ELSEVIER

Clinical Neurophysiology 112 (2001) 555–562

www.elsevier.com/locate/clinph

Weighted averaging of steady-state responses

M. Sasha John, Andrew Dimitrijevic, Terence W. Picton*

Rotman Research Institute, Baycrest Centre for Geriatric Care, University of Toronto, 3560 Bathurst Street, Toronto, Ontario, Canada M6A 2E1

Accepted 18 December 2000

Abstract

Objective: To compare weighted averaging and artifact-rejection to normal averaging in the detection of steady-state responses.

Methods: Multiple steady-state responses were evoked by auditory stimuli modulated at rates between 78 and 95 Hz. The responses were evaluated after recording periods of 3, 6 and 10 min, using 5 averaging protocols: (1) normal averaging; (2) sample-weighted averaging; (3) noise-weighted averaging; (4) amplitude-based artifact-rejection; and (5) percentage-based artifact rejection. The responses were analyzed in the frequency domain and the signal-to-noise ratio was estimated by comparing the signals at the modulation-frequencies to the noise at adjacent frequencies.

Results: Weighted averaging gave the best signal-to-noise ratios. Artifact-rejection was better than normal averaging but not as good as weighted averaging. Responses that were not significant with normal averaging became significant with weighted averaging much more frequently than vice versa. False alarm rates did not significantly differ among the protocols. The advantage of weighted averaging was especially evident when stimuli were presented at lower intensities or when smaller amounts (e.g. only 3 or 6 min) of data were evaluated. Weighted averaging was most effective when the background noise levels were variable. Weighted averaging underestimated the amplitude of the responses by about 2%.

Conclusion: Weighted averaging should be used instead of normal averaging for detecting steady-state responses.

Keywords: Signal-to-noise ratio; Weighted averaging; Artifacts; Steady-state responses; Frequency analysis; Objective audiometry

1. Introduction

Sensory evoked potentials (signals) recorded from the human scalp are usually unrecognizable in the background EEG (noise). Averaging is commonly used to increase the signal-to-noise ratio so that evoked potentials can be detected and measured. If (i) the signal remains constant from trial to trial, (ii) the noise on any one trial is uncorrelated with the noise on other trials, and (iii) the noise statistics remain stationary from trial to trial, averaging increases the signal-to-noise ratio by a factor equal to the square root of the number of trials averaged. If the noise varies from one trial to the next, averaging is less effective. Two techniques can then be used to improve the process. The first is to reject from the averaging those trials wherein the noise is higher than some criterion – 'artifact-rejection' (Picton et al., 1983; Pantev and Khvoles, 1984). The second is to weight the recorded data according to its variance prior to summation, and then to divide by the sum of the weights – 'weighted averaging' (Hoke et al., 1984; Lütkenhöner et al., 1985).

Steady-state responses occur when the frequency constituents of a response are stable in amplitude and phase (Regan, 1989). These responses are usually evoked by periodic stimuli and measured at the frequency of stimulation or one of its harmonics. Although the responses are most effectively measured in the frequency-domain, the time-domain waveforms are often averaged prior to conversion to the frequency-domain. Dobie and Wilson (1994) showed that weighted averaging improved the detection of auditory steady-state responses to 40-Hz stimuli compared to normal averaging.

Our present paper investigates the use of weighted averaging for recording auditory steady-state responses at faster stimulus rates (78–95 Hz), and compares weighted averaging to artifact-rejection. We used two kinds of weighting – one based on the variance of the time-domain waveform ('sample weighting', i.e. the whole sample including both the responses and the noise) and the other based on the power in the spectrum but excluding the power at the signal frequencies ('noise weighting'). In addition we used two kinds of artifact rejection – one based on an absolute ampli-

* Corresponding author. Tel.: +1-416-785-2500 ext 3505; fax: +1-416-785-2862.

E-mail address: picton@psych.utoronto.ca (T.W. Picton).

1388-2457/01/$ - see front matter CLINPH 2000149

PII: S1388-2457(01)00456-4

tude criterion and the other based on a percentage of the trials with the highest amplitudes. We based both the weighting factors and the artifact-rejection criteria on the frequencies in the recorded data (70–110 Hz) that were close to the response frequencies.

2. Methods

2.1. Steady-state responses

We recorded the steady-state responses using the MASTER system (John and Picton, 2000; see also www.hearing.cjb.net). The system presented 8 simultaneous tones, which were modulated at rates between 78 and 95 Hz, using either amplitude-modulation (AM) or mixed modulation (MM). In MM both amplitude-modulation and frequency-modulation occur at the same modulation rate and therefore produce a single steady-state response. The stimuli are described in greater detail in previous papers (e.g. Cohen et al., 1991; Lins et al., 1996; John and Picton, 2000; 2001). For the purpose of this paper all that the reader needs to know is that the stimuli were periodic and continuous.

