Available online at www.sciencedirect.com

ScienceDirect

Medical Engineering & Physics 29 (2007) 191–198

Medical Engineering & Physics

www.elsevier.com/locate/medengphy

Objective detection of evoked potentials using a bootstrap technique

Jing Lv*, David M. Simpson, Steven L. Bell

Institute of Sound and Vibration Research, University of Southampton, Highfield, Southampton, SO17 1BJ, UK

Received 27 July 2005; received in revised form 17 February 2006; accepted 3 March 2006

Abstract

Evoked potentials are usually evaluated subjectively, by visual inspection, and considerable differences between interpretations can occur. Objective, automated methods are normally based on calculating one (or more) parameters from the data, but only some of these techniques can provide statistical significance (*p*-values) for the presence of a response. In this work, we propose a bootstrap technique to provide such *p*-values, which can be applied to a wide variety of parameters. The bootstrap method is based on randomly resampling (with replacement) the original data and gives an estimate of the probability that the response obtained is due to random variation in the data rather than a physiological response.

The method is illustrated using auditory brainstem responses (ABRs) to detecting hearing thresholds. The flexibility of the approach is illustrated, showing how it can be used with different parameters, numbers of stimuli and with user-defined false-positive rates. The bootstrap method provides a new, simple and yet powerful means of detecting evoked potentials, which is very flexible and readily adapted to a wide variety of signal parameters.

Keywords: Evoked potentials; Auditory brainstem response; Bootstrap; Signal processing

1. Introduction

Evoked potentials are the electrical signals generated by the brain in response to sensory stimuli, most commonly from the ears, eyes or the somatosensory system. Usually, the responses are interpreted subjectively, by visual inspection. However, this requires well trained professionals, and is strongly dependent on the experience of the observer. Objective, automated methods for detecting responses are clearly desirable, especially for screening (e.g. neonatal hearing tests) and monitoring (e.g. during surgery).

One of the most widely used evoked potentials is the auditory brainstem response (ABR) used extensively to determine hearing thresholds in patients that are unable or unwilling to cooperate with behavioural testing. We will illustrate the proposed method using this application, for which we give some detailed results. However, the technique proposed could also readily be applied in other modalities.

The conventional way to analyze and interpret the ABR is visual inspection by experienced audiologists, who usually identify significant peaks (the most important in the ABR are denoted with roman numerals I, III and V – see Fig. 1). However, this identification is subjective, and considerable inconsistency has been found between different experienced professionals in estimating hearing thresholds [1,2] from the ABR. As a result of this, a number of methods and algorithms for automated ABR identification and detection have been described in the literature. Some of these identify the highest amplitudes in latency regions where peaks are expected to occur in normal subjects [3–5]. Others are based on different statistical properties, either in the time-domain (e.g. F_{sp} [6] and ± *difference* [7]), or in the frequency domain (e.g. magnitude-squared coherence (MSC) [8], phase coherence [9], spectral *F*-test [10]). Some of these methods provide an exact statistical criterion (*p*-value) when a response can be considered to be significant, others do not. The advantage of the former is that the false-positive-rate provides a clearly defined criterion for detecting responses, whereas for the latter empirically derived threshold criteria are used, so it becomes difficult to compare

[*] Corresponding author. Tel.: +44 23 80593221; fax: +44 23 8059 3190.
E-mail address: jl2@isvr.soton.ac.uk (J. Lv).

1350-4533/$ – see front matter © 2006 IPEM. Published by Elsevier Ltd. All rights reserved.
doi:10.1016/j.medengphy.2006.03.001

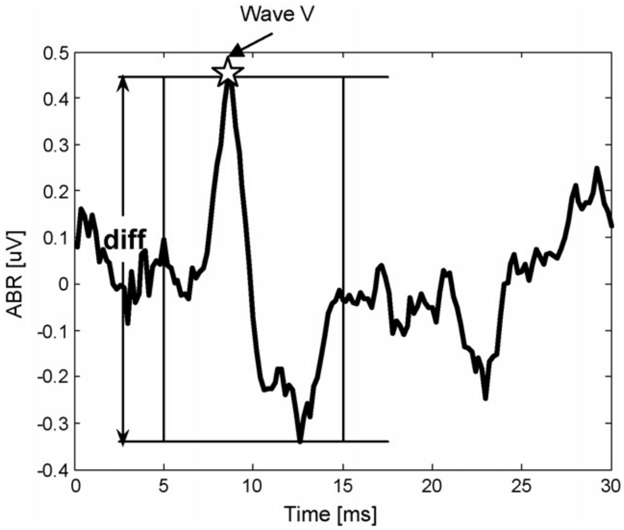

Fig. 1. The ABR for one subject (click stimulation at 30 dB SL). The vertical lines at 5 and 15 ms show the region of the response that was used in analysis. The parameter *diff* gives the range of the ABR within this interval. The symbol ☆ indicates wave V.

techniques based on the trade-off between sensitivity and specificity.

