

The use of phase in the detection of auditory steady-state responses

Terence W. Picton*, Andrew Dimitrijevic, M. Sasha John, Patricia Van Roon

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

Accepted 24 May 2001

Abstract

Objective: To investigate how phase measurements might facilitate the detection of auditory steady-state responses.

Methods: Multiple steady-state responses were evoked by auditory stimuli modulated at rates between 78 and 95 Hz and with intensities between 50 and 0 dB SPL. The responses were evaluated in 20 subjects after 1, 2, 4, and 6 min. The responses were analyzed in the frequency domain using 4 different detection protocols: (1) phase-coherence, (2) phase-weighted coherence, (3) *F* test for hidden periodicity, and (4) phase-weighted *t* test. The phase-weighted measurements were either based on the mean phase of a group of normal subjects or derived for each subject from the phase of the response at higher intensities.

Results: Detection protocols based on both phase and amplitude (*F* test and phase-weighted *t* test) were more effective than those based on phase alone (phase coherence and phase-weighted coherence) although the difference was small. Protocols using phase-weighting were more effective than those without phase-weighting. The lowest thresholds for the steady-state responses were obtained using the phase-weighted *t* test.

Conclusion: Threshold detection can be improved by weighting the detection protocols toward an expected phase, provided that the expected phase can be reliably predicted. © 2001 Elsevier Science Ireland Ltd. All rights reserved.

Keywords: Signal detection; Phase-weighting; Weighted averaging; Steady-state responses; Frequency analysis; Objective audiometry

1. Introduction

Steady-state responses occur when the frequency constituents of a response remain stable in amplitude and phase over time (Regan, 1989). These responses are usually evoked by periodic stimuli and measured in the frequency domain at the frequency of stimulation or one of its harmonics. The responses are two-dimensional and can be described in terms of their amplitude and phase. Phase is related to latency although converting phase to latency involves solving some ambiguities deriving from the circular nature of phase measurements (John and Picton, 2000b).

Auditory steady-state responses can be recorded at many different stimulus rates (Rickards and Clark, 1984; Rees et al., 1986). Responses recorded at stimulus rates near 40 Hz (Galambos et al., 1981; Rodriguez et al., 1986; Kuwada et al., 1986) and at rates near 80 Hz (Cohen et al., 1991; Rance et al., 1995; Lins et al., 1996) have been used to estimate hearing thresholds. These responses can provide audiometric information without requiring a patient to respond

behaviorally to sounds. This is clinically important when evaluating the hearing of infants, young children, cognitively impaired adults and patients who may have a functional hearing loss. However, objective audiometry with the steady-state responses is only possible if the responses can be reliably detected.

Like all scalp-recorded evoked potentials, steady-state responses are intermixed with noise. Detecting whether a response is present therefore requires some assessment of whether the recorded signal is significantly different from what might be expected from the noise alone. There are basically two kinds of tests for whether a response is present: one based on the similarity of a measurement across replications and the other based on the difference between a measurement at the frequency of stimulation (signal) and other measurements (noise) in the spectrum. Similarity in phase across replications can be assessed using phase coherence (Lord Rayleigh, 1880; Stapells et al., 1987). Similarity in both phase and amplitude can be assessed using either the T^2 test (Hotelling, 1931; Picton et al., 1987; Victor and Mast, 1991) or magnitude squared coherence (Dobie and Wilson, 1989, 1994a). Comparing the response (both amplitude and phase) at the frequency of stimulation to measurements at other frequencies in the spectrum is performed

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

E-mail address: tpicton@rotman-baycrest.on.ca (T.W. Picton).

using the F test (Schuster, 1898; Zurek, 1992). Although measurements that combine both phase and amplitude are theoretically more powerful than phase measurements alone (Dobie and Wilson, 1993), the different tests are often equally effective in detecting responses in real data (Picton et al., 1987; Valdes et al., 1997).

If the phase of the response is likely to be a particular value, one can bias the detection procedures towards recognizing as significant responses with phases that are close to this expected value. Dobie and Wilson (1994b) used a cosine-squared function to weight responses with phases that were within 90° of an expected phase. They found that such ‘phase weighting’ of the results improved the detection of responses using magnitude squared coherence. Lins et al. (1996) used a simple cosine function to bias their data over the full 360° . They also proposed a combined weighting approach, whereby the response was compared to how close it was in both amplitude and phase to an expected response. Both techniques improved the detection of responses without changing the false alarms. Simple phase weighting was a little more effective than amplitude–phase weighting, but the difference between the techniques was not significant. The techniques of Dobie and Wilson, and Lins et al. required empirical adjustment of the decision criteria to prevent the procedure from recognizing as significant too many trials in which there were no responses.

Dobie (personal communication) also suggested the possibility of converting the Hotelling’s T^2 test into a simple t test by projecting the two-dimensional data onto a one-dimensional line oriented at the expected phase. We have applied this idea to the F test. Instead of projecting the replicated responses onto the expected phase, we projected the measurement at the stimulus frequency and the measurements at adjacent frequencies onto the same phase. Then we compared the projected amplitude at the stimulus frequency to the distribution of the projected noise. This new ‘phase-weighted t test’ is simple to implement. Furthermore, it does not require setting an empirical decision criterion. The confidence limits for the noise can be determined using the statistical distribution of Student’s t .

The present study evaluated the use of phase measurements in detecting the auditory steady-state responses. The steady-state responses were recorded using weighted averaging (Lütkenhöner et al., 1985; John et al., 2001) at stimulus modulation frequencies between 75 and 100 Hz. The overall goal was to determine the most effective way of detecting these responses. Three main questions were addressed. The first was whether detection protocols based on phase information alone recognized responses as accurately as protocols using both phase and amplitude. The second was whether the detection procedures could be facilitated by biasing the protocols to recognize responses with phases similar to what was expected. The third question was whether these procedures would alter the threshold at which the responses were recognized.

2. Methods

2.1. Subjects

Twenty subjects (ten females, mean age 30 and age range 23–48 years) participated in this study. Behavioral thresholds were obtained using a Grason Stadler Model 16 audiometer, which was also used to present the stimuli during the experimental protocols. During the recording of the auditory steady-state responses, the subjects slept in a reclining chair.

2.2. Stimuli

Eight tones were presented simultaneously, 4 to the left ear and 4 to the right ear, with each tone having a different carrier frequency and a specific rate of amplitude modulation (AM) between 78 and 95 Hz. The modulation envelopes were based on an exponential sine function using a power of two and a modulation depth of 100% (John et al., submitted). Such stimuli have a broader frequency spectrum with 4 sidebands (two on either side of the carrier) rather than the two sidebands of sinusoidal AM. Probably because of their more rapid increase in amplitude, these stimuli elicit larger responses than sinusoidal AM. In the left ear, the carrier frequencies were 750, 1500, 3000, and 6000 Hz and the modulation frequencies were 80.08, 84.96, 89.84, and 94.73 Hz, respectively. In the right ear, the carrier frequencies were 500, 1000, 2000, and 4000 Hz and the modulation frequencies were 78.13, 83.01, 86.91, and 91.80 Hz. The modulation frequencies were selected so that there was an integer number of cycles of modulation within 1.024 s. For example, for the 500 Hz carrier, 80 cycles of modulation occurred within an epoch, giving a modulation frequency of $80/1.024$ or 78.125 Hz.

Stimuli were presented using Etymotic ER-2 insert earphones at intensities of 50, 40, 30, 20, 10, and 0 dB SPL. In the 0 dB condition, the insert earphones were actually withdrawn from the ears and taped shut to ensure that no sound was heard. This condition could then serve as a control for checking the false alarm rates of the tests. Although the stimulus intensity was therefore much lower, we shall use the ‘0 dB SPL’ nomenclature for simplicity. The intensities were based on root-mean square amplitudes of the individual stimuli and the combination of 4 stimuli had an overall intensity about 5 dB higher than a single stimulus. Intensities were calibrated using a Brüel and Kjaer 2230 sound level meter with a 2 cc DB 0138 coupler and were accurate within ± 3 dB.

2.3. Steady-state responses

The steady-state responses were recorded using the MASTER system (John and Picton, 2000a). Responses were recorded between Cz and the neck (ground on the right clavicle) with an analog/digital (A/D) conversion rate of 1000 Hz. The analog filter bandpass for recording these data was 1–300 Hz. 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 FFT provides a spectrum of real and imaginary values at each of 8192 frequencies between 0 and 500 Hz (resolution of 0.061 Hz).

