINALARBEIT

# Single-sweep-based methods to improve the quality of auditory brain stem responses Part II: Averaging methods

*Helmut Riedel, Michael Granzow, Birger Kollmeier*
AG Medizinische Physik, Carl von Ossietzky Universität Oldenburg, D-26111 Oldenburg

**Abstract** *This study is a systematic evaluation of the influence of different averaging methods on the quality of recorded auditory brain stem responses performed on a single sweep basis, i.e., on a post-hoc analysis of all unaveraged single epochs. The question of an optimal averaging method is addressed. Single sweeps of auditory brain stem responses were recorded for monaural and binaural click stimuli at levels of 20, 40, and 60 dB nHL. Recording sites were both mastoids and the forehead (Fz), with the vertex (Cz) serving as the common reference electrode. Five averaging methods were applied to the same set of data to compare their capability of reducing residual noise. In addition, the method of iterative averaging was introduced, which relies on an improved estimation of the noise of single epochs. Simulation allowed us to verify the quality of the different averaging methods as well as the estimators for signal, residual noise, and signal-to-noise ratio provided by these methods. Single-sweep-based estimation of residual noise and signal-to-noise ratio was shown to be superior to average-based estimation. Weighted averaging with one iteration step is the most favourable averaging method with regard to minimum residual noise and a valid estimation of the signal. For reliable quality estimation, single-sweep-based methods are preferable. Weighted averaging using iteration not only provides reliable signal and noise estimates, but also overcomes the arbitrariness of an artifact threshold.*

*Key words:* *auditory brain stem responses*
*averaging methods*
*single sweeps*
*residual noise section*

Corresponding author: Prof. Dr. med. Dr. rer. nat. Birger Kollmeier
AG Medizinische Physik
Universität Oldenburg
D-26111 Oldenburg
Phone +49 441 7985466
Fax +49 441 7983698
E-mail: birger.kollmeier@uni-oldenburg.de

# Einzelepochenbasierte Methoden zur Verbesserung der Qualität früher akustisch evozierter Potentiale Part II: Mittelungsmethoden

*Helmut Riedel, Michael Granzow, Birger Kollmeier*
AG Medizinische Physik, Carl von Ossietzky Universität Oldenburg, D-26111 Oldenburg

**Zusammenfassung** *In dieser Studie wird der Einfluss verschiedener Mittelungsmethoden auf die Qualität früher akustisch evozierter Potenziale untersucht, die auf Einzelepochenbasis geschätzt wird. Es wird die Frage nach einer optimalen Mittelungsmethode untersucht. Einzelepochen früher akustisch evozierter Potenziale wurden für monaural und binaural präsentierte Clicks bei den Pegeln 20, 40 und 60 dB nHL aufgezeichnet. Die Positionen der aktiven Elektroden waren beide Mastoide und die Stirn (Fz), die Referenzelektrode wurde am Vertex (Cz) geklebt. Fünf Mittelungsmethoden werden auf den gleichen Datensatz angewendet, um ihr Vermögen zur Reduktion des Restrauschens zu vergleichen. Zusätzlich wird die Methode des iterierten Mittelns eingeführt, die auf einer verbesserten Schätzung der Rauschleistung der Einzelepochen beruht. Eine Simulation erlaubt eine Überprüfung der Qualität der verschiedenen Mittelungsmethoden und der zugehörigen Schätzer für Signal, Restrauschen und Signal-Rausch-Verhältnis. Einzelepochenbasierte Schätzung des Restrauschens ist der mittelwertbasierten Schätzung überlegen. Iteriertes, gewichtetes Mitteln ist die beste Methode im Hinblick auf minimales Restrauschen und zuverlässige Schätzung des Signals. Für eine verlässliche Qualitätsschätzung sind einzelepochenbasierte Methoden vorzuziehen. Gewichtetes Mitteln mit Iteration bietet nicht nur zuverlässige Signal- und Rauschschätzungen, sondern überwindet auch die Beliebigkeit einer Artefaktschranke.*

*Schlüsselwörter:* *frühe akustische evozierte Potenziale*
*Mittelungsmethoden*
*Einzelepochen*
*Restrauschen*

## 1 Introduction

In the last two decades, recording of auditory brain stem responses (ABR) has developed into a standard method of performing differential diagnostics for the auditory pathway. Hence, high quality requirements have to be fulfilled within a minimum time to measure the responses. In the clinical routine, artifact rejection is the common technique used to increase the signal-to-noise ratio (SNR) of the measurement (*Gevins* and *Rémond* 1987, chapter 5).

The quality of ABR measurements is often only assessed by visual inspection of the responses averaged into two buffers. SNR estimates are provided using the sum and the difference of two averages as the signal and noise estimates, respectively (*Schimmel* 1967; *Wong* and *Bickford* 1980). However, estimation of the residual noise can be improved by calculating the so-called single-point variance, i.e., variance across sweeps for a fixed time point (*Elberling* and *Don* 1984; *Don* et al. 1984). Recently, *Cebulla* et al. (2000), using Monte-Carlo simulations, showed that the estimate of residual noise can be further improved by calculating the single-point variance for every time sample and averaging over time if the number of sweeps is low.

In this paper, the quality of ABR is investigated when not only one or two averages, but all single epochs[1] of a recording are available. Although such methods imply higher computational costs, they provide the possibility of using optimized post-hoc criteria to decide whether each individual epoch should be included or excluded from the average and what kind of weighting should be used. With the increased availability of high computing capacities, we do not regard computational costs or online realization as an issue here.

Two methods are typically used to improve the SNR of the recorded signals: filtering and averaging. Digital linear-phase filters are investigated in a companion paper (*Granzow* et al. 2001). It is demonstrated there that the estimation of residual noise based on single sweeps is superior to an estimation based on two averages. In the present study we will experimentally show that this also holds true for all the different averaging methods.

In addition to the commonly-used averaging method involving an artifact criterion, alternative techniques are investigated: sorted averaging (*Mühler* and *von Specht* 1997; *Mühler* and *von Specht* 1999), weighted averaging (*Hoke* et al. 1984; *Lütkenhöner* et al. 1985) and averaging by means of Bayesian inference (*Elberling* and *Wahlgreen* 1985), which we refer to as block-weighted averaging.

Using an artifact threshold has one major drawback: the threshold must be known prior to measurement. In actual practice, researchers and clinicians rely on their experience and choose a value that has proven reasonable in the past. However, since background noise varies significantly between subjects, it is not always possible (and certainly not reasonable) to use a fixed artifact threshold. The researcher will notice after a while whether the value chosen is adequate for the subject under study and will adjust it accordingly. Clearly, this is a very unsatisfactory situation, as the value of the artifact threshold is arbitrary and the result is not reproducible.

*Mühler* and *von Specht* (1999) have suggested the method of sorted averaging on a single sweep basis so as to determine the sweeps entering an average *a posteriori*. Before averaging, the recorded epochs are sorted according to their (estimated) contamination by noise. Only sweeps containing less than a certain degree of noise are included in the average.

The method of weighted averaging (*Hoke* et al. 1984; *Lütkenhöner* et al. 1985) is an extension of the former technique and allows the assignment of continuous positive weightings to individual epochs. The weightings are chosen according to the extent of contamination of the sweeps by noise. This could be useful in cases where the EEG background noise is not stationary, which can arise, for example, when the subject moves or his muscular activity changes.

The scheme of block-weighted averaging (*Elberling* and *Wahlgreen* 1985) tries to circumvent the problem of estimating the noise power of a single sweep by forming blocks of sweeps. From these blocks a more accurate noise power estimate can be obtained. This is important since a good estimate of the noise power of a single epoch (or a block of epochs) is essential for the weighted averaging schemes.

The goal of an improved noise estimate led to the new technique of iterative averaging, which is introduced and assessed in this paper and compared to the techniques mentioned so far. The idea is to subtract the current signal estimate from each epoch in order to estimate the respective noise component. This estimate is used in the subsequent iteration step for a weighted average and results in a more accurate signal estimate.

Finally, the properties of the investigated averaging methods and the SNR estimates are validated by a simulation study using a known signal and recordings of noise in the no-stimulus condition.

[1] We use the terms »sweep« and »epoch« interchangeably.

## 2 Methods

### 2.1 Subjects, stimuli, and recordings

Nine male subjects aged from 25 to 35 years participated voluntarily in this study. They were clinically classified as normal hearing and had no history of audiological or neurological problems. ABR were recorded from the left mastoid (M1), the right mastoid (M2), and the forehead (Fz) with respect to the common reference electrode at the vertex (Cz). Responses to monaural left, monaural right, and binaural stimulation at levels of 20, 40, and 60 dB normal hearing level were recorded for all subjects. For the simulation study, no-stimulus recordings were also made. For every stimulus condition $J$ = 10.000 individual sweeps were collected and stored to hard disk. The experimental setup and recording procedure is described in detail the companion paper (*Granzow* et al. 2001).

### 2.2 Averaging

The processing of the raw data primarily comprises linear filtering and averaging. If no weighting or artifact rejection is applied, the order of filtering and averaging is interchangeable due to the linearity of both operations. Since a high DC value or drift of the epochs can thwart any meaningful weighting, all single epochs were filtered before a decision about exclusion or assignment of weightings was made. An FIR bandpass-filter with 200 taps – designed with the window design method using a Hamming window – with corner frequencies 50 and 1500 Hz was used (see *Granzow* et al. 2001).

The averaging methods considered here can commonly be expressed as yielding a signal estimate $s(t)$ by forming a weighted average of $J$ (filtered) epochs $x_j(t)$ where $t$ denotes the time:

$$s(t) = \frac{\sum_{j=1}^{J} w_j x_j(t)}{\sum_{j=1}^{J} w_j} \,. \qquad (1)$$

The averaging methods differ in the strategy of assigning the weightings $w_j$ to the epochs $x_j(t)$. The most simple average is obtained by setting $w_j$ = 1 for all epochs. We call this the conventional average $s_c(t)$:

$$s_c(t) = \frac{1}{J} \sum_{j=1}^{J} x_j(t) \,. \qquad (2)$$

#### 2.2.1 Averaging using an artifact criterion

According to this strategy, epochs $x_j(t)$ with a peak-to-peak voltage $A_j$ larger than a certain threshold value $A$ (which has to be specified in advance) are considered non-physiological and are excluded from the average ($w_j$ = 0). The remaining $J_a \leq J$ epochs enter the average with $w_j$ = 1.

$$s_a(t) = \frac{1}{J_a} \sum_{j=1}^{J_a} x_{j, A_j \leq A}(t) \,. \qquad (3)$$

$J_a$ and the average itself are critically dependent on the choice of the artifact threshold $A$.

#### 2.2.2 Sorted averaging

The idea of the sorted averaging method (*Mühler* and *von Specht* 1999) is to sort the sweeps according to their contamination with noise and to classify them into two groups: sweeps with a small amount of noise are included in the average ($w_j$ = 1), while sweeps with high noise values are excluded ($w_j$ = 0). The critical noise value, which separates accepted and rejected epochs, is derived from the following consideration: because the single sweep SNR is very low (-20 to -30 dB) in ABR recordings, one can approximate

$$x_j(t) = S(t) + N_j(t) \approx N_j(t) \,, \;\; j = 1 \ldots J \,. \qquad (4)$$

The capital letters $S$ and $N$ are used to denote »true« signal and noise quantities respectively, in contrast to estimates which are designated by lower case letters ($s$ and $n$). Eq. 4 states that the measured and filtered signal primarily consists of noise.

The power[2] $P$ of any discrete signal $x(t)$ of length $T$ is defined as

$$P(x) = \frac{1}{T} \sum_{t=1}^{T} (x(t))^2 \,. \qquad (5)$$

Now the epochs are sorted in order of increasing power $P(x_j(t)) \equiv P_j$. The noise value dividing accepted and rejected sweeps is determined by minimizing the power of the mean cumulative normalized noise

$$P_{\text{cum}}(J') = \frac{1}{J'(J'-1)} \sum_{j'=1}^{J'} P(N_{j'}) \approx \frac{1}{J'(J'-1)} \sum_{j'=1}^{J'} P_{j'} \,, \qquad (6)$$

where $j'$ stands for the index of the sorted epochs. If all $J'$ terms of the sum roughly have the same magnitude, the numerator will increase in proportion to $J'$, whereas the denominator increases with $J'^2$. Consequently, the sum decreases proportionally to $1/J'$. In the case of non-stationary noise, however, not all the terms in the sum in eq. (6) have the same magnitude. If the cumulative noise for the inclusion of one sweep after the other is computed, i.e., if $J'$ is increased, a minimum can be found for a certain number of sweeps. This only holds if the increase of the noise power caused by the inclusion of a given sweep outweighs the increase of the denominator caused by raising $J'$ by one.

[2] Power is the variance across time. We reserve the term »variance« for the statistical variance over the ensemble of epochs.

For most practical ABR measurements, such a minimum can in fact be found. Hence an optimal number $J_s$ of sweeps can be determined. The sorted average therefore becomes

$$s_s(t) = \frac{1}{J_s} \sum_{j'=1}^{J_s} x_{j'}(t) \; . \qquad (7)$$

### 2.2.3 Weighted averaging

*Hoke* et al (1984) have shown that the highest SNR is obtained if the inverse power of the noise of an epoch is assigned as weighting $w_j$ to sweep $x_j(t)$.

$$w_j = \frac{1}{P(N_j(t))} \approx \frac{1}{P_j} = \frac{T}{\sum_{t=1}^{T} x_j^2(t)} \; , \qquad (8)$$

where again the approximation of eq. (4) has been used. The weighted average over $J$ sweeps is then

$$s_w(t) = \frac{\sum_{j=1}^{J} \frac{x_j(t)}{P_j}}{\sum_{j=1}^{J} \frac{1}{P_j}} \; . \qquad (9)$$

Weighted averaging has the advantage that it is not necessary to set a somewhat arbitrary artifact threshold.

