Noise and signal power and their effects on evoked potential estimation

B.I. Turetsky, J. Raz and G. Fein
University of California, San Francisco, and San Francisco VA Medical Center, San Francisco, CA 94121 (U.S.A.)

(Accepted for publication: 13 November 1987)

Summary  Signal power, noise power and their ratio (SNR) are important variables underlying estimation of evoked potential signals, yet, they are rarely explicitly considered in the design or analysis of EP experiments. A model is developed which relates the reliability of the average evoked potential (AEP) wave form to signal power, noise power, SNR, and the number of single trials included in the average. Measurements taken from auditory and visual EP experiments in elderly subjects show that noise power is highly reliable across experimental conditions and probably reflects global CNS anatomic or physiologic factors. In contrast, signal power and SNR are variable across conditions and sensory modalities, but are stable across replications. Thus signal power reflects CNS processes specific to the experimental paradigm. These results have importance for EP estimation. The expected reliability of the AEP cannot be adequately predicted from estimates of a subject’s noise power, or from SNR estimated under different experimental conditions. These findings suggest the need for on-line estimation of SNR during data acquisition to ensure adequate reliability of AEPs.

Key words: Average evoked potentials; Signal-to-noise ratio; Statistical analysis

The technique of averaging electrical potentials across multiple trials to estimate the evoked potential (EP) to a stimulus is based on 2 assumptions: (1) that the signal does not change across trials, and (2) that the background electrical activity has no time-locked relationship to the stimulus and is a random process with a mean potential of zero. Even when one accepts the sometimes questionable assumption that the underlying signal is homogeneous across trials, the average evoked potential (AEP) remains only an estimate of the true signal. The AEP still includes some residual EEG ‘noise’ as well as signal. The power of the signal, the power of the noise, and the signal-to-noise ratio (SNR), are potentially important variables reflecting CNS function and are critical for determining the reliability of AEP data. These variables, however, are rarely explicitly considered in the design or analysis of evoked potential experiments.

For a fixed number of trials in the average, the AEP is a more accurate reflection of the true EP when the EP signal is strong or when the power of the background EEG is small. However, if a particular subject, experimental protocol, or electrode location produces either a smaller evoked response or more power in the background EEG, then the AEP will contain relatively more noise and will less accurately estimate the underlying EP.

Averaging a uniform number of trials for each subject in an experiment assumes that the SNR is relatively constant for the experimental paradigm, and that the AEP is a reasonable portrayal of the underlying signal across subjects and electrode locations. However, before drawing conclusions from measures of the latency or amplitude of particular component wave forms, it is important to consider the reliability of the measures. If a measure is taken at a fixed latency (e.g., 300 msec), then the reliability depends on the standard deviation (SD) of the measure only for noise measure depends on the standard deviation of the measure (e.g., between 250 and 350 msec), the measure is a function of the AEP wave shape and mean AEP. If the reliability of the measure is increased, the statistical power of a test is increased.

We begin by developing noise power SNR and AEP wave shape data from a group of experiments. Separate EP experiments were conducted across subjects, different stimulus paradigms, and stimulus parameters to suggest how this model may be applied to data acquisition and analysis of EPs.

Methods

Theoretical considerations

Consider an experiment to estimate the average evoked potential 

\[ X_j(t) = \mu(t) + N_j(t) \]

where \( \mu(t) \) is the unknown signal and \( N_j(t) \) is random noise in trials \( j \). The AEP activity in each trial is assumed to be a random number of EPs, which is a function of the noise in any other trial. We define the total signal-to-noise ratio as

\[ \text{SNR} = \frac{\sigma^2}{\sigma_N^2} \]

where \( \sigma^2 = \frac{1}{T} \sum_{t=1}^{T} \mu(t)^2 \) is the measure of the standard deviation of the mean EP and \( \sigma_N^2 \) is the measure of the standard deviation of the noise in any other trial. The signal-to-noise ratio is used to determine the reliability of the AEP data. The variables, however, are rarely explicitly considered in the design or analysis of evoked potential experiments.
NOISE AND SIGNAL POWER IN EP MEASUREMENT

msec), then the reliability of the measure only depends on the standard error of the AEP, which is a function only of noise power. However, if the measure depends on the AEP wave shape to determine its latency (e.g., the peak amplitude between 250 and 350 msec), then the reliability of the measure is a function of both the reliability of the AEP wave shape and the standard error of the AEP. If the reliability is attenuated, a greater amount of random error is present in the data and the statistical power of any test of the experimental hypothesis is reduced.