In this work, we describe a bootstrap technique, which allows the statistical significance to be estimated for a wide range of different parameters used in the objective detection of evoked responses, and does so in an easy and very flexible manner. The bootstrap method was introduced by Efron [11–13] as an approach to calculate confidence intervals for parameters in circumstances where standard methods cannot be easily applied. The bootstrap has subsequently become well established as a powerful statistical tool, in which complex mathematical analysis is replaced by intensive computational load. In recent years bootstrap methods have also been extensively used in biomedical signal processing [14,15]. To the best of our knowledge this technique has not previously been used in the detection of evoked potentials, though other uses of the bootstrap in evoked potentials have been reported [16–18].

In the following, we first describe how we recorded our ABR data. Next, we described the bootstrap approach. We then provide the results from a Monte-Carlo study carried out to evaluate the performance of the technique in well-controlled conditions simulating no stimulus response. Hearing thresholds detected by ABRs are then found in 12 subjects using the bootstrap method, and compared to those determined by experienced professionals through visual analysis of the evoked potentials. Finally, we discuss the results, and other potential applications of the proposed method, and some of its limitations.

2. Data

ABRs were recorded from 12 normal-hearing adults subjects (6 males and 6 females), who were aged between 18 and 30 years. The ABR was recorded between the vertex and the nape of neck, with a frontal electrode serving as ground. The auditory stimulation was a rectangular click stimulus with a duration of 100 μs delivered by ER-2 insert phones (Etymotic, USA), at a click rate of 33.3 Hz. Stimulation started at 50 dB sensation level (SL), decreasing in 10 dB steps to 0 dB SL. Here, ‘dB SL’ refers to the stimulus level above the auditory threshold level of the subject, as determined from conventional audiometry. The insert phones and associated cables were screened to minimize electromagnetic artefacts. The number of stimuli contributing to each coherent averaged response was $K \approx 2000$. Two recordings were made at each stimulus intensity, in each subject. The acquired raw signals were band-pass filtered between 30 and 2100 Hz in order to emphasize wave V – which is the most important feature of ABRs (Fig. 1). In addition a notch filter (50 Hz) was applied to remove mains noise. The signal was sampled at 5 kHz. The ABR was then obtained by coherently averaging the ensemble of data segments following each stimulus. The bootstrap method then uses both the averaged waveforms and the raw recorded signal, prior to averaging. The latter, containing spontaneous background cerebral activity, and noise as well as the ABR, will be referred to as the ‘EEG’.

3. Methods

From the evoked potential, we calculated parameters that quantify the strength of stimulus response – these will be described first. Their statistical significance is subsequently tested using the bootstrap method.

3.1. Parameters used in detecting ABRs

For click stimuli in adults, a time window of 10 ms or 12 ms is usually sufficient to record the ABR, because wave V occurs in normal individuals within 5–6 ms of the stimulus at high intensities and within 8–9 ms for intensities near threshold [19]. We kept the analysis window from 5 to 15 ms, which should in all cases include wave V. The four parameters described below were then calculated from the ABRs. Each of these provides a measure of the strength of the stimulus response, and is calculated over the time-interval 5–15 ms.

- *diff* [20], is the difference between the maximum and minimum value of the ABR, as shown in Fig. 1;
- power is the mean *power* of the ABR:

$$\text{power} = \frac{1}{M}\sum_{i=1}^{M} x[i]^2 \tag{1}$$

where $x[i]$ is the amplitude of each sample in the ABR signal and M is the number of samples in the analysis window 5–15 ms ($M = 50$). Clearly, when a strong stimulus response is present, the power of the coherent average will increase.

- F_{sp} is an estimate of the signal-to-noise ratio of the evoked potential, which has been used extensively in detecting ABRs [6,21]:

$$F_{\text{sp}} = \frac{\text{var}(\overline{\text{ABR}})}{\text{var(SP)}/K} \tag{2}$$

  where var($\overline{\text{ABR}}$) is the variance of the coherently average ensemble between 5 and 15 ms after the onset of the stimulus, and var (SP) is the variance of the ensemble of K (≈2000 in our application) stimulus-responses at a single point. Thus, var (SP) is obtained from the ensemble of signals before averaging, and represents the power of the noise (background activity), and var($\overline{\text{ABR}}$) is found from the coherent average and corresponds to the power of the ABR. Since the variance of the EEG can be assumed to be constant over the interval between stimuli, the single point can be chosen arbitrarily; we chose 10 ms, others have chosen 6 ms [6].
- ±*difference* is an alternative estimate of the signal-to-noise ratio [7] and is found by first allocating the even-numbered stimulus responses to one ensemble, and the odd-numbered ones to another. The coherent average of each of these two ensembles is then found. Hence,

$$\pm\text{difference} = \frac{\text{std(Sum)}}{\text{std(Diff)}} \tag{3}$$

  where the numerator refers to the standard deviation of the sum of the two averages, calculated over the time-window from 5–15 ms following the stimuli, and the denominator to the standard deviation of the difference of the two averages. Clearly, if there is a strong stimulus-response, the sum of the averages will be much larger than their difference (where stimulus responses are cancelled), leading to relatively large ± differences.