### 2.2.4 Block-weighted averaging

The block-weighted averaging technique, or method of Bayesian inference, was introduced to the field of ABR analysis by *Elberling* and *Wahlgreen* (1985). They showed that the power of the noise can better be estimated on the basis of a group or block of sweeps rather than from a single sweep. In our study we compared block sizes $\beta = 1, 2, 4, ..., 256$ with $\beta = 1$ representing the weighted average. A block of $\beta$ consecutive sweeps is averaged conventionally resulting in intermediate »sweeps« $x_{(b\beta)}(t)$:

$$x_{(b_\beta)}(t) = \frac{1}{\beta} \sum_{j=(b-1)\beta+1}^{b\beta} x_j(t) \; , \qquad (10)$$

with the block size $\beta$. The various blocks are identified by a subscript $b$, $b = 1 ... B$, and their respective average powers are

$$P_{(b_\beta)} = \frac{1}{\beta} \sum_{j=(b-1)\beta+1}^{b\beta} P_j \; . \qquad (11)$$

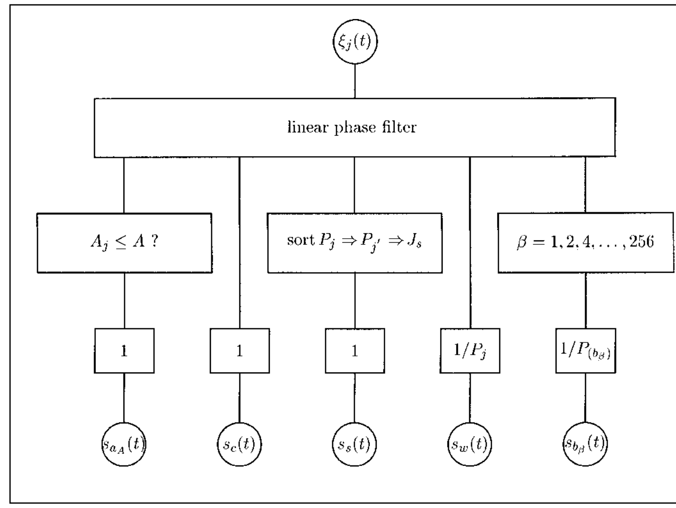



*Fig. 1: Flow diagram illustrating the averaging methods: The epochs $\xi_j(t)$ are passed through an FIR filter and then processed as indicated in the boxes in the second row ($A_j$ : peak-to-peak-voltage of sweep j, A artifact criterion, $P_j$ powers of the epochs, $P_{j'}$ sorted powers, $J_s$ number of sweeps entering the sorted average, $\beta$ block size of the block weighted averaging scheme). The weighting of the processed sweeps is shown in the third row ($P_j$ powers of the epochs, $P_{(b\beta)}$ powers of blocks of epochs). The resulting averages are depicted in the circles at the bottom of the figure ($s_{aA}(t)$ average using artifact rejection with threshold A, $s_c(t)$ conventional average, $s_s(t)$ sorted average, $s_w(t)$ weighted average, $s_{b\beta}(t)$ block weighted average with block size $\beta$.*

*Abb. 1: Flußdiagramm der Mittelungsmethoden: Die gemessenen Epochen $\xi_j(t)$ werden FIR-gefiltert und dann weiterverarbeitet wie in der zweiten Reihe angezeigt ($A_j$ : Maximalspannung (Spitze zu Spitze) in der Epoche j, A Artefaktschranke, $P_j$ Leistungen der Epochen, $P_{j'}$ sortierte Leistungen, $J_s$ Zahl der Epochen, die ins sortierte Mittel eingehen, $\beta$ Blockgröße bei der block-gewichteten Mittelungsmethode). Die Gewichtung der prozessierten Epochen zeigt die dritte Reihe ($P_j$ Leistungen der Epochen, $P_{(b\beta)}$ Leistungen der Blöcke aus Epochen). Die resultierenden Mittel sind unten in den Kreisen dargestellt: ($s_{aA}(t)$ Mittel unter Verwendung der Artefaktschranke A, $s_c(t)$ konventionelles Mittel, $s_s(t)$ sortiertes Mittel, $s_w(t)$ gewichtetes Mittel, $s_{b\beta}(t)$ block-gewichtetes Mittel mit der Blockgröße $\beta$.*

The inverse powers of the quantities $x_{(b\beta)}(t)$ serve as weightings for a weighted average of blocks of sweeps:

$$s_{b_\beta}(t) = \frac{\sum_{b=1}^{B} \frac{x_{(b_\beta)}(t)}{P_{(b_\beta)}}}{\sum_{b=1}^{B} \frac{1}{P_{(b_\beta)}}} . \qquad (12)$$

The number of epochs entering this average is $J_b = \beta * B$.

The block diagram in Fig. 1 gives an overview of the averaging methods used in this paper.

### 2.3 Estimation of the residual noise

The single-sweep-based estimate for the residual noise of the conventional average is the standard error over the epochs:

$$\sigma_c(t) = \sqrt{\frac{1}{J(J-1)} \sum_{j=1}^{J} (x_j(t) - s_c(t))^2} . \qquad (13)$$

For averaging using an artifact criterion and sorted averaging, $\sigma_a(t)$ and $\sigma_s(t)$ are defined similarly. The standard error of the weighted average is defined as

$$\sigma_w(t) = \sqrt{\frac{\sum_{j=1}^{J} \frac{1}{P_j}(x_j(t) - s_w(t))^2}{(J-1)\sum_{j=1}^{J} \frac{1}{P_j}}} . \qquad (14)$$

The residual noise estimate for the block-weighted averaging scheme is defined analogously by

$$\sigma_b(t) = \sqrt{\frac{\sum_{b=1}^{B} \frac{1}{P_{(b_\beta)}\beta} \sum_{j=(b-1)\beta+1}^{b\beta} (x_{(b_\beta)}(t) - s_b(t))^2}{(J_b - 1)\sum_{b=1}^{B} \frac{1}{P_{(b_\beta)}}}} . \qquad (15)$$

As in *Granzow* et al. (2001), the rms values of signal and noise estimates ($s$ and $\sigma$) are used for the data analysis. Their quotient $\gamma = s/\sigma$ serves as an SNR estimate based on single epochs. As proven in *Granzow* et al. (2001), single-sweep-based estimation of data quality is superior to average-based estimation. However, to compare both estimation methods, the average-based noise estimate $\sigma_{oe}(t)$ was also computed for all averaging methods. $\sigma_{oe}(t)$ is half the difference between two sub-averages. For sorted and block-weighted averaging, it is not possible to apply a recording technique using alternating buffers. Therefore, we split each raw epoch file into two files, one containing $J/2$ sweeps with odd, the other $J/2$ sweeps with even epoch numbers. Afterwards all averaging methods were applied to both files, but only the averages were used in the subsequent analysis. The average-based estimate of the residual noise was defined as $\sigma_{oe}$ = rms($\sigma_{oe}(t)$).

### 2.4 Iterative averaging

The iterative averaging technique was developed to avoid problems with weighted averaging, as described, e.g., by *Lütkenhöner* et al. (1985). The correct weighting of an epoch for the weighted average is the inverse power of its noise. However, with the approximation of eq. (4), the weighting is determined as the inverse power of the measured epoch that consists of signal and noise. This leads to an undesirable underestimation of the overall magnitude of the signal. With the approximation of eq. (4) an estimate of the noise in a single epoch is defined as

$$n_j^{(0)}(t) = x_j(t) . \qquad (16)$$

As indicated by the superscript in parentheses, we call this quantity the noise estimate to the order of zero.

Since an estimate of the noise in the single sweep is the basis for all averaging schemes except for the conventional average, an improvement of this estimate will affect the sorted, weighted, and block-weighted average as well as the average involving an artifact criterion. We only present the equations for the case of weighted averaging here, because it is straightforward to apply the following considerations to the other methods:

The residual noise of the weighted average to the order zero[3] is given by

$$\sigma_w^{(0)}(t) = \sqrt{\frac{\sum_{j=1}^{J} \frac{1}{P_j}(x_j(t))^2}{(J-1)\sum_{j=1}^{J} \frac{1}{P_j}}} . \qquad (17)$$

The signal estimate corresponding to eq. (16) is

$$s^{(0)}(t) \equiv 0 \qquad (18)$$

In the computation of the weighted average, this approximation was used to determine the weightings (cf. eq. (8)). The result of this computation, however, is a better signal estimate $s_w(t) = s_w^{(1)}(t)$ of the first order (cf. eq. (9)), which in turn can be used to improve the noise estimate of the single sweeps:

$$n_{j,w}^{(1)}(t) = x_j(t) - s_w^{(1)}(t) \qquad (19)$$

[3] This quantity is a noise estimate of the measurement rather than a standard error.

If we use the inverse power of the noise estimate $n^{(1)}_{j,w}(t)$ as the new weighting

$$w_j^{(1)} = \frac{1}{P(n_{j,w}^{(1)}(t))} = \frac{1}{\sum_{t=1}^{T} (x_j(t) - s_w^{(1)}(t))^2}\,, \qquad (20)$$

we can now calculate an improved signal estimate, the weighted average of the second order

$$s_w^{(2)}(t) = \frac{\sum_{j=1}^{J} w_j^{(1)} x_j(t)}{\sum_{j=1}^{J} w_j^{(1)}}\,. \qquad (21)$$

The residual noise of the improved weighted average is defined analogously as

$$\sigma_w^{(2)}(t) = \sqrt{\frac{\sum_{j=1}^{J} w_j^{(1)} (x_j(t) - s_w^{(2)}(t))^2}{(J-1)\sum_{j=1}^{J} w_j^{(1)}}}\,. \qquad (22)$$

Of course it is possible to repeat the process by stating that, if $s^{(2)}_w(t)$ is a better signal estimate than $s^{(1)}_w(t)$, then $n^{(2)}_{j,w}(t) = x_j(t) - s^{(2)}_w(t)$ should also be a better noise estimate than $n^{(1)}_{j,w}(t)$. We therefore call this method iterative weighted averaging.

One can easily generalize the process of iteration in the case of block-weighted averaging. On the other hand, it is not so obvious that iteration can also be applied in the cases of averaging with artifact criteria and sorted averaging. However, if we recall that these methods can also be considered as weighted averaging schemes with only the weightings zero and one allowed, then improving the noise estimate of the single epochs will affect the result. Given the better noise estimate $n^{(1)}_{j,aA}(t) = x_j(t) - s^{(1)}_{j,aA}(t)$, one has to reinvestigate the artifact criterion $A$ for $n^{(1)}_{j,aA}(t)$ instead of the $x_j(t)$. The same holds true for sorted averaging of course. Only conventional averaging is not influenced by iteration, since no rejection or weighting takes place.

### 2.5 Simulations

The properties of the respective averaging methods can only be assessed on the basis of *estimates* of the »true« signal and noise components, rather than being directly based on the components themselves. Hence, the apparent advantage of one of the methods tested above in comparison to another might perhaps be due to an overly optimistic signal estimate or a too low residual noise estimate. To overcome this problem, we performed simulations using *a priori* known signal and noise components as follows: from each subject, sweeps were recorded without presenting a stimulus. To each of these sweeps a known signal $S(t)$ was added. (We chose the mean over subjects of one of the previously calculated averages.) All averaging methods described above were applied to these derived test signals. From the results the following questions can be addressed:

1. Which averaging method is superior to the others in terms of the best reconstruction $s(t)$ of the true signal $S(t)$?

2. Which averaging method yields the best estimation of the signal, the residual noise and the SNR of the average it produces, i.e., is most reliable in assessing its own performance?

The true residual noise $\Sigma$ is defined as the rms value of the difference between $s(t)$ and $S(t)$. Note that rms($s(t)$) = rms($S(t)$) does not imply $\Sigma = 0$ because the true and estimated signals may have the same rms value without being identical. Therefore, the quality of the averaging methods must be evaluated by computing $\Sigma$. However, the comparison of $s$ and $S$, $\sigma$ and $\Sigma$ as well as $\gamma$ and $\Gamma$, i.e., of the estimated and true quantities, provides the answers to the second question.

## 3. Results

### 3.1 Averaging methods

Fig. 2 gives an example of the different approaches underlying the single-sweep-based and average-based signal and noise estimates. Data from binaural stimulation at 60 dB nHL for one subject are shown. The time interval from zero to ten ms after stimulus onset was chosen to determine all quantities described in the following.

In the upper left graph of Fig. 2, the conventional average (10000 sweeps) $s_c(t)$ is depicted as a solid line, while the dotted lines refer to $s_c(t) \pm \sigma_c(t)$, i.e., signal estimate $\pm$ residual noise estimate. The upper right graph shows the same data averaged into two buffers in an alternating way. Both averages result from 5000 sweeps, the first from the sweeps with odd sweep numbers, the second from the sweeps with even sweep numbers.

In the lower left graph, the time dependent standard error $\sigma_c(t)$ and its rms value $\sigma_c$ are depicted. In the lower right graph the average-based estimate of the residual noise $\sigma_{oe}(t)$, equal to half the difference between the subaverages (see Sect. 2.3), and its rms value $\sigma_{oe}$ are shown.