The averaging process used a weighting procedure (Lütkenhöner et al., 1985; John et al., 2001). The weighting factor for each epoch was based upon the frequencies near those of the responses. Accordingly, we initially filtered each sweep of data using a digital second-order Butterworth filter with a bandpass of 70–110 Hz. The weighting factor was then the reciprocal of the variance of the filtered activity over each epoch. The unfiltered data for each epoch was then multiplied by the weighting factor. 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. This procedure calculates the weights on the basis of activity in a selected bandpass but applies the weights to the unfiltered data so as not to distort the signals or the detection protocols (John et al., 2001). In order to ensure that the weighting procedure improved the efficiency of signal detection, we also recorded responses using normal averaging.

Signal amplitudes were calculated as the square root of the sum of the squares of the real and imaginary components provided by the FFT at each of the resolved frequencies and a cosine onset phase was calculated from the arctangent of the real and imaginary components. When combining phases across subjects, we used vector averaging with each subject contributing equally, provided the response was judged significant. When combining amplitudes, we simply averaged the individual amplitude measurements (whether or not they were significant). The standard deviations (SD) for amplitude were calculated conventionally. The standard deviations for the phase measurements can be calculated in several ways (reviewed by Zar, 1999, p. 604). We chose to use the technique of Mardia (1972) since it appears most similar to the linear standard deviation. A value R was calculated as described in Section 2.4.1 and the circular standard deviation (CSD) in degrees was estimated as

$$\frac{180}{\pi} \sqrt{-2 \log_e(R)}$$

where $\log_e(R)$ is the natural logarithm of R .

2.4. Signal-detection protocols

The responses were evaluated at each of the 8 stimulus frequencies and at 4 other frequencies (75.20, 82.03, 92.77, and 96.68 Hz) that served as arbitrary controls to assess the

incidence of false detections. The responses were assessed using 4 different measurements, described in the following paragraphs. Two of these measurements involved checking for a response at an expected phase. The expected phase was determined in two ways. For the responses to the 50 dB SPL stimuli, the expected phases were estimated from pilot data for the responses at each of the carrier frequencies. These phase data were determined from the results of 5 subjects. The data were close to but not equivalent to the actual mean phases later obtained. (We therefore also re-analyzed the results using the actual mean data.) For intensities lower than 50 dB SPL, the phase was calculated as the phase of the response for each subject at each carrier frequency at 50 dB SPL intensity less than the normal change in phase that occurs with decreasing intensity. This change ($-2.8^\circ/\text{dB}$) was also estimated from the pilot data. This slope was kept constant over all carrier frequencies and not estimated specifically for each.

2.4.1. Phase coherence

This test derives from the work of Lord Rayleigh (1880). After each epoch of recording (1.024 s), an FFT was performed and the phase of the response was measured at each of the stimulus-frequencies. The sine and the cosine of these phases (θ) were then added separately to running sums. The phase coherence (R) after N epochs could then be calculated as:

$$R = \frac{1}{N} \sqrt{\left(\sum_{i=1}^N \cos \theta_i \right)^2 + \left(\sum_{i=1}^N \sin \theta_i \right)^2}$$

This measurement varies between 0 and 1, with higher values indicating a lower probability that the phase is changing randomly from epoch to epoch (Lord Rayleigh, 1880; Fisher, 1993; Zar, 1999; Cohen et al., 1991). The significance of the result was assessed using the equations described by Fisher (1993). This calculation was based on the phases estimated from each epoch ('epoch-based') and did not benefit from the weighted averaging. We therefore also estimated the phase coherence from the 16-epoch sweep obtained after weighted averaging. In this evaluation, N was always 16.

2.4.2. Phase-weighted coherence

If the distribution of phases is tested for an expected phase (θ_e) rather than for any departure from random phase, a modified measurement (R_0) can be calculated:

$$R_0 = R \cos(\bar{\theta} - \theta_e)$$

where $\bar{\theta}$ is the mean phase of the sample (calculated from the arctangent of the sums of the cosines and the sines of the individual phase). The significance of this result was assessed differently from R , once again using equations described by Fisher (1993). As well as calculating the phase-weighted coherence on an epoch-by-epoch basis,

we also made the calculations using the 16 epochs within the weighted-average sweeps.

2.4.3. *F test for hidden periodicity*

This test derives from the initial description of Schuster (1898) and further work by Fisher (1929). 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 can be obtained from frequencies where no stimulus occurred. We estimated the signal-to-noise ratio (SNR) by comparing the power at each stimulus-frequency (a_s^2), equivalent to the sum of the squares of the real and imaginary parts of the FFT, to the average power at 120 nearby frequencies (60 above and 60 below the stimulus-frequency), excluding the frequencies where there were other stimuli:

$$\frac{120a_s^2}{\sum_{\substack{i=s-60 \\ i \neq s}}^{i=s+60} a_i^2}$$

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, 2000a).

2.4.4. *Phase-weighted t test*

A two-dimensional vector can be projected onto an expected phase to give a one-dimensional measurement (Strasburger, 1987). Basically, the amplitude of each vector is multiplied by the cosine of the difference in phase between it and the expected phase (cf. the weighted phase coherence measurement). We used a similar approach to convert the *F* test for hidden periodicity into a *t* test. Each of the amplitudes (for the signal and the 120 adjacent frequencies) was projected onto the expected phase and the significance of the ratio between the projected amplitude and the projected noise was assessed using a *t* test with 119 degrees of freedom. Fig. 1 shows a graphic representation of this procedure. We made the figure legible by using only 16 adjacent bins (rather than 120) for the noise estimate.

2.5. *Simulations*

In order to ensure that the statistical analyses (and their instantiation in our software) were performing correctly, we evaluated the protocols using simulated signals and noise. The noise was normally distributed and the signal was a sine

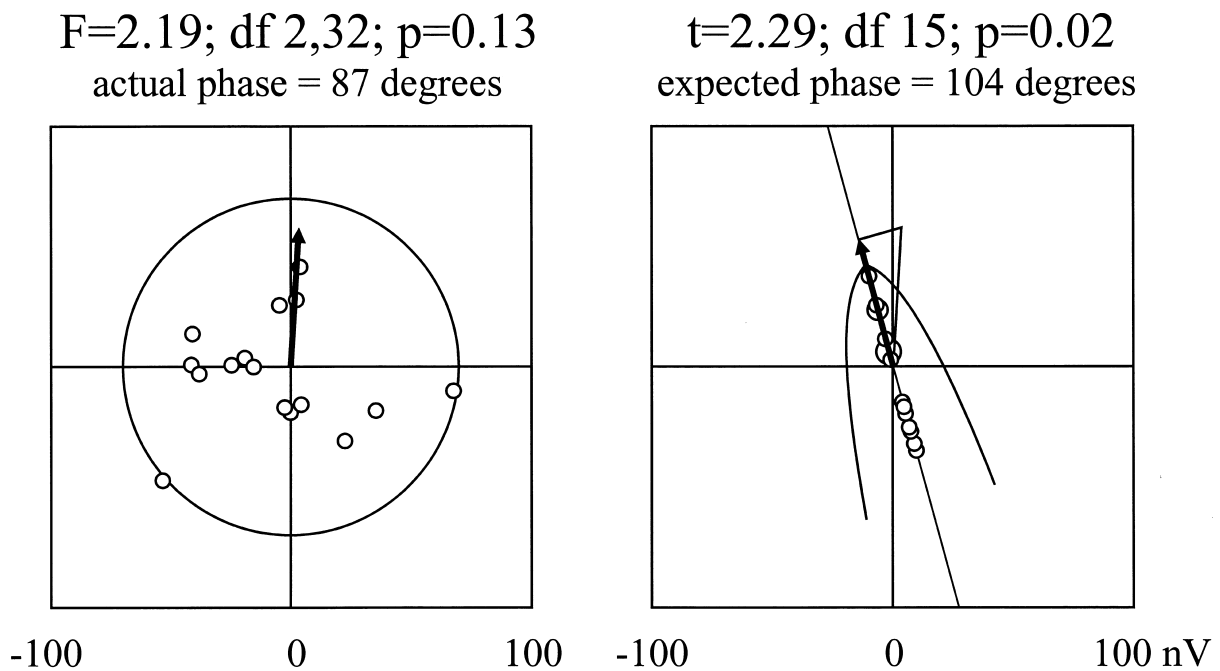


Fig. 1. Projecting data onto an expected phase. These data were obtained from one subject and show a response to a 50 dB SPL 2 kHz carrier frequency modulated at 83 Hz. In order to make the diagram legible, the response is compared to 16 adjacent noise-bins (8 on each side of the modulation frequency) rather than the usual 120. On the left diagram, the $P < 0.05$ confidence-limits for the noise are shown with the circle. The response plotted with an arrow is not significantly different from noise at $P < 0.05$ since it is not larger than the radius of the circle. On the right, the data have been projected onto an expected phase of 104°. The actual process of projection is shown only for the response. The larger circles indicate the superimposition of two or 3 projected data-points. The parabola shows the $P < 0.05$ confidence limits for the one-tailed *t* test. The projected response (arrow) is significantly different from the noise at $P < 0.05$ because it goes beyond these limits.

