informa
healthcare

Original Article

International Journal of Audiology 2009; 48: 833–842

Maryanne Golding
Harvey Dillon
John Seymour
Lyndal Carter

National Acoustic Laboratories, Chatswood, Sydney, New South Wales, Australia
The Hearing Cooperative Research Centre, East Melbourne, Victoria, Australia

The detection of adult cortical auditory evoked potentials (CAEPs) using an automated statistic and visual detection

Key Words

Hotelling's T2
Cortical auditory evoked potentials (CAEP)
Visual detection
Sensitivity
Specificity

Abbreviations

CAEP: Cortical auditory evoked potentials
ABR: Auditory brainstem response
SNR: Signal-to-noise ratio
RMS: Root-mean-square (voltage)
d': d prime sensitivity index
EEG: Electroencephalographic (brain activity)
GA: Gestational age
ROC curve: Receiver operating characteristic curve
CI: Confidence interval
SE: Standard error

Abstract

The detection of adult cortical auditory evoked potentials (CAEPs) can be challenging when the stimulus is just audible. The effectiveness of a statistic compared with expert examiners in (1) detecting the presence of CAEPs when stimuli were present, and (2) reporting the absence of CAEPs when no stimuli were present, was investigated. CAEPs recorded from ten adults, using two speech-based stimuli, five stimulus presentation levels, and non-stimulus conditions, were given to four experienced examiners who were asked to determine if responses to auditory stimulation could be observed, and their degree of certainty in making their decision. These recordings were also converted to multiple dependent variables and Hotelling's T2 was applied to calculate the probability that the mean value of any linear combination of these variables was significantly different from zero. Results showed that Hotelling's T2 was equally sensitive to the best of individual experienced examiners in differentiating a CAEP from random noise. It is reasonable to assume that the difference in response detection for a novice examiner and Hotelling's T2 would be even greater.

Sumario

La detección de los potenciales evocados auditivos corticales (CAEP) del adulto puede ser un desafío cuando el estímulo es apenas audible. Se investigó la efectividad de una comparación estadística con examinadores expertos para (1) detectar la presencia del CAEP cuando el estimulo estuvo presente, y (2) reportar la ausencia del CAEP cuando ningún estímulo estaba presente. Los CAEP registrados para diez adultos, usando dos estímulos con base en lenguaje, cinco niveles de presentación del estímulo, y condiciones sin estímulo, fueron presentados a cuatro examinadores con experiencia, a quienes se solicitó que determinaran si las respuestas a estímulos auditivos podían ser observadas y su grado de certeza en la toma de sus decisiones. Estos registros fueron también convertidos a múltiples variables dependientes y se aplicó el T2 de Hotelling para calcular la probabilidad de que el valor medio de una combinación lineal de estas variables fuera significativamente diferente de cero. Los resultados mostraron que el T2 de Hotelling era igualmente sensible que lo mejor de los examinadores experimentados individuales para diferenciar un CAEP de un ruido aleatorio. Es razonable asumir que la diferencia en detección de respuestas para un examinador novato y el T2 de Hotelling serían aún mayor.

The recording of cortical auditory evoked potentials (CAEPs) has a long history. As early as the 1960s, studies reported the application of this technique in estimating auditory threshold to frequency specific stimuli in adults who were unable to participate in normal behavioral testing (Coles & Mason, 1984; Cone-Wesson & Wunderlich, 2003; Davis, 1965; Hyde et al, 1986; Rickards et al, 1996). CAEP testing did not, however, maintain its place as a key electrophysiological measure after the discovery of the relatively more stable auditory brainstem response (ABR) (Jewett & Williston, 1971), even though threshold detection at the level of the brainstem does not guarantee cortical detection.

The physiological mechanisms of the CAEP are complex and have been the subject of much study and debate (Hall, 2007). Much of what is understood regarding the mechanisms of cortical processes in humans has been inferred from animal studies. Several comprehensive texts can provide the reader with an overview of this debate (e.g. Burkard et al, 2007; Hall, ibid). Eggermont (in Burkard et al, 2007) describes the auditory cortex as a hierarchical structure, with three broad subdivisions, referred to as the primary, secondary, and tertiary cortex. The cortex as a whole is organized in terms of six laminae, or layers, parallel to the folded cortex surface (Katz, 1985).

The origins of the voltages observed on the scalp can be understood as follows (Burkard et al, ibid). The auditory cortex receives excitatory input, initially from specific auditory areas in the thalamus. Axons from thalamic neurons communicate synaptically with the long fibre-like dendrites of cortical neurons. These dendrites run perpendicular to the cortical laminae and cortical outer surface. The inflow of positive ions that occurs following excitation at a synapse creates a positive-going, post-synaptic potential within the dendrite (i.e. depolarization). As a result of the inflow, the extracellular environment local to the inflow position becomes more negative.

ISSN 1499-2027 print/ISSN 1708-8186 online
DOI: 10.3109/14992020903140928
© 2009 British Society of Audiology, International Society of Audiology, and Nordic Audiological Society

Received:
May 1, 2008
Accepted:
June 23, 2009

Lyndal Carter
National Acoustic Laboratories, 126 Greville St., Chatswood,
Sydney, NSW 2067, Australia.
E-mail: lyndal.carter@nal.gov.au

Extra-cellular current then flows throughout the cortex from positively to negatively charged regions to equalize the extra-cellular potential throughout the cortex. If the synaptic excitation (and hence depolarization) happens relatively deep below the cortical surface, for instance at layer IV, the extra-cellular currents predominantly flow from superficial layers to the deeper layers. Positions on the scalp closest to the superficial layer therefore acquire a positive voltage relative to positions on the scalp closer to the deeper layers. Because the superficial layer of the auditory cortex lies mostly on the superior surface of Hechl's gyrus within the Sylvian fissure, most auditory dendritic fibers run tangential to the scalp. Consequently the upper half of the scalp, including the vertex, becomes positive relative to the lower half of the scalp, including the mastoid. This is thought to be the origin of the P1 peak, and to be one of the origins of the P2 peak. Conversely, if the original synaptic excitation and consequent depolarization occur in a more superficial layer (e.g. layer II), then the extra-cellular current flows from deeper to more superficial layers, and the vertex will become negative with respect to the mastoid. This is thought to be the origin of the N1 response. As dendrites and synaptic connections do not form in the superficial layers until after about five years of age, this is consistent with the absence of an N1 response in infants (Kushnerenko, 2001).

Amplitudes of auditory evoked potentials, including the CAEP, are thought to be affected by the summation of contributions from multiple sources that may have opposite polarities and hence have partially cancelling effects (Burkard et al, ibid).

Multiple mechanisms contribute to the latency of the cortical responses, including the cochlear travelling wave (which is more relevant to the low frequency components), and the time taken for the neural activity to travel through the brainstem to the cortex. However, the latencies of the P1 peak (60 ms in adults), N1 peak (100 ms in adults), and the P2 peak (180 ms in adults) are far too long to be fully accounted for by these factors, particularly the latter two peaks. The additional delay may be accounted for by neural activity forming synchronized circuits between different layers and areas of cortex, and between the cortex and thalamus (Burkard et al, ibid). EEG activity in the cortex can be detected because the dendrites predominantly run in 'vertical' columns (i.e. at right angles to the cortical surface). Because they are parallel, the extra-cellular currents they induce sum coherently, and hence can be detected at a distance, on the scalp. Conversely, dendrites in the thalamus are not arranged in a parallel structure, so the activity cannot be detected at the head surface, perhaps accounting for the periods between the peaks of the CAEP waveform.