While $\sigma_c(t)$ only shows a small variation over time, $\sigma_{oe}(t)$ exhibits large fluctuations. Note that the voltage scale for $\sigma_c(t)$ is 20 times smaller than for $\sigma_{oe}(t)$. Since $\sigma_{oe}(t)$ vanishes where the sub-averages intersect, it does not provide a realistic time course

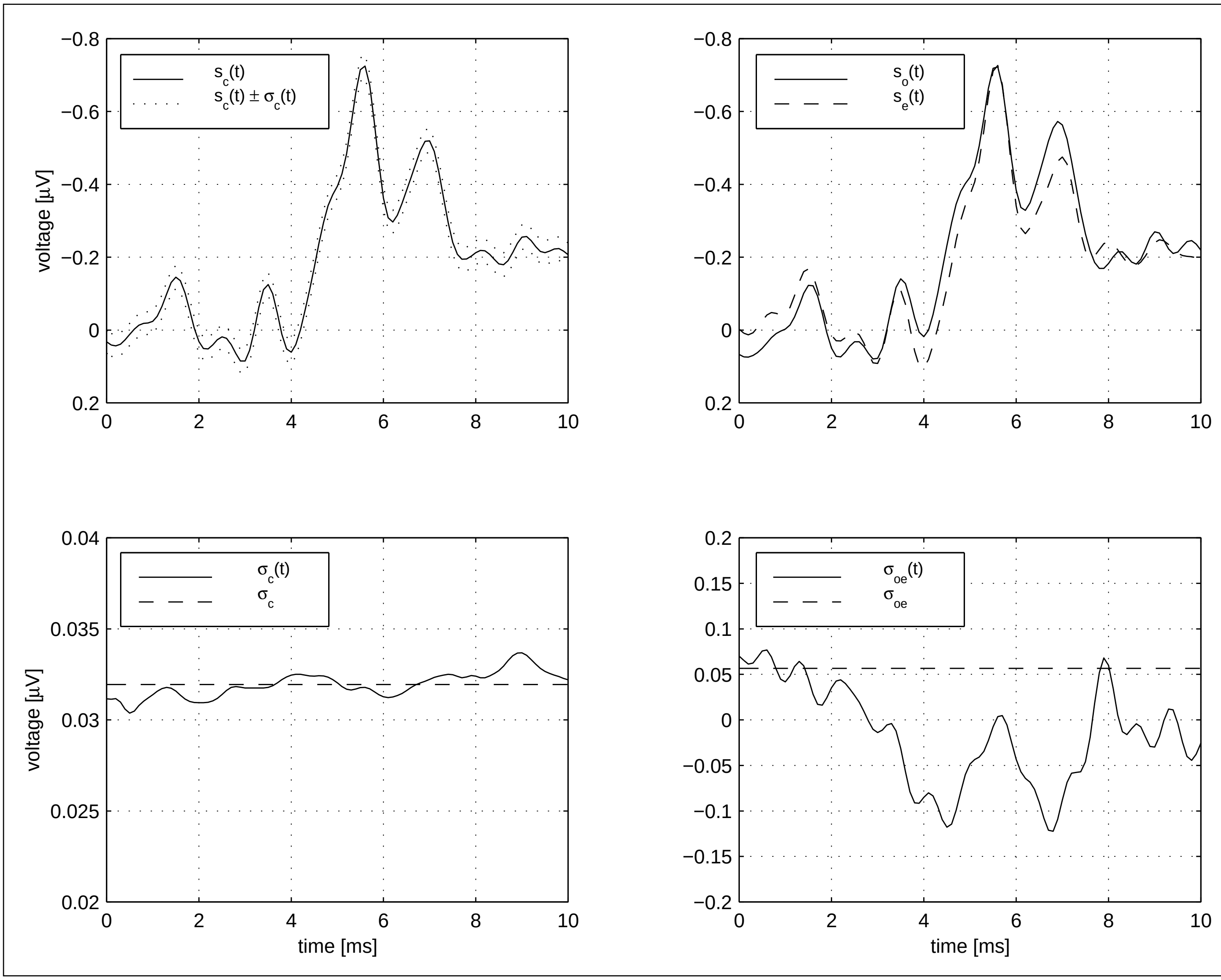



*Fig. 2: Comparison between single-sweep-based and average-based signal and noise estimates, example for binaural stimulation at 60 dB nHL for subject vk.* ***Top row****: Signal estimates.* ***Upper left graph****: Conventional average of 10000 epochs with one standard error.* ***Upper right graph****: Two sub-averages of 5000 epochs each.* ***Bottom row****: Corresponding noise estimates.* ***Lower left graph****: The single-sweep-based time dependent estimate of the standard error $\sigma_c(t)$ and its rms value $\sigma_c$.* ***Lower right graph****: The average-based time dependent noise estimate $\sigma_{oe}(t)$ and its rms value $\sigma_{oe}$.*

*Abb. 2: Vergleich der einzelepochenbasierten und mittelwertbasierten Signal- und Rauschschätzung, Beispiel für binaurale Stimulation bei 60 dB nHL für eine Versuchsperson (vk).* ***Oben****: Schätzungen der Signale.* ***Oben links****: Konventionelles Mittel aus 10000 Epochen mit einem Standardfehler.* ***Oben rechts****: Zwei Teilmittelwerte aus je 5000 Epochen.* ***Unten****: Schätzungen des entsprechenden Restrauschens.* ***Unten links****: Der zeitabhängige Standardfehler auf Basis von Einzelepochen $\sigma_c(t)$ und sein RMS-Wert $\sigma_c$.* ***Unten rechts****: Die zeitabhängige mittelwertbasierte Schätzung des Restrauschens $\sigma_{oe}(t)$ und ihr RMS-Wert $\sigma_{oe}$.*

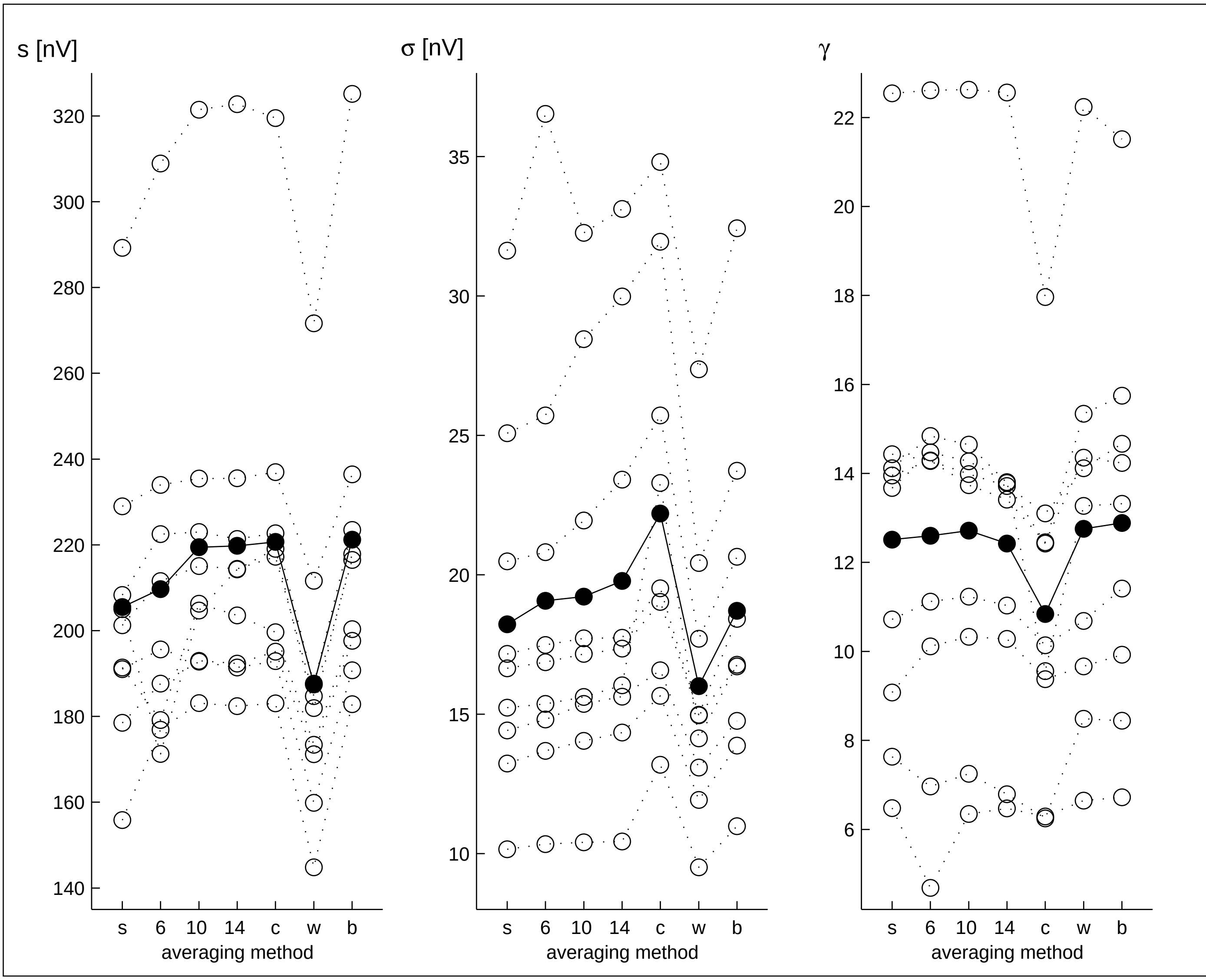



*Fig. 3: Signal rms estimate s (left), residual noise rms estimate σ (middle), and SNR estimate γ (right) for various averaging methods: s: sorted averaging, 6,10,14: different artifact thresholds, c: conventional averaging without artifact criterion, w: weighted average, b: blockweighted average with block size 256. Data from individual subjects are connected by dotted lines; solid lines with filled symbols represent the mean across subjects. Data from diotic stimulation at 60 dB normal hearing level for channel M2 (right mastoid versus vertex).*

*Abb. 3: Schätzungen des RMS-Werts des Signals s (links), des RMS-Werts des Restrauschens σ (Mitte) und des SNR γ (rechts) für verschiedene Mittelungsmethoden: s: sortiertes Mitteln, 6,10,14: verschiedene Artefaktschranken, c: konventionelles Mitteln ohne Artefaktschranke, w: gewichtetes Mitteln, b: block-gewichtetes Mitteln mit der Blockgröße 256. Daten der einzelnen Versuchspersonen sind durch gestrichelte Linien verbunden, gefüllte Symbole und durchgezogene Linien kennzeichnen Mittelwerte über Versuchspersonen. Daten für diotische Stimulation bei 60 dB nHL, Kanal M2 (rechtes Mastoid gegen Vertex).*

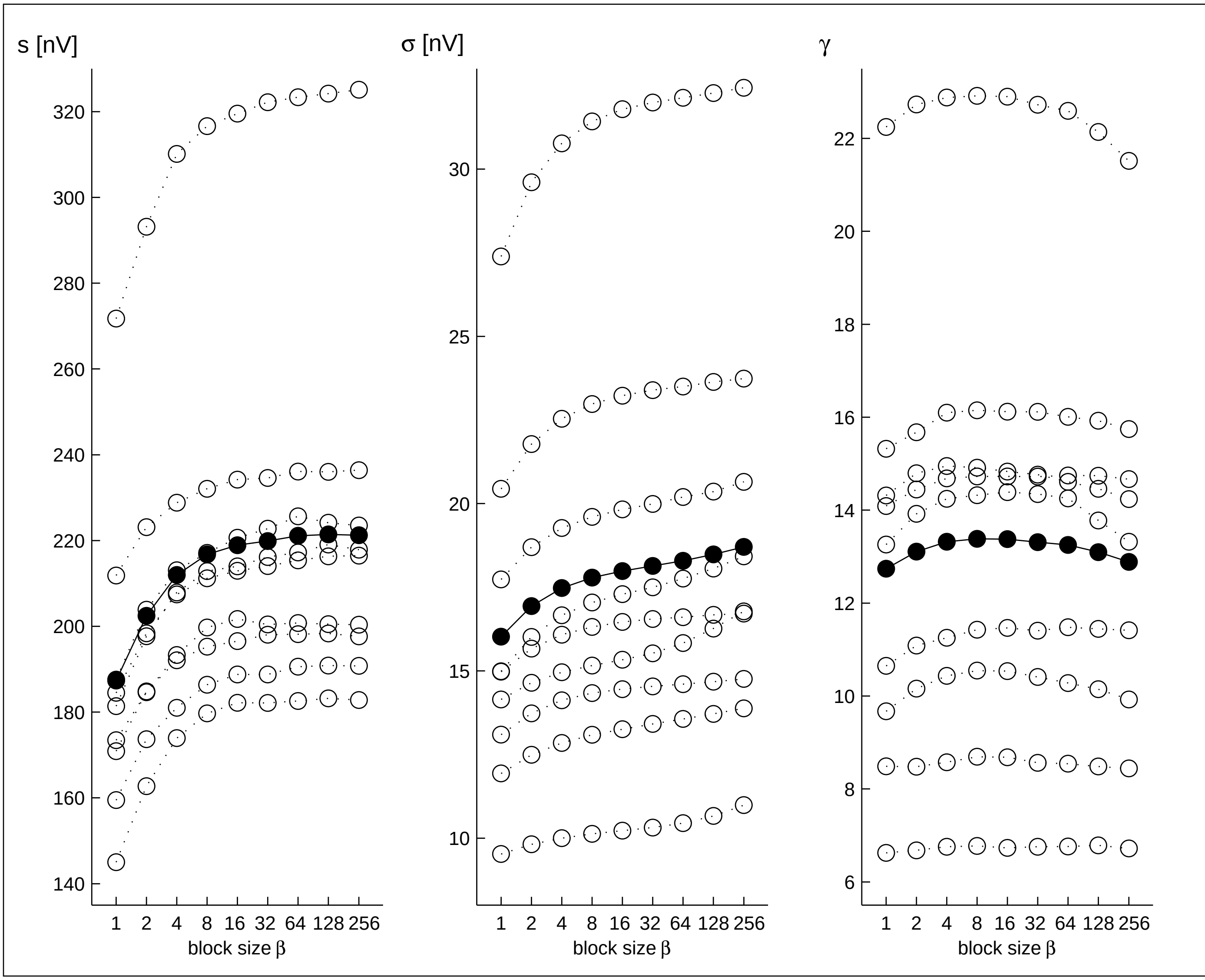



*Fig. 4: Signal rms estimate s (left), residual noise rms estimate $\sigma$ (middle), and SNR estimate γ (right) as a function of the block size $\beta$ for the block weighted averaging method. Data from individual subjects are connected by dotted lines; solid lines with filled symbols represent the mean across subjects. Data from diotic stimulation at 60 dB normal hearing level for channel M2 (right mastoid versus vertex).*

*Abb. 4: Schätzungen des RMS-Werts des Signals s (links), des RMS-Werts des Restrauschens $\sigma$ (Mitte) und des SNR γ (rechts) als Funktion der Blockgröße $\beta$ für das block-gewichtete Mittelungsverfahren. Daten der einzelnen Versuchspersonen sind durch gestrichelte Linien verbunden, gefüllte Symbole und durchgezogene Linien kennzeichnen Mittelwerte über Versuchspersonen. Daten für diotische Stimulation bei 60 dB nHL, Kanal M2 (rechtes Mastoid gegen Vertex).*

for the noise. Therefore, only the rms value $\sigma_{oe}$ can be considered as a meaningful noise estimate. Thus, the higher accuracy of $\sigma_c$ in comparison to $\sigma_{oe}$, as theoretically derived in the companion paper (*Granzow* et al. 2001), is experimentally supported here.