Following the calculation of these parameters, the statistical significance of each is tested against the null-hypothesis (H0) of no stimulus-response.

3.2. Bootstrap test

The bootstrap method [11–13,22,23] is based on repeatedly drawing random samples (with replacement) from the original data. The parameter of interest is then calculated from these 'resamples', building up an estimate of the sampling distribution of the estimated parameter (we use symbol θ to denote any of the four parameters described above). The bootstrap method allows confidence limits of the estimate to be determined, or the statistical significance (with respect to some null hypothesis) to be tested – as in the current application.

First the coherent average of the EEG is calculated by averaging the K stimulus-responses, from which the parameters (to generalize, these will be denoted by θ) are found. We then apply the bootstrap test, by selecting K random points anywhere throughout the recorded raw signal, and use these as starting points in obtaining an ensemble of K corresponding segments with same length as the stimulus-response. Thus, at this step, we thus ignore the actual timing of the stimuli and use random 'trigger points'. A uniform distribution of starting points covering the entire length of the recorded data is used. The new ensemble of K segments is averaged to form an 'incoherent average' (because it is not synchronized with the stimulus-timing), for which the parameter θ is again calculated. The parameter, from the 'bootstrap' resample, will be denoted as θ^*. The bootstrap resampling process is then repeated $L = 499$ times, and a 'bootstrap distribution' of θ^* is obtained. This provides an estimate of the sampling distribution of the parameter θ^* as would be expected if there is no stimulus response present (H0). By comparing θ with the cumulative probability distribution of θ^* (see Fig. 2), we find the fraction of θ^* that are larger than θ: this is the estimated p-value. If this is smaller than some chosen significance level α (say $\alpha = 5\%$), we reject the null-hypothesis of no response (Fig. 2, right) and consider the value of θ to be statistically significant, i.e., a response has been detected. If all $\theta^* < \theta$, we say $p < 1/\text{L}$ (i.e., $p < 0.002$ in our case of $L = 499$). If $p > \alpha$ and θ is towards the left of the distribution of θ^* (Fig. 2, left plot) we accept the null hypothesis of no response.

3.3. Monte-Carlo simulation

In order to test the proposed methods, we first carried out a Monte-Carlo study, simulating signals with no stimulus response. The aim was to determine whether the selected false positive rate ($\alpha = 5\%$) is actually obtained, when no response is present. We used simulated signals in this task, in order to obtain the large amount of well-controlled test data required for this task. We used an autoregressive (AR) model to simulate the EEG signals. The AR parameters were estimated from one of the recorded signals by the Yule–Walker method. We selected this 'model' signal by identifying the EEG recording whose power spectrum was closest to the median power spectrum of the 12 recordings at 0 dB SL. This may be considered the most 'typical' of all the recorded signals. We estimated the AR model order of this signal according to the Final Prediction Error (FPE) [24], from which an order of 16 was selected. It was found that the FPE did not give a minimum but showed an initial sharp decrease, and after a 'knee' an almost flat section, where higher orders would lead to minimal improvements in FPE. The order chosen corresponds to the point just after the 'knee'. The estimated AR power spectrum does not show spectral peaks at the stimulus frequency or its harmonics. We then simulated 500 EEG signals. All four parameters (*diff*, *power*, F_{sp}, ±*difference*) were calculated from the coherent average of these signals (with trigger points at 30 ms intervals, and analysing the time-interval from 5–15 ms following each stimulus) and tested the significance (with $\alpha = 5\%$) using the bootstrap method. Since this signal does not contain a stimulus response, false-positive detection of a stimulus response is expected in approximately 5% of cases. The false-positive rate from the simulation study gives