CAEP recordings are still regarded by some as the technique of choice for a number of clinical applications, in particular for estimating the pure tone audiogram in cases where a frequency-specific, non-behavioural method is required (Hyde, 1997).

A resurgence of interest in CAEPs has occurred in recent years, associated with the need to verify hearing-aid fittings in very young infants diagnosed with hearing loss through newborn screening programs. The recording of ABR thresholds to brief-tone stimuli (Stapells & Kurtzberg, 1991; Stelmachowicz, 1999), or auditory steady state responses (ASSR) to frequency modulated stimuli (Perez-Abalo et al, 2001; Picton et al, 2002) provides valuable diagnostic information and threshold estimates for hearing-aid prescription methods (Dillon, 2001) in most cases. Recent research has also suggested that sub-cortical structures such as the brainstem are affected by disruptions to normal cortical function and hence electrophysiological studies of brainstem timing deficits may also provide useful diagnostic information (Abrams et al, 2006; Song et al, 2008). The recording of CAEPs alone, however, verifies the detection of the speech stimuli at the level of the cortex, and therefore are potentially influenced by all parts of the auditory system. The technique therefore has the potential to be used in verifying hearing aid fittings.

The CAEP, when recorded in awake adults, generally consists of gently sloping and broad components, which contrast to the sharp and narrowly defined components of the shorter latency responses such as ABR (Hall, 2007). The adult response (i.e. over 20 years of age), is dominated by a negativity known as N1 with a latency of 80–120 ms that is preceded and followed by positive components (i.e. P1 at 50–70 ms, and P2 at 150–200 ms) (Davis, 1965). The amplitudes and latencies of CAEP components can be highly stimulus-dependent (Gage & Roberts, 2000; Hyde, 1997; Martin & Boothroyd, 1999). Speech sounds, for example, that are dominated by high frequency energy tend to elicit smaller amplitude N1 and P2 components than those elicited by speech sounds with spectral emphasis in the lower frequencies (Agung et al, 2006). The response amplitude, latency, and wave morphology will also vary substantially between and within subjects, with (1) varying levels of alertness (Hyde, 1997; Wunderlich & Cone-Wesson, 2006), (2) an inadequate signal to noise ratio (SNR) arising from an inadequate numbers of epochs within the averaged response (Molfese, 1978), or (3) heightened levels of background noise from a participant's restless state (Hyde, 1994). These scenarios lead to uncertainty in visual response detection methods.

The common method of response detection relies on superimposing two or more averaged CAEP responses that are visually examined for reliability (Coles & Mason, 1984; Hoth, 1993; Yeung & Wong, 2007). Various criteria to identify the existence of a true response have been reported. In one study where stimuli had variable presentation levels, the cortical response at reduced presentation levels was accepted as present if the largest amplitude negative peak (N1) or the largest amplitude positive peak (P2) within the recording period, had latencies consistent with the latencies observed using the same stimulus presented at high stimulus levels (Rickards et al, 1996). Yeung & Wong (2007) reported that responses were detectable if the difference in peak latency for two or more recordings at any one presentation level, was less than or equal to 10 ms. When multiple electrode recording sites were used, Korczak et al (2005) based response detection on two criteria. First, the peak amplitude had to be larger than the pre-stimulus baseline, and second, the responses that were generated at different sites on the scalp followed known scalp topography rules (e.g. N1 should be greater in amplitude at fronto-central sites than at parietal sites). It is apparent from studies such as these that the method used to detect CAEP responses varies considerably across experiments and experimenters. It is also reasonable to assume that even when specified criteria are applied, human judgment is required to determine if the criteria are met.

The difficulties of electrophysiological response recognition by visual detection methods increase substantially when the stimulus is just audible, and hence the response has a poor SNR (Hoppe, 2001; Hoth, 1993; Schimmel, 1974). The rules used in the detection process can become quite complex. Rickards et al (1996) defined threshold as the lowest level at which the cortical response was detected, or 5 dB less if the N1 component had a reduction in amplitude of at least 50% compared with that observed using high stimulus levels, and the latency was within 20 ms of the response that was 10 dB above it. Yeung & Wong (2007) defined threshold as the lowest stimulus level at which N1 could be visually detected by two independent examiners. Lightfoot & Kennedy (2006) and Coles & Mason (1984) estimated threshold based on empirically derived expectations of amplitude for the frequency under test. As an example, the N1-P2 amplitude was expected to be 5 uV for 1 k, and 3 uV for higher frequencies at CAEP threshold. If the amplitude was larger, then threshold was equal to the stimulus presentation level minus 5 dB (Lightfoot & Kennedy 2006).

Given the challenges of electrophysiological response recognition by visual detection methods, automated or machine scoring techniques have been applied. Such techniques have been employed with relative frequency in ABR testing and include (1) attempts to estimate the SNR and hence the likelihood of response detection (Don et al, 1984; Elberling & Don, 1984, 1987; Wong & Bickford, 1980), (2) the application of software classification models that are generated from artificial neural networks and decision tree classifiers using some combination of time, frequency and cross-correlation measures (Davey et al, 2007; Delgado & Ozdamar, 1994), and (3) the application of statistical tests on either the distribution of the phase angle of the Fourier harmonics in a sample of stimulus-generated epochs alone or in combination with spectral amplitudes (Sturzebecher & Cebulla, 1997).

Automated techniques have also been described with specific application to CAEP testing, although the reports are fewer. Various forms of machine-based scoring, which provide a mechanism for extracting quantitative measures, have been suggested. Some rely on feature extraction using either a template of the normal adult CAEP response to identify a particular point or points within the EEG at which the voltage should be measured (Hoth, 1993; Schimmel, 1967; Schimmel et al, 1974), or discrete wavelet transformations (Hoppe et al, 2001). Where response detection is based on feature extraction, however, difficulties arise when the response under review is distorted or the component features of interest are small or missing (Picton, 1987). Any algorithm that assumes the signal of interest has predictable latency, amplitude, or phase with respect to the stimulus will not be very successful in CAEP detection (Wong & Bickford, 1980) given the wide variation in the response that exists between and within participants.

Machine-based scoring methods that do not rely on a template have also been suggested. Hoth (1993) and Wicke et al (1978) calculated the ratio of the root-mean-square (RMS) voltage of the evoked response to that of the background noise where the voltage of the background noise was determined from the pre-stimulus region within the epoch. Schimmel (1967) and Schimmel et al (1974) estimated a mean background noise component based on the alternate addition and subtraction of epochs and divided by the number of epochs (i.e. ± reference). It is however impossible to identify the magnitude of the true response in complete isolation from background noise (Elberling & Don, 1987) making these estimates less than ideal.

The application of statistical tests in CAEP detection is not new. Mason et al (1977) calculated the correlation coefficient and resulting p value for a series of voltage measures taken at specified sampling points across two averages; each based on 32 epochs. However, the number of degrees of freedom that should be applied to this correlation is signal dependent. Hoth (1993) calculated the covariance of two partial averages divided by the square root of the product of the variances of the first and second partial averages. The resulting measure of wave reproducibility was used in conjunction with a measure of the response's resemblance to the normal adult waveform and a measure of the SNR to determine CAEP detection. Salomon (1974) applied the non-parametric Friedman's rank correlation statistic (Siegel, 1956) to the voltage measurements taken at sampling points within the analysis period of each epoch, converted to ranked values. The sampling points within the response region of interest were weighted differently to those in the pre- and post-stimulus region, and the independence of each sampling point was improved by deriving amplitude values from the difference in amplitude between each consecutive pair of sample points. The consequent sampled values may have shown satisfactory independence, but the removal of absolute amplitude may decrease the efficiency with which genuine responses can be detected, also, the potential for a correlation between sequential sampling points must still exist and confound the statistical outcome.