Fig. 3 shows the estimates of signal rms *s*, noise rms $\sigma$ and their quotient $\gamma$, which serves as SNR estimate, depending on the various averaging methods for channel M2 (right mastoid versus vertex). Data for individual subjects is plotted with open circles and connected by broken lines. Mean values across subjects are represented by solid circles and connected by solid lines.The indices at the abscissa refer to the averaging methods as explained in section 2.2 and are also used as labels in the subsequent figures. The variance of *s*, $\sigma$ and $\gamma$ across subjects is considerably higher than across averaging schemes. From the left graph in Fig. 3 it can be seen that the signal estimates for weighted averaging $s_w$ (label »w«) and sorted averaging $s_s$ (label »s«) are the lowest, whereas the value of $s_b$ for the block-weighted averaging scheme with the block size 256 is similar to $s_a$ (i.e., employing different artifact criteria) and $s_c$ (i.e., conventional averaging). The middle graph shows that estimates of the noise rms values $\sigma$ increase with an increasing artifact threshold. The only exception – for one subject the artifact criterion ±6 $\mu$V results in the highest $\sigma$ value – is due to the small number of accepted epochs in that particular average. Except for this case, conventional averaging leads to the highest levels of residual noise. Hence, switching off the artifact rejection or choosing a too strict criterion for it raises the noise level. $\sigma_s$ and $\sigma_w$ are the lowest noise values that reflect the known underestimation of signal and noise in weighted averaging and suggest that there is a similar, but weaker effect for sorted averaging. The noise level of block-weighted averaging $\sigma_b$ is in the same range as for averaging with artifact criteria.

The right graph in Fig. 3 indicates that the estimated SNR values $\gamma$ are lowest in the case of conventional averaging, but no method can be identified unambiguously as being superior to the others. For some subjects the non-weighted averaging schemes give higher $\gamma$ values, for other subjects the weighted averaging schemes are advantegous.

Fig. 4 demonstrates the effect of the different block sizes used in the block-weighted averaging method on *s*, $\sigma$ and $\gamma$. For each subject, the same 9984 (39 x 256) epochs entered the nine different averages. Block sizes were chosen in steps of powers of two, ranging from 1 to 256. Averages corresponding to block sizes $\beta = 1$ and $\beta = 256$ already appeared in Fig. 3 with the labels »w« and »b«. Fig. 4 therefore provides a fine resolution between weighted averaging and block-weighted averaging, using the largest block size available. Both *s* and $\sigma$ increase monotonically with increasing block size. This reflects the monotonic decrease of the underestimation of signal and residual noise. Because $\sigma$ increases faster with $\beta$ than *s*, a maximum of $\gamma$ is found at block size eight for the average across subjects. Note that the resulting value ($\gamma_{b8} = 13.4$) is larger than that obtained for all averaging methods considered in the previous figure.

In Fig. 5 the estimates of residual noise based on single epochs and based on two sub-averages are compared for the various averaging methods. Data is normalized to the individual estimate obtained for each individual subject using averaging with an artifact criterion of ±10 $\mu$V. This eliminates the high variance between subjects. The average-based noise estimate depicted in the right graph of Fig. 5 shows noticeably higher variation than the single-sweep-based noise estimate.

Monaural stimulation results in *s* values of about half the magnitude of binaural stimulation. Since the residual noise estimates remain almost unaffected by the mode of stimulation, the $\gamma$ values are also halved in the case of monaural stimulation. Lowering the presentation level did not affect the $\sigma$ values either, but lowered the *s* and $\gamma$ values instead. For the stimulus levels of 40 and 20 dB normal hearing level, mean SNR estimates of about 10 and 6, respectively, were obtained. Inspection of channel M1 (left mastoid versus vertex) and channel FZ (forehead versus vertex) shows a similar dependence on averaging schemes. For channel M1, the *s* , $\sigma$, and $\gamma$ values are nearly identical to the values for channel M2. In the case of channel FZ, the *s* values range at around one third of the values of channels M1 and M2. Residual noise level is at about 75 % of that of the mastoidal channels, resulting in $\gamma$ values of about 45 % of the values of channels M1 and M2.

### 3.2 Iterative averaging

To demonstrate the effect of iteration, an example of weighted averaging is shown for the case of diotic stimulation at 60 dB nHL on one subject in Fig. 6. In the upper graph the non-iterated signal estimate $s_w^{(1)}(t)$ is compared to the iterated estimates $s_w^{(2)}(t)$ and $s_w^{(3)}(t)$ as well as to the trivial signal estimate $s_w^{(0)}(t) = 0$. Apparently, $s_w^{(1)}(t)$ has a smaller amplitude than the other two curves, demonstrating the underestimation of the signal by weighted averaging, which is eliminated by the iteration process. The fact that $s_w^{(2)}(t)$ and $s_w^{(3)}(t)$ are nearly identical shows that the iteration process leads to a significant change in the waveform only in the first step, resulting in $s_w^{(2)}(t)$. Similar results from the data for the other subjects show that the iteration procedure converges quickly and is stable.

In the lower graph, the corresponding residual noise estimates $\sigma_w^{(0)}(t)$, $\sigma_w^{(1)}(t)$, $\sigma_w^{(2)}(t)$ and $\sigma_w^{(3)}(t)$ are depicted. The dashed line representing $\sigma_w^{(0)}(t)$ clearly contains signal information because the measured signal is considered as noise alone (cf. eq. (4)). Apparently, a more accurate weighted average is achieved after at least one iteration step. Further iteration steps do not significantly lower the estimated residual noise.

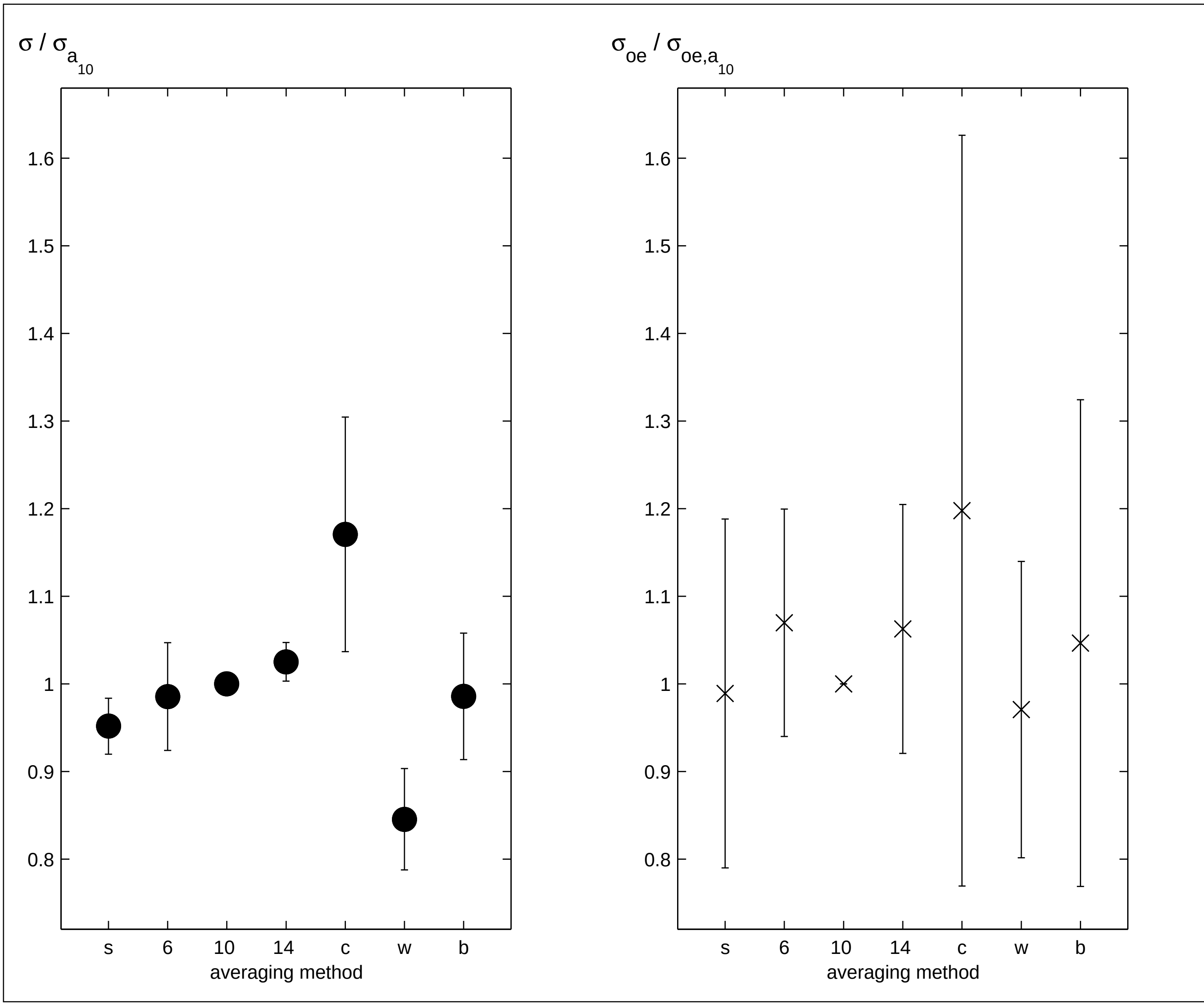



*Fig. 5: Comparison between single-sweep-based and average-based noise estimation for the various averaging methods. To reduce the high interindividual variance, data is normalized to the averaging method using an artifact criterion of ±10 µV.* ***Left graph****: noise estimate σ based on single sweeps.* ***Right graph****: noise estimate $\sigma_{oe}$ based on two averages. The standard deviation (error bars) of $\sigma_{oe}$ is significantly higher than that of σ for all averaging methods.*

*Abb. 5: Vergleich der einzelepochenbasierten und mittelwertbasierten Rauschschätzung für die verschiedenen Mittelungsmethoden. Zur Reduktion der hohen interindividuellen Varianz wurden die Daten normiert auf die Werte zur Mittelungsmethode mit der Artefaktschranke 10 µV.* ***Links****: Rauschschätzung σ auf Basis von Einzelepochen.* ***Rechts****: Rauschschätzung $\sigma_{oe}$ auf Basis von zwei Mittelwerten. Für alle Mittelungsmethoden ist die Standardabweichung (Fehlerbalken) von $\sigma_{oe}$ signifikant größer als die Standardabweichung von σ.*

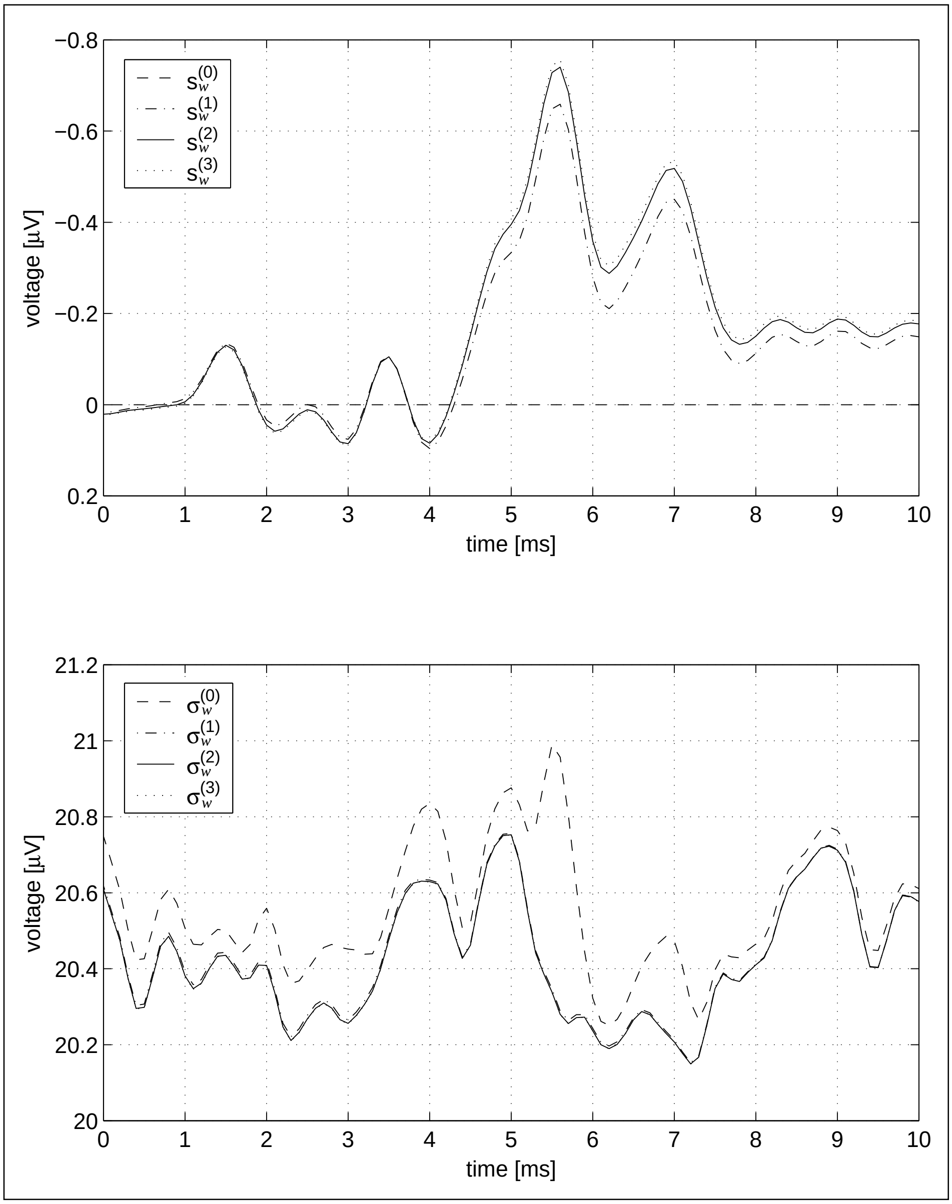



*Fig. 6: Iterative averaging:* ***Upper graph****: Signal estimates* $s_w^{(0)}(t) \equiv 0$, $s_w^{(1)}(t)$, $s_w^{(2)}(t)$ *and* $s_w^{(3)}(t)$. $s_w^{(1)}$ *has a smaller magnitude than* $s_w^{(2)}$. *The difference between* $s_w^{(2)}$ *and* $s_w^{(3)}$ *is negligible.* ***Lower graph****: Noise estimates* $\sigma_w^{(0)}(t)$, $\sigma_w^{(1)}(t)$; $\sigma_w^{(2)}(t)$ *and* $\sigma_w^{(3)}(t)$. $\sigma_w^{(0)}$ *contains signal components, which disappear in the iterated estimates. The differences between* $\sigma_w^{(1)}$, $\sigma_w^{(2)}$ *and* $\sigma_w^{(3)}$ *are very small, indicating the quick convergence of the iteration process. Data is taken from diotic stimulation at 60 dB nHL for one subject.*