wave with a constant phase for each epoch. Fig. 2 shows sample results from 1000 simulations. For each simulation, the data array (1024 values) was filled with noise consisting of normally distributed numbers with a mean of zero and a standard deviation of 2. On the left are shown the results of testing for a signal when none was present. The probability histogram is plotted using bins equal to a 1% change in probability. Thus the first bin shows the incidence of evaluations showing $0 < P < 0.01$ and the next bin shows the incidence of evaluations showing $0.01 < P < 0.02$ and so on. The probability that a signal was incorrectly detected was constant across the different probabilities. The cumulative probability was such that the incidence of $P < 0.01$ results was indeed near 1.0% (as indicated with arrows). On the right are shown the histograms of the probabilities when a small signal (a cosine wave with an amplitude of ± 0.1 and a frequency of 32/1024 points in the array) was added to the noise. The weighted analyses were performed using an expected phase equal to the actual phase of the simulated signal (45°). The histograms tilt toward the lower probabilities with more of the responses judged

significant using criteria such as $P < 0.05$ or $P < 0.01$ than can be attributed to chance alone. This effect is greater for the measurements based on phase and amplitude (F test and phase-weighted t test) than for those based on phase alone and greater for those measurements using phase-weighting (when the expected phase is equal to the actual phase) than without phase-weighting.

2.6. Threshold estimations

Thresholds were estimated for each of the carrier-frequencies according to the following rules. If the response at 50 dB SPL was not significant at $P < 0.05$, the threshold was arbitrarily set at 60 dB SPL. Otherwise, the threshold at a particular carrier-frequency was the lowest intensity at which a response was detected as significant when all responses at higher intensity were also significant. Thresholds were determined using the data obtained with the 4 different protocols using mean phases at 50 dB SPL and mean changes in the phase with intensity estimated from pilot data. In addition, we repeated the evaluation using

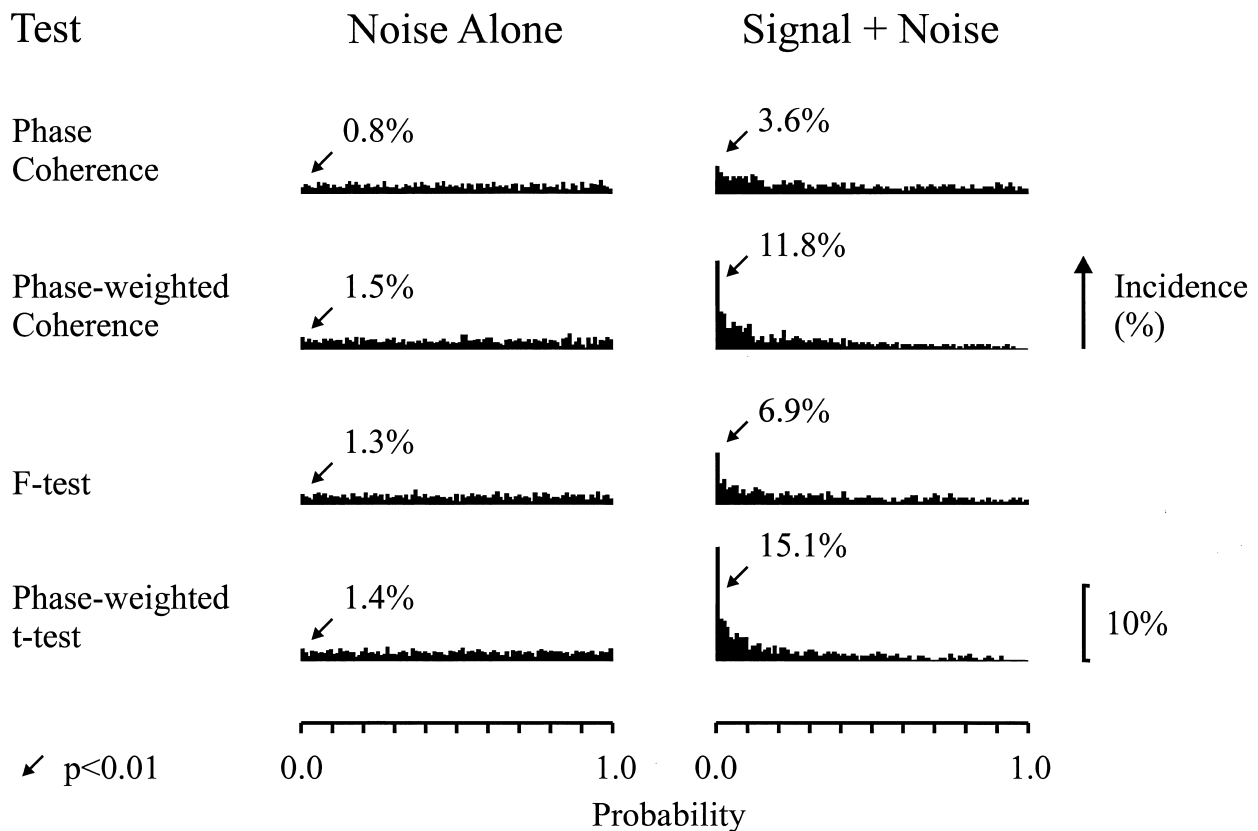


Fig. 2. Detection protocols using simulated data. The simulated data was normally distributed noise with a standard deviation of 2. The results on the left were obtained with the noise alone. The results on the right were obtained when a sine-wave signal with an amplitude of ± 0.1 was added to the noise. One thousand simulations were performed and the incidences of the probability values from each of the tests were plotted in histograms using one bin for each 0.01 increment in the probability. The numbers with the arrows show the incidence of recognizing the signal at $P < 0.01$ in each of the different conditions using each of the different protocols. If the test is performed correctly, these histograms should be flat across the range of probabilities when noise alone was analyzed. This indeed occurred for all the tests. The incidence of tests showing $P < 0.01$ is indicated with the arrows. This should be close to 1.0%. When a signal was present, the incidence of recognizing it as significantly different from the noise should show up as an increase in the lower probabilities in the histogram. This clearly occurred for all tests with the phase-weighted t test showing the greatest effect.

the actual mean phases at 50 dB SPL over the 20 subjects and the actual mean slopes of the phase change with intensity for each carrier frequency in order to see whether the phase-weighting protocols could be improved by more accurate phase data.

2.7. Receiver operating curves

In order to evaluate how well the techniques were recognizing responses, we plotted receiver-operating curves (ROC) for each subject in each condition. The probability of true positives (*y*-axis) was the percentage of detections at the 8 stimulus frequencies and the probability of false positives (*x*-axis) was the percentage of detections at the 4 control frequencies. Points on the ROC were calculated at protocol decision criteria of $P < 0.05, 0.10, 0.20, 0.30, 0.40, 0.60,$ and 0.80 . The area under the curve (*A*), calculated by joining these points was used as a measure of detection accuracy (Swets, 1988; Swets et al., 2000).

2.8. Statistical analyses

In order to determine whether weighted averaging improved the detection of responses over normal averaging, we compared the *F*-ratios obtained after the two procedures. The *F*-ratio provides a way to assess the SNR. The *F*-ratio is actually a measure of the signal-plus-noise to the noise. When the SNR approaches 0, the *F*-ratio will approach one. Other than this, the *F*-ratio will be affected by the experimental manipulations in a similar way to the SNR. 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 SNR. The effects of weighted versus normal averaging were assessed using an analysis of variance (ANOVA) with repeated measures across subjects. We used a 3-way ANOVA (protocol \times time \times carrier-frequency) and repeated this ANOVA at 50 dB and at 0 dB SPL. The time variable was equivalent to the number of sweeps averaged prior to analysis (4, 8, 16, 24 sweeps lasting approximately 1, 2, 4, and 6 min, respectively). A second set of ANOVAs was performed on the amplitude measurements to see if weighted averaging caused any significant changes in the responses. Greenhouse–Geisser corrections for the probability levels were used when appropriate.

Unfortunately, there is no easy way to compare the SNR across the different detection-protocols since the noise was differently estimated in each protocol. We therefore assessed the effects of the protocols upon the incidence of significant responses (at $P < 0.05$) using the McNemar test (McNemar, 1947; Siegel, 1956). Basically, if *L* is the number of responses that change from significant to non-significant between the different detection protocols and *M* the number of responses that change from non-significant to significant, the value

$$\frac{(|L - M| - 1)^2}{L + M}$$

is distributed as χ^2 with one d.f. Since multiple tests could be performed in our experimental design, we performed these tests in a hierarchical manner, only checking further if global effects were significant.

The areas under the ROC plots were evaluated using an ANOVA design. Other studies have used this approach to ROC areas (Thompson and Zucchini, 1989; Song, 1997).