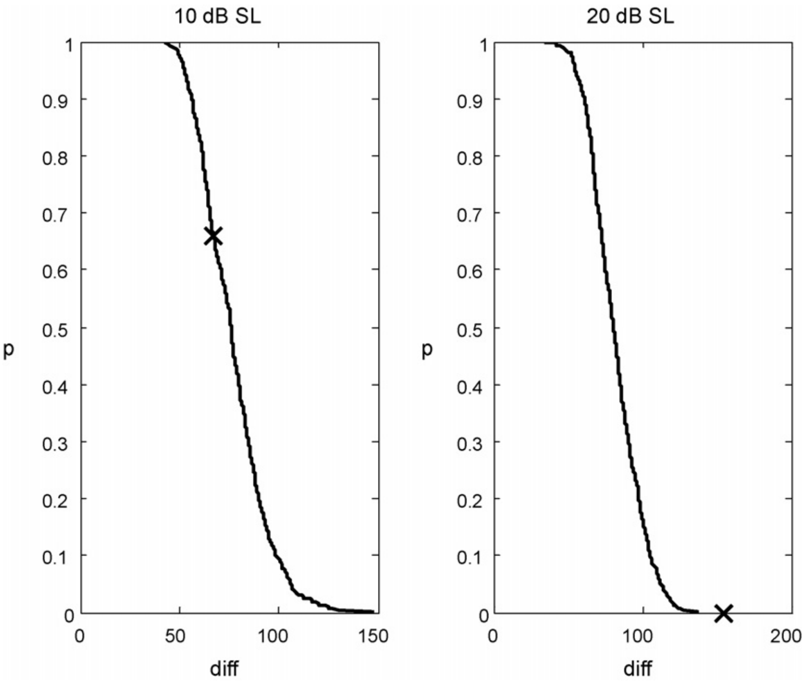

Fig. 2. Bootstrap distribution of *diff** from one subject at two different stimulus intensities. The *p*-value gives the fraction of cases (out of $L = 499$) which were larger than a given value of *diff*. The × marks the value of *diff* obtained from the original data, and the corresponding *p*-value gives the statistical significance of that value. The example on the left did not give a statistically significant response ($p = 0.65$), but for the one on the right, a response is detected ($p < 0.002$).

an indication of the 'coverage error' of the bootstrap method when applied to each of the parameters.

We then investigated the power of the proposed method to detect responses when present. To this end we simulated ABR data by adding a 'response' to a random background EEG signal. The stimulus response used corresponds to the coherent average from one of the signals recorded in a normal subject at 40 dB SL, which was then multiplied by a gain factor to obtain the desired SNR. We did this for nine different signal-to-noise ratios (SNR = −20 dB to 20 dB in the steps of 5 dB, calculated on the averaged signals, corresponding to −53 to −13 dB in the raw data). The background EEG signals were obtained by the same AR process used above. At each SNR, 500 simulated ABR data were generated. As before, the four parameters were calculated and their significance tested using the bootstrap method. The fraction of these 500 signals, at which $p < 0.05$ was determined, indicating the power of the method.

3.4. Application to recorded signals

The bootstrap tests were then applied to the data recorded from the normal subjects, and hearing thresholds were found for each of the four parameters. The threshold was defined as the minimum stimulus intensity at which $p < 0.05$ (with $p < 0.05$ for all higher stimulus intensities also). We also show the change in hearing threshold when $p < 0.01$ is used. These thresholds were compared to those determined by three experienced audiologists, who independently inspected the ABRs visually. Furthermore, inter-observer reliability for the visual inspections in this experiment was measured by Cohen's Kappa statistic [25]. Kappa is defined as the 'proportion of observed agreement after correction for chance agreement'. Its value is between 0 and 1, which accounts for the range from poor to excellent reliability.

Finally, in order to show how the bootstrap method can be applied with varying numbers of stimuli, and how this affects the detection of responses, we broke each recording (roughly 2000 stimuli) into blocks of $n = 100$ stimuli with no overlaps between blocks. Then we extracted the parameters θ and applied the bootstrap test to every block, and thus obtained a *p*-value for each. We then found the fraction of blocks (over all 12 recordings) in which the response could be detected, at each of the six stimulus levels (0–50 dB SL in steps of 10 dB). We then repeated this process for $n = 200, 300 \ldots 2000$ stimuli. This provides a quantitative measure of the improvement in performance, as more stimuli are averaged.

4. Results

4.1. Monte Carlo simulations

The percentages of false positives in the simulated data without a stimulus-response were 4.0% for *diff*, 3.4% for *power*, 4.4% for F_{sp} and 6.0% for ± *difference*. These values are all close to the expected value of $\alpha = 5\%$, and within the acceptable range of 3.2–6.8% given by the binomial probability distribution of 500 trials with probability of 'success' equal to 5% (95% confidence limits). Note that the four

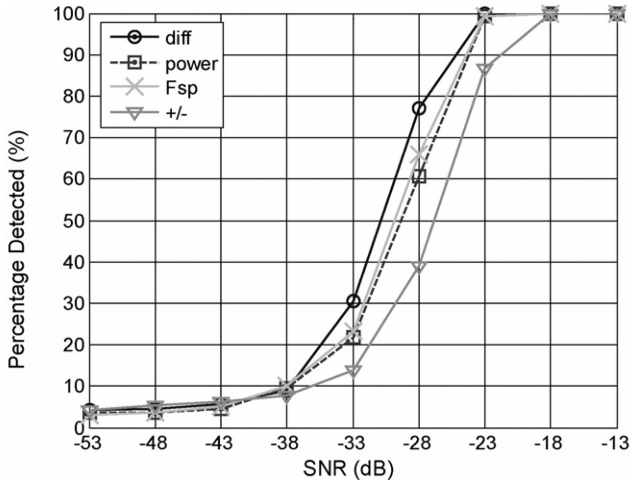

Fig. 3. Percentage of responses detected as a function of signal-to-noise ratio (SNR) of the raw data. Results correspond to $K = 2000$ averages (SNR = −20 to 20 dB in the coherent average).

parameters were all calculated from the same set of 500 simulated signals.