Responses were recorded between Cz and the neck with an AD conversion rate of 1000 Hz. An electrode placed over the left clavicle served as ground. The analog filter bandpass for recording these data was 1–300 Hz. Epochs of 1024 data points (1.024 s) were rejected if the amplitude at any point in the epoch exceeded ± 80 μV. As well as evaluating the data on-line, the MASTER system stored the data in continuous disk files. The stored data were analyzed off-line using MATLAB programs. Sixteen individual data epochs of 1024 points each were collected and linked together into a sweep lasting 16.384 s. As each sweep was completed, it was added to a running average, and the final average sweep was transformed into the frequency domain by means of a fast Fourier transform (FFT).

The data set contained two recording conditions during which subjects with normal hearing were presented with 8 different stimuli (binaural stimulation, 4 to each ear) all of which were either AM or MM (John et al., 2001). Each of the two conditions contained 3 separate recordings (12 sweeps each), which were obtained at each of 3 intensities (30, 40 or 50 dB SPL). Since both AM and MM responses should be similarly affected by the averaging protocols, we collapsed the data across stimulus type, ear of presentation, and carrier-frequency. We evaluated the responses at each intensity after combining 1, 2 and 3 recordings, i.e. after 12, 24 and 36 sweeps. This was equivalent to evaluating the responses after 3.2, 6.4 and 9.6 min of recording. This approach allowed us to examine the effects of weighted averaging and artifact-rejection as the analysis included increasing amounts of data (and decreasing amounts of background noise). Twenty-two separate recording sessions were analyzed giving us 178 separate recordings.

Since most of our subjects were able to sleep during the recording period, the noise levels were quite low. The data were only occasionally contaminated by high noise due to movement or ongoing activity of the scalp-muscles. However, some subjects awoke more frequently during the recording period or tended to make movements more than others. Even a relatively small duration of high noise can cause responses, that were significant early during the recording, to become non-significant at the end of the testing period.

2.2. Averaging protocols

2.2.1. Protocol 1 – normal averaging

Sweeps were formed by concatenating 16 adjacent epochs of data. An average sweep was created by summing together N individual sweeps and dividing these values by N.

2.2.2. Protocol 2 – sample-weighting

The weighting factor was based upon the frequencies near those of the responses, rather than upon the higher-amplitudes which occur in the lower frequencies of the EEG. Accordingly, we initially filtered each sweep of data using a digital first-order Butterworth filter with a bandpass of 70–110 Hz applied in both forward and backward directions (giving a final filter slope of 12 dB/octave) to prevent phase-distortion. Each data epoch (1.024 s) within a filtered sweep (16.384 s) was then weighted by dividing all the values in the epoch by the estimated variance of that epoch. As in the case of normal averaging, the weighted epochs were linked together to form sweeps, and the sweeps were then added together to form a summed sweep. Each epoch of the final summed sweep was then divided by the sum of the weights of the epochs that had been combined to form that particular epoch. The formulae (adapted from Lütkenhöner et al., 1985) were as follows:

$$a(i) = \sum_{j=1}^{N} w_j x_j(i) \tag{1}$$

where a is the weighted average waveform across the time-points (i) of the epoch, N is the number of epochs being summed together, and w_j is the weighting factor for the jth epoch:

$$w_j = \xi_j^{-2} \left(\sum_{k=1}^{N} \xi_k^{-2}\right)^{-1} \tag{2}$$

where ξ^2 is an estimate of the variance of the epoch:

$$\xi_k^2 = \left(\sum_{i=1}^{M} x_i^2\right)/M \tag{3}$$

where M is the number of points in the epoch and the mean of the epoch was zero because of the filtering.

2.2.3. Protocol 3 – noise-weighting

In this process, each unfiltered epoch was transformed to the frequency-domain using an FFT. We then computed the average power between 70 and 110 Hz after removing the power at the 8 frequencies at which responses occurred (and 4 other control frequencies). Whereas sample-weighting was based on both the signal and the noise, this noise-weighting protocol was based only on the noise and was uncontaminated by any signal. The time-domain epoch was then weighted and concatenated with the preceding epochs to form sweeps and the final average sweep computed as was done for the sample-weighted average.

2.2.4. Protocol 4 – amplitude-rejection

In the fourth protocol we used fixed amplitude-level for rejecting trials from the averaging process. Epochs in which the amplitudes of the unfiltered data had exceeded ± 80 μV had already been rejected. This protocol (and the next) rejected further epochs on the basis of their amplitudes within the 70–110 Hz frequency-range. The sweeps were filtered (as described for the sample-weighting protocol) and then the root-mean-square value of the waveform was calculated for each epoch. An epoch was rejected if this value exceeded 1.8 μV. This value was chosen after visually examining histograms of the noise levels of a large subset of recordings and selecting the point at which the values appeared to deviate from a normal distribution. If an epoch of data was rejected, its place was taken by the next acceptable epoch. Although the rejection criterion was based on the amplitudes of the filtered data, unfiltered data were used for averaging. The final sweep was averaged on an epoch-by-epoch basis since the number of epochs later in the sweep could be one less than at the beginning of a sweep. For example, if the total number of recorded epochs was 192 (sufficient for 12 full sweeps), and if 30 of these were rejected, the final sweep would be the sum of 11 for the first two epochs and the sum of 10 for the subsequent epochs.