Thresholds were compared using ANOVAs. Physiological thresholds (using weighted averaging over 24 sweeps and the phase-weighted *t* test detection protocol) were compared to behavioral thresholds using a two-way (protocol \times carrier-frequency) ANOVA with repeated measures across subjects. Thresholds among the different detection protocols were compared using a 3-way (protocol \times time \times carrier-frequency) ANOVA.

3. Results

3.1. Illustrative data

Fig. 3 shows data recorded from one subject as the number of sweeps increased from 4 to 24 and as the intensity decreased from 50 to 10 dB SPL. The steady-state responses are displayed in the frequency-domain after weighted averaging. The average amplitudes and phases of the responses across all the subjects are plotted in Fig. 4. Table 1 gives estimates of the variance across subjects.

3.2. Effect of weighted averaging

The weighted-averaging protocol improved the detection of signals over the normal-averaging protocol. The ANOVA of the square root of the *F* ratios at 50 dB SPL showed a significant main effect of protocol ($F = 12.0$; d.f. = 1,19; $P < 0.01$). As expected, the SNR was greater after a longer period of analysis ($F = 45.6$; d.f. = 3,57; $P < 0.001$) and differed across carrier-frequency ($F = 4.5$; d.f. = 7,133; $P < 0.01$). There was a significant interaction between protocol and time ($F = 6.1$; d.f. = 3,57; $P < 0.05$) with the protocol effect being larger after a longer period of analysis. There were no significant effects in the ANOVA of the results at 0 dB SPL. The ROC analysis (Fig. 5) showed a greater amplitude under the curve for the weighted averaging, particularly after a higher number of sweeps, but neither the main effect nor the interaction reached significance on the ANOVA.

The averaging protocols decreased the measured amplitude of the response as the number of sweeps increased. The ANOVA conducted on the amplitudes of the responses at 50 dB SPL showed a significant decrease with an increase in the number of sweeps ($F = 7.6$; d.f. = 3,57; $P < 0.01$), and a significant effect of carrier frequency ($F = 7.4$; d.f. = 7,133; $P < 0.001$). There was a small decrease in

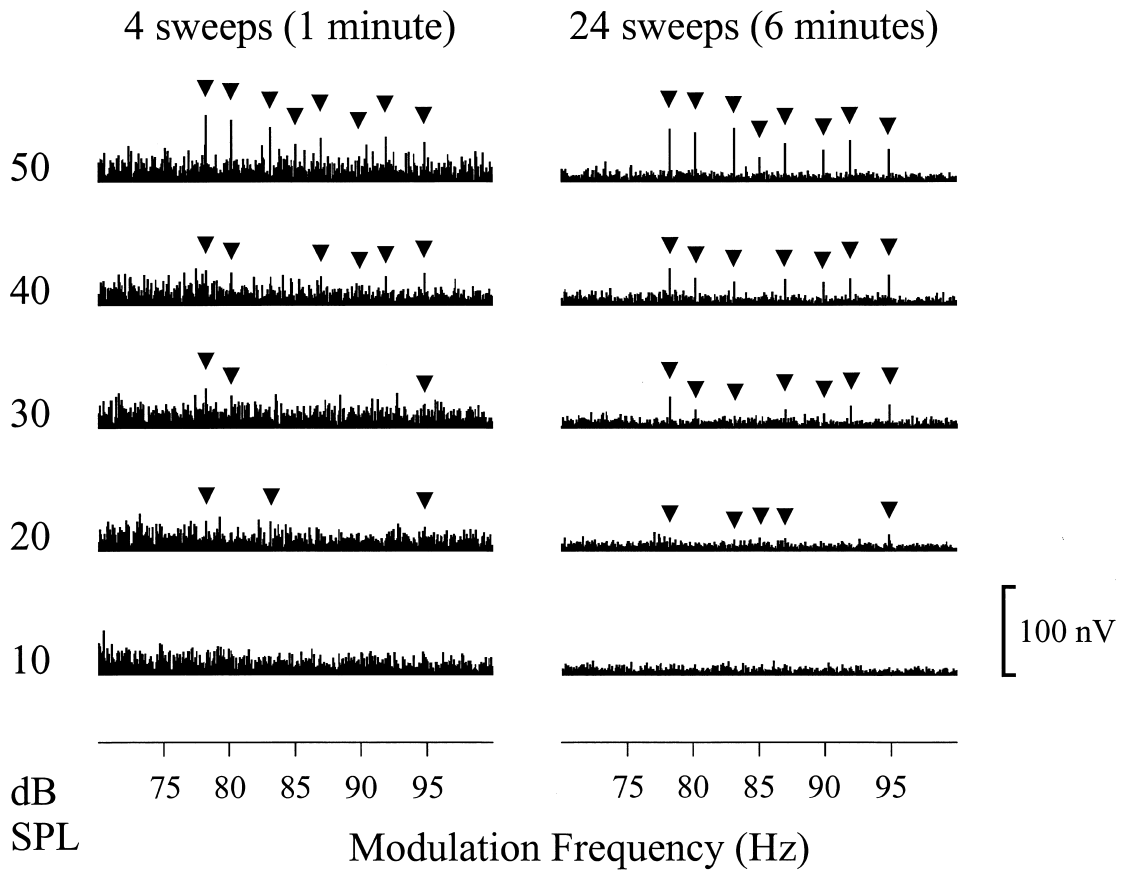


Fig. 3. Auditory steady-state responses at different intensities. This figure shows the responses to all 8 stimuli in one subject as the stimulus intensity is decreased from 50 to 10 dB SPL. On the left are shown the responses after averaging 4 sweeps and on the right, after averaging 24 sweeps. The arrowheads indicate responses that were recognized as significant using the *F* test at $P < 0.05$.

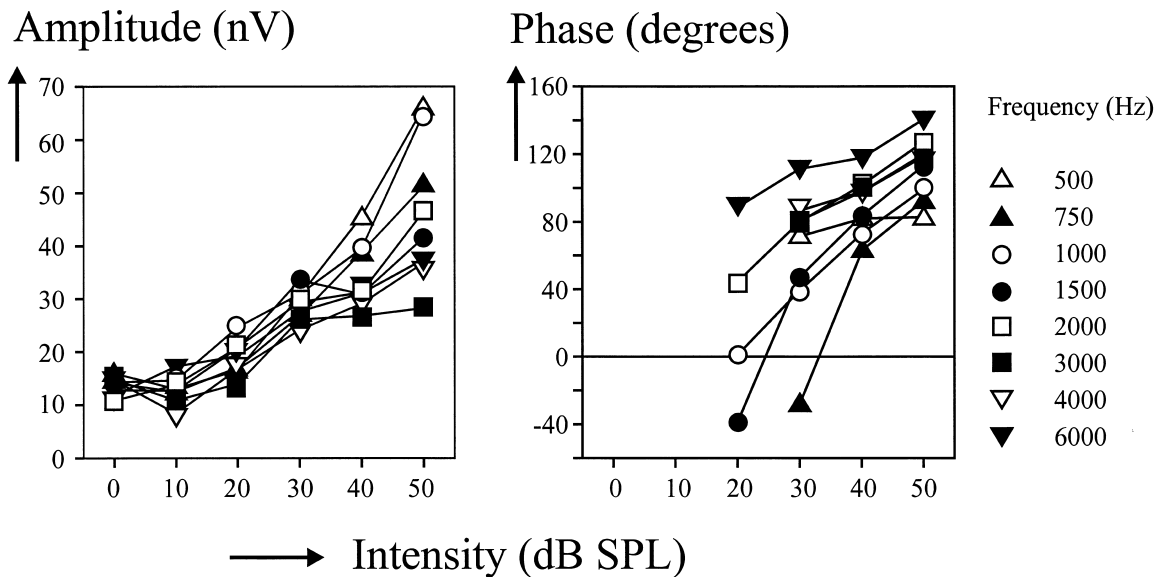


Fig. 4. Effects of intensity on amplitude and phase. This figure shows how the amplitudes and phases of the responses change with stimulus intensity. The amplitudes were normally averaged over all 20 subjects whether or not the responses were recognized as significant. The amplitudes at the 0 dB SPL intensity therefore indicate the noise levels of the recording. The phases were vector-averaged across subjects only when the responses were significant. Data were then only plotted if 5 or more subjects contributed to the average.