The results of the simulation with added responses are shown in Fig. 3. As expected, the percentage of detected responses consistently increases with the increase of the SNR levels for all four parameters. For all parameters (*diff*, *power*, F_{sp} and ± *difference*) results converge to 100% detection at high SNR, and to the expected $\alpha = 5\%$ at low SNR. At mid-range SNRs, there is no significant difference between results for F_{sp} and *power* (t-test, $p > 0.05$), but *diff* and ± *difference* are better and worse, respectively (t-test, $p < 0.05$).

4.2. Recorded ABRs

Subjective inspections: The three experienced audiologists determined the hearing thresholds by comparing the two replicate coherent averages of ABR data at the same stimulus intensity, and then finding the minimal stimulus level at which a consistent response was obtained. The results are given in Fig. 4, showing quite large variations between raters, consistent with the observations in [3], and underlining the need for objective methods for response detection.

Inter-observer reliability was measured by the Kappa statistic [25]. The values of Kappa for all three pairs of judges were shown in Table 1. The common interpretation of the reliability is that Kappa should be no less than 0.90 to be regarded as high [1], i.e., for there to be a good agreement between judges. Clearly, this is not the case in Table 1.

Table 1
Kappa values for all possible pairs of the judges

Judges	Kappa
A & B	0.70
B & C	0.63
C & A	0.81

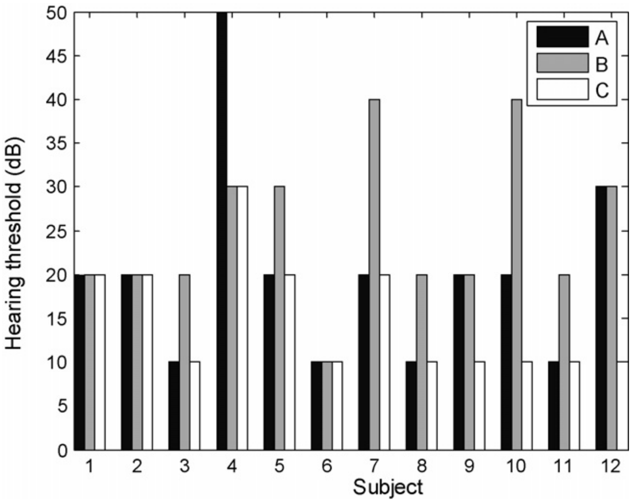

Fig. 4. Hearing thresholds for 12 normal hearing subjects, as determined from the ABR by three experienced audiologists (A, B, and C) through visual inspection. For each subject, the three bars represent the hearing threshold estimate of A, B and C respectively.

The hearing threshold was then estimated for each subject, using the bootstrap technique. The p-values calculated for each of the four parameters, at each of the stimulus intensities are shown in Table 2 for one subject. The minimum stimulus intensity at which a significant response ($p < 0.05$) is consistently obtained, is considered the hearing threshold. For example, for *diff*, a response was detected from 10 dB ($p < 0.05$) upward. For higher stimulus-intensities, the results are also significant. So the hearing threshold for *diff*, F_{sp} and ± *difference* is considered to be 10 dB in this case, and that of power 0 dB. In general, the higher stimulus intensities provide stronger responses and lower p-values. However, there were a number of exceptions to this (not shown), and visual inspection of responses confirms that in some recordings the responses are somewhat less evident at slightly higher stimulus intensities. Note that for these cases we define hearing threshold to be the lowest stimulus intensity at which $p < \alpha$, and for which all higher stimulus intensities also showed a significant response.

Fig. 5 shows the average hearing threshold given by the three observers (subjectively) and compares these to the results obtained from the bootstrap method. Results are given

Table 2
Examples of p-values for the four parameters at different stimulus intensities, for one subject

Stimulus intensity (dB)	diff_p	power_p	F_{sp}_p	±_p
0	0.236	**0.006**	0.144	0.744
10	**0.002**	0.004	**<0.002**	**0.026**
20	<0.002	<0.002	<0.002	0.004
30	<0.002	<0.002	<0.002	<0.002
40	<0.002	<0.002	<0.002	<0.002
50	<0.002	<0.002	<0.002	0.002

p-values are obtained from the bootstrap test using roughly 2000 stimuli. The p-values marked in bold indicate the hearing threshold.