The aim of the present study was to investigate the effectiveness of an automated statistic that did not rely on matching an assumed template, and that capitalized on the correlation between adjacent sampling points in (1) detecting the presence of CAEPs when stimuli were present, and (2) reliably reporting the absence of CAEPs when no stimuli were present. The effectiveness of this technique was compared with detection by expert examiners for the same adult-generated CAEPs, for two different sized data sets.

A rating scale, rather than a simple forced-choice 'presence/absence' of the response was used and results examined by calculating the area under the receiver operating characteristic (ROC) curve and d prime (d') that is derived from calculations of the mean of the signal distribution (i.e. stimulus present condition) and the mean of the noise distribution (i.e. no stimulus condition) (Korczak et al, 2005; McNicol, 1972; Oates et al, 2002). In essence, d' is a sensitivity index that provides a summary of the CAEP detection hit-rate (i.e. sensitivity) and false-alarm rate (i.e. specificity) in a single metric.

Method

Stimuli

The speech stimuli /g/ (duration of 20 ms) and /t/ (duration of 30 ms) were extracted from continuous discourse spoken by a female with an average Australian accent. The recording was sampled at a rate of 44.1 kHz and filtered to closely match the international long-term average speech spectrum (ILTASS) (Byrne et al, 1994). The stimuli include very little of the vowel transition. An additional high-pass filter of 250 Hz was applied to both stimuli to remove additional unwanted low frequency noise. These essentially vowel-free stimuli had a spectral emphasis in the mid- and high frequency regions as shown in Figure 1. The stimuli were presented with an alternating onset polarity to prevent any response contamination by stimulus artifact, though this is strictly unnecessary in that the stimulus ends before the response interval of interest commences. The inter-stimulus interval was 1125 ms.

Participants

CAEPs were recorded from four males and six females aged 24–61 years (mean 39.1, SD 10.9). Participants had a mean four frequency

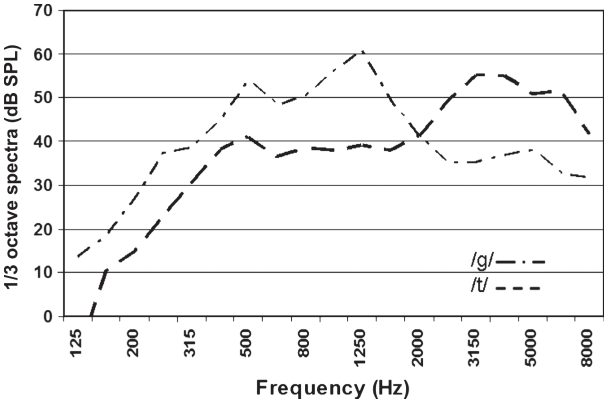

Figure 1. Spectral analysis of the two stimuli /g/ and /t/ showing the primary energy of /g/ between 800 and 1600 Hz, and /t/ above 3000 Hz.

average (500 to 4000 Hz) for the right ear of 15.5 dB HL (SD 19.08), and for the left ear it was 19.5 dB HL (SD 27.42). Prior to commencement of CAEP testing, /g/ and /t/ stimuli were presented in the free-field with a loud speaker positioned one metre distant and at 45º azimuth to the test-side. The test-side was chosen to be the same as the participant's reported better ear. If the participant reported no difference between the ears, the test-side was randomly assigned. Behavioral sound-field thresholds for both stimuli were found using a standard Hughson-Westlake threshold seeking technique. The mean free-field behavioral threshold for /g/ was 27.6 dB SPL (SD 14.4) and for /t/ it was 23.1 dB SPL (SD 14.6). These SPL values refer to the level recorded using the impulse/maximum-hold setting of the sound level meter. This setting detects the maximum short-term rms level measured with a 35-ms time constant.

Procedure

CAEP recording was conducted in the free-field with speech stimuli being delivered at five sensation levels (−10 dB, 0 dB, +10 dB, +20 dB, +30 dB SL) relative to each participant's behavioral threshold for /g/ and /t/, and a non-stimulus condition was also added. There were therefore six stimulus conditions and two stimuli presented to each of 10 adult participants. This made a total of 120 CAEP responses.

During CAEP testing, participants were alert and watched silent DVDs. Their brain electrical activity was recorded using the Neuroscan^{TM} system with electrodes positioned at Cz, C3, and C4 referenced to the right mastoid with forehead as ground. Individual sweeps of the electroencephalographic (EEG) activity were amplified and analog filtered online at 0.1–100 Hz with a sampling rate of 1000 Hz. The recording window consisted of a 100 ms pre-stimulus baseline and a further 600 ms. Artifact reject was set at ±150 μV for all electrodes. Each stimulus and stimulus condition was presented in blocks until 100 artifact-free EEG samples were acquired. Each block of stimuli was presented on two occasions to form paired (i.e. 100 x 2) data sets, with the stimulus order and condition order randomized for each participant.

These raw EEG files were baseline corrected using the averaged pre-stimulus data points, low-pass filtered at 30 Hz using a 24 dB/octave slope zero-phase filter, averaged, and then prepared for visual examination by four clinicians who each had more than five years experience in identifying adult CAEPs. The same EEG files were also transferred to MATLAB^{TM}, baseline corrected and filtered before undergoing off-line statistical analysis. A preliminary review of the replicated and overlaid waveforms for each stimulus and condition by electrode site, showed very little difference in the waveform morphology, amplitude or latency across sites and therefore responses recorded at Cz only were used in the analysis.

To create a second smaller 30 × 2 data set, the 30th to 59th artifact-free raw EEGs were extracted from each of the 100 EEG data sets and processed following the same procedure detailed for the larger 100 × 2 data set. This particular subset was selected to ensure that the averaged response was free of influence from any large amplitude initial response.

An 'InSeries' and a 'NonSeries' dataset was prepared for examiners with the former simulating the normal clinical practice of viewing responses generated in order from high to low-stimulus intensities, and the latter representing a complete randomization of paired waveforms. The 30 × 2 (i.e. 60 epochs) and 100 × 2 (i.e. 200 epochs) data sets were therefore presented to the examiners as, (1) paired responses in order of decreasing stimulus intensity (i.e. 'InSeries'), and (2) paired responses to a single stimulus and condition (i.e. 'NonSeries') as shown in Figure 2. This division between 'InSeries' and 'NonSeries' was naturally unnecessary for the statistical analysis as each analysis included data from only one stimulus and condition.

Response detection by the examiners

The paired waveforms were divided into eight tasks for the examiners to review and each task had to be completed and returned before the next was issued. The eight tasks were: (1) 'InSeries' (60 epoch) CAEP responses from the first five participants, (2) 'InSeries' (60 epoch) CAEP responses from the second five participants, (3) 'NonSeries' (60 epochs) CAEP responses from the first five participants, (4) 'NonSeries' (60 epochs) CAEP responses from the second five participants. The remaining four tasks were identical but used the larger 200 epoch data sets. All examiners were given the 60 epoch data sets before the larger 200 epoch data sets but the task order was otherwise randomized across examiners.