*Abb. 6: Iterierte Mittelung:* ***Oben****: Die Signalschätzungen* $s_w^{(0)}(t) \equiv 0$, $s_w^{(1)}(t)$, $s_w^{(2)}(t)$ *und* $s_w^{(3)}(t)$. $s_w^{(1)}$ *hat eine kleinere Amplitude als* $s_w^{(2)}$. *Die Differenz zwischen* $s_w^{(2)}$ *und* $s_w^{(3)}$ *ist vernachlässigbar.* ***Unten****: Die Rauschschätzungen* $\sigma_w^{(0)}(t)$, $\sigma_w^{(1)}(t)$, $\sigma_w^{(2)}(t)$ *und* $\sigma_w^{(3)}(t)$. $\sigma_w^{(0)}$ *enthält Signalanteile, die in den iterierten Schätzungen verschwinden. Die Unterschiede zwischen* $\sigma_w^{(1)}$, $\sigma_w^{(2)}$ *und* $\sigma_w^{(3)}$ *sind sehr klein und zeigen die schnelle Konvergenz des Iterationsprozesses. Die Daten sind für diotische Stimulation einer Versuchsperson bei 60 dB nHL.*

In Fig. 7 the effect of iteration on the estimates is shown for the various averaging methods. Again data is normalized to the individual estimate obtained for each individual subject using averaging with an artifact criterion of $\pm 10\ \mu$V. For each averaging method a pair of values is shown: on the left, the non-iterated quantities (order 1) by means of a triangle pointing to the right, on the right, the iterated quantities (order 2) by means of a triangle pointing to the left. From the left graph in Fig. 7 it can be seen that the $s^{(2)}$ values are almost independent of the averaging method. A pronounced increase in signal power due to a single iteration is observed for $s_w^{(2)}$, $s_s^{(2)}$ and $s_6^{(2)}$, respectively. This underlines the notion that the underestimation of the signal in the case of weighted and sorted averaging can be overcome by one iteration step.

The middle graph shows that the residual noise estimates $\sigma^{(1)}$ and $\sigma^{(2)}$ do not differ significantly. This was already observed in the lower graph of Fig. 6 for one subject. The residual noise estimates $\sigma^{(0)}$ – not shown here – are slightly higher than $\sigma^{(1)}$ and $\sigma^{(2)}$ for all averaging methods. This reflects the exclusion of signal components from the noise estimates if iteration is used.

The SNR estimates depicted in the right graph indicate that weighted averaging with iteration is superior to all the other methods. Sorted averaging is also improved by iteration. Except for the strictest artifact criterion, iteration does not greatly affect the other averaging methods.

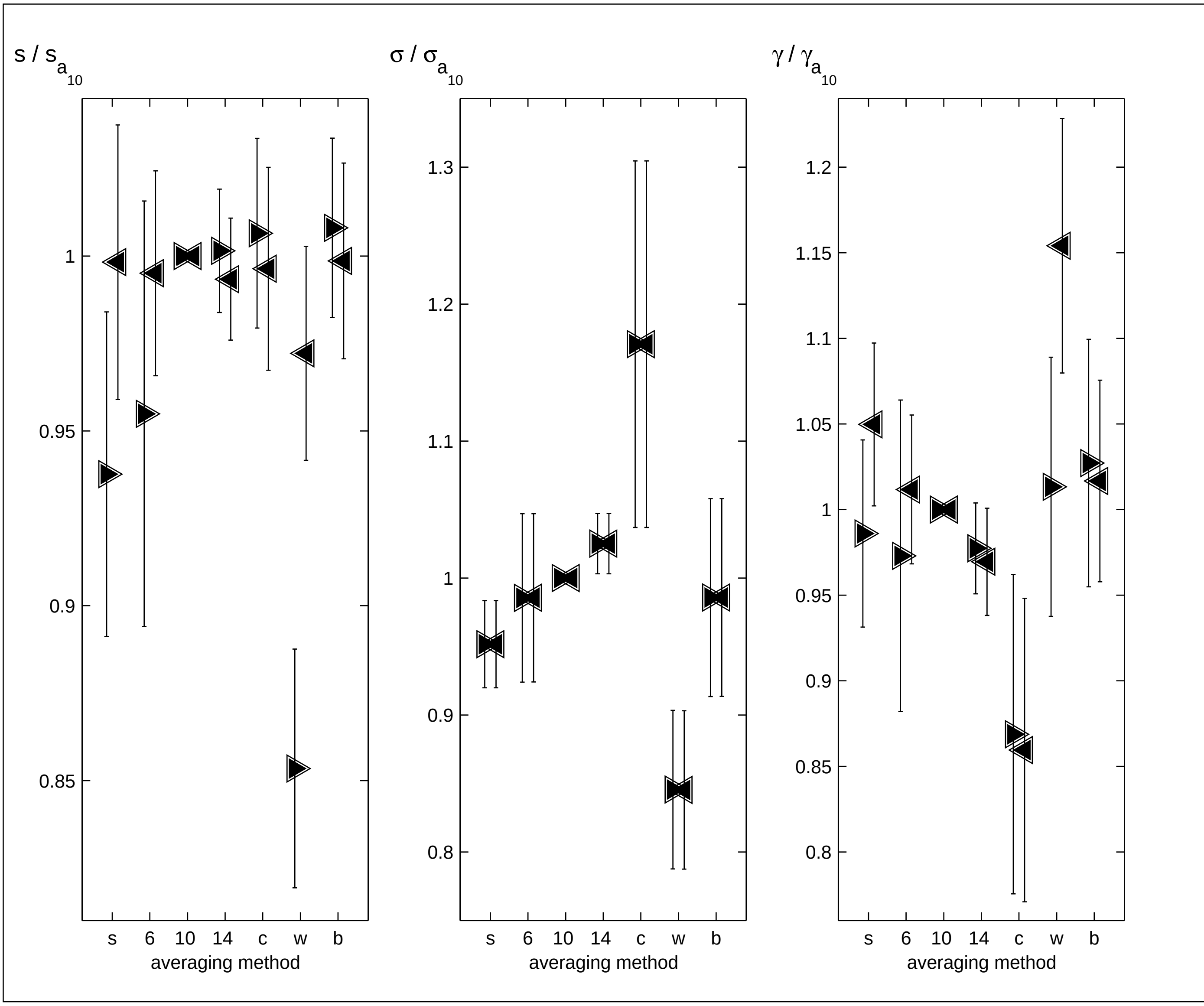



*Fig. 7: Mean and interindividual standard deviation of the relative signal rms estimate $s/s_{a10}$ (left), residual noise rms estimate $\sigma/\sigma_{a10}$ (middle) and SNR estimate $\gamma/\gamma_{a10}$ (right) for various averaging methods (labelled as in Fig. 3). Data is individually normalized to the values for averaging using an artifact threshold of ±10 µV. For each method a pair of data is depicted: the symbols on the left represent the non-iterated estimates, the symbols on the right the estimates after one iteration step. Data is from diotic stimulation at 60 dB normal hearing level for channel M2 (right mastoid versus vertex).*

*Abb. 7: Mittelwerte und interindividuelle Standardabweichungen der Schätzer für die normierten RMS-Werte des Signals $s/s_{a10}$ (links), des Restrauschens $\sigma/\sigma_{a10}$ (Mitte) und des Signal-Rausch-Abstands $\gamma/\gamma_{a10}$ (rechts) für die verschiedenen Mittelungsmethoden (Bezeichnung wie in Abb. 3). Die Daten sind normiert auf die Werte zur Mittelungsmethode mit der Artefakt-schranke 10 µV. Für jede Methode ist ein Datenpaar eingezeichnet, die Symbole links gelten für die nicht iterierten Schätzungen, die Symbole rechts für die Schätzungen nach einem Iterationsschritt. Die Daten sind für diotische Stimulation bei 60 dB nHL und Kanal M2.*

The effect of iteration on block-weighted averaging is shown in Fig. 8. As before, data is normalized to the individual data using averaging from an artifact threshold of ±10 $\mu$V. Iteration raises the *s* values significantly for small block sizes, but its effect decreases as the block size increases. As with all the other averaging schemes, the $\sigma$ values do not vary because of iteration. The »optimal« block size with highest SNR estimate decreases from eight in the non-iterated case to one when iteration is applied. $\gamma$ values in iteration decrease monotonically as the block size increases. Hence, the optimal averaging scheme in the sense of the highest γ is weighted averaging with one iteration.[4]

### 3.3 Simulations

In Fig. 9 the true residual noise rms value Σ, defined as the rms value of the difference between true signal *S* and estimate *s*, is plotted for each subject (open symbols) for the various averaging methods, whether non-iterated or iterated. Again, mean values (solid symbols) and standard deviation across subjects are given. The resulting residual noise Σ shows greater values for non-iterated than for iterated averaging in the case of weighted averaging, to a lesser extent also for sorted averaging and averaging with the artifact threshold ± 6 $\mu$V.

Except for conventional averaging, the mean true residual noise after one iteration $\Sigma^{(2)}$ ranges from 15 to 19 nV. Of the averaging methods shown in Fig. 9, the lowest value is obtained for block-weighted averaging with the largest block size ($\Sigma^{(2)}_{b256}$ = 15.0 nV), followed by averaging with artifact threshold ± 10 $\mu$V ($\Sigma^{(2)}_{a10}$ = 15.9 nV) and weighted averaging ($\Sigma^{(2)}_{w}$ = 16.4 nV).

The rms values of the true residual noise Σ are plotted for the block-weighted averaging method in Fig. 10 as a function of the block size. Without iteration, the residual noise values $\Sigma^{(1)}$ for small block sizes are comparatively high. For block sizes greater than eight, $\Sigma^{(1)}$ stays almost constant. $\Sigma^{(2)}$ values (iterated case) remain almost constant for all block sizes, and there is a shallow minimum for block size four with $\Sigma^{(2)}_{b4}$ = 13.9 nV. For block sizes greater than 16, the difference between $\Sigma^{(1)}$ and $\Sigma^{(2)}$ are negligible.

Fig. 11 gives the quotients of the estimated and true quantities for the various averaging methods with and without iteration. The left graph shows that, without iteration, weighted averaging underestimates the signal by 15 %, while the effect on sorted averaging (-7 %) and averaging with artifact threshold ± 6 $\mu$V (-4 %) is not so great. Increasing the artifact threshold ± 10 $\mu$V or ± 14 $\mu$V results in a slight overestimation of *S*.

Iteration considerably improves sorted and weighted averaging: signal estimates deviate by less than 1 % above and 2 % below the true value S. In conclusion, the use of iteration causes signal estimation to work well for all methods. Without iteration, the classical technique using a not too strict artifact criterion provides good signal-rms estimates.

With regard to the residual noise estimation depicted in the middle graph of Fig. 11, rather large deviations between the non-iterated and iterated cases can be seen. Without iteration, residual noise is heavily underestimated by weighted averaging (38 % of the true value) and sorted averaging (69 %). On the other hand, averaging using artifact criteria and block-weighted averaging overestimate Σ by up to 17 %. Surprisingly, the best non-iterated residual-noise estimation is provided by conventional averaging and averaging with the strictest artifact criteria. This results in an overestimation of only about 2 %.

Iteration does not correct the overestimation in those methods that overestimate the residual noise before iteration. However, the huge underestimation in sorted and weighted averaging is reduced to 5 % and 8 % respectively. This indicates that sorted and weighted averaging in iteration are capable of providing accurate estimates of residual noise.

In the right graph of Fig. 11, the mean and standard deviations of the normalized SNR estimates across subjects are given. Without iteration, the SNR is overestimated by sorted averaging by 45 % and by more than a factor of two by weighted averaging. Averaging using artifact rejection yields estimates close to the true Γ, but both signal and residual noise are overestimated. The overestimation of Γ by sorted and weighted averaging is greatly reduced by iteration and amounts to 9 % and 13 %, respectively. The other averaging methods are almost unaffected by iteration.

In Fig. 12 the same type of plot is presented as in the previous figure, but for different block sizes of the block-weighted averaging method. Generally, the effect of iteration is negligible for block sizes greater than eight. There is a trend to overestimate the signal as the block size increases. Non-iterated signal estimators underestimate *S* for block sizes smaller than 32. For larger block sizes and for all iterated estimators, signal rms *S* is well estimated within a range of ± 2 %.

Again, the effect of iteration is larger for the residual noise estimates $\sigma$ than for the signal estimates. Iteration increases $\sigma$ values for block sizes up to $\beta$ = 8. For larger block sizes, Σ is overestimated by about 15 % in both the non-iterated and iterated cases. Without iteration, Σ is underestimated for $\beta$ =1 by 62 %, for $\beta$ = 2 by 35%, and best estimated for $\beta$ = 4 and $\beta$ = 8. With iteration, $\beta$ =1 (weighted averaging) fits the true Σ best (-9 %), while bigger block sizes overestimate Σ by at least 10 %.