2.2.5. Protocol 5 – percentage-rejection

The final method was the same as that described for amplitude-rejection, except that we varied the criterion for artifact-rejection for each recording so that the 25% of the epochs with the highest root-mean-square values were rejected. This protocol required calculating the root-mean-square values for all epochs, determining the 25% criterion and then only including those epochs under this criterion in the final average sweep.

2.3. Evaluation of the signal-to-noise ratio

The amplitude spectrum of the final sweep showed the steady-state responses at the frequencies equal to the modulation rates of the carrier-frequencies. An estimate of the background noise was obtained from frequencies where no stimulus occurred. We estimated the signal-to-noise ratio by comparing the power at each stimulus-frequency to the power at 120 nearby frequencies (60 above and 60 below the stimulus-frequency). Since the spectra were derived from a sweep lasting 16.384 s, power measurements were available at a resolution of 1/16.384 or 0.061 Hz. The noise estimates therefore came from 3.7 Hz (i.e. 0.061×60) above and below the frequency at which the steady-state signal appeared. The significance of this ratio can be assessed through the F-distribution with 2 and 240 degrees of freedom (Zurek, 1992; John and Picton, 2000). Prior to statistically comparing these ratios across the different protocols, we normalized the ratios by taking their square root, effectively using an amplitude-based rather than power-based signal-to-noise ratio (SNR). In addition to checking the responses at the stimulus-frequencies, we also evaluated the significance of the responses at 4 'control' frequencies at which no stimuli occurred to obtain an estimate of false positive rates for each averaging protocol.

2.4. Statistical analyses

The effects of the different protocols on the signal-to-noise ratios were assessed using an ANOVA with repeated measures across subjects. We used a 3-way ANOVA (protocol X time X intensity). The time variable was equivalent to the number of sweeps available for analysis (12, 24, 36 sweeps) although in the rejection protocols, the number of sweeps actually used in the analysis was reduced. Since occasional data were missing, the degrees of freedom were reduced. Greenhouse–Geisser corrections for the probability levels were used when appropriate.

We calculated a rough measure of the positive skewness of the distribution of the epoch-amplitudes by dividing the difference between the 90th percentile and the median by the difference between the median and the 10th percentile. We then correlated this measure of the 'noisiness' of the recording with the SNR improvement caused by the sample-weighted averaging (a ratio of the SNR for weighting to the SNR for normal averaging) using a Pearson product-moment correlation coefficient.

Finally, we assessed the effects of the protocols upon the incidence of significant responses using χ^2 statistics. To test that incidences were equal (e.g. that the number of responses becoming significant was the same as the number losing their significance), we used simple observed-minus-expected calculations. To compare incidences between protocols, we used 2×2 contingency tables with appropriate adjustments of the χ^2 calculations.

3. Results

3.1. Illustrative data

Figs. 1 and 2 show data recorded from two recording sessions in one subject. Fig. 1 shows the histograms of the

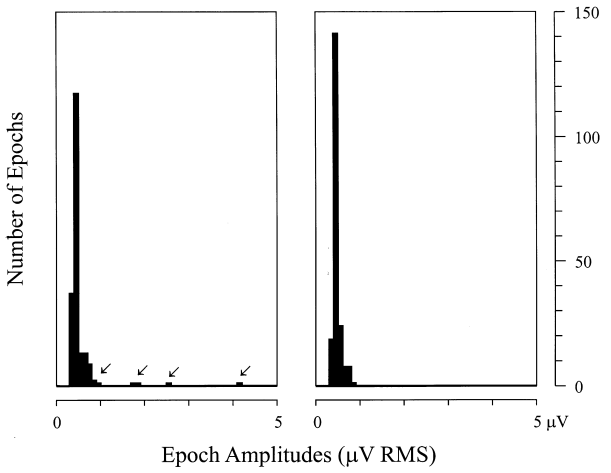

Fig. 1. Variation in noise levels over epochs. This figure shows the analysis of two recording sessions from the same subject, presented with the same set of stimuli (eight MM tones at 50 dB SPL). The sessions were analyzed separately, rather than cumulatively as in the data for the statistical analysis, to provide a comparison of the analysis protocols with all things constant (subject, stimuli, recording time) except the noise in the recordings. The figure shows histograms of the amplitudes of the recorded data in the epochs for the two recording sessions. The first recording contained occasional epochs with high levels of noise (arrows). The second recording did not contain any high-noise trials.

root-mean-square amplitudes for the filtered epochs (70–110 Hz). The data from the first recording session (on the left) contained epochs that were intermittently contaminated with high-amplitude noise (arrows). The distribution is skewed toward the higher amplitudes. The subject was more consistently quiet during the second recording. The histogram shows that the epoch amplitudes are normally distributed, Fig. 2 displays the steady-state responses in the frequency-domain for 3 of the protocols for the two recording sessions with the epoch amplitudes shown in Fig. 1. Both sample-weighting and amplitude-rejection increased the number of significant responses (indicated by the filled triangles) compared to normal averaging in the first recording session (left). The 3 protocols performed similarly for the second session (right).