Examiners were asked to study each paired waveform and determine if a cortical response to auditory stimulation could be observed or not. They were not provided with any specific criteria to use in making their judgment but they were asked to rate their degree of certainty in making their decision using a five-point scale:

1. You are certain that a cortical response is absent
2. It is more likely that a response is absent but you are not certain
3. It is equally likely that a response is present or absent
4. It is likely that a response is present but you are not certain
5. You are certain that there is a cortical response present

Statistical method

Hotelling's T2, which is the multi-dimensional equivalent of the (squared) univariate t-statistic (Flury & Riedwyl, 1988) and a special case of MANOVA (Tabachnick & Fidell, 2001), was selected as the statistical method. It is suited to applications where there are multiple dependent variables that are likely to be correlated. The single-sample Hotelling's T2, tests the overall hypothesis that the population means of the outcome measures are each identical

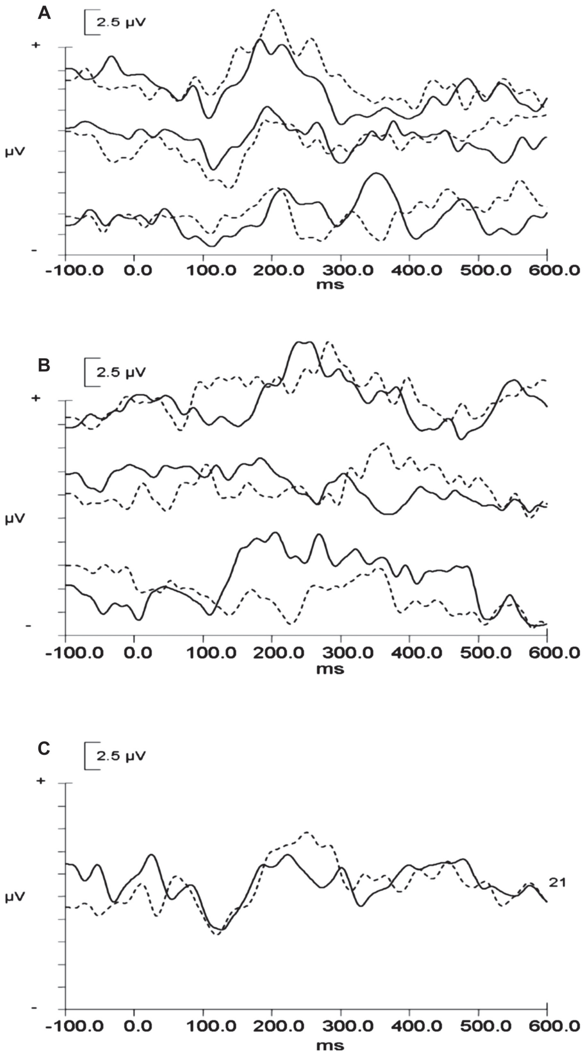

Figure 2. An 'InSeries' 60 epoch data set based on six stimulus conditions (A to F) are shown across panel A and B; a single 'NonSeries' 60 epoch data set is shown in panel C.

to a specified value that is independent of observation (Flury & Riedwyl, 1988; Harris, 2001). The test is suited to situations where each subject receives more than one experimental condition or they are measured in the same way on several occasions. The single-sample form provides a valuable tool for analysing repeated measures designs and it requires much less stringent assumptions than does the more common analysis of variance (Harris, 2001; Kinnear & Gray, 2000). As the averaged evoked response, with its multiple sampling points and multiple epochs across time, is a multivariate measure, the Hotelling's T2 statistic appears to be an appropriate choice in this context. The choice of test statistic is, however, only part of the process. The selection and manipulation of sampling points to produce the data used by the Hotelling's T2 analysis is critical. The total analysis window must be wide enough to encompass all the waveform portions likely to lead to detection. Each bin (within which the sampling points are averaged) must be sufficiently narrow that it does not encompass both a negative and a positive component or else their effects will cancel each other. Both of these considerations suggest the need for many bins, however test sensitivity will decrease as the number increases (due to a greater opportunity for chance to affect the outcome) unless the additional bins contain significant new information.

For the statistical analysis in this experiment, each accepted and epoched EEG file was divided into nine data-bins that cover an analysis period of 450 ms, from 50 ms to 500 ms post stimulus onset. Within each 50 ms bin, the multiple sampling points, that reflect the amplitude of the response, were reduced by averaging to form one variable per bin. In this way, nine variables for each epoch generated were created and entered to analysis. This particular array was chosen based on earlier data, with the aim of optimizing the trade-offs described previously. There were no subsequent manipulations of the analysis period or width of the data-bins. Having created a nine variable and 60 or 200 row data set, Hotelling's T2 was applied. Effectively, the statistic determines the probability that any linear combination of the nine variables has a mean value significantly different from zero. Because the statistic automatically applies the set of weights that best separates the weighted sum from zero, it subsequently tends to give the greatest weight to those time bins where the mean response has the greatest amplitude. A resulting p value less than or equal to 0.05 suggests that a CAEP response is likely to be present, while a p value greater than 0.05 suggests that there is no evidence that a CAEP response is present.

Calculating the d' sensitivity index

There were four cut-off criteria selected for the Hotelling's T2 generated p values (i.e. <0.005, <0.01, <0.02, <0.05) and the proportion of responses meeting each of these criteria was calculated for all stimulus conditions. These proportional values were then converted to z scores using the cumulative normal distribution. Although a number of different cut-off criteria for the Hotelling's T2 generated data could have been used, these four were consistent with commonly applied alpha levels and they were not severely influenced by extreme proportional values equal to 0 or 1 for which z scores cannot be calculated directly. These extreme values occurred when the examiners or statistical method *always* identified the presence of a response in which case the proportion would equal 1, or the examiners or statistical method *never* said there was a response for a non-stimulus trial in which case the proportion would equal 0. When the use of these floor and ceiling values could not be avoided completely, proportions equal to 0 were replaced by 0.375/R, and proportions equal to 1 were replaced by 1 – 0.375/R (where R was the number of observations on which the proportions were based) (Dillon, 1984). Although these were approximations, it was preferable to the deletion of all the data where performance was extremely good. In order to calculate d', the z score for the noise condition (i.e. non-stimulus condition) was subtracted from the z score for each stimulus present condition to form a series of difference z scores. Finally, the difference z score was averaged across all cut-off criteria to form a single d' value for each stimulus present condition.

Similarly, there were four cut-off criteria generated from the five-point scale given to examiners (i.e. >1, >2, >3, >4). The proportion of responses meeting each of these criteria was calculated for all stimulus conditions, and converted to z scores from which d' was calculated following the procedure outlined above. d' was calculated for each stimulus present condition and each examiner. To ensure

Table 1. The area under the receiver operating characteristic (ROC) curve for extreme sensation levels of –10 dB and +30 dB.

	60 epoch data set			*200 epoch data set*		
	Area	*SE**	*p*	*Area*	*SE**	*p*
–10 dB SL						
Examiner 1	0.53	0.09	0.77	0.42	0.09	0.41
Examiner 2	0.45	0.09	0.58	0.44	0.09	0.53
Examiner 3	0.41	0.09	0.31	0.53	0.09	0.78
Examiner 4	0.65	0.09	0.11	0.44	0.09	0.52
Hotelling's T^2	0.53	0.09	0.79	0.59	0.09	0.35
30 dB SL						
Examiner 1	0.97	0.03	<0.001	1	0.01	<0.001
Examiner 2	0.98	0.02	<0.001	1	<0.01	<0.001
Examiner 3	0.99	0.02	<0.001	1	<0.01	<0.001
Examiner 4	1	0.01	<0.001	1	<0.01	<0.001
Hotelling's T^2	0.99	0.02	<0.001	1	<0.01	<0.001

*SE estimated assuming bi-negative distribution.

that a comparison between the performance of four examiners and Hotelling's T2 could be based on equal sized sets of data, d' was also calculated for a 'composite' examiner by averaging the ratings from all examiners before calculating d' for each stimulus present condition.