[4] Once again the analysis of the other channels recorded showed analogous dependencies of the estimates on averaging methods and iteration. For channel M1 (left mastoid versus vertex), the magnitudes of the data are nearly identical, while for channel Fz (forehead versus vertex) the *s* and γ values are lower than in the other channels.

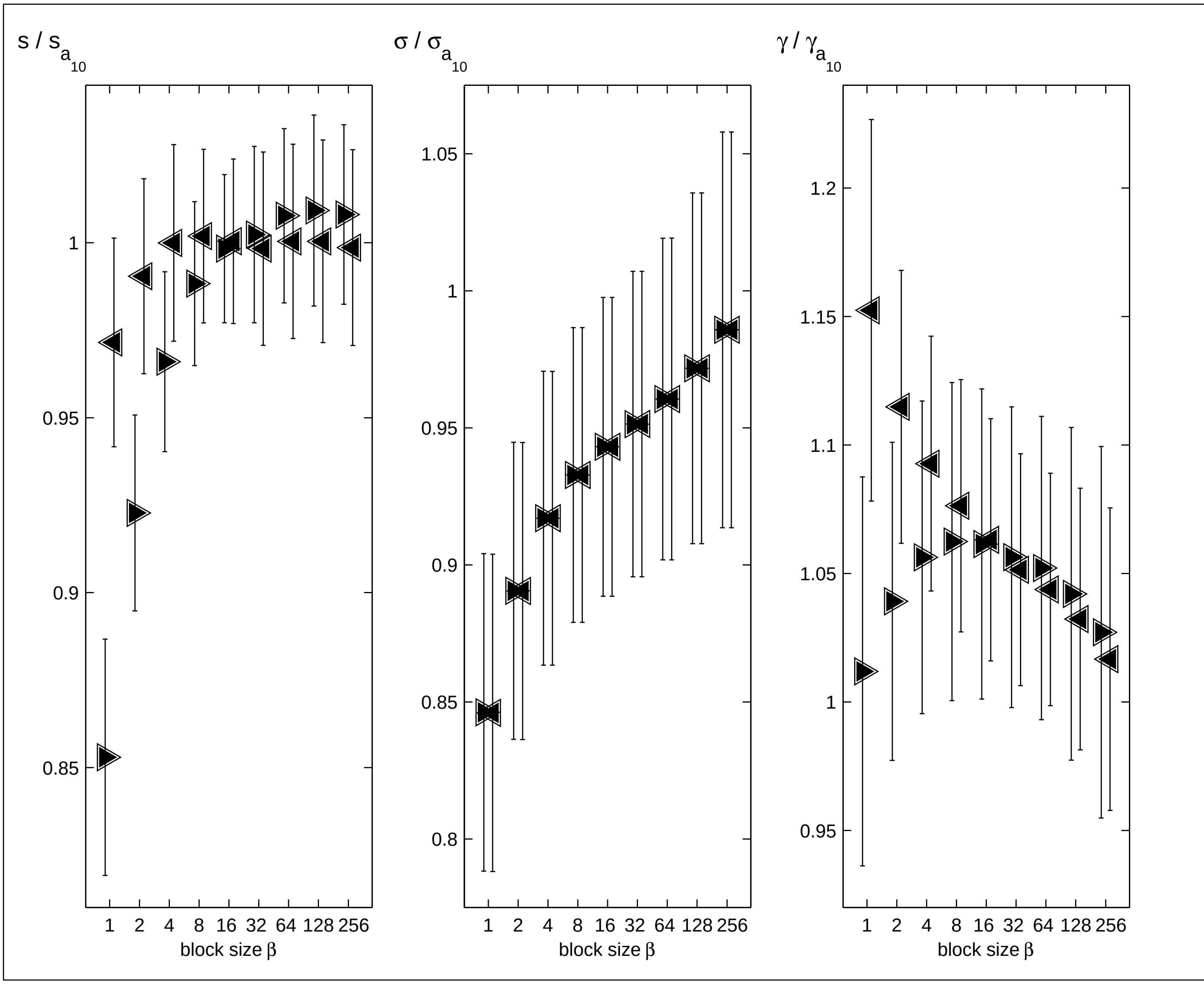



*Fig. 8: Mean and interindividual standard deviation of the relative signal rms estimate* $s/s_{a10}$ *(left), residual noise rms estimate* $\sigma/_{a10}$ *(middle) and SNR estimate* $\gamma/\gamma_{a10}$ *(right) as a function of the block size* $\beta$ *for the block-weighted averaging method. Data is divided by the values for averaging using an artifact threshold of* ±10 μV *to eliminate the variance across subjects. For each method a pair of data is depicted: the symbols on the left represent the non-iterated estimates, the symbols on the right the iterated estimates. Data is from diotic stimulation at 60 dB normal hearing level for channel M2 (right mastoid versus vertex).*

*Abb. 8: Mittelwerte und interindividuelle Standardabweichungen der Schätzer für die normierten RMS-Werte des Signals* $s/s_{a10}$ *(links), des Restrauschens* $\sigma/\sigma_{a10}$ *(Mitte) und des Signal-Rausch-Abstands* $\gamma/\gamma_{a10}$ *(rechts) als Funktion der Blockgröße* $\beta$ *für das block-gewichtete Mittelungsverfahren. Die Daten sind normiert auf die Werte zur Mittelungsmethode mit der Artefaktschranke 10* μV*. Für jede Methode ist ein Datenpaar eingezeichnet, die Symbole links gelten für die nicht iterierten Schätzungen, die Symbole rechts für die Schätzungen nach einem Iterationsschritt. Die Daten sind für diotische Stimulation bei 60 dB nHL und Kanal M2.*

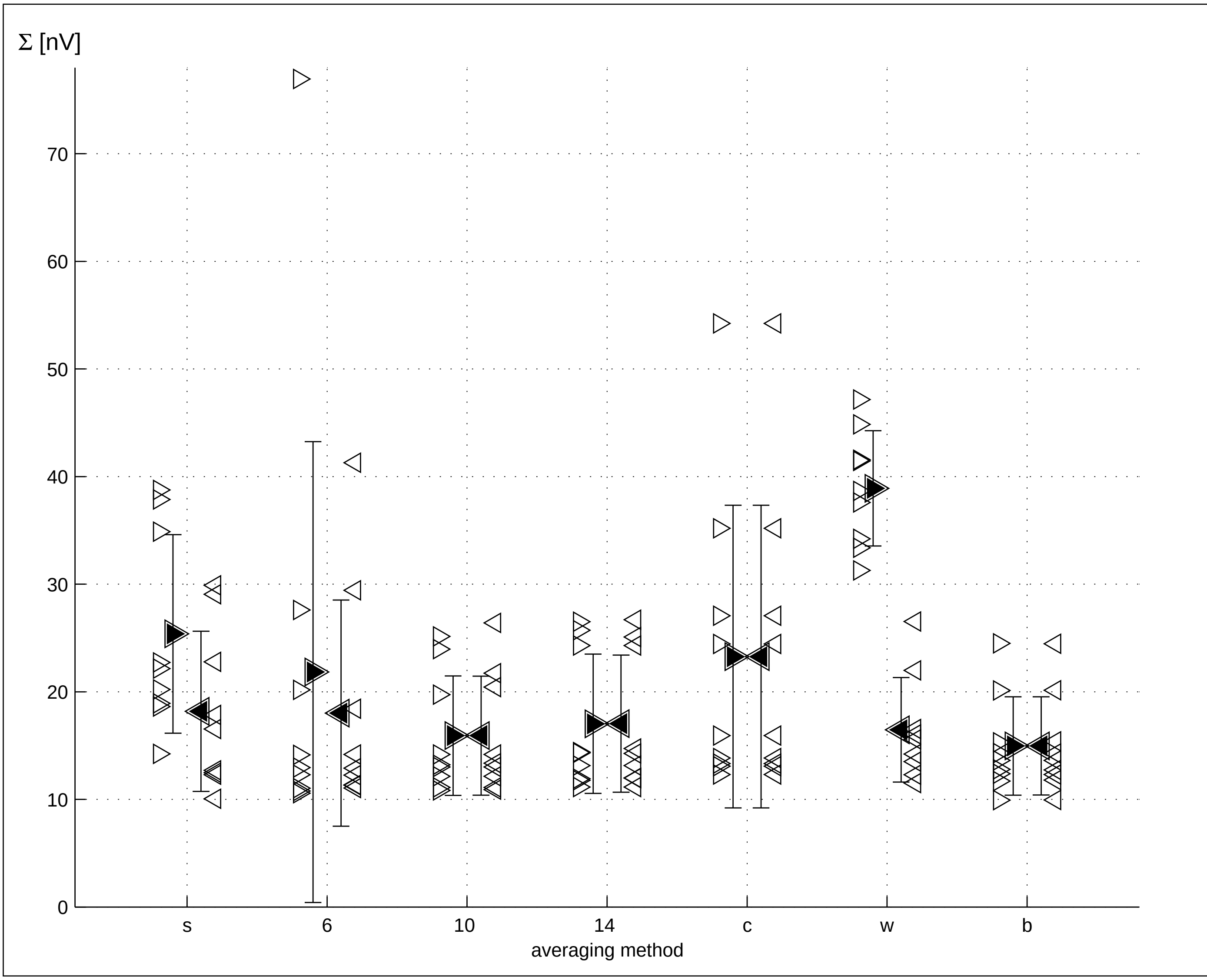



*Fig. 9: Simulation results. True residual noise Σ for various averaging methods (labelled as in Fig. 3) and iterations. Mean values (over subjects) are shown with filled symbols and error bars indicating ± one standard deviation. The open symbols represent the data for individual subjects. Non-iterated data is plotted on the left of the vertical grid lines, iterated data on the right. Data is from diotic stimulation at 60 dB normal hearing level for channel M2 (right mastoid versus vertex).*

*Abb. 9: Simulationsergebnisse. Wahres Restrauschen Σ für die verschiedenen Mittelungsmethoden (Bezeichnung wie in Abb. 3) und Iterationen. Mittelwerte (über die Versuchspersonen) sind mit gefüllten Symbolen und Fehlerbalken (± eine Standardabweichung) gezeichnet, die offenen Symbole repräsentieren die Daten für die einzelnen Versuchspersonen. Nicht iterierte Daten sind links, iterierte Daten rechts gezeichnet. Die Daten sind für diotische Stimulation bei 60 dB nHL und Kanal M2.*

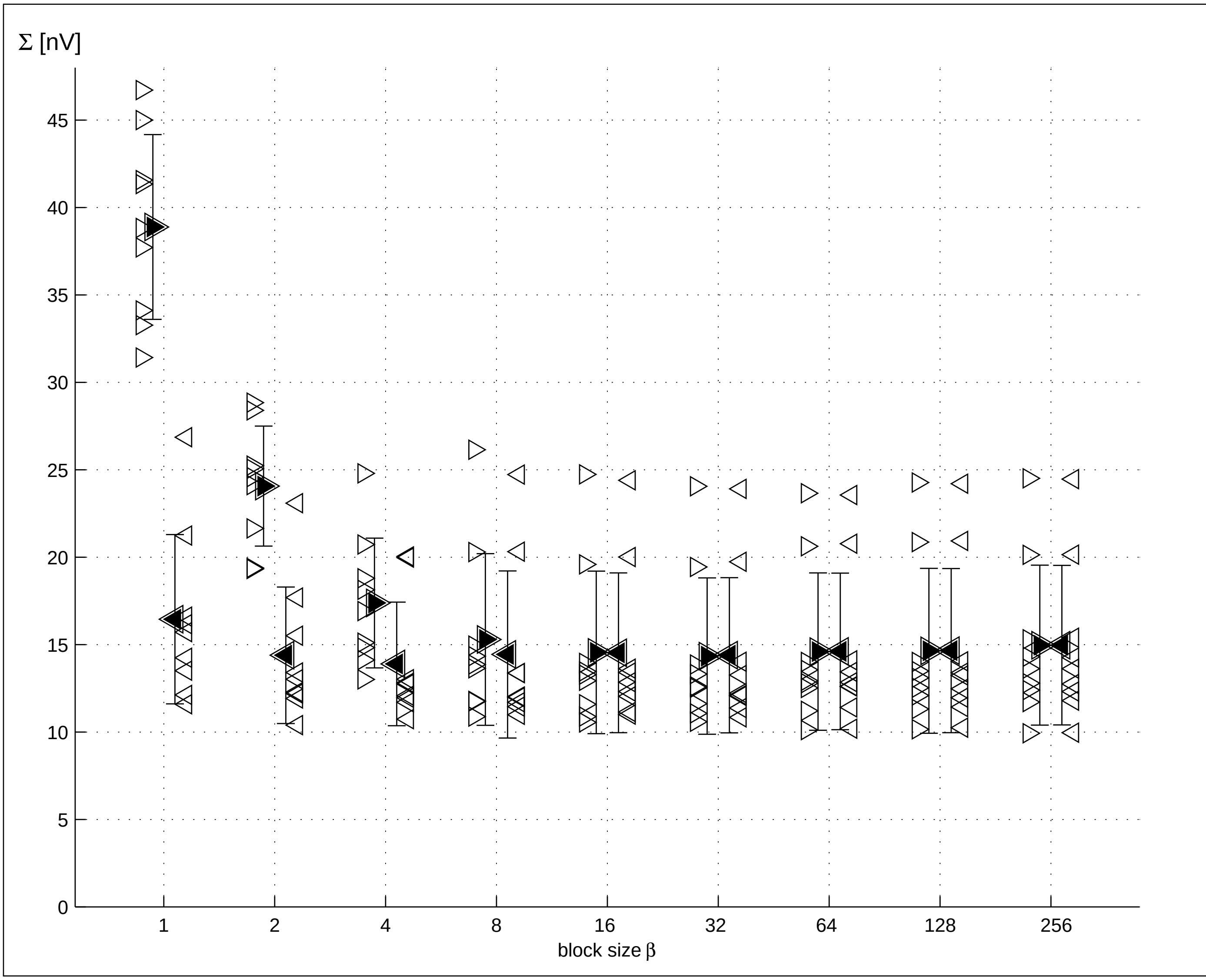



*Fig. 10: Simulation results. True residual noise Σ as a function of the block size β for the block-weighted averaging method. Mean values (over subjects) are shown with filled symbols and error bars indicating ± one standard deviation. The open symbols represent the data for individual subjects. Non-iterated data is plotted on the left of the vertical grid lines, iterated data on the right. Data is from diotic stimulation at 60 dB normal hearing level for channel M2 (right mastoid versus vertex).*