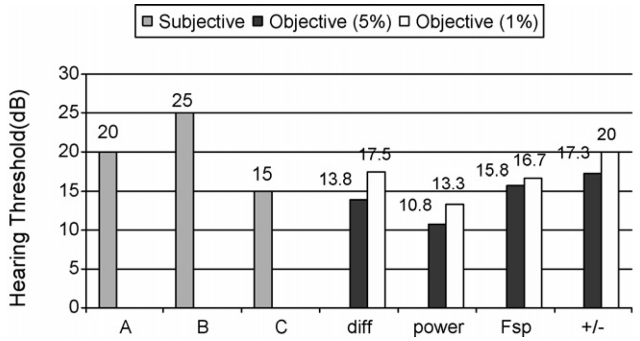

Fig. 5. Mean hearing thresholds of 12 subjects for 3 subjective visual inspections (A, B, C) and 4 objective parameters (*diff*, *power*, F_{sp}, ±*difference*, respectively). The objective hearing thresholds are determined with significance levels $\alpha = 5\%$ and 1%.

for $\alpha = 5\%$ and $\alpha = 1\%$. The parameter *power* appears to be the most sensitive, finding responses at lower stimulus intensities than any of the human examiners ($p < 0.05$, sign test, *power* at $\alpha = 5\%$ compared against examiners A and B, but not C). The parameter *power* was also found to give a significantly ($p < 0.05$, sign test) lower thresholds than ± *difference*; no significant difference was found between the remaining parameters. As expected, the thresholds for $\alpha = 1\%$ are higher than those for $\alpha = 5\%$.

We also investigated the effect of the number of epochs (stimulus responses) recorded, on the ability to detect a response using the bootstrap approach. We therefore applied the bootstrap tests to progressively increasing numbers of stimuli. Fig. 6 illustrates the results for the parameter power. As expected, the fraction of cases in which the ABR is detected increases with increasing stimulus intensity and also with the number of sweeps. At 40 and 50 dB SL, 800 stimuli were enough to detect the response in all of the 12 subjects with the parameter *power* (see Fig. 6); 1100 stimuli were required for *diff* and F_{sp}. For ± *difference*, 2000 stimuli at 50 dB were required to achieve 100% detection.

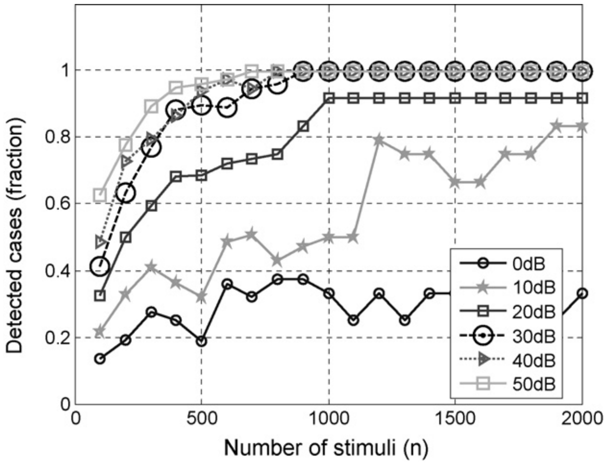

Fig. 6. The fraction of cases in which the ABR was detected is shown as a function of number of epochs (stimuli) and the parameter *power*. The bootstrap method ($p < 0.05$) was applied with increasing numbers of stimuli, and stimulus intensities between 0 and 50 dB SL. Note that for this result the signals were broken down into non-overlapping blocks of n stimuli, such that for example at $n = 100$ each of the 12 subjects provided 20 blocks, but at $n = 2000$, only a single block.

5. Discussion and conclusion

The need for objective methods to detect evoked responses was clearly illustrated by the example of the ABR presented here. There was considerable disagreement between the subjectively selected hearing thresholds given by the three experienced audiologists (A, B, C) and this was reflected in the relatively low values of Kappa. Techniques for the automated detection of evoked responses usually involve the calculation of a parameter, for which a threshold is then selected, above which the response is deemed to be present. The selection of this threshold may be based on experience and experimental work e.g. [26]. The bootstrap technique presents a very attractive and flexible alternative, by providing a simple means of estimating the statistical significance (p-value) of a parameter. It does so by comparing the parameter-value to that expected under the null-hypothesis of no stimulus response.