3.2. Signal-to-noise ratios

Fig. 3 shows the average signal-to-noise ratios after analyzing the equivalent of 12, 24 or 36 recordings (collapsed across intensity level) for the 5 different averaging protocols. In relation to the major experimental hypothesis, the ANOVA showed a significant main effect of averaging protocol ($F = 20.4$; d.f. $= 4, 72$; $P < 0.001$). As expected, the signal-to-noise ratio was greater after a longer period of analysis ($F = 95.5$; d.f. $= 2, 36$; $P < 0.001$) and at higher intensity ($F = 22.5$; d.f. $= 2, 36$; $P < 0.001$). There were significant interactions between protocol and time (the protocol effects being larger after a longer period of analysis) and between intensity and time (the intensity effects being larger after a longer period) but the other interactions were not significant. Post hoc evaluations for the protocol effect showed that sample-weighting

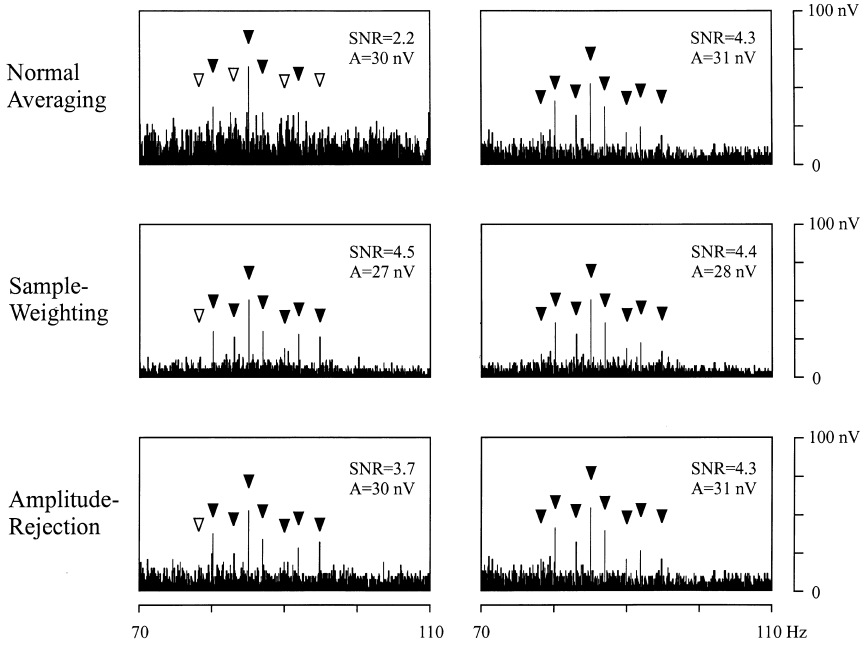

Fig. 2. Weighted averaging and artifact rejection. This figure shows the responses obtained during the two recording sessions described in Fig. 1. The responses are plotted in the frequency domain for 3 of the analysis protocols. In the first recording session where there were occasional high-noise trials, both sample-weighting and amplitude-rejection (at 1.8 μV) improved the signal-to-noise ratio and increased the number of responses recognized as significant (filled triangles as opposed to open triangles). The average amplitudes (A) and signal-to-noise ratio (SNR) across the 8 stimuli are given together with the graphical plot of the spectrum. In the second recording session (right column), all 3 protocols performed similarly.

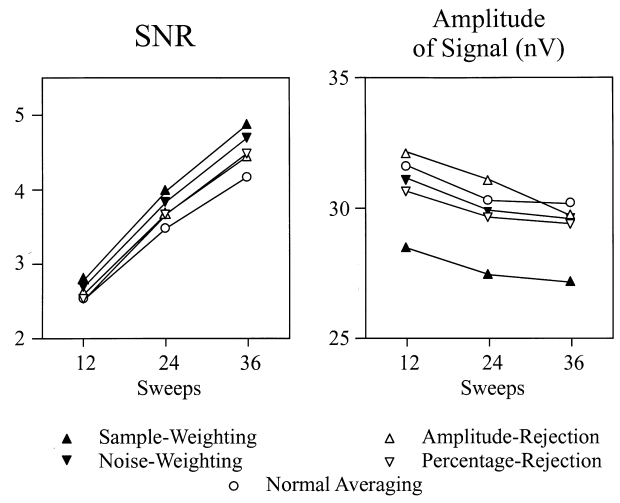

Fig. 3. Signal-to-noise ratios and signal amplitudes. The left graph shows the increase in the signal-to-noise ratio (SNR) obtained using each analysis protocol after the equivalent of 12, 24 and 36 sweeps were analyzed. The data have been collapsed across stimulus intensity. A clear advantage emerges for the weighting protocols over normal averaging with artifact rejection falling between. The right graph shows the estimated signal amplitude. This decreases slightly with increasing time and is consistently smaller for the sample-weighting protocol than the others.