We begin by developing a model for the relationship of noise power ($\sigma_N^2$), signal power ($\sigma_S^2$), SNR and AEP wave shape reliability. Then, using data from a group of elderly subjects tested in 3 separate EP experiments, we examine these measures across subjects, differing experimental conditions and stimulus modalities. Finally, we suggest how this model may lead to improved AEP data acquisition and analysis.

Methods

Theoretical considerations

Consider an experiment in which electrical potentials $X_j(t)$ are recorded at times $t = 1, 2, \ldots, T$ in each of $j = 1, 2, \ldots, J$ trials. We assume that:

$$X_j(t) = \mu(t) + N_j(t)$$

where $\mu(t)$ is the unknown true signal at time $t$ and $N_j(t)$ is random noise arising from background EEG activity and measurement error. We assume that the noise in any trial $j$ is a realization of a stationary random process with mean zero and power (i.e., variance) $\sigma_N^2$, and is independent of the noise in any other trial.

We define the total signal power across the $T$ time points as

$$\sigma_S^2 = \frac{1}{T} \sum_{t=1}^{T} \mu(t)^2$$

and the signal-to-noise ratio as

$$\text{SNR} = \frac{\sigma_S^2}{\sigma_N^2}.$$
function only of the stimulus conditions, then $J$ can be chosen based on the expected SNR for the stimulus condition and the desired reliability of the AEP wave form. Subjects and experimental conditions can then be compared using equally reliable estimates of the signal in all cases. However, if there is variation in SNR between subjects, electrode placements, or experimental conditions, the reliability of the AEP wave form based on a fixed value of $J$ and the power of statistical tests based on this standard AEP will be affected by these parameters. Further, for subjects or conditions in which the SNR is large, the EP wave form could be reliably estimated using relatively fewer trials, possibly allowing for simpler experimental protocols or for the successful acquisition of data from subjects who would otherwise have to be excluded. Conversely, for subjects or conditions where the SNR is small, EP measures may be unreliable unless $J$ is increased dramatically. Knowledge of the value of $J$ necessary to achieve reliable AEPs may be particularly relevant for studies of groups such as children, the elderly, or psychiatric patients where $J$ is limited by the subject's ability to cooperate with the experimental demands for extended periods of time.

Subjects

Twenty male and 18 female subjects between the ages of 65 and 70 were recruited for a study of neuropsychological and electrophysiologic correlates of normal aging. Subjects were screened on the basis of IQ, visual acuity, and medical and psychiatric history prior to data collection.

Procedures

Data were collected according to 3 separate protocols over the course of approximately 2 h.

(1) Visual pattern reversal evoked potentials (PREPs). Checkerboard-reversal visual stimuli with individual squares measuring 1.24 cm per side were presented at a viewing distance of 127 cm, each check subtending a visual angle of 36°. The full-field checkerboard pattern measured 21.2 cm × 13.6 cm, subtending 9.5° × 6.1° of visual angle. The average luminance of the black squares was 15.1 cd/m² and the average luminance of the white squares was 346.3 cd/m². The checkerboard pattern reversed once every second.

Electrodes were applied at O1, O2, Oz, and Cz referenced to Fz. EEG and electro-oculogram (EOG) were recorded from gold-cup electrodes and amplified with a Grass Model 7B polygraph. Horizontal eye movement greater than 1.4° of visual angle, eye blinks, or A/D converter saturation ($+/−125$ μV) resulted in trial rejection. Each channel of EEG was sampled every 2 ms, from 20 ms pre-stimulus to 380 ms post-stimulus. ERPs were computed separately for reversals from pattern A to pattern B and from pattern B to pattern A. Stimuli were presented until 80 trials without EOG artifact or A/D saturation were recorded for each condition.

(2) Two-tone auditory oddball paradigm. Auditory tones randomly selected to be either 1000 Hz or 2000 Hz were presented binaurally through headphones. Interstimulus interval varied between 1.5 and 1.6 sec. Stimulus duration was 40 ms; intensity was 50 dB. The probability of occurrence of the 1000 Hz standard tone was 0.8; the probability of the 2000 Hz tone was 0.2. Subjects were instructed to attend to the infrequent 2000 Hz target.

Data were recorded from Fz, Cz, Pz, and Oz referenced to linked ears. EOG was monitored as in the visual pattern-reversal experiment, and trials with excessive eye movement artifact or A/D saturation were excluded. Each channel of EEG was sampled every 4 ms, from 40 ms pre-stimulus to 960 ms post-stimulus on each trial. Stimuli were presented until 30 good target trials and a minimum of 80 good standard trials were collected.