Automated statistic and examiner detection of CAEPs

The area under the receiver operating characteristic (ROC) curve was calculated for each examiner using the five-point grading scale. The calculation of area provided a means of examining the sensitivity/specificity pairs across the complete range of decision thresholds (Zweig & Campbell, 1993). This area value is a one-statistic summary of the ROC curve and indicates the probability that a randomly chosen examiner's grade for a non-stimulus trial will exceed that of a randomly chosen examiner's grade for a stimulus-present trial (Hanley & McNeil, 1982).

To create a five-point scale for Hotelling's T2 results that was comparable to the examiners' scaling, the criterion p-levels were converted to z scores. The range of z scores was divided into five equal-divisions. The different categories thus represented different degrees of conservatism in the acceptance criteria, just as for the human expert examiners.

Results

Series and size of data set

To evaluate whether the examiners' sensitivity index in detecting CAEPs differed by series (i.e. 'InSeries' and 'NonSeries') or the size of the data set (i.e. 30 × 2 and 100 × 2), a repeated measures factorial ANOVA of d' values with series and data set as the repeated factors was performed for each sensation level. Results showed that there were no significant differences in the sensitivity index by series (i.e. −10 dB SL (F (1,3) = 1.13, p = 0.37); 0 dB SL (F (1,3) = 3.4, p = 0.16); 10 dB SL (F (1,3) = 7.34, p = 0.07); 20 dB SL (F (1,3) = 7.18, p = 0.08); 30 dB SL (F (1,3) = 2.39, p = 0.22). As would be expected, there were some significant differences in the sensitivity index by data set when the sensation level was ≥ 10 dB SL (i.e. 10 dB SL (F (1,3) = 15.92, p = 0.03); 20 dB SL (F (1,3) = 44.11, p <0.01); 30 dB SL (F (1,3) = 21.39, p = 0.02) but not at −10 dB SL (F (1,3) = 0.36, p = 0.59) or 0 dB SL (F (1,3) = 2.17, p = 0.24).

As the presentation of traces 'InSeries' and 'NonSeries' made no significant difference to examiners and there are occasions in clinical practice where judgments need to be made without reference to responses at higher presentation levels, results for 'NonSeries' only were used for further analysis. As the examiner's had significantly higher sensitivity index for detection of CAEPs for the larger data set for all the supra-threshold sensation levels, results for 30 × 2 and 100 × 2 averaged epochs were analysed separately.

The areas under the ROC curves that were generated for −10 dB SL, ranged from 0.41 to 0.65 (60 epochs) and 0.42 to 0.59 (200 epochs). When the Wilcoxon statistic was applied to test the hypothesis that examiners/Hotelling's T2 had successfully distinguished between the stimulus present and stimulus absent conditions (i.e. their determination was above chance; the area was insignificantly different from 0.500), the outcomes were not significant as shown in Table 1. This table also shows that the areas under the ROC curves that were generated for 30 dB SL (60 and 200 epochs) were all = 0.97 and significantly different from the chance level of 0.50 (p < 0.001) suggesting extremely high sensitivity/specificity for all examiners and Hotelling's T2 at this highest sensation level.

Figures 3 and 4 show the area under the curve for the remaining 0 dB, 10 dB, and 20 dB SL, for the 60 epoch data set (Figure 3) and for the 200 epoch data set (Figure 4). To investigate whether differences in performance across examiners/Hotelling's T2 exist, the standard error (SE) was estimated using the negative exponential distribution model for both the stimulus present and stimulus absent conditions (Hanley & McNeil, 1982) and the 95% confidence interval (CI) was calculated as shown.

At 0 dB SL, the majority of examiners (and Hotelling's T2) performed no better than chance in successfully distinguishing between the stimulus present and stimulus absent condition when the data set was small (i.e. 60 epoch data set) but examiners 3, 4, and Hotelling's T2 demonstrate moderate sensitivity/specificity with areas under the curve at 0.7–0.85 when the data set is larger. At 10 dB SL, the

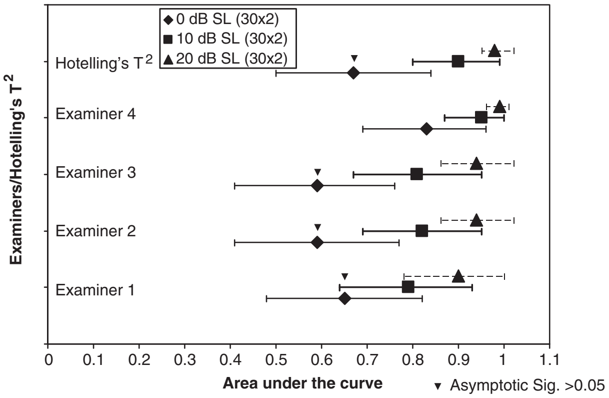

Figure 3. The area under the receiver operating characteristic (ROC) curve and 95% CI for the 60 data set are shown for all examiners and Hotelling's T2.

areas under the curve demonstrate higher sensitivity/specificity than that seen for 0 dB SL although there is overlap in the CIs for the smaller data set in particular. When comparing the areas under the curve across examiners/Hotelling's T2, there is substantial overlap of CIs. At 20 dB SL, all examiners/Hotelling's T2 achieve areas under the curve of 0.9 to 1.0, indicating near perfect sensitivity/specificity. Examiner 4 shows the highest sensitivity/specificity and smallest CI of all the examiners when the data set is small. Hotelling's T2 achieves comparable sensitivity/specificity to that of examiner 4 when the sensation level is 10 dB or above. At the highest sensation level, especially for the larger data set, mistakes by either the statistic or the examiners are so rare that performance reaches a ceiling. Area under the ROC curve approaches 1.0, and it is difficult to calculate the d' score meaningfully, so it too reaches a ceiling.

Figures 5 and 6 show d' as a function of sensation level for the 'composite' examiner and Hotelling's T2, for 200 epoch data set and 60 epoch data set respectively. It is clear that the sensitivity index increased for the 'composite' examiner and for Hotelling's T2 as the sensation level increased. It is also clear that the sensitivity index increased when the size of the data set increased, for all sensation levels but −10 dB SL.

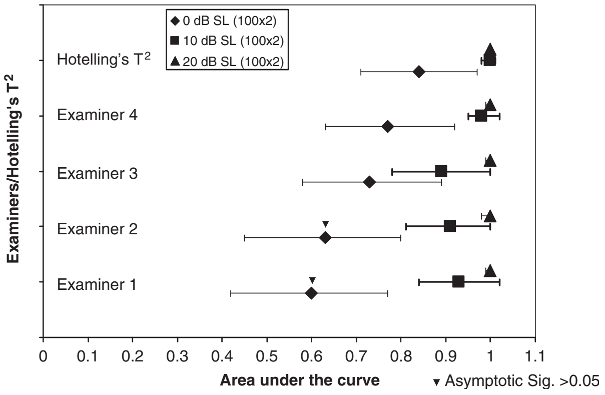

Figure 4. The area under the receiver operating characteristic (ROC) curve and 95% CI for the 200 data set are shown for all examiners and Hotelling's T2.