The use of the bootstrap method circumvents potentially intractable statistical analysis, which would otherwise have to be carried out, in order to obtain a closed-form solution for the statistical analysis of each parameter. Such analyses would also usually involve assumptions regarding signal statistics, which it may be difficult to justify or test, for each recording. Conventional statistical analysis is also complicated in this work by the autocorrelation of the signals, such that successive samples are not independent. This, for example, is the reason why the F_{sp} does not correspond to the F-statistic, with the degrees of freedom corresponding to the number of samples analysed [6]. One approach to overcome this is to find thresholds based on 'worst case' assumptions for the degrees of freedom, from which a critical (threshold) value for the F_{sp} of 2.25 ($\alpha = 5\%$) [6] has been suggested, when using 250 sweeps. We compared this to the mean value of the 95th percentile (corresponding to $\alpha = 5\%$) of the bootstrap distribution of the F_{sp}, which gives 1.81 (mean value from the 12 subjects over all stimulus intensities). It would thus appear that the threshold of $F_{\text{sp}} = 2.25$ is too high for $\alpha = 5\%$ - in accordance with the worst-case assumptions made in deriving it. When the number of stimuli was increased to 2000, the bootstrapped critical value was 1.75. As expected the increased degrees of the freedom of the denominator lead to tighter bounds on the F_{sp}. Furthermore, it was found that the critical values ($\alpha = 5\%$) varied quite considerably between individuals, indicating that universally valid threshold values for F_{sp} probably cannot be justified. The criterion for ± *difference* [7] was also tested with 2000 stimuli. This was found to be 3.19 (mean value) when using the bootstrap distribution, which is considerably higher than the value of 2, given by Wong and Bickford [7] based on experimental studies.

The bootstrap analysis thus shows that the threshold values for the parameters depend on the number of stimuli used (for a given level α), and vary quite considerably between individuals. The latter is probably due to varying signal and noise ratios, and band-widths of each recording. Thus, clearly any fixed threshold for parameters such as F_{sp} or ± *difference* would lead to false-positive rates that differ between subjects.

The bootstrap method makes few assumptions about the data, which is one of its main benefits. A 'significant' response to stimulation may be considered to be one in which the parameter θ of the coherent average has 'surprisingly' large (or small) values. The bootstrap method allows this to be tested directly, by comparing the θ from the coherent average, to the θ^* of the incoherent averages obtained from the same data. If there is nothing 'special' about the signal segments following the stimulation, θ and θ^* would be similar; if θ is very different to θ^*, there is clear evidence of signal component that is time-locked to the stimulus. The bootstrap method is thus intuitive in testing for a significant response, and does so without assuming a statistical distribution for the samples in the signals. It does assume that the signal is ergodic, such that samples drawn randomly from the recording represent the 'random process' generating the data.

In using the bootstrap method, the significant level (α) has to be chosen. In this work we used $\alpha = 5\%$. If this is reduced to 1%, the hearing threshold for the case illustrated in Table 2 would remain the same for *diff*, *power* and F_{sp}, but increase to 20 dB for ± *difference*. Overall, the increase in threshold is small (Fig. 5). Clearly the drawback of reducing the false-positive rate is the concomitant increase in false-negatives. Which of these errors is more important depends on the application: for example, in monitoring depth of anaesthesia [27] a significant mid-latency auditory evoked responses may indicate that the patient is awakening, which might require prompt intervention by the anaesthetist. Thus high sensitivity to the presence of a response (and hence high α) is desirable. On the other hand, in screening tests for hearing loss, a false positive response may lead to missing a hearing impairment, and a low false-positive rate is desirable.

The bootstrap technique can deal with varying numbers of stimuli, while maintaining pre-defined false-positive rates. In Fig. 6, it is evident that at 40 and 50 dB SL, 800 stimuli were enough to detect the response, which is rather less than the 2000 recommended in the literature. Thus, in normal hearing subjects, at these levels of stimulus the duration of the test could be considerably reduced, as already indicated by Don et al. [21].

In this work we illustrated the bootstrap approach using ABRs. It also can be applied in other modalities (e.g. visual, somatosensory, and event-related). Bootstrap methods have been used previously in finding confidence limits for the SNR and inter-ocular amplitude ratio in visual evoked potentials [18], and various parameters in somatosensory evoked potentials [16], as well as in assessing ROC curves [28] for steady-state auditory evoked potentials. However, it does not appear to have been used previously for detecting the presence of an evoked response. In the current work we are not proposing that the bootstrap method should replace established statistical criteria for detecting responses [29,30]. However, the bootstrap method can be applied in testing the significance of parameters that are not readily analysed by conventional statistical approaches, such as the F_{sp} or ± *difference*.

The bootstrap approach presented allows the statistical significance of arbitrary signal parameters to be assessed and thus provides a very powerful tool for the future development of evoked-response analysis, including the selection of new and optimized parameters for response detection. It allows responses to be detected at a user-defined false-positive rate, for an arbitrary number of stimuli, and takes the statistical characteristics of each individual recorded signal into account. In addition to evoked responses, it could also be applied to other applications in which coherent averaging is used, such as high-frequency ECG analysis.

Acknowledgements

The authors are grateful for ORS (UK) and Rayleigh (ISVR/University of Southampton) scholarships that supported this work.

References

[1] Arnold SA. Objective versus visual detection of the auditory brain stem response. Ear Hearing 1985;6(3):144–50.