(3) Three-tone auditory oddball paradigm. The protocol was identical to that of the 2-tone experiment, except that a third tone of 500 Hz was also presented to the subjects. The relative probabilities of the 3 tones were 0.70 for the 1000 Hz tone, and 0.15 for each of the other two. Eighteen of the subjects were asked to respond to the 2000 Hz tone, as in the 2-tone experiment, while for the other 20, the 500 Hz tone was the designated target. Data were collected until there was a minimum of 30 good trials in each of the rare conditions and 80 good tone.

Computations

For the visual experiment to the 400 msec of data on each of 80 trials, the epoch from msec post-stimulus, with the principal N100–P100 standard tone, was an vertex electrode. In ac target condition were epochs, from stimulus stimulus (N100–P200 msec post-stimulus, P300 signal was expre restricted to the first 8 the first 30 targets, sc tal condition, all subj number of trials for $r$.

In an experimental power are known; estimators are commc

$$\hat{\delta}_N^2 = \frac{1}{I(J-1)} \sum_{i=1}^{J} \left( \bar{X}_i^2 - \frac{1}{J} \hat{\delta}_S^2 \right).$$

These estimators by:

$$\hat{\delta}_N^2 = \frac{\hat{\delta}_S^2}{S^2}$$

and predicting the $\hat{\rho}$

$$\hat{\rho} = \frac{1}{1 + \frac{1}{J(S\hat{\delta}_N^2)}.$$
3.1. TURETSKY ET AL.

The luminance of the background was 0.2. The checkerboard stimulus was presented to the left or right of the fixation point on each trial.

The checkerboard stimulus was a black-and-white square pattern with a size of either 1 or 2 cycles per degree (cpd) and a contrast of 0.8. The stimulus was presented for 2000 msec post-stimulus, which was expected to include the principal N100-P200 response to the 1000 Hz standard tone, was analyzed using data from the vertex electrode. In addition, vertex data from the target condition were examined in two separate epochs, from stimulus onset until 280 msec post-stimulus (N100-P200), and also from 280 to 560 msec post-stimulus, where the late endogenous P300 signal was expected. All computations were restricted to the first 80 artifact-free standards and the first 30 targets, so that within any experimental condition, all subjects would have an identical number of trials for reliability estimates.

In an experimental setting the signal and noise power are unknown. The following unbiased estimators are commonly used (Möcks et al. 1984):

\[
\hat{\delta}_N = \frac{1}{T(J-1)} \sum_{j=1}^{T} \left( \sum_{i=J+1}^{T} (X_{ij} - \overline{X}_j)^2 \right)
\]

\[
\hat{\delta}_S = \frac{1}{T} \sum_{i=1}^{T} \overline{X}_i^2 - \frac{1}{J} \hat{\delta}_N
\]

These estimators suggest estimating the SNR by:

\[
\text{SNR} = \frac{\hat{\delta}_S}{\hat{\delta}_N}
\]

and predicting the reliability of the AEP by:

\[
\hat{\rho} = \frac{1}{1 + \frac{\text{SNR}}{J(S&R)}}
\]

Since our data sets included replications of subject responses to certain experimental conditions, we were also able to measure the reliability of the AEP by computing the cross-correlation between the two replications.

Results

We first assessed the validity of our model by comparing the predicted AEP reliability, \(\hat{\rho}\), with the cross-correlation computed from the reliability of both the auditory standard and visual checkerboard conditions. The scatterplots of predicted and measured reliabilities for visual and auditory experiments are presented in Fig. 1. For the visual paradigm, the \(z^2\)-transformed predicted and measured reliabilities were correlated at 0.75; for the auditory experiment the corresponding correlation was 0.72. The model thus provided a close approximation to the real data. We therefore applied the model to the analysis of signal and noise variability across subjects and conditions.
Table I presents the means and standard deviations of noise power ($\sigma^2$) and signal power ($\delta^2$) separately for males and females for each condition.

**Power of the noise**

We examined the variability of $\sigma^2$ across experiments and were surprised by its marked stability within an individual subject, regardless of the stimulus modality or cognitive task. Fig. 2 presents scatter plots that compare $\sigma^2$ for subjects under different experimental conditions. The correlation between the 2 auditory standard conditions (Fig. 2a) was 0.97, but even under the most disparate circumstances — when an Oz recording in response to a visual stimulus was compared to a vertex recording for an auditory stimulus (Fig. 2c) — the correlation of $\sigma^2$ was 0.86; subject differences accounted for fully 74% of the variance in $\sigma^2$. This suggests that $\sigma^2$ is primarily a function of certain fixed physiologic or anatomic characteristics of the individual, such as skull thickness, rather than a product of experimental variables. Further, there was a high intersubject variance in $\sigma^2$, with maximum values exceeding minimum values by a factor of 9. There was also a sex difference in $\sigma^2$; the males in our sample had significantly lower $\sigma^2$ than the females across all conditions ($P < 0.05$).