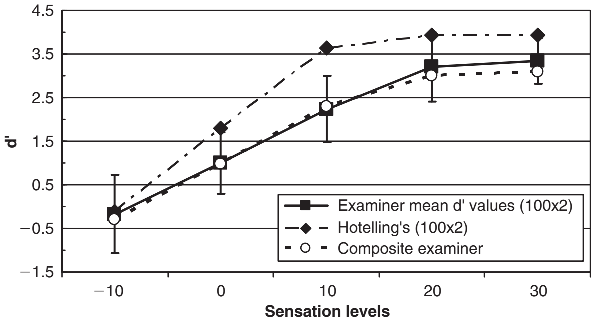

Figure 5. d prime (d') as a function of sensation levels is shown for the 'composite' examiner, the mean of the examiners, and Hotelling's T2 for the 200 epoch data set. The error bars indicate the expected d' range for a population of examiners with similar experience.

Finally, as the calculation of d' values is not a linear process, the mean d' of the four examiners was calculated and compared with Hotelling's T2, and these are also shown in Figures 5 and 6. Despite the non-linearity of the statistical process, the d' values for the composite examiner are almost identical to the average of the d' values calculated separately for each examiner. The expected d' range for a population of examiners with similar experience is also shown where the error bars were calculated by the formula:

$$\text{Mean } d' \pm t\,(3) * \text{SD}$$

Results show that the expected d' range for a population of examiners is much greater for the smaller 60 epoch data set than the 200 epoch data set. The Hotelling's T2 sensitivity index is either within the upper boundary of that range or above the expected range for examiners, using both the 60 and 200 epoch data sets.

Discussion

This experiment has used a quantitative analysis of the accuracy of detecting an electrophysiological response (i.e. analysis of specificity and sensitivity combined in the receiver operating curve)

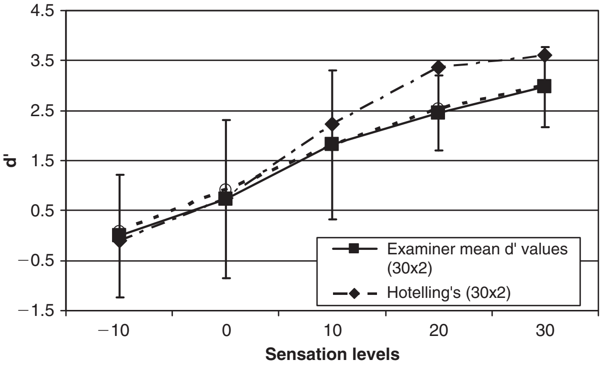

Figure 6. d prime (d') as a function of sensation levels is shown for the 'composite' examiner, the mean of the examiners, and Hotelling's T2 for the 60 epoch data set. The error bars indicate the expected d' range for a population of examiners with similar experience.

by both expert human observers and a novel statistical technique. While electrophysiological responses provide an objective measurement tool in the sense that the participant is not required to respond, there remains considerable subjectivity in determining the presence/absence of a response when visual detection methods are used. Inter-observer and intra-observer variations, particularly when specific criteria for response detection are not provided, have long been reported (Cohen et al, 1971; Rose et al, 1971). How effective the individual examiner is in response detection will be determined by their degree of experience. If highly specific rules are provided, the novice and the expert should arrive at the same outcome, but everything that experience teaches must be captured in the rules, or the benefits of experience are lost. When the response deviates from the expected average pattern (which is often the case with CAEPs) or when responses are generated at near-threshold stimulation, even experienced examiners will not always agree (Davey et al, 2007; Pratt et al, 1995; Schimmel et al, 1974).

When experienced examiners are asked to stipulate the rules that they use in determining the presence/absence of an electrophysiological response, it appears that they rely on some combination of response replication and response tracking as the stimulus presentation level is reduced. Where possible, agreement between multiple observers may also be required before a response or response feature can be considered present/absent (Elberling & Don, 2007; Garinis & Cone-Wesson, 2007; Oates et al, 2002). In our study, the examiners were not supplied with any specific detection rules nor did we give them an opportunity to cross-check their answers with each other. We did, however, provide our examiners with replicated data sets that were presented 'InSeries' and 'NonSeries' format. We were therefore able to determine the importance of response tracking in their decision process. We predicted, based on our own observations and the literature (Delgado & Ozdamar, 1994; Elberling & Don, 2007), that presenting replicated responses 'InSeries' should increase the examiner's response detection sensitivity index. Our results, however, showed that as a group there was a small but insignificant increase in the detection sensitivity index when the replicated responses were presented 'InSeries', and this was the case for both the small and larger data sets. It may be that the 'response tracking' rule which relies on identifying a particular response morphology for each individual cannot be as readily applied to CAEP detection as in the detection of the earlier potentials. This is possibly due to the small number of relatively broad peaks in the CAEP response with poorly defined latencies. Without the aid of response tracking, CAEP detection is likely to be more challenging than the detection of ABR even for the most experienced examiner.

All examiners and Hotelling's T2 showed increasing detection sensitivity index as the sensation level (above threshold) increased. This finding is not surprising as increasing the sensation level should increase the response amplitude, which increases the SNR, and hence should facilitate identification whether by human examiner or statistical technique.

It has been reported that a CAEP average consisting of 20–50 epochs, and replicated (Garinis & Cone-Wesson, 2007; Thornton, 2007; Tomlin et al, 2006), should provide an adequate SNR for visual response detection and that averaging an excessive number of epochs can be detrimental to response amplitude (Hyde, 1997; Walter, 1964). By contrast, other studies have suggested that extended averaging should not impair response amplitude (Lightfoot & Kennedy, 2006) as the greatest decrease in amplitude occurs from the first to the second epoch after which amplitude stabilizes (Budd et al, 1998; Henry & Teas, 1968; Woods & Elmasian, 1986). We examined the detection sensitivity index using data sets that varied in size. The smaller 60 epoch data set was consistent with recommendations made for CAEP data collection in adults and we also chose a larger size data set, which would not be routinely applied in a clinical setting, but provided a means of testing whether an increase in the number of epochs affected the detection sensitivity index for the examiners and/or Hotelling's T2. Our results showed that the examiners' detection sensitivity was significantly greater for the larger data set than for the smaller data set, for each of the sensation levels above threshold. This finding was demonstrated by the 'composite' examiner and the predicted population of examiners (Figures 5 and 6) but it was less clear for individual examiners (Figures 3 and 4). When Hotelling's T2 was applied, there was the same notable increase in the detection sensitivity index for all but −10 dB SL. At 10 dB SL and above, confidence in the accuracy of the detection sensitivity index was much greater when the size of the data set was larger. Our results therefore suggest benefit for both the examiner and the statistic in using the larger data set whenever the sound is above threshold. This occurred, no doubt, from the impact of additional averaging where reduced electrophysiological noise is expected to lead to an increase in the SNR. This supports the hypothesis that extended averaging, at least to 200 epochs, facilitates response determination.

The main aim of this study was to compare the sensitivity/specificity of CAEP response detection for experienced examiners and an automated statistic (i.e. Hotelling's T2). We compared the performance of this statistic with our examiners as individuals, as a 'composite' examiner, and with an estimate of a population of similar examiners. Our results demonstrated the effectiveness of Hotelling's T2 in detecting CAEPs in several ways. First, the Hotelling's T2 statistic was more able than the composite examiner to differentiate a CAEP from random electrical activity at sensation levels of 10 dB or more. Second, it has a sensitivity index that was equal to the best of the individual examiners. Third, it may be slightly better in response detection than a population of similar examiners when the sensation level is low (i.e. 10 dB SL) and the variance is reduced by increasing the size of the data set. Fourth, it was no better or worse in response detection than the majority of examiners when stimuli were presented at threshold and the data set was small. There is, however, a possible measurement limitation which may have impacted these results and their interpretation. The range in the detection sensitivity index, as illustrated by the d' data, was reduced because of restrictions imposed by the conversion of floor and ceiling proportional data to z scores. This range restriction applied to data from both the examiners and the statistic, but more often for the statistic, and may have masked more substantial differences between them. We chose to employ experienced examiners in this study but it seems reasonable to assume that novice examiners, working without imposed rules, would achieve lower detection sensitivity/specificity than these experienced examiners. This being the case, the performance differences between inexperienced examiners and Hotelling's T2 could only increase.