Table 1
Variability of the responses at 50 dB SPL^a

Carrier (Hz)	500	750	1000	1500	2000	3000	4000	6000
Modulation (Hz)	78	80	83	85	87	90	92	95
Ear	R	L	R	L	R	L	R	L
Amplitude (nV)	66	51	65	42	47	28	37	38
SD	90	49	46	33	54	11	33	42
Phase (°)	83	92	100	113	127	119	120	141
CSD	60	44	37	31	28	46	39	42
Pilot phase	120	74	104	77	120	87	138	113

^a The CSD is the 'circular standard deviation'. The 'pilot phase' derives from pilot experiments and was used as the expected phase for the assessment of the responses at 50 dB SPL.

amplitude with weighted averaging and this decreased with an increase in the number of sweeps – from 12% after 4 sweeps to 3% after 24 sweeps. However, these changes did not reach significance in the ANOVA. At 0 dB SPL, where there was no response and the measurements were effectively evaluating noise, the ANOVA showed a significant decrease in amplitude (42% after 4 sweeps and 35% after 24 sweeps) with weighted averaging ($F = 5.4$; d.f. = 1,19; $P < 0.05$) as well as a more significant decrease with an increase in the number of sweeps.

3.3. Comparison of detection protocols

Phase-weighting increased the number of responses that were recognized as significant. This effect depended upon whether the recorded response had a phase that was similar to the expected phase. The changes occurring with phase-weighting are illustrated in Fig. 6, which shows the operation of the F test and the phase-weighted t test on some data from a single subject. After 8 sweeps had been averaged, both techniques showed significant responses. However, after 4 sweeps only, the phase-weighted t test detected the response as significantly different from noise.

We first evaluated whether the tests gave the expected number of false-positive detections. At a significance level of $P < 0.05$, the incidence of false positives should be 5%.

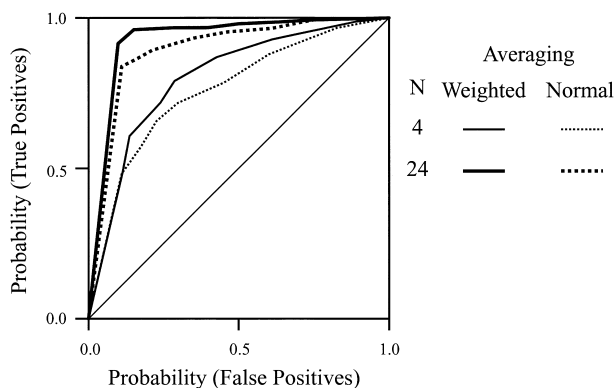


Fig. 5. Weighted averaging. This figure compares the ROC for detecting responses after weighted averaging or after normal averaging. The figure also shows the effects of combining 4 or 24 sweeps prior to the analysis.

Two sets of data can be used for this evaluation. The first is the incidence of detections at frequencies where there was no stimulus, i.e. at the 4 control frequencies at all intensities. The second is the incidence of detections when there was no stimulus, i.e. at the stimulus frequencies when the intensity was 0 dB SPL. The results are shown in Table 2. None of the incidences was significantly different using χ^2 from the expected level of 5%.

The incidence of response detection varied with the number of sweeps analyzed and with the intensity. This is shown diagrammatically in Fig. 7. As can be seen, the phase–amplitude protocols (F test and phase-weighted t test) generally detected more responses than the phase-alone protocols (phase coherence and phase-weighted coherence). In addition, the phase-weighted protocols (phase-weighted coherence and t test) detected more responses than the protocols that were independent of phase (F test, coherence). The phase-alone protocols based on the weighted averaging were not significantly different from the same protocols based on unweighted data.

The McNemer test basically compares the incidence of tests that become significant with a change in the testing protocol to the incidence of tests that lose significance, using the null hypothesis that these incidences are equal. Our initial McNemer analysis compared pairs of tests using all the data across intensity and frequency. Subsequent specific comparisons to determine whether phase alone or phase and amplitude gave more detections were between phase coherence and the F test and between phase-weighted coherence and the phase-weighted t test. The comparisons to determine whether phase-weighting improved the detection were between phase coherence and phase-weighted coherence, and between the F test and the phase-weighted t test. These are shown in Table 3.

Whereas the McNemer test compares the detection of responses at one criterion, the ROC area uses all criteria. Using all of the ROC data across detection protocol, intensity, and number of sweeps, we found the expected significant effects of intensity ($F = 46.5$; d.f. = 5,95; $P < 0.001$) and number of sweeps ($F = 13.4$; d.f. = 3,57; $P < 0.001$). In addition, there was a protocol vs. number of sweeps interaction ($F = 2.3$; d.f. = 9,171; $P < 0.05$). Post hoc testing showed that the phase-weighted protocols performed better than the unweighted protocols and that this effect was greater for the lower numbers of sweeps. Table 4 gives the mean areas under the curve for the different protocols and Fig. 8 shows the ROCs plotted from data combined across all subjects.

3.4. Thresholds

The mean thresholds for the different techniques across the 20 subjects are shown in Fig. 9. The phase-weighted t test detected responses at lower levels than the other tests. The ANOVA considered the two kinds of phase coherence and phase-weighted coherence separately, so that there were

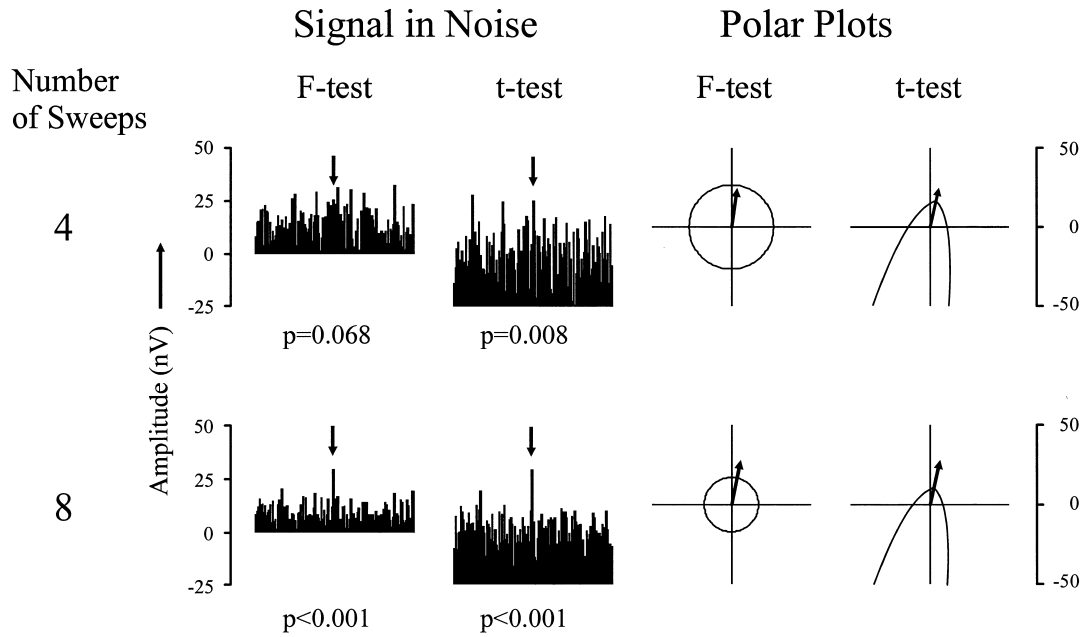


Fig. 6. Effects of phase-weighting. This figure compares the *F* test and the *t* test during the recognition of a response to a 1500 Hz tone modulated at 85 Hz and presented at 50 dB SPL. On the left are plotted the signals with the adjacent noise measurements (60 bins on either side of the signal, with the signal bin indicated by the arrow). For the *F* test, all of the measurements are necessarily above zero. The response is not recognizable at $P < 0.05$ after 4 sweeps have been averaged but is recognizable after 8 sweeps have been averaged. For the *t* test, the data are projected onto the expected phase and vary above and below zero. The response is clearly recognizable after 4 sweeps and even more prominently after 8 sweeps. On the right, the data are plotted in polar form. The noise data are not plotted individually, but the $P < 0.05$ confidence limits for the noise are plotted together with the response.

6 different protocols. The ANOVA showed a significant effect of the number of sweeps ($F = 5.1$; d.f. = 3,57; $P < 0.05$) with the thresholds being lower after more sweeps had been combined and a significant effect of carrier frequency ($F = 4.0$; d.f. = 7,133; $P < 0.01$) with the thresholds being lowest for carrier frequencies of 1500 Hz. There was also a significant effect of protocol ($F = 66.6$; d.f. = 5,95; $P < 0.001$) with the lowest thresholds occurring with the phase-weighted *t* test and an interaction between protocol and carrier frequency ($F = 2.8$; d.f. = 35,665; $P < 0.01$) with the protocol effect being less evident at 500 and 750 Hz. Table 5 compares the differences between the physiological thresholds and the behavioral thresholds.