**Power of the signal**

We compared the estimates of $\delta^2$ across the various experimental paradigms and found that intersubject differences in $\delta^2$ within an experimental paradigm were highly reliable but that $\delta^2$ varies within a subject across paradigms. The high reliability of $\delta^2$ within an experimental paradigm is illustrated by intra-class correlations of $\delta^2$ of 0.95 across the 2 visual PREP replications, and 0.92 across the 2 auditory standard replications. In contrast to the results obtained for $\sigma^2$, the cross-paradigm correlations of $\delta^2$ were very low. As an example, the largest intra-class correlation of $\delta^2$ between visual and auditory experiments was only 0.16; the comparable correlation for $\sigma^2$ was 0.91. Within the auditory modality, the situation was similar. For the 2-tone oddball target condition, the intra-class correlation of $\delta^2$ between the N100–P200 and P300 intervals was 0.09, while the corresponding correlation for $\sigma^2$ was 0.39. As a direct consequence of the across-paradigm stability of $\sigma^2$ and variability of $\delta^2$, the relationship between $\delta^2$ and $\sigma^2$ across subjects varied from paradigm to paradigm. The correlation between $\sigma^2$ and $\delta^2$ ranged from 0.57 for the 3-tone, 2 replications; (b) 0.75 for the 2-tone, 2 replications; (c) auditory standard to 0.01 for a given subject. The results above (SNR) will vary for a given subject,
The results above imply that the observed SNR (S/N) will vary across experimental conditions for a given subject, because of changing $\delta^2_N$ while $\delta^2_S$ remains stable. The variability of S/N across subjects within an experiment is due to variation in both $\delta^2_S$ and $\delta^2_N$. This is illustrated by the plots in Fig. 3, which show that S/N was very stable for each subject during the replication of an experiment (Fig. 3a), but only modestly associated with the S/N under other experimental conditions (Fig. 3b, c). As the S/N directly determines the minimum number of trials that are needed to achieve any specified reliability of the AEP waveform, that number of trials will vary across paradigms for any particular subject.

APE wave shape reliability

We selected an AEP reliability of $\hat{\rho} = 0.90$ and estimated the number of single trials ($J$) required...
to obtain an AEP with this reliability for each subject and condition. Table II presents the mean and standard deviation of $J$, separately for males and females, for each stimulus. The results indicate the large variance of $J$ across subjects in each condition. For some subjects, highly reliable AEPs were obtained after only very few single trials, while other subjects required literally hundreds of single trial exposures to the same stimulus to achieve a predicted reliability of 0.90 in the AEP wave form. As a direct consequence of the $\delta^2$ results presented above, the largest correlation between the minimum number of trials across paradigms was only 0.23, implying that different subjects were producing the more unreliable wave forms under different conditions. Even within the same auditory target condition, there was no correlation between the number of trials needed to achieve a predicted reliability of 0.90 in the N100–P200 and P300 intervals.

To determine the degree to which prior knowledge of an individual's noise power alone could be utilized to improve the estimated AEP reliability, we reanalyzed the data as follows: we first used the mean $\delta^2_J$ and $\delta^2_N$ for each experiment to determine the minimum number of trials needed to produce a predicted AEP reliability of 0.90. Assuming an experimental protocol with this fixed number of single trials, we then calculated the expected AEP reliability for each subject and condition, using the subject's $\delta^2_J$ and $\delta^2_N$. To remove variation due to noise power, we repeated this analysis, but used each individual's own $\delta^2_J$, and the mean $\delta^2_N$, to select the number of single trials for each subject. We then recalculated the predicted reliability of the AEP for each subject based on that subject's number of trials and examined the variability of the estimated AEP reliabilities. For the PREP paradigm, the use of an individual's noise power to choose $J$ decreased the variance of the estimated AEP wave form reliability by 18.5%. For the N100–P200 auditory data, the comparable reduction was only 7.5% for the standard tone, and 8.0% for the target. However, for the P300 data, choosing $J$ based on an individual's noise power had no impact on the variance of the estimated AEP wave form reliability (i.e., a reduction of only 0.7%). These results demonstrate that controlling for variation in $\delta^2_N$ alone does not offer much improvement in the AEP wave form reliability.

### Discussion

Estimation of SNR in EP data has two sources of error. First, there is the usual sampling error that arises from choosing a particular set of trials from the population of all possible trials. Second, there is error in partitioning the data into signal and noise. This second source of error is increased when the EP is heterogeneous over trials. Despite these two sources of error, our predicted reliability was highly correlated with the actual cross-correlation computed from experimental replications, providing evidence for the validity of our model of AEP wave shape reliability.