significantly and consistently provided higher signal-to-noise ratios than all of the other protocols ($P < 0.01$ for noise-weighting; $P < 0.001$ for the others). Noise-weighting was significantly better than the artifact-rejection protocols and normal averaging. Neither of the artifact-rejection protocols was better than normal averaging after the shortest period of analysis but both became better after the second and third time period.

The amount of increase in the signal-to-noise ratio brought about by the sample-weighting protocol was significantly correlated with the noisiness of the recordings as assessed using our rough measure of the positive skewness for the epoch histograms ($r = 0.43$, $t = 27.3$, d.f. $= 57$, $P < 0.001$).

3.3. Amplitudes

Fig. 3 also demonstrates that the amplitude of the responses varied significantly with protocol ($F = 18.5$; d.f. $= 4, 72$; $P < 0.001$). Post-hoc testing showed that the amplitude with the sample-weighting protocol was smaller by about 10% compared to the amplitude with the other protocols. The increase in the signal-to-noise ratio with sample-weighting was therefore due to a larger decrease in the noise than in the signal. The amplitude showed the expected increase with increasing intensity ($F = 77.9$; d.f. $= 2, 36$; $P < 0.001$). There was also a decrease in the estimated signal amplitude with increasing time for analysis ($F = 5.5$; d.f. $= 2, 36$; $P < 0.05$).

3.4. Response detection

When the signal-to-noise ratio increases, more responses can be recognized as significantly different from noise. The number of responses which shifted from non-significant to significant (using a $P < 0.05$ criterion) and vice versa when using the weighting or rejection protocols rather than normal averaging are shown in Table 1. A total of 339 responses became significant while 74 that were significant became not significant ($\chi^2 = 170.0$, d.f. $= 1$, $P < 0.001$). Each of the protocols showed significantly more responses becoming significant than losing significance. The two weighting conditions were not significantly different from each other but were significantly better than the two artifact rejection protocols ($\chi^2 = 14.6$, d.f. $= 1$, $P < 0.001$). In

Table 1
Detection of significant responses[a]

Number of sweeps		12		24		36	
Intensity	Protocol	S → N	N → S	S → N	N → S	S → N	N → S
30 dB SPL	SW	1	13	1	23	3	13
	NW	1	17	1	20	3	9
	AR	0	6	1	12	3	5
	PR	5	5	6	14	5	11
40 dB SPL	SW	2	12	1	12	0	15
	NW	5	11	3	13	0	14
	AR	0	8	1	9	1	6
	PR	2	9	2	10	1	8
50 dB SPL	SW	2	13	0	3	1	4
	NW	2	13	0	4	2	4
	AR	4	7	1	1	4	1
	PR	8	10	0	3	2	1

[a] This table shows the number of responses that became either significant (N → S) or insignificant (S → N) by applying the weighted averaging or artifact-rejection protocols compared to normal averaging for the first data set. The protocols are sample-weighting (SW), noise-weighting (NW), amplitude-rejection (AR) and percentage-rejection (PR). The total number of responses detected as significant with normal averaging increased with intensity and with the number of sweeps analyzed. The number of significant responses (out of a possible 176 – 8 stimuli, 22 recording sessions) was 62 (35%) at 30 dB after 12 sweeps and 157 (89%) at 50 dB after 36 sweeps.

general, the number of responses becoming significant is larger with weighted averaging than with artifact-rejection, is larger when fewer sweeps have been analyzed (i.e. when the noise is still high) and is larger when the intensity is low (i.e. when the responses are smaller).

The rate of false positive results was assessed by determining the significance of measurements at 4 frequencies where there were no stimuli. The overall rates on the first data set were 4.3, 6.3, 5.3, 5.8 and 5.2% for the protocols 1–5, respectively. None of these rates were significantly different from the expected rate of 5% (e.g. for the sample-weighting data, $\chi^2 = 2.87$, d.f. $= 1$, $0.05 < P < 0.10$). However, the difference between rates for sample-weighting (protocol 2) and normal averaging (protocol 1) was borderline significant ($\chi^2 = 3.72$, d.f. $= 1$, $P = 0.054$). When we examined the rates for each of the 4 control frequencies tested in the sample-weighting protocol, we found that the greatest rates occurred at the lowest and highest of the control frequencies (75 and 105 Hz).