The phases expected on the basis of pilot experiments were not equal to the mean phases obtained from the actual experiment, with differences ranging from -36 to $+37^\circ$ at 50 dB SPL. We therefore recalculated the thresholds using the actual mean phases at 50 dB SPL and the mean phase-change with intensity for each carrier frequency. The new

thresholds for the phase-weighted *t* test were on an average less than 1 dB better than those calculated on the basis of the pilot data. However, this difference was not statistically significant.

4. Discussion

4.1. Weighted averaging

In keeping with our previous findings (John et al., 2001), the present results show clearly that weighted averaging improves the SNR compared to normal averaging. This effect was caused by a significant reduction in the noise amplitude with weighted averaging. The estimated response amplitude also decreased – partly due to the reduction in the noise and partly due to the weighted averaging itself – but this was much less than the reduction in the noise and was not significant on statistical testing. Since our subjects were relatively quiet during the recording sessions, the effect of

Table 2
Incidence of false-positive detections^a

Test	Phase coherence	Phase-weighted coherence	<i>F</i> test	Phase-weighted <i>t</i> test
Non-stimulus frequencies	4.4	4.7	5.5	5.9
Stimulus frequencies 0 dB SPL	3.9	4.7	4.7	4.7

^a Incidence is in percentage. The total number of tests for the non-stimulus frequencies was 1920 (20 subjects, 4 times, 4 frequencies, 6 intensities) and for the 0 dB SPL was 640 (20 subjects, 4 times, 8 carrier-frequencies). The expected incidence is 5%.

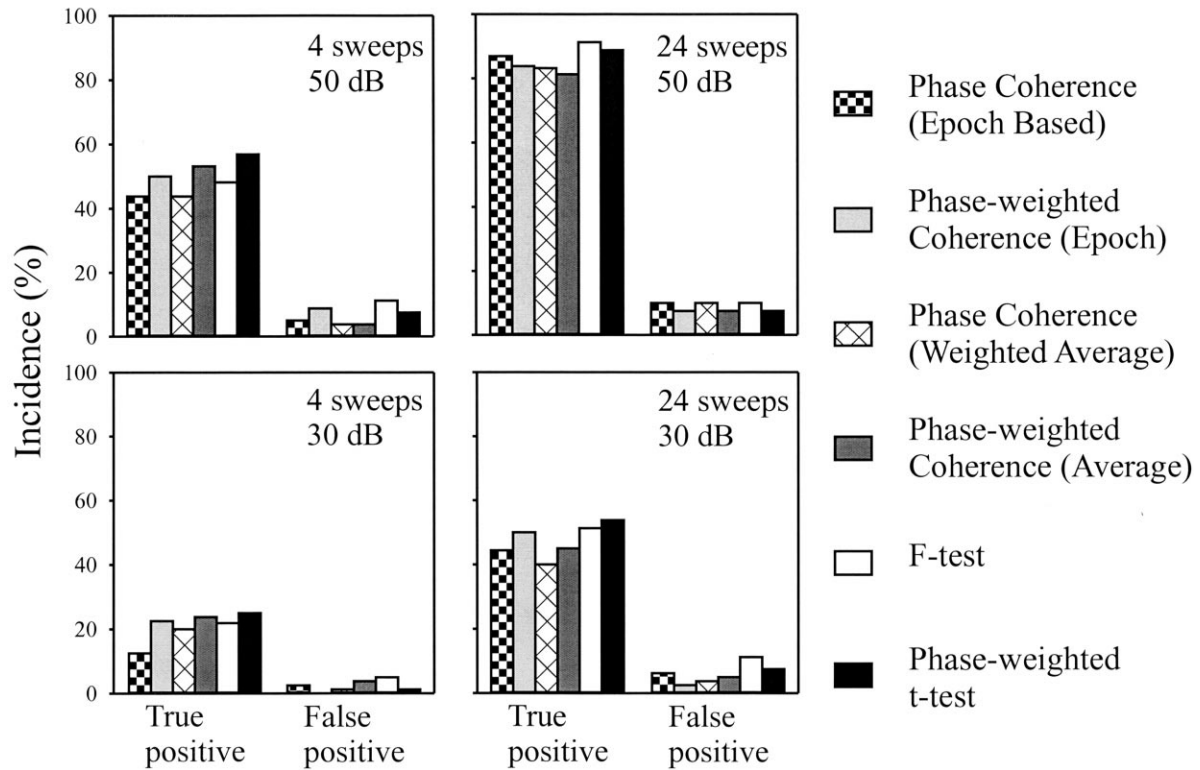


Fig. 7. Response detection. This figure plots the incidence of detected responses using the different protocols at two different intensities (50 and 30 dB SPL) after combining either 4 or 24 sweeps. The false-positive detections do not show any clear pattern. The true-positive detections are generally larger for the phase-weighted protocols (darker shading) than for the unweighted protocols and generally larger for the amplitude-phase protocols (*F* test and *t* test) than for the protocols based on phase alone. The data from two versions of the phase-based measurements are included: one based on an epoch-by-epoch analysis and one following weighted averaging.

Table 3
Differences between tests^a

X	Y	Sweeps	L(X + & Y -)	M(X - & Y +)	Significance
Phase coherence	<i>F</i> test	4	21	72	***
		8	25	78	***
		16	20	74	***
		24	20	65	***
Phase-weighted coherence	Phase-weighted <i>t</i> test	4	26	56	**
		8	22	79	***
		16	21	65	***
		24	15	60	***
Phase coherence	Phase-weighted coherence	4	23	75	***
		8	43	57	
		16	43	62	
		24	48	71	*
<i>F</i> test	Phase-weighted <i>t</i> test	4	41	72	**
		8	47	65	
		16	48	57	
		24	46	69	*

^a The column labeled L gives the number of tests (out of a total of 960 – 20 subjects, 6 intensities, 8 carrier frequencies) that were positive for test X and negative for test Y. The column labeled M gives the number of tests that were negative for test X and positive for test Y. Significance from the McNemar test is given as: *, *P* < 0.05; **, *P* < 0.01 and ***, *P* < 0.001.

Table 4
ROC areas

Intensity (dB SPL)	Number of sweeps	Detection protocol			
		Phase coherence	Phase-weighted coherence	<i>F</i> test	Phase-weighted <i>t</i> test
50	4	0.77	0.85	0.81	0.84
	8	0.84	0.88	0.87	0.87
	16	0.90	0.91	0.90	0.88
	24	0.92	0.90	0.92	0.91
40	4	0.69	0.81	0.71	0.82
	8	0.72	0.81	0.76	0.84
	16	0.78	0.84	0.83	0.86
	24	0.86	0.89	0.88	0.89
30	4	0.64	0.71	0.64	0.75
	8	0.64	0.75	0.64	0.75
	16	0.72	0.76	0.71	0.77
	24	0.73	0.76	0.75	0.77
20	4	0.55	0.49	0.56	0.53
	8	0.56	0.53	0.60	0.53
	16	0.60	0.60	0.61	0.60
	24	0.67	0.62	0.65	0.63

weighted averaging was small and did not show up as significant in either the McNemer comparisons or the ROC area measurements.

4.2. Phase and amplitude in the detection of responses

We found that using both phase and amplitude (*F* test) rather than phase alone (phase coherence) led to higher levels of response detection. This effect is small and shows up most clearly on the McNemer tests. The ROC data show similar effects but these do not reach significance. The fact that the differences are small explains why previous studies (e.g. Picton et al., 1987; Valdes et al., 1997) have not found significant effects. The small difference indicates that the phase of the response may be more reliable than ampli-

tude, in terms of the variability between subjects. However, such a comparison is difficult to evaluate because of the circular nature of phase.

4.3. Phase-weighting

Weighting the data so that the detection protocols favor responses with phases similar to those that are expected on the basis of prior knowledge improves the accuracy of detection. Phase-weighting has a similar effect for both phase-amplitude assessments and simple phase coherence.