[2] Vidler M, Parker D. Auditory brainstem response threshold estimation: subjective threshold estimation by experienced clinicians in a computer simulation of the clinical test. Int J Audiol 2004;43:417–29.

[3] Mason SM. On-line computer scoring of the auditory brainstem response for estimation of hearing threshold. Audiology 1984;23:277–96.

[4] Ozdamar O, Delgado RE, Eilers RE, Urbano RC. Automated electrophysiologic hearing testing using a threshold-seeking algorithm. J Am Acad Audiol 1994;5:77–88.

[5] Pool KD, Finitzo T. Evaluation of a computer-automated program for clinical assessment of the auditory brainstem response. Ear Hearing 1989;10:304–10.

[6] Elberling C, Don M. Quality estimation of averaged auditory brainstem responses. Scand Audiol 1984;13(3):187–97.

[7] Wong PKH, Bickford RG. Brain stem auditory evoked potentials: the use of noise estimate. Electroencephalogr Clin Neurophysiol 1980;50:25–34.

[8] Dobie RA, Wilson MJ. Analysis of auditory evoked potentials by magnitude-squared coherence. Ear Hearing 1989;10:2–13.

[9] Jerger J, Chmiel R, Frost JD, Coker N. Effect of sleep on the auditory steady state evoked potential. Ear Hearing 1986;7(4):240–5.

[10] Zurek PM. Detectability of transient and sinusoidal otoacoustic emissions. Ear Hearing 1992;13:307–10.

[11] Efron B. Bootstrap methods. Another look at the Jackknife. Ann Stat 1979;7:1–26.

[12] Efron B. Computers and the theory of statistics: thinking the unthinkable. SIAM Rev 1979;4:460–80.

[13] Efron B. Nonparametric Standard Errors and Confidence Intervals (with discussion). Can J Stat 1981;9:1–26.

[14] Haynor DR, Woods SD. Resampling estimates of precision in emission tomography. IEEE Trans Med Imaging 1989;8:337–43.

[15] Simpson DM, Panerai RB, Ramos EG. Assessing blood flow control through a bootstrap method. IEEE Trans Biomed Eng 2004;51(7): 1284–6.

[16] Adams HP, Kunz S. Inter- and intraindividual variability of posterior tibial nerve somatosensory evoked potentials in comatose patients. J Clin Neurophysiol 1996;13:84–92.

[17] Darvas F, Rautiainen M, Pantazis D, Baillet S, Benali H, Mosher JC, et al. Investigations of dipole localization accuracy in MEG using the bootstrap. NeuroImage 2005;25:355–68.

[18] Fortune B, Zhang X, Hood DC, Demirel S, Johnson CA. Normative ranges and specificity of the multifocal VEP. Documenta Ophthalmol 2004;109:87–100.

[19] Hood LJ. Clinical applications of the auditory brainstem responses. Aan Diego London: Singular Publishing Group, Inc; 1998.

[20] Lv J, Bell SL, Simpson DM. A statistical test for the detection of auditory evoked potentials. IPEM meeting: signal processing applications in clinical neurophysiology, Feb. 2004.

[21] Don M, Elberling C, Waring M. Objective detection of averaged auditory brainstem responses. Scand Audiol 1984;13:219–28.

[22] Efron B, Gong G. A leisurely look at bootstrap, the jackknife, and cross-validation. Am Stat 1983;37(1):36–48.

[23] Zoubir AM, Boashash B. "The bootstrap and its application in signal processing," 15th ed. 1998; pp. 56–76.

[24] Marple SL. Digital spectral analysis with applications. Prentice-Hall; 1987.

[25] Altman DG. Practical statistics for medical research. London: Chapman & Hall; 1991.

[26] Ozdamar O, Delgado RE, Eilers RE, Widen JE. Computer methods for on-line hearing testing with auditory brain stem responses. Ear Hearing 1990;11(6):417–29.

[27] Aceto P, Valente A, Gorgoglione M, Adducci E, De Cosmo G. Relationship between awareness and middle latency auditory evoked responses during surgical anaesthesia. Br J Anaesthesia 2003;90(5): 630–5.

[28] Valdes JL, Perez-Abalo MC, Martin V, Savio G, Sierra C, Rodriguez E, et al. Comparison of statistical indicators for the automatic detection of 80 Hz auditory steady state responses. Ear Hearing 1997;18: 420–9.

[29] Mauricio A, Miranda de Sa FL, Infantosi AF, Simpson DM. A statistical technique for measuring synchronism between cortical regions in the EEG during rhythmic stimulation. IEEE Trans Biomed Eng 2001;48(10):1211–5.

[30] Simpson DM, Tierra-Criollo CJ, Leite RT, Zayen EJB, Infantosi AFC. Objective response detection in an electroencephalogram during somatosensory stimulation. Ann Biomed Eng 2000;28: 691–8.