4. Discussion

The results show clearly that weighted averaging and artifact-rejection improve the signal-to-noise ratio compared to normal averaging when recording human auditory steady-state responses to multiple stimuli presented at rates between 80 and 100 Hz. The improvement in the signal-to-noise ratio with weighted averaging varies with the positive skewness of the distribution of the root-mean-square amplitudes of the epochs being analyzed. Weighted averaging has little effect when the noise-distribution is not skewed.

Weighted averaging protocols were also significantly better than artifact-rejection protocols. This depends upon the distribution of the noise from epoch to epoch. If the noise were consistently either very large or very small, the protocols would probably perform similarly, since the high-noise trials would be reduced by the weighting to an extent that their contribution to the final result would be the same as if they were rejected. However, this does not usually occur when recording auditory steady-state responses. In this case, the noise is often more widespread in its amplitude-distribution and weighted averaging performs better than artifact-rejection. For our particular data set, epochs with very high amplitude noise had already been rejected prior to weighting or further artifact-rejection.

In general, epochs that are rejected from averaging by rejection protocols would be those that are most attenuated by the weighting protocols. However, rejection protocols often operate on the basis of different rules than the variance of an epoch. For example, many rejection protocols are based on the maximum amplitude within an epoch. In this case, weighted averaging and artifact rejection might have differed more than they did in the present study.

One of the difficulties with artifact-rejection is selecting which trials to reject. We used two approaches – rejecting epochs with amplitudes higher than an absolute criterion, and rejecting the epochs with the highest amplitudes relative to that particular recording session. In both cases we calculated the amplitudes after filtering the data to eliminate frequencies irrelevant to the signals we were seeking. Rejecting data on the basis of the EEG amplitudes at the lower frequencies is not optimal when evaluating the auditory steady-state responses. Mühler and von Specht (1999) suggest sorting the recorded epochs from low to high noise-amplitudes and then averaging the trials in order of the amount of noise until the signal-to-noise ratio of the recording starts to decrease when additional trials are included. This elegant approach works well when a set number of recorded epochs have been collected and the computer performs an offline analysis, but would be computationally very demanding if used online. Weighted averaging has a clear advantage over this approach and over the percentage-rejection procedure (our protocol 5) in that in can easily be performed online.

Weighted averaging requires choosing a weighting factor. The variance of a recording epoch is most commonly used. We restricted the variance estimate to the frequency-range that we were interested in (70–110 Hz) rather than the frequencies that we recorded (1–300 Hz). The contribution of the signal to this variance may be removed so that the weighting is only based on noise. Gerull et al. (1996) proposed a nice technique of subtracting one epoch from another to eliminate the signal and leave an estimate of the noise for the paired epochs (cf. the (±) reference of Schimmel, 1967). Eliminating the response prior to determining the weighting factor is easier for a steady-state response than for transient responses since one just has to eliminate a specific component (or set of components) from the spectrum.

Dobie and Wilson (1994) found that eliminating the signal from the estimate used to determine the weighting factors did not improve the beneficial effect of the weighting. This is probably related to the low signal-to-noise ratio in unaveraged recordings of the auditory steady-state response. If the signal were large and/or variable, it would be worthwhile trying to eliminate it before determining the weighting factor. For the auditory steady state responses the signals are small and the overall variance of the epoch is not significantly affected by the presence or absence of the signal. For our data, sample-weighting and noise weighting should therefore be roughly equivalent in terms of their effect. However, sample-weighting protocol is computationally less demanding than the noise-weighting protocol.

Unlike Dobie and Wilson (1994), our noise-weighting protocol did not perform quite as well as the sample-weighting protocol in increasing the signal-to-noise ratio. Since the two techniques performed similarly in improving response-detection (Table 1), the difference in the signal-to-noise ratio was unexpected. The main reason was that we filtered the data prior to the analysis in the sample-weighting proto-

col. This was necessary since we wished to calculate a weighting factor based on the frequencies in the recording that were near the frequencies of stimulation rather than the higher-amplitude lower frequencies present in the recording. The sloping edge of the filter (bandpass 70–110 Hz) near the highest and lowest response frequencies (78 and 95 Hz) would have altered the signal-to-noise ratios at these frequencies by attenuating the energy in the adjacent bins more than the energy at the response-frequency. This would not have occurred for the noise-weighting since the data were not filtered, and the weighting factor was determined by an exact selection of frequencies in the spectrum.