We derived the expected phase from two kinds of prior knowledge. To evaluate responses at 50 dB SPL, we used the phases of other subjects studied in pilot experiments. To evaluate responses at lower intensity, we used the phase of

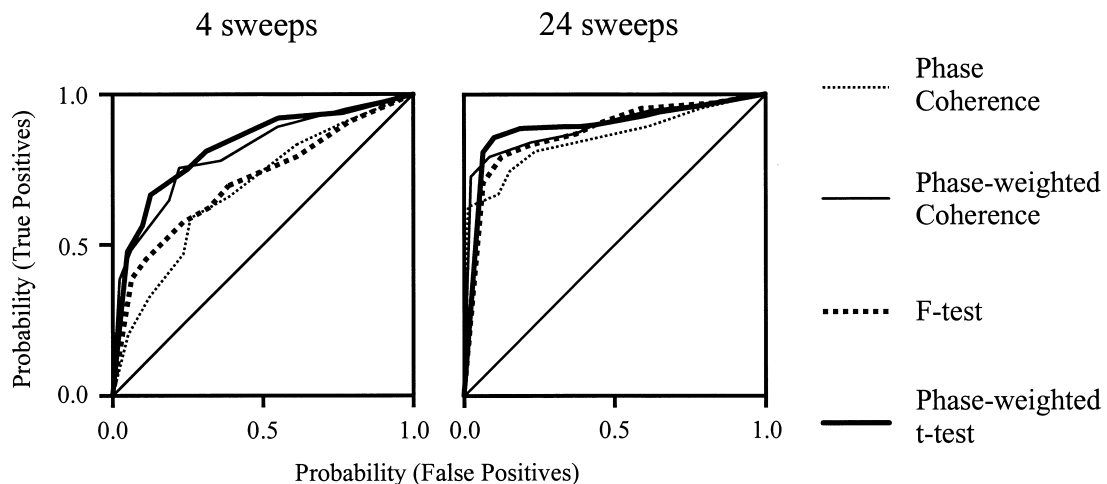


Fig. 8. ROC analysis for different detection protocols. This figure plots the ROC data using the 4 different protocols after 4 or 24 sweeps were averaged. The advantage of phase-weighted protocols over unweighted protocols is more obvious after 4 sweeps than after 24.

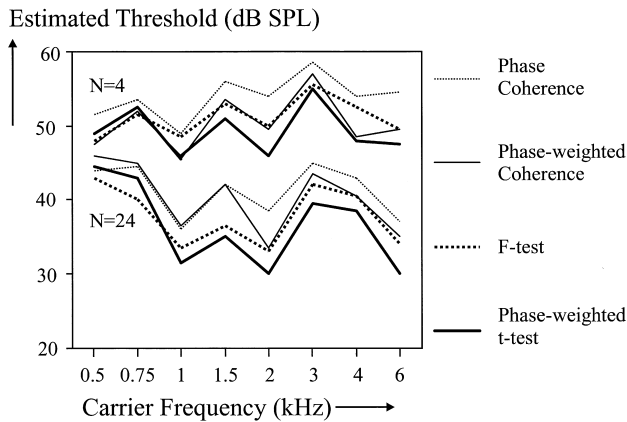


Fig. 9. Threshold estimations. This figure plots the physiological thresholds for each of the carrier frequencies according to the type of detection protocol and whether the number of sweeps (*N*) averaged prior to analysis was 4 or 24.

each subject recorded to each of the stimuli at 50 dB adjusted by an expected change in phase with intensity derived from the pilot data. The second approach was more effective, probably because it eliminated much of the inter-subject variance. Other sources of prior knowledge might be used to set an expected phase. For example, if several of the responses are recognized as significant during a multiple stimulus protocol, an expected phase for the responses that are not yet significant can be extrapolated from the phases of the recognized responses.

In general, phase-weighting worked better when the response was not quite recognizable using normal techniques. This can be seen in the top part of Fig. 7 where there is a clearly beneficial effect of phase-weighting after 4 sweeps but not after 24 sweeps.

We used a simple cosine weighting function. This gentle weighting function worked reasonably well even when the expected phases from our pilot data were not completely accurate. Other weighting functions, such as the cosine-

squared function of Dobie and Wilson (1994b), might be more effective in some situations. For example, one might adjust the ‘tightness’ of the weighting function to the degree of normal inter-subject variance. Whatever the weighting function, one must ensure that applying the function does not distort the probability estimates (cf. Fig. 2).

The present set of experiments looked at the effects of phase-weighting on auditory steady-state responses recorded in sleeping adults using stimulus rates of 78–95 Hz. Phase-weighting should also improve threshold estimation in infants and young children, but this is yet to be evaluated. Phase-weighting may also facilitate threshold estimation for steady-state responses near 40 Hz. Indeed, phase-weighting might be more effective at these stimulus-rates because of the lower inter-subject variance of phase.

4.4. Thresholds for the steady-state responses

The response thresholds for recognizing the auditory steady-state responses using phase-weighting were on an average 21 dB above those obtained behaviorally (Table 5). This difference is higher than the 13 dB difference reported by Lins et al. (1996), the 11 dB reported by Herdman and Stapells (2001), and the 12 dB reported by Perez-Abalo et al. (2001), all studies using similar multiple-stimulus protocols. The difference likely depends on the strict threshold-seeking algorithms used in the present study, and the short recording periods at near-threshold levels (discussed in a subsequent paragraph). Protocols using single-stimulus auditory steady-state responses in normal subjects have reported thresholds that are on an average 28–34 dB (Aoyagi et al., 1994) and 17–35 dB (Rance et al., 1995) above normal behavioral thresholds (HL). These comparisons are not really the same as ours since the subjects tested may not have had hearing thresholds at 0 dB HL, either from the variability of individual thresholds or the variability of the acoustic noise levels during testing. We did not test in a properly sound-attenuated room and our subject’s thresholds were 6–17 dB above normal HL thresholds.

The steady-state responses at stimulus rates of 75–100 Hz can provide thresholds that are similar to those obtained with stimulus rates near 40 Hz. The average differences between the thresholds for the 40 Hz responses and behavioral thresholds vary between 3 and 18 dB (Szyfter et al., 1984; Rodriguez et al., 1986; Aoyagi et al., 1993). The advantages of using the faster stimulus rates is that the faster responses are not affected by sleep (Cohen et al., 1991; Lins and Picton, 1995), can be recorded in infants (Rickards et al., 1994; Lins et al., 1996) and can be recorded using multiple simultaneous stimuli (Lins and Picton, 1995; John et al., 1998). These advantages are particularly important when testing the hearing of newborn infants and young children.

In patients with hearing loss, thresholds for the auditory steady-state responses are lower relative to behavioral

Table 5
Thresholds^a

Carrier (Hz)	500	750	1000	1500	2000	3000	4000	6000
Behavioral	18	13	10	13	18	21	19	11
SD	6	5	7	7	6	8	6	9
HL (inserts)	9		4		7		2	-2
<i>F</i> test threshold	43	40	34	37	33	42	41	34
<i>F</i> difference	25	27	24	24	15	21	22	23
SD	13	10	13	14	9	11	13	15
<i>t</i> Test threshold	45	43	32	35	30	40	39	30
<i>t</i> Difference	27	30	21	22	12	19	20	19
SD	14	11	14	15	9	13	11	16

^a Thresholds are expressed in dB SPL for behavioral, HL, *F* test and *t* test. The HL levels are from Wilber (1994). *F* difference and *t* difference are the differences between the *F* test threshold and the behavioral threshold and between the *t* test threshold and the behavioral threshold. Differences are in dB.

thresholds than in normal subjects (Rance et al., 1995; Lins et al., 1996; Aoyagi et al., 1999; Perez-Abalo et al., 2001). This will show up as a slope of greater than 1.0 when regressing the physiological thresholds (x) against the behavioral thresholds (y). For example, Rance et al. (1995) found a regression equation for the thresholds at 1000 Hz of

$$y = 1.18x - 26.1$$

and Lins et al. (1996) found an equation over the frequencies between 500 and 4000 Hz of

$$y = 1.3x - 30$$

We had expected to obtain thresholds that were closer to the behavioral thresholds. Two reasons may have explained the discrepancy between our expectations and our results. First, our threshold detection algorithm was very conservative (perhaps Draconian) in its requirement that all supra-threshold responses be significant. When testing patients, we have sometimes noted that a response may be insignificant at one intensity and then appear consistently at lower intensities. Just as there is a small chance that a response may be falsely recognized when it is not there, there is also a small chance that a response may not be recognized when it is there. Since there is always a 1 in 20 chance of a false detection when using a $P < 0.05$ criteria, we clearly cannot just conclude that threshold is at the lowest level at which a response is recognized. However, we could consider a decision rule along the lines that threshold is the lowest intensity at which a response is detected, provided that it was also detected at an intensity 10 dB higher (Picton et al., 1998). Deciding upon the most effective threshold-seeking algorithm is not simple and will require further research.

The second reason for the discrepancy between behavioral and physiological thresholds depends on when the recording stops. A recording should stop as soon as all of the responses are recognized or until the noise has been attenuated below the level at which a response might be recognized. Although at high intensity we could have stopped the recording earlier, at low intensities we could have allowed the recording to continue much longer. Herdman and Stapells (2001), for example, recorded for up to 12 min. when the stimuli were near threshold.

Although using different threshold-seeking algorithms and increasing the time for recording near-threshold responses could have lowered our physiological thresholds, the effects of the different testing protocols on threshold-estimation were still clear. First, phase-weighting significantly reduced the estimated thresholds, thus making them closer to the behavioral thresholds. Second, thresholds were even more reduced by increasing the amount of averaging prior to response evaluation.