*Abb. 10: Simulationsergebnisse. Wahres Restrauschen Σ als Funktion der Blockgröße β für das block-gewichtete Mittelungsverfahren. Mittelwerte (über die Versuchspersonen) sind mit gefüllten Symbolen und Fehlerbalken (± eine Standard-abweichung) gezeichnet, die offenen Symbole repräsentieren die Daten für die einzelnen Versuchspersonen. Nicht iterierte Daten sind links, iterierte Daten rechts gezeichnet. Die Daten sind für diotische Stimulation bei 60 dB nHL und Kanal M2.*

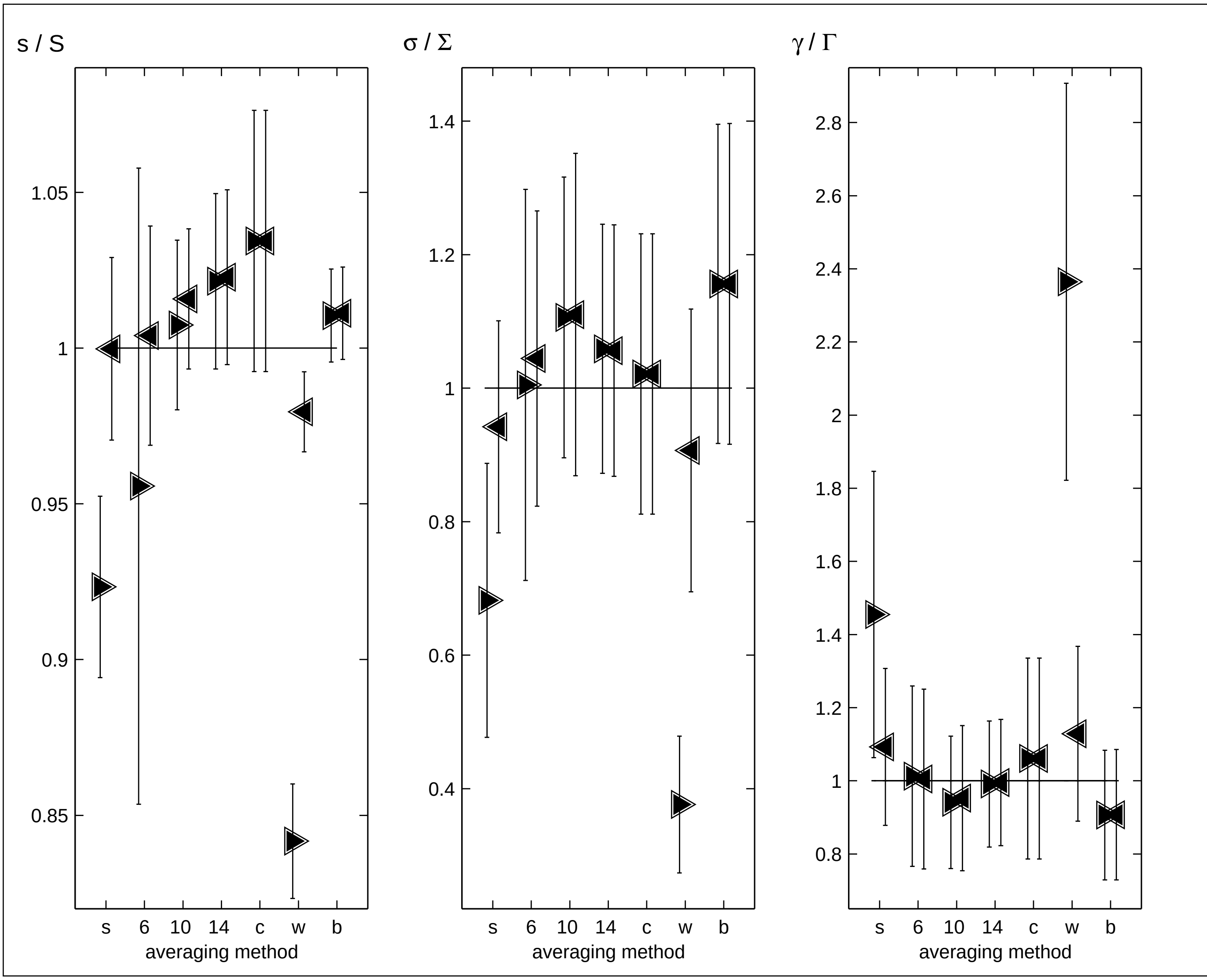



*Fig. 11: Quotients of estimated signal rms s and true signal rms S (left), estimated residual noise rms σ and true residual noise Σ (middle) and estimated SNR γ and true SNR Γ (right) for various averaging methods. Labels of averaging methods are the same as in Fig. 3. As in the previous figure, the data left of the respective abscissa value is not iterated, while the data to the right is iterated. Triangles indicate mean values, error bars denote one standard deviation. Data is from diotic stimulation at 60 dB normal hearing level for channel M2 (right mastoid versus vertex).*

*Abb. 11: Quotienten aus geschätztem und wahren RMS-Wert des Signals (links), des Restrauschens (Mitte) und des Signal-Rausch-Verhältnisses (rechts) für die verschiedenen Mittelungsmethoden (Bezeichnung wie in Abb. 3). Nicht iterierte Daten sind links, iterierte Daten rechts gezeichnet (Mittelwert ± eine Standardabweichung). Die Daten sind für diotische Stimulation bei 60 dB nHL und Kanal M2.*

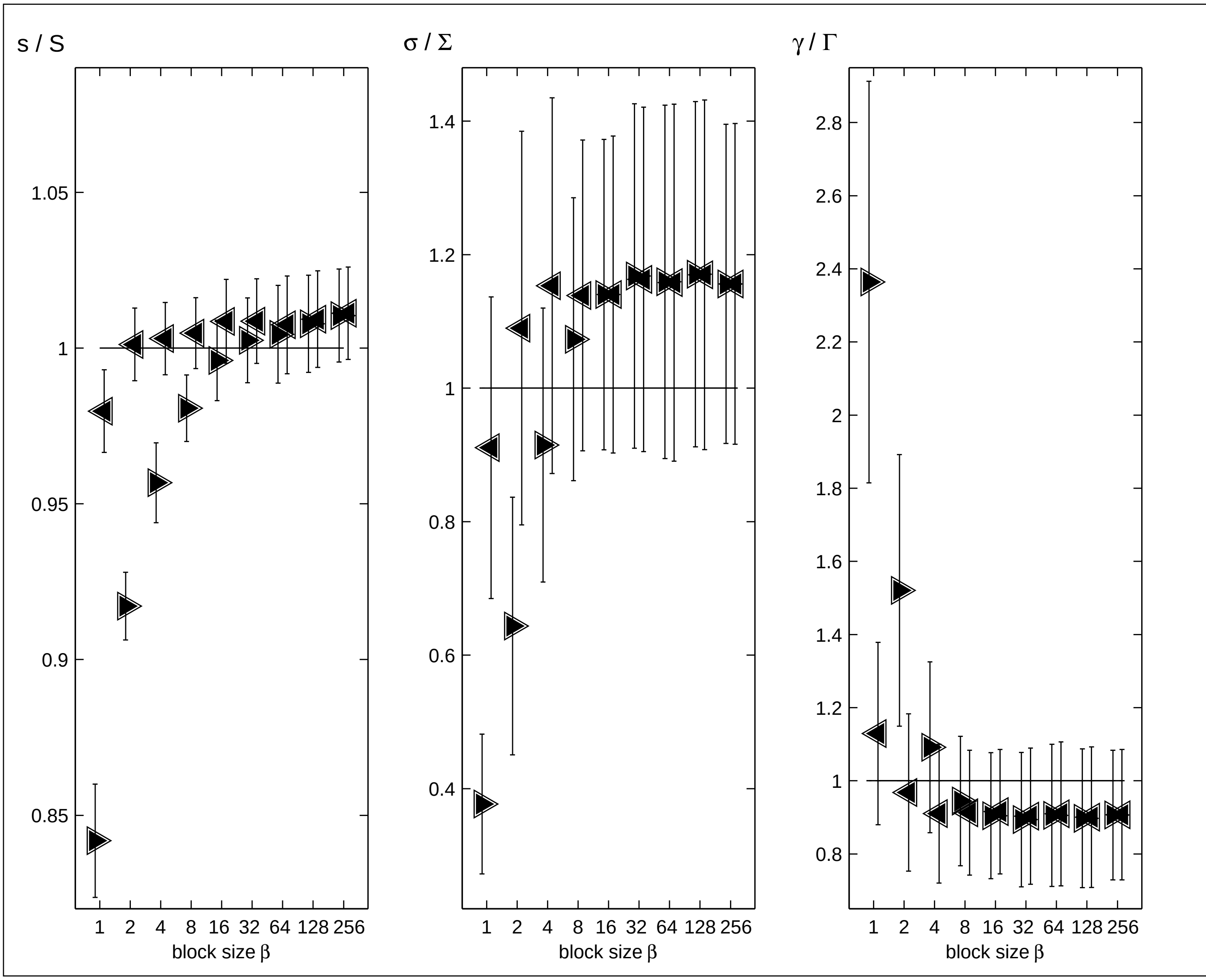
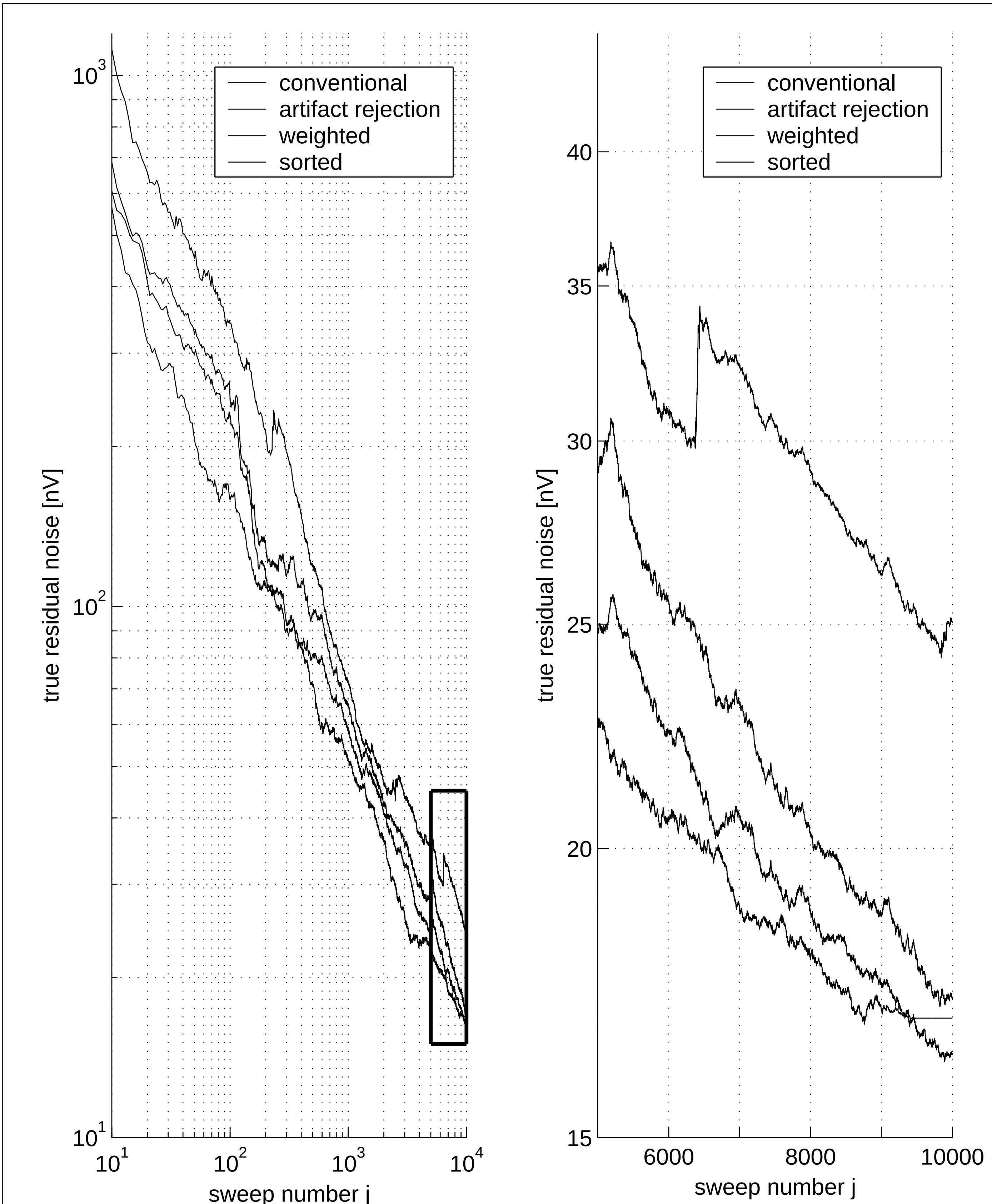



*Fig. 12: Quotients of estimated signal rms s and true signal rms S (left), estimated residual noise rms σ and true residual noise Σ (middle) and estimated SNR γ and true SNR Γ (right) as a function of the block size β for the block-weighted averaging method. To facilitate comparison, presentation and scales are identical to the previous figure. Data is from diotic stimulation at 60 dB normal hearing level for channel M2 (right mastoid versus vertex).*

*Abb. 12: Quotienten aus geschätztem und wahren RMS-Wert des Signals (links), des Restrauschens (Mitte) und des Signal-Rausch-Verhältnisses (rechts) als Funktion der Blockgröße β für das block-gewichtete Mittelungsverfahren. Zum einfacheren Vergleich sind Präsentation und Skalierung identisch zur vorherigen Abbildung gehalten. Die Daten sind für diotische Stimulation bei 60 dB nHL und Kanal M2.*


conventional
artifact rejection
weighted
sorted
true residual noise [nV]
sweep number j
$10^3$
$10^2$
$10^1$
$10^1$
$10^2$
$10^3$
$10^4$
conventional
artifact rejection
weighted
sorted
true residual noise [nV]
sweep number j
40
35
30
25
20
15
6000
8000
10000

The quality of the SNR estimates is again shown in the right graph. In the non-iterated case, a considerable overestimation (by more than the factor two and 52 %) occurs for block size one (weighted averaging) and two, respectively. In the iterated case, Γ is best estimated by weighted averaging ($\beta = 1$) and $\beta = 2$. A considerable underestimation of Γ, independent of an iteration of about 10 %, is observed for block sizes larger than four.