Although it became smaller, the difference between sample-weighting and noise weighting persisted if we removed the filtering. Another possible reason for the difference stems from the fact that the FFT analysis of the epochs (lasting only 1.024 s) had much less resolution than the analysis of the full sweep (lasting 16.384 s). Removing the spectral bin containing the signal response also removed frequencies that were close to the signal. These frequencies would be assessed as noise on the full-sweep analysis, since they would fall into adjacent frequency bins. This effect would be compounded by the fact that we were using multiple stimuli. We limited our noise estimate to between 70 and 110 Hz. In the FFT of the full sweep (16.384 s) this would contain 655 bins (resolution 0.061 Hz). However, in the FFT of each epoch (1.024 s) there are only 41 bins in this frequency range, and 12 of these would be removed (for the 8 signals and the 4 control frequencies). Our estimate of the epoch noise could therefore have been less accurate than the sample-weighting estimate. However, since a post-hoc computation of the noise-weighting data without removing the signal frequencies did not significantly alter the results, the decreased resolution of the weighting estimate for epochs could not explain the differences between sample-weighting and noise-weighting. The average of the spectral amplitudes is not directly related to the root-mean-square amplitude in the time-domain waveform, since different frequencies can partially cancel each other in the time domain and this cancellation depends on their phases. Whatever the differences, our time-domain weighting-factor caused a slightly better signal-to-noise ratio than the frequency-domain weighting-factor.

Several other weighting factors may be used instead of the inverse of the epoch variance. Several groups (Gasser et al., 1983; Davila and Mobin, 1992; Bezerianos et al., 1995) have suggested using the covariance between the epoch and the average of the other epochs in the data set, so that the weighting is correlated with the signal strength rather than the noise level. These procedures effectively weight each epoch on the basis of how similar it is to an estimated signal. This similarity would vary directly with the amplitude of the signal in the epoch and inversely with the amount of noise. The approach of weighting by signal strength is more appropriate for recordings where the signal is of the same order of magnitude as the noise. Since the auditory steady state responses are significantly smaller than the noise, this approach would probably not be helpful. The calculation of the weighting factor may also be improved by 'pre-whitening' (Gasser et al., 1983; Davila et al., 1997). However, for the analysis of the auditory steady-state responses at stimulus rates of 80–100 Hz, where the noise is already relatively homogeneous (i.e. white) across the frequency range of the signals (John and Picton, 2000), pre-whitening would likely not lead to significant improvement. Özdamar and Kalayei (1999) have shown that median averaging attenuates the effects of high-noise trials, since the median is much less affected by outliers than the mean. However, this procedure is much more computationally demanding for online use than weighted averaging.

We found that normal averaging resulted in a false-alarm rate of 4.3% that was slightly lower than the expected 5.0% level. This result is the same as the 4.3% that we have reported previously with other data (John and Picton, 2000). The background EEG noise falls off slightly with increasing frequency, and this probably makes the *F*-test slightly more conservative than its nominal values. The borderline increase in the false-alarm rate with the sample-weighting protocol compared to normal averaging probably results from the filtering of the data. We used 4 control frequencies to test the false alarm rate. The lowest and highest control frequencies (75 and 105 Hz), which were just beyond the frequency range of the responses, showed greater false alarm rates than the middle two frequencies. This could have been due to the *F*-test using frequency-bins that were beginning to be slightly attenuated by the filter (−6 dB points at 70 and 110 Hz) in its denominator. The weighting protocol did not significantly change the false-alarm rate within the range of the stimulus frequencies.

Although it significantly increased the signal-to-noise ratio, our sample-weighting protocol also under-estimated the amplitude of the responses compared to normal averaging. This is an acknowledged drawback of weighted averaging (Lütkenhöner et al., 1985). The amount of reduction will depend on the range of the weighting factors used. In our sample-weighting protocol, the reduction was larger (approximately 10%) than could be explained by weighted averaging. A major factor was the filtering used in the sample-weighting protocol, which accounted for about 7% of the reduction in the signal amplitude, and which mainly affected the responses with the highest and lowest frequencies. This effect could be reduced by increasing the bandwidth or changing the nature of the filter. Compromises have to be struck between computational time, selecting weighting factors based on the frequencies near the response-frequencies, and the accuracy of response estimation. One approach would be to compensate the amplitudes for the known reduction by the filter. A more efficient approach would be to base the weighting-factor on the filtered data but then to weight and analyze the unfiltered

data. This is the procedure that we currently recommend. (Using this procedure in a post hoc analysis of our data, the estimated signal-amplitude did indeed become only 3% smaller with sample-weighting as opposed to normal averaging.)

Another factor might also lead to the under-estimation of the signal amplitude. It is possible that the amplitude-estimate was less affected by the residual background noise when this was significantly reduced by weighted averaging. When it is measured from a combination of signal and noise, the signal amplitude is over-estimated by an amount that varies with the amount of noise (Strasburger, 1987). However, this could only explain a small part of the change. The estimated signal amplitude decreases by about 4% when normal averaging is based on 36 rather than 12 sweeps (Fig. 2). With this increase in analysis time, the signal-to-noise ratio increases much more (actual 1.65, expected 1.73 according to the square root rule) than it does when changing from normal averaging to sample-weighting (a 1.17 increase). The change in the noise level could therefore be only a small part (perhaps 1%) of the 10% decrease in the estimated signal amplitude with our sample-weighting protocol. Given the effects of filtering and residual noise we would estimate the reduction in the signal amplitude due to signal averaging as about 2% (similar to the 1.5% effect of noise-weighting).