The analysis of sensitivity and specificity has been applied previously for evaluating evoked potential detection methods, though not for detection of cortical responses. When the sensitivity/specificity for visual detection of ABR responses was compared with three different objective methods; namely correlation, variance ratio, and multiple pre-post test z score tests, visual detection gave the highest sensitivity index (Arnold, 1985). However, while visual scoring was statistically the most sensitive, the practical difference between measures was reportedly small. Valdes-Sosa et al (1987) reported that their chosen statistical methods; namely correlation, the ratio of the averaged response to an estimate of the noise, and Hotelling's T2 computed for discrete Fourier transforms of response sub-averages, provided a 'hit rate' that was equal to the rate for visual detection or, in the case of Hotelling's T2, slightly superior. Our CAEP experiment also suggested that response detection using Hotelling's T2 was at least equal to that of a good examiner with several years experience. Although several methodological differences exist between the two ABR studies and our CAEP study, it may be that this statistic is particularly suited to response detection in evoked potential studies because of the multivariate nature of the responses and the capacity to compare these with zero (Picton, 1987). One substantial difference between the two ABR studies was in the origins of the ABR waveforms. In the former study, adult-generated responses were used while in the latter study, where the automated statistic was at least equal to the examiners in the detection rate, infant-generated recordings (aged 40 to 55 weeks GA) were used. The difference in study outcomes may therefore reflect the inherent difficulties posed by detecting infant responses, which are far less robust than those from adults. If this is the explanation, then the application of statistical detection to infant CAEP testing may prove more useful than we have already seen in adult testing.

Acknowledgements

Selected results from this paper were presented at Audiology Now! in 2006 and at the XX IERASG Biennial Symposium in 2007. The authors would like to acknowledge the instructive comments of Bram Van Dun, regarding the physiological mechanisms of the CAEP.

Declaration of interest: The authors report no conflicts of interest. The authors alone are responsible for the content and writing of the paper.

References

Abrams, D.A., Nicol, T., Zecker, S.G. & Kraus, N. 2006. Auditory brainstem timing predicts cerebral asymmetry for speech. *J Neuroscience*, 26 (43), 11131–11137.

Agung, K., Purdy, S.C., McMahon, C.M., Newall, P. 2006 The use of cortical auditory evoked potentials to evaluate neural encoding of speech sounds in adults. *J Am Acad Audiol*, 18 (8), 559–72.

Arnold, S.A. 1985. Objective versus visual detection of the auditory brain stem response, *Ear Hear*, 6 (3), 144–150.

Budd, T.W., Barry, R.J., Gordon, E., Rennie, C. & Michie, P.T. 1998. Decrement of the N1 auditory event-related potential with stimulus repetition: Habituation vs. refractoriness. *Int J Psychophysiol*, 31, 51–68.

Byrne, D., Dillon, H., Tran, K., Arlinger, S., Bamford, J. et al. 1994, An international comparison of long-term average speech spectra. *J Acoust Soc Am*. 96 (4) 2108–20.

Cohen, M.M., Rapin, I., Lyttle, M. & Schimmel, H. 1971. Auditory evoked response (AER). *Arch Otolaryngol*, 94, 214–219.

Coles, R.A. & Mason, S.M. 1984. The results of cortical electric response audiometry in medico-legal investigations. *Br J Audiol*, 18, 71–78.

Cone-Wesson, B. & Wunderlich, J. 2003. Auditory evoked potentials from the cortex: Audiology applications. *Curr Opin Otolaryngol Head Neck Surg*, 11 (5) 372–377.

Davey, R., McCullagh, P., Lightbody, G. & McAllister, G. 2007. Auditory brainstem response classification: A hybrid model using time and frequency features. *Artif Intell Med*, 40, 1–14.

Davis, H. 1965. Slow cortical responses evoked by acoustic stimuli. *Acta Otolaryngol*, 59, 179–185.

Delgado, R.E. & Ozdamar, O. 1994. Automated auditory brainstem response interpretation. *IEEE Eng Med Biol Mag*, April/May, 227–237.

Dillon, H. 1984. *A procedure for subjective quality rating of hearing aids.* National Acoustic Laboratories, Commonwealth Department of Health Report number 100. Canberra, Australia: Australian Government Publishing Service.

Dillon, H. 2001. *Hearing Aids.* New York: Thieme.

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

Eggermont, J.J. In: R.F. Burkard, M. Don & J.J. Eggermont. 2007 *Auditory Evoked Potentials. Basic Principles and Applications*, Philadelphia: Lippincott, Williams, and Wilkins.

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

Elberling, C. & Don, M. 1987. Detection functions for the human auditory brainstem response. *Scand Audiol*, 16, 89–92.

Elberling, C. & Don, M. 2007. Detecting and assessing synchronous neural activity in the temporal domain (SNR, response detection). In: R.F. Burkard, M. Don & J.J. Eggermont (eds.) *Auditory Evoked Potentials.* Baltimore, MD: Lippincott, Williams & Wilkins, pp. 102–123.

Flury, B. & Riedwyl, H. 1988. *Multivariate Statistics: A Practical Approach.* London: Chapman and Hall.

Gage, N.M. & Roberts, T.P.L. 2000. Temporal integration: Reflections in the M100 of the auditory evoked field. *Neuroreport* 11 (12), 2723–26.

Garinis, A.C. & Cone-Wesson, B.K. 2007. Effects of stimulus level on cortical auditory event-related potentials evoked by speech. *J Am Acad Audiol*, 18 (2), 107–116.

Hall III, J.W. 2007. *New Handbook of Auditory Evoked Responses*. Pearson, Boston.

Hanley, J.A. & McNeil, B.J. 1982. The meaning and use of the area under a receiver operating characteristic curve. *Radiology*, 143, 29–36.

Harris, R.J. 2001. *A Primer of Mulitvariate Statistics.* (3rd ed.) Mahwah, NJ: Lawrence Erlbaum Associates.

Henry, G.B. & Teas, D.C. 1968. Averaged evoked responses and loudness: Analysis of response estimates. *J Speech Hear Res*, 11, 334–342.

Hoppe, U., Weiss, S., Stewart, R.W. & Eysholdt, U. 2001. An automated sequential recognition method for cortical auditory evoked potentials. *IEEE Trans Biomed Eng*, 48 (2), 154–164.

Hoth, S. 1993. Computer-aided hearing threshold determination from cortical auditory evoked potentials. *Scand Audiol*, 22, 165–177.

Hyde, M. 1994. The slow vertex potential: Properties and clinical applications. In: J.T. Jacobson (ed.) *Principles and Applications in Auditory Evoked Potentials.* Needham Heights MA: Allyn & Bacon, pp. 179–218.

Hyde, M. 1997. The N1 response and its applications. *Audiol Neurootol*, 2, 281–307.

Hyde, M., Alberti, P.W., Matsumoto, N. & Li, Y.L. 1986. Auditory evoked potentials in audiometric assessment of compensation and medicolegal patients. *Ann Otol Rhinol Laryngol*, 95, 514–519.

Jewett, D. & Williston, J. 1971. Auditory-evoked far-fields averaged from the scalp of humans. *Brain*, 94, 681–696.

Katz, J. 1985. *Handbook of Clinical Audiology; Third Edition*. Baltimore: Williams & Wilkins.