4.5. Estimating behavioral thresholds

The lowest intensity at which a physiological response can be recognized will always be higher than the subject's

behavioral threshold, provided that the subject is following instructions correctly during the behavioral testing. Physiological responses are difficult to record at near-threshold intensity because the amplitudes of the responses are nearer to the noise levels of the physiological recording and because the response may vary from moment to moment and therefore not show up in averaged recordings.

One way to make the estimate of threshold more accurate would therefore be to subtract the expected difference between physiological and behavioral thresholds from the physiological threshold. However, patients with a hearing loss often have a physiological threshold that is closer to the behavioral threshold than normal hearing subjects. This is likely related to recruitment. Simply subtracting a constant amount from the physiological threshold might therefore underestimate the behavioral thresholds in patients with hearing-loss. This problem could be circumvented by using a regression between physiological and behavioral thresholds (cf. Rance et al., 1995).

5. Conclusions

The use of both phase and amplitude in detection protocols recognizes more auditory steady-state responses than using phase data alone. Further improvement comes from weighting the detection protocols to recognize responses with a phase near that which is expected.

Acknowledgements

This research was funded by the Canadian Institutes of Health Research. The authors thank James Knowles, the Baycrest Foundation, and the Catherall Foundation for their support. Malcom Binns provided advice and suggestions on the statistics.

References

- Aoyagi M, Kiren T, Kim Y, Suzuki Y, Fuse T, Koike Y. Frequency-specificity of amplitude-modulation-following response detected by phase spectral analysis. *Audiology* 1993;32:293–301.
- Aoyagi M, Kiren T, Furuse H, Suzuki Y, Yokota M, Koike Y. Pure-tone threshold prediction by 80-Hz amplitude-modulation following response. *Acta Otolaryngol (Stockh)* 1994;511:7–14.
- Aoyagi M, Suzuki Y, Yokota M, Furuse H, Watanabe T, Ito T. Reliability of 80-Hz amplitude-modulation-following response detected by phase coherence. *Audiol Neurootol* 1999;4:28–37.
- 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.
- Dobie RA, Wilson MJ. Analysis of auditory evoked potentials by magnitude-squared coherence. *Ear Hear* 1989;10:2–13.
- Dobie RA, Wilson MJ. Objective detection in the frequency domain. *Electroencephalogr Clin Neurophysiol* 1993;88:516–524.
- Dobie RA, Wilson MJ. Objective detection of 40 Hz auditory evoked potentials: phase coherence vs. magnitude-squared coherence. *Electroenceph clin Neurophysiol* 1994a;92:405–413.

- Dobie RA, Wilson MJ. Phase weighting: a method to improve objective detection of steady-state evoked potentials. *Hear Res* 1994b;79:94–98.
- Fisher NI. *Statistical analysis of circular data*, Cambridge, MA: Cambridge University Press, 1993.
- Fisher RA. Tests of significance in harmonic analysis. *Proc R Soc Lond Ser A* 1929;125:54–59.
- Galambos R, Makeig S, Talmachoff PJ. A 40 Hz auditory potential recorded from the human scalp. *Proc Natl Acad Sci USA* 1981;78:2643–2647.
- Herdman AT, Stapells DR. Thresholds determined using the monotic and dichotic auditory steady-state response technique in normal-hearing subjects. *Scand Audiol* 2001;30:41–49.
- Hotelling H. The generalization of Student's ratio. *Ann Math Stat* 1931;2:360–378.
- John MS, Picton TW. MASTER: a Windows program for recording multiple auditory steady-state responses. *Comp Methods Prog Biomed* 2000a;61:125–150.
- John MS, Picton TW. Human auditory steady-state responses to amplitude-modulated tones: Phase and latency measurements. *Hear Res* 2000b;141:57–79.
- John MS, Dimitrijevic A, Picton TW. Auditory steady-state responses to exponential modulation envelopes. *Ear Hear*, submitted.
- John MS, Lins OG, Boucher BL, Picton TW. Multiple auditory steady state responses (MASTER): Stimulus and recording parameters. *Audiology* 1998;37:59–82.
- John MS, Dimitrijevic A, Picton TW. Weighted averaging of steady-state responses. *Clin Neurophysiol* 2001;112:555–562.
- Kuwada S, Batra R, Maher VL. Scalp potentials of normal and hearing-impaired subjects in response to sinusoidally amplitude-modulated tones. *Hear Res* 1986;21:179–192.
- Lins OG, Picton TW. Auditory steady-state responses to multiple simultaneous stimuli. *Electroenceph clin Neurophysiol* 1995;96:420–432.
- Lins OG, Picton TW, Boucher B, 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.
- Mardia KV. *Statistics of directional data*, New York, NY: Academic Press, 1972.
- McNemar Q. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika* 1947;12:153–157.
- Perez-Abalo MC, Savio G, Torres A, Martin V, Rodriguez, Galan L. Steady state responses to multiple amplitude modulated tones: and optimized method to test frequency specific thresholds in hearing impaired children and normal subjects. *Ear Hear* 2001 (in press).
- Picton TW, Vajsar J, Rodriguez R, Campbell KB. Reliability estimates for steady state evoked potentials. *Electroenceph clin Neurophysiol* 1987;68:119–131.
- Picton TW, Durieux-Smith A, Champagne SC, Whittingham J, Moran LM, Giguère C, Beauregard Y. Objective evaluation of aided thresholds using auditory steady-state responses. *J Am Acad Audiol* 1998;9:315–331.
- Lord Rayleigh. On the resultant of a large number of vibrations of the same pitch and of arbitrary phase. *Philos Mag* 1880;10:73–78.
- Rance G, Rickards FW, Cohen LT, De Vidi S, Clark GM. The automated prediction of hearing thresholds in sleeping subjects using auditory steady-state evoked potentials. *Ear Hear* 1995;16:499–507.
- Rees A, Green GGR, Kay RH. Steady-state evoked responses to sinusoidally amplitude-modulated sounds recorded in man. *Hear Res* 1986;23:123–133.
- Regan D. *Human brain electrophysiology: evoked potentials and evoked magnetic fields in science and medicine*, Amsterdam: Elsevier, 1989.
- Rickards FW, Clark GM. Steady state evoked potentials to amplitude-modulated tones. In: Nodar RH, Barber C, editors. *Evoked potentials II*, Boston, MA: Butterworth, 1984. pp. 163–168.
- Rickards FW, Tan LE, Tan Cohen LT, Wilson OJ, Drew JH, Clark GM. Auditory steady-state evoked potential in newborns. *Br J Audiol* 1994;28:327–337.
- Rodriguez R, Picton T, Linden D, Hamel G, Laframboise G. Human auditory steady state responses: effects of intensity and frequency. *Ear Hear* 1986;7:300–313.
- Schuster A. On the investigation of hidden periodicities with application to a supposed 26 day period of meteorological phenomena. *Terrestr Magnet Atmos Electr* 1898;3:13–41.
- Siegel S. *Nonparametric statistics for the behavioral sciences*, New York, NY: McGraw-Hill, 1956.
- Song HH. Analysis of correlated ROC areas in diagnostic testing. *Biometrics* 1997;53:370–382.
- Stapells DR, Makeig S, Galambos R. Auditory steady-state responses: threshold prediction using phase coherence. *Electroenceph clin Neurophysiol* 1987;67:260–270.
- Strasburger H. The analysis of steady-state evoked potentials revisited. *Clin Vision Sci* 1987;1:245–256.
- Swets JA. Measuring the accuracy of diagnostic systems. *Science* 1988;240:1285–1293.
- Swets JA, Dawes RM, Monahan J. Psychological science can improve diagnostic decision making. *Psychol Sci Public Interest (Suppl Psychol Sci)* 2000;1:1–26.
- Szyfter W, Dauman R, Charlet de Sauvage R. 40 Hz middle latency responses to low frequency tone pips in normally hearing adults. *J Otolaryngol* 1984;13:275–280.
- Thompson ML, Zucchini W. On the statistical analysis of ROC curves. *Stat Med* 1989;8:1277–1290.
- Valdes JL, Perez-Abalo MC, Martin V, Savio G, Sierra C, Rodriguez E, Lins O. Comparison of statistical indicators for the automatic detection of 80 Hz auditory steady state responses. *Ear Hear* 1997;18:420–429.
- Victor JD, Mast J. A new statistic for steady-state evoked potentials. *Electroenceph clin Neurophysiol* 1991;78:378–388.
- Wilber LA. Calibration, puretone, speech and noise signals. In: Katz J, editor. *Handbook of clinical audiology*, 4th ed. Baltimore, MD: Williams and Wilkins, 1994. pp. 73–94.
- Zar JH. *Biostatistical analysis*, 4th ed. Upper Saddle River, NJ: Prentice-Hall, 1999.
- Zurek PM. Detectability of transient and sinusoidal otoacoustic emissions. *Ear Hear* 1992;13:307–310.