The power of the noise is a stable attribute of an individual during a single testing session, even when one varies the sensory modality of stimulus presentation, the particular stimulus or cognitive conditions, or the time window relative to stimulus onset. While our data cannot address the ques-
of trials needed to achieve reliability of 0.90. We calculated the reliability of the individual's own \( \sigma_n^2 \), and the reliability of single trials based on the prediction of the subject's own \( \sigma_n^2 \), and the reliability of single trials based on the prediction of the subject's own \( \sigma_n^2 \). We examined the reliability of our model of individual's noise power stability across testing sessions. The finding that \( \sigma_n^2 \) is significantly higher for females across all conditions possibly reflects gender differences in these underlying anatomic or physiologic factors.

In contrast to noise power, signal power varies depending on the particular modality of stimulation or the task demands that are involved. Our estimate of \( \sigma_i^2 \) had a low correlation across the different protocols. However, within an experimental paradigm, \( \sigma_i^2 \) is highly reliable across replications, suggesting a relatively invariant subject response to consistent stimulus parameters and task demands. Signal power must therefore be determined, at least in part, by specific processing variables rather than by global aspects of CNS activity. These variables are presumably different for auditory and visual stimuli, as well as for stimuli with and without target value. In our sample, males and females differ in signal power only in response to visual, but not to auditory stimuli. This gender difference in the EP to visual pattern reversal signal power persists even after controlling for the variability of noise power argues against attributing this sex difference to anatomic factors such as scalp musculature or skull thickness as others have done (Halliday 1982). A detailed analysis of our data from the visual pattern reversal experiment has been presented elsewhere (Fein and Brown 1987).

Intersubject differences in \( \sigma_n^2 \) and \( \sigma_i^2 \) directly affect the estimation of evoked potential wave forms. The wide variation in SNR means that for a fixed number of trials, the AEP wave form will better estimate the true signal for some subjects than for others. This insight would enable one to identify individuals or classes of subjects for whom the AEP is a poorer estimate, and to consider this fact when assessing the results of hypothesis tests. Before concluding that 2 cohorts did not differ in their evoked responses to a stimulus, it would be important to know how representative the individual AEPs are of the underlying EP signals. If, for a particular experimental condition, time window, or subject group, the AEPs are relatively unreliable estimates, then there will be more error in the selection of component latencies and amplitudes for each subject, and it is less likely that real group differences in the EPs would be detected. The exclusion or inclusion of data of low reliability could alter test conclusions. One approach would be to weight each subject's data in the analysis either inversely by \( \sigma_n^2 \), or directly by SNR, depending on whether the dependent variable was measured at a fixed latency or at a latency derived from the AEP wave shape.

Ideally, the question of expected AEP reliability should be addressed prior to data acquisition, so that an experimental protocol can be modified or extended as needed to assure adequate data. Our results, though, suggest that SNR cannot be adequately estimated based on such demographic variables as age or sex. Neither can it be predicted from the power of an individual subject's background EEG (i.e., noise power), or from prior estimation of SNR obtained under different experimental conditions. We are in the process of developing a method for the on-line estimation of SNR during the computer acquisition of single trial electrical potentials. Such a procedure would allow an experiment to continue until a specified minimum AEP reliability has been obtained. It could also be adapted to focus on the predicted reliability of particular time segments or conditions, such as the P300 response to oddball targets.

Appendix

Cross-correlation of AEPs approximates the intraclass correlation coefficient

Suppose an evoked potential experiment is performed twice, using J trials each time. Let \( \mu_j(t) \) be the noise at time t in trial j in experiment i, for \( i = 1, 2 \). We assume that the signal \( \mu(t) \) is identical in the 2 experiments and that the noise in experiment 1 is independent of the noise in experiment 2. Let \( \bar{X}_i(t) \) denote the average of the trials in
experiment i. We define the cross-correlation between $x_1(t)$ and $x_2(t)$ as:

$$r = \frac{\sum_{t=1}^{T} \bar{x}_1(t) \bar{x}_2(t)}{\sqrt{\sum_{t=1}^{T} \bar{x}_1(t)^2} \sqrt{\sum_{t=1}^{T} \bar{x}_2(t)^2}}.$$ 

Thus $E(r)$ is approximately equal to

$$\frac{\sigma_S^2}{\sigma_S^2 + \frac{\sigma_N^2}{J}},$$

which we defined as the intra-class correlation coefficient $\rho$.

References