Kinnear, P.R. & Gray, C.D. 2000. *SPSS for Windows Made Simple.* Hove UK: Psychology Press, Taylor & Francis Group.

Korczak, P.A., Kurtsberg, D. & Stapells, D.R. 2005. Effects of sensorineural hearing loss and personal hearing aids on cortical event-related potential and behavioural measures of speech-sound processing. *Ear Hear*, 26 (2), 165–185.

Kushnerenko, E., Ceponiene, R., Balan, P., Fellman, V., Huotilainen, M. et al. 2002. Maturation of the auditory event-related potentials during the first year of life. *Neuroreport*, 13 (1), 47–51.

Lightfoot, G. & Kennedy, V. 2006. Cortical electric response audiometry hearing threshold estimation: Accuracy, speed, and the effects of stimulus presentation features. *Ear Hear*, 27 (5), 443–456.

Martin, B.A. & Boothroyd, A. 1999. Cortical, auditory, event-related potentials in response to periodic and aperiodic stimuli with the same spectral envelope. *Ear Hear*, 20 (1), 33–44.

Mason, S.M., Su, A.P. & Hayes, R.A. 1977. Simple online detector of auditory evoked cortical potentials. *Med Biol Eng Comput*, 15, 641–647.

McNicol, D. 1972. *A Primer of Signal Detection Theory*. London: George Allen & Unwin Ltd.

Molfese, D.L. 1978. Neuroelectrical correlates of categorical speech perception in adults. *Brain Lang*, 5 (1), 25–35.

Nunez, P.L. & Srinivasan, R. 2006. *Electric Fields of the Brain: The Neurophysics of EEG*. Oxford University Press.

Oates, P.A., Kurtzberg, D. & Stapells, D.R. 2002. Effects of sensorineural hearing loss on cortical event-related potential and behavioural measures of speech-sound processing. *Ear Hear*, 23 (6), 399–415.

Perez-Abalo, M.C., Savio, G., Torres, A., Martin, V., Rodriguez, E. et al. 2001. Steady state responses to multiple amplitude-modulated tones: An optimizing method to test frequency-specific thresholds in hearing-impaired children and normal-hearing subjects. *Ear Hear*, 22 (3), 200–211.

Picton, T. 1987. The recording and measurement of evoked potentials. In: A.M. Halliday, S.R. Butler & R. Paul (eds.) *A Textbook of Clinical Neurophysiology*. New York: John Wiley & Sons, pp. 23–40.

Picton, T., Dimitrijevic, A. & John, M.S. 2002. Multiple auditory steady-state responses. Ann Otol Rhinol Laryngol, *Suppl.*, 189, 16–21.

Pratt, T.L., Olsen, W.O. & Bauch, C.D. 1995. Four-channel ABR recordings: Consistency in interpretation. *Am J Audiol*, 4, 47–54.

Rickards, F.W., DeVidi, S. & McMahon, D.S. 1996. Cortical evoked response audiometry in noise induced hearing loss claims. *Aust J Otolaryngol*, 2(3), 237–241.

Rose, D.E., Keating, L.W., Hedgecock, L.D., Schreurs, K.K. & Miller, K.E. 1971. Aspects of acoustically evoked responses. *Arch Otolaryngol*, 94, 347–350.

Salomon, G. 1974. Electric response audiometry (ERA) based on rank correlation. *Audiol*, 13, 181–194.

Schimmel, H. 1967. The (±) reference: Accuracy of estimated mean components in average response studies. *Science*, 157 (3784), 92–94.

Schimmel, H., Rapin, I. & Cohen, M.M. 1974. Improving evoked response audiometry with special reference to the use of machine scoring. *Audiol*, 13, 33–65.

Siegel, S. 1956. *Nonparametric Statistics: For the Behavioural Sciences.* Tokyo: McGraw-Hill Book Company Inc.

Song, J.H., Banai, K. & Kraus, N. 2008. Brainstem timing deficits in children with hearing impairment may result from corticofugal origins. *Audiol Neuro,* (13), 335–344.

Stapells, D.R. & Kurtzberg, D. 1991. Evoked potential assessment of auditory system integrity in infants. *Clin Perinatol*, 18 (3), 497–518.

Stelmachowicz, P. 1999. Hearing aid outcome measures for children. *J Am Acad Audiol*, 10, 14–25.

Sturzebecher, E. & Cebulla, M. 1997. Objective detection of auditory evoked potentials: Comparison of several statistical tests in the frequency domain on the basis of near-threshold ABR data. *Scand Audiol*, 26 (7), 7–14.

Tabachnick, B.G. & Fidell, L.S. 2001. *Using Multivariate Statistics.* (4th ed.) Needham Heights MA: Allyn & Bacon.

Thornton, A.R.D. 2007. Instrumentation and recording parameters. In: R.F. Burkard, M. Don & J.J. Eggermont (eds.) *Auditory Evoked Potentials.* Baltimore: Lippincott, Williams & Wilkins, pp. 73–101.

Tomlin, D., Rance, G., Graydon, K. & Tsialios, I. 2006. A comparison of 40 Hz auditory steady-state response (ASSR) and cortical auditory evoked potential (CAEP) thresholds in awake adult subjects. *Int J Audiol*, 45, 580–588.

Valdes-Sosa, M.J., Bobes, M.A., Perez-Abalo, M.C., Perera, M., Carballo, J.A. et al. 1987. Comparison of auditory-evoked potential detection methods using signal detection theory. *Audiol*, 26, 166–178.

Walter, W.G. 1964. Retrospective summary of definitive tests for hearing in young children. *Acta Otolaryngol (suppl)*, 206, 162–172.

Wicke, J.D., Goff, W.R., Wallace, J.D. & Allison, T. 1978. On-line statistical detection of average evoked potentials: Application to evoked audiometry (ERA). *Electroenceph Clin Neurophysiol*, 44, 328–343.

Wong, P.K.H. & Bickford, R.G. 1980. Brain stem auditory evoked potentials: The use of noise estimates. *Electroencephalog Clin Neurophysiol*, 50, 25–34.

Woods, D.L. & Elmasian, R. 1986. The habituation of event-related potentials to speech sounds and tones. *Electroencephalog Clin Neurophysiol*, 65 (6), 447–459.

Wunderlich, J. & Cone-Wesson, B. 2006. Maturation of CAEP in infants and children: A review. *Hear Res*, 212, 212–223.

Yeung, K.N.K. & Wong, L.L.N. 2007. Prediction of hearing thresholds: Comparison of cortical evoked response audiometry and auditory steady state response audiometry techniques. *Int J Audiol*, 46, 17–25.

Zweig, M.H. & Campbell, G. 1993. Receiver-operating characteristic (ROC) plots: A fundamental evaluation tool in clinical medicine. *Clin Chem*, 39 (4), 561–577.

Appendix

The Hotelling's formula is:

$$T^2 = n(\bar{x} - \mu_0)\mathrm{S}^{-1}\,(\bar{x} - \mu_0)$$

where

$$\bar{x} = \begin{bmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \vdots \\ \bar{x}_p \end{bmatrix} \quad \mu_0 = \begin{bmatrix} \mu_1^0 \\ \mu_2^0 \\ \vdots \\ \mu_p^0 \end{bmatrix}$$

S is the covariance matrix of the p variables (the voltage within the nine time bins in our case). n is the number of observations (the number of epochs in our case), μ_0 is the hypothesized array of values (all zero in our case) against which the variables x_i, are tested, and $\bar{x}_i$ is the mean value across epochs of variable x_i.