With the simulation technique employed here, the true residual noise Σ can be computed as a function of the number of sweeps included in the average. This was done by calculating the rms value of the difference between the known signal $S(t)$ and the average after the inclusion of $j$, $j = 1...J$ sweeps. In Fig. 13 the average $\Sigma_j^{(2)}$ across subjects is shown for four averaging methods using iteration: conventional averaging, averaging using the ±10 $\mu$V artifact criterion, weighted and sorted averaging. For a given number of sweeps, $\Sigma_j^{(2)}$ decreases in the order of the above list, i.e., conventional averaging yields the highest residual noise values for a given number of sweeps. The classical method using the ±10 $\mu$V artifact threshold considerably lowers residual noise. A further improvement is possible with weighted averaging. Sorted averaging has the lowest residual noise values over a wide range of sweep numbers. However, we must not forget that $\Sigma_{j'}$ is shown for sorted averaging, i.e., the residual noise as a function of the epoch index after sorting. Hence, the sweeps with the lowest noise contamination enter the average first. This results in a lower residual noise at the beginning than for the other methods.

🡐

*Fig. 13: **Left graph**: True residual noise averaged over subjects after one iteration step ($\Sigma^{(2)}$), dependent on the sweep number j (j´ for sorted averaging) for four averaging methods. From top to bottom: conventional ($\Sigma_c^{(2)}$), artifact rejection ± 10 $\mu$V ($\Sigma_{a10}^{(2)}$), weighted ($\Sigma_w^{(2)}$), and sorted ($\Sigma_s^{(2)}$). Data is from diotic stimulation at 60 dB normal hearing level for channel M1 (left mastoid versus vertex). **Right graph**: Magnification of the left graph for sweep numbers 5.000 to 10.000.* At about *j = 9400, iterated sorted averaging cannot further decrease residual noise and is outperformed by iterated weighted averaging.*

*Abb. 13: **Links**: Das über die Versuchspersonen gemittelte wahre Restrauschen nach einem Iterationsschritt ($\Sigma^{(2)}$) in Abhängigkeit von der Epochennummer j (j´ für sortiertes Mitteln) für vier Mittelungsmethoden. Von oben nach unten: konventionell ($\Sigma_c^{(2)}$), Artefaktschranke ± 10 $\mu$V ($\Sigma_{a10}^{(2)}$), gewichtet ($\Sigma_w^{(2)}$) und sortiert ($\Sigma_s^{(2)}$). Die Daten sind für diotische Stimulation bei 60 dB nHL und Kanal M1. **Rechts**: Vergrößerung des linken Teilbilds für Epochennummern 5000 bis 10000. Bei etwa 9400 Epochen kann iteriertes sortiertes Mitteln das Restrauschen nicht weiter reduzieren und wird vom iterierten gewichteten Mitteln übertroffen.*

*Tab. 1: Sweep numbers necessary to reach a given residual noise criterion for four averaging methods using the iteration procedure proposed here, i.e., conventional averaging ($J_c^{(2)}$), averaging using an artifact threshold of* ±10 $\mu$V ($J_{a10}^{(2)}$), *weighted averaging ($J_w^{(2)}$) and sorted averaging ($J'^{(2)}_s$). Mean over subjects. Channel M1 (left mastoid versus vertex).*

*Tab. 1: Anzahl der Epochen, die notwendig ist, um ein gegebenes Kriterium für das Restrauschen zu erreichen. Vier Mittelungsmethoden einschließlich der Iterationsprozedur werden verglichen: Konventionelles Mitteln ($J_c^{(2)}$), Mitteln mit Artefaktschranke* ±10 $\mu$V ($J_{a10}^{(2)}$), *gewichtetes Mitteln ($J_w^{(2)}$) und sortiertes Mitteln ($J'^{(2)}_s$). Mittelwerte über Versuchspersonen, Kanal M1 (linkes Mastoid gegen Vertex).*

| Σ[nV] | $J_c^{(2)}$ | $J_{a10}^{(2)}$ | $J_w^{(2)}$ | $J'^{(2)}_s$ |
|---|---|---|---|---|
| 50 | 1775 | 1586 | 1229 | 1069 |
| 45 | 2101 | 1864 | 1746 | 1239 |
| 40 | 3621 | 2132 | 2092 | 1679 |
| 35 | 5290 | 3196 | 2498 | 2084 |
| 30 | 6248 | 3817 | 3420 | 2466 |
| 25 | 9513 | 6052 | 4683 | 3114 |
| 20 | > 10000 | 8039 | 7215 | 6449 |

However, for sweep numbers over approx. 9400, $\Sigma_s$ remains constant. Approximately 600 sweeps are rejected by this averaging scheme, because their inclusion would raise $\Sigma_{j'}$ again. Weighted averaging, on the other hand, can further decrease Σ by using all sweeps, i.e., also those greatly contaminated by noise and to which small weightings are assigned.

Fig. 13 also shows how many epochs have to be included in an average in order to bring the residual noise below a given criterion. Tab. 1 lists those values for residual noise criteria between 20 and 50 nV.

Although the number of sweeps required is lowest the case of sorted averaging, this does not allow a reduction of the number of epochs that have to be recorded before sorting because all the sweeps have to be collected. To reach the 25-nV criterion, for example, artifact rejection reduces the number of epochs to 85 % and weighted averaging to 76 %, compared to conventional averaging.[5]

[5] The results of the simulations look very similar if channel M1 (left mastoid versus vertex) is considered. For the third channel with poor signal, of course, true and estimated signal and SNR are lower, but the dependence of the averaging methods and iteration is the same.

## 4 Discussion

Various averaging methods known from the literature were applied and compared with respect to their ability to estimate ABR waveforms accurately. It was shown, both theoretically in the companion paper (*Granzow* et al. 2001) and empirically here, that single-sweep-based estimation of signal and residual noise is superior to average-based estimation. The new concept of iterative averaging was investigated for all averaging methods. The iteration technique does not strongly influence the results for conventional averaging or averaging using an artifact criterion. However, the improved estimation of the power of a single epoch results in much better signal, noise and SNR estimates in the case of sorted and weighted averaging. Table 1 shows that there is a considerable advantage to iterated weighted averaging. For all residual noise criteria, the number of sweeps that have to be included is lowest with this method ($J_w^{(2)} < J_{a10}^{(2)} < J_c^{(2)}$). For subjects with strongly inhomogeneous noise (background EEG), the advantage of the iterated weighted averaging scheme is more pronounced, while the difference from the classical method using the $\pm 10\ \mu$V artifact criterion becomes negligible for subjects with more homogeneous noise.

The first approach to estimating SNR and residual noise on a single-sweep basis was the single-point variance introduced by *Elberling* and *Don* (1984) and *Don* et al. (1984). Fig. 6 of the present paper shows that the standard error $\sigma^{(1)}(t)$, i.e., the residual noise, does not vary much over time, i.e., within the epoch. There are only small differences in the residual noise estimate if analysis is performed at different instances of time. Therefore, the approach underlying the single-point variance method can be justified on the basis of our data. However, on the basis of Monte-Carlo simulations, *Cebulla* et al. (2000) showed that a residual noise estimate based on all samples of the epochs is advantageous, especially for small numbers of epochs entering the average.

Estimation of residual noise using single sweep information has a large impact on the accuracy of peak detection. The typical clinical question is to decide if a given ABR component is a response or not. If we assume Gaussian measurement errors, the residual noise based on single sweeps, i.e., the standard error $\sigma$, allows for far more precise statements about the significance of peaks than the residual noise estimated on averages.

The SNR improvement of weighted averaging was investigated by *Lütkenhöner* et al. (1985). Their conclusion was that, for a homogeneous ensemble of epochs, weighted averaging did not improve the SNR because both signal and noise were reduced by the same factor. They showed that, for inhomogeneous sweep ensembles, weighted averaging was superior to conventional averaging in terms of a better SNR, but led to a systematic underestimation of the signal. To overcome this problem and maintain the advantage of weighted averaging, the present study demonstrates that undesirable underestimation is effectively inhibited by the iteration procedure employed here (cf. Fig. 11).

*Don* and *Elberling* (1994) analyzed the effect of varying the block size $\beta$ in the block-weighted averaging scheme. They compared the residual variance using 256, 128, 64 and 32 sweeps per block and 1, 2, 4, and 8 »single« time points per sweep, respectively, which always yielded 256 data points to estimate the variance/power of a block. The number of time points per sweep used for noise estimation was not increased above 8, because they argued that there is only a limited number of degrees of freedom in a single epoch, i.e., a small number of independent samples in the band-limited noise signal.

In accordance with their results, our findings for the block-weighted averaging method confirm that a block-weighted average with a small number of epochs is not ideal for estimating of the weightings. Without iteration, the lowest acceptable block size was found to be eight in our study. However, using the iterated noise estimation, we observed that the true residual noise (see Fig. 10) as well as the estimated signal and noise (Fig. 12) did not greatly depend on the block size. In contrast to *Elberling* and *Don*, we therefore conclude that the main error in estimating residual noise is due to a bias produced by the »desired« signal, which is removed by the iteration procedure. The small remaining dependence of residual noise on the block size may be due to the small number of degrees of freedom, although a second iteration step seems to further diminish the dependence of residual noise on the block size (data not shown).

With iterative averaging, the noise in a single sweep can therefore be estimated more accurately than without iteration. Hence, it is not necessary to form a block-weighted average in order to improve the estimate of the noise from a block of epochs.

In the case of ABR, the approximation of eq. (4) is very good since the SNR of single epochs is very low. For evoked and event-related potentials generated in cortical areas, the above approximation is worse since the SNR of the single epochs is generally higher than in ABR recordings. It is therefore to be expected that more pronounced differences between iterated and non-iterated averages will be seen in these cases. Iterative averaging should produce better signal and noise estimates. Additionally, due to the smaller number of sweeps required for late evoked potentials, the computational costs for the single-sweep-based methods are at least an order of magnitude smaller than for ABR.

Taking into account the following three aspects – elimination of the arbitrariness of the artifact threshold, low residual noise (cf. Fig. 9, Tab. 1) and good estimation properties (Fig. 11) – weighted averaging in iteration appears to be the most favorable averaging method.

## 5 Conclusions

- For accurate estimation of signal and residual noise, the methods based on single sweeps are shown to be superior to those based on averages.
- Weighted averaging avoids the arbitrariness of the choice of the artifact threshold.
- The effect of underestimation of signal and noise of weighted averaging can be overcome by the use of the iteration procedure.
- For a given number of recorded sweeps, iterative weighted averaging provides the most reliable estimates of the signal and the residual noise, while iterated sorted averaging appears to be the second best method.

## Acknowledgements

The present work was supported by the *Deutsche Forschungsgemeinschaft* through the *Sonderforschungsbereich Neurokognition* (SFB 517).

## Literatur

*Cebulla M, Stürzebecher E, Wernecke KD* (2000) Untersuchung verschiedener SNR-Schätzer für den Nachweis von biologischen Reizantworten im Rauschen. Z Audiol 39 (1), 14–22

*Don M, Elberling C* (1994) Evaluating residual background noise in human auditory brainstem responses. J Acoust Soc Am 96 (5, Pt.1), 2746–2757, November

*Don M, Elberling C, Waring M* (1984) Objective detection of averaged auditory brainstem responses. Scand Audiol 13, 219–228

*Elberling C, Don M* (1984) Quality estimation of averaged auditory brainstem responses. Scand Audiol 13, 187–197

*Elberling C, Wahlgreen O* (1985) Estimation of auditory brainstem response, ABR, by means of Bayesian inference. Scand Audiol 14, 89–96

*Gevins AS, Rémond A* (1987) Methods of Analysis of Brain Electrical and Magnetic Signals, volume 1 of Handbook of EEG and Clinical Neurophysiology. Elsevier, Amsterdam

*Granzow M, Riedel H, Kollmeier B* (2001) Single-sweep-based methods to improve the quality of auditory brain stem responses. Part I: optimized linear filtering. Z Audiol 40 (1), 32–44

*Hoke M, Ross B, Wickesberg R, Lütkenhöner B* (1984) Weighted averaging – theory and application to electric response audiometry. Electroenceph Clin Neurophysiol 57, 484–489

*Lütkenhöner B, Hoke M, Pantev C* (1985) Possibilities and limitations of weighted averaging. Biol Cybern 52, 409–416

*Mühler R, von Specht H* (1997) Reduction of background noise in human auditory brainstem response by means of classified averaging. In: *Syka J* (ed.) International Symposium on Acoustical Signal Processing in the Central Auditory System, September 4–7, 1996, Prague, Czech Republic, pages 599–604. Plenum Press, New York

*Mühler R, von Specht H* (1999) Sorted averaging – principle and application to auditory brainstem responses. Scand Audiol 28 (3), 145–149

*Schimmel H* (1967) The (±) reference: accuracy of estimated mean components in average response studies. Science 157 (July), 92–93

*Wong PKH, Bickford RG* (1980) Brain stem auditory evoked potentials: the use of noise estimate. Electroenceph Clin Neurophysiol 50, 25–34