Although it would not affect the findings if one is making a yes-no decision about whether a response is present or not, this slight decrease in signal amplitude might cause concern if the amplitude of the response is being compared across conditions or between subjects. Since it will equally affect the real and imaginary components of the response, weighted averaging should not affect phase.

Weighted averaging attenuates the effect of high-noise epochs on the final response. Everyone who has recorded average evoked potentials has experienced the situation when responses that are just beginning to look significant after a period of averaging sadly vanish with the subsequent occurrence of a few trials with higher noise. Weighted averaging should prevent this from happening. Weighted averaging has an advantage over other techniques, in that it can easily be performed online.

Because weighted averaging enhances the signal-to-noise ratio, it will detect responses more quickly than normal averaging. In evoked potential audiometry, time is particularly important when evaluating the hearing of babies or when monitoring the integrity of the auditory system during anesthesia. In these situations, the examiner wishes to gain as much information as possible in as little time as possible – before the baby wakes up and makes further testing impossible, or before transient intra-operative dysfunction becomes permanent. Weighted averaging should prove very helpful in these clinical situations.

Acknowledgements

This research was funded by the Canadian Institutes of Health Research. The authors also thank James Knowles, the Catherall Foundation and the Baycrest Foundation for their support. Malcom Binns provided advice on statistical strategies, and Patricia van Roon assisted with the actual statistical analyses.

References

Bezerianos A, Laskaris N, Fotopoulos S, Papathanasopoulos P. Data dependent weighted averages for recording of evoked potentials signals. Electroencephalogr Clin Neurophysiol 1995;96:468–471.

Cohen LT, Rickards FW, Clark GM. A comparison of steady-state evoked potentials to modulated tones in awake and sleeping humans. J Acoust Soc Am 1991;90:2467–2479.

Davila CE, Mobin MS. Weighted averaging of evoked potentials. IEEE Trans Biomed Eng 1992;39:338–345.

Davila CE, Srebo R, Ghaleb IA. Optimal detection of multiharmonic visually evoked potentials. IEEE/EMBS Proceedings of the 19th International Congress 1997;4:1514–1517.

Dobie RA, Wilson MJ. Objective detection of 40 Hz auditory evoked potentials: phase coherence vs. magnitude-squared coherence. Electroencephalogr Clin Neurophysiol 1994;92:405–413.

Gasser T, Möcks J, Verleger R. SELAVCO: a method to deal with trial-to-trial variability of evoked potentials. Electroencephalogr Clin Neurophysiol 1983;55:717–723.

Gerull G, Graffunder A, Wernicke M. Averaging evoked potentials with an improved weighting algorithm. Scand Audiol 1996;25:21–27.

Hoke M, Ross B, Wickesberg R, Lütkenhöner B. Weighted averaging – theory and application to electric response audiometry. Electroencephalogr Clin Neurophysiol 1984;57:484–489.

John MS, Picton TW. MASTER: a Windows program for recording multiple auditory steady-state responses. Comp Methods Prog Biomed 2000;61:125–150.

John MS, Dimitrijevic A, van Roon P, Picton TW. Multiple auditory steady-state responses to AM and FM stimuli. Audiol Neuro-Otol 2001 (in press).

Lins OG, Picton TW, Boucher BL, Durieux-Smith A, Champagne SC, Moran LM, Perez-Abalo MC, Martin V, Savio G. Frequency-specific audiometry using steady-state responses. Ear Hear 1996;17:81–96.

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

Mühler R, von Specht H. Sorted averaging–principle and application to auditory brainstem responses. Scand Audiol 1999;28:145–149.

Özdamar Ö, Kalayei T. Median averaging of auditory brain stem responses. Ear Hearing 1999;20:253–264.

Pantev C, Khvoles R. Comparison of the efficiency of various criteria for artifact rejection in the recording of auditory brain-stem responses (ABR). Scand Audiol 1984;13:103–108.

Picton TW, Linden RD, Hamel G, Maru JT. Aspects of averaging. Semin Hear 1983;4:327–341.

Regan D. Human brain electrophysiology: evoked potentials and evoked magnetic fields in science and medicine, Amsterdam: Elsevier, 1989.

Schimmel H. The (±) reference: accuracy of estimated mean components in average response studies. Science 1967;157:92–94.

Strasburger H. The analysis of steady-state evoked potentials revisited. Clin Vision Sci 1987;1:245–256.

Zurek PM. Detectability of transient and sinusoidal otoacoustic emissions. Ear Hear 1992;13:307–310.