A NOVEL STATISTICAL TEST TO DETECT AUDITORY EVOKED POTENTIALS

J. Lv, D. M. Simpson, and S. L. Bell

ISVR, University of Southampton, Southampton, UK

jl2@isvr.soton.ac.uk, ds@isvr.soton.ac.uk, slb@isvr.soton.ac.uk

Abstract: Auditory evoked potentials (AEPs) are usually evaluated subjectively, by visual inspection, and considerable differences between interpretations commonly appear. Objective, automated methods are available in the literature, 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 on auditory brainstem responses (ABRs) to detecting hearing thresholds. Analysis of a set of 72 recordings from 12 subjects found hearing thresholds tended to be lower than those found by visual inspection by 3 experienced audiologists, and that fewer stimuli may be required than in conventional clinical procedures. The bootstrap method provides a new, simple and yet powerful means of detecting AEPs, which is also very flexible and readily adapted to a wide variety of signal parameters.

Introduction

Auditory evoked potentials (AEPs) measure the response of the hearing systems to acoustic stimulation. The response is extracted by presenting a series of clicks or tone-burst to the ears, while simultaneously recording the electroencephalogram (EEG). AEPs are often classified by their response time relative to the onset of a stimulus, and three main types are defined: (1) auditory brainstem response (ABR) is a series of five to seven peaks arising from auditory nerve and brainstem structures and occurring within 10 ms of the onset of the stimulus; (2) middle latency response (MLR) occurs with latencies ranging from roughly 12 to 75 ms and is generated from the thalamus and auditory cortex; (3) slow vertex potential (SVP) occurs beyond 75 ms. AEPs have many applications in clinical settings, such as hearing screening and threshold prediction, intraoperative monitoring and anaesthetic depth measurement. The most common conventional method used for recovering the AEP signal from the raw 'EEG' signal recorded in the patient is coherent averaging, in which the ensemble of signal segments following the stimuli ('single trials')is first obtained, and the sample mean then calculated [1].

In clinical applications, an objective automated method to detect the AEP is desirable, because the traditional approach, which involves visual inspection of the waveform by experts who detect the presence or absence of a response can lead to great differences between audiologists [2]. Many automated AEP detection methods are described in the literature. These usually calculate one or more parameters from the AEP and then compare these to some 'threshold value', beyond which a response is deemed to be present. However, it is difficult to compare the different estimated measures and determine a suitable criterion as to when a response is present. An estimate of the statistical significance of the observed response is desirable, since this p-value is equivalent to the false-positive rate, i.e. how often a response would be detected when in fact there is no response present. In this work, we will use a novel approach, based on the bootstrap technique, to estimate the statistical significance of the response [3]. The bootstrap method was introduced by Efron [4] to solve some complex statistical problems, replacing possibly intractable mathematical analysis by computationally intensive resampling methods. We describe how the bootstrap technique can be used to provide a very flexible means of assessing the statistical significance of a wide range of different parameters for detecting the response.

During the period of collecting AEP data, sometimes muscle activity creates a large-amplitude artefact in the recorded signal. In coherent averaging, automated artefact rejection schemes are often employed. We present a means by which this can also be taken into account in estimating the p-values through the bootstrap method.

Finally, in order to reduce the time for AEP recording, it is desirable to reduce the number of sweeps required for detecting the response. We investigate the minimum number of sweeps required at varying stimulus intensities, when using the bootstrap statistical test.

ISSN: 1727-1983

Materials and Methods

ABR data

ABRs were recorded from 12 normal-hearing adults subjects (6 males and 6 females), who were aged between 18 and 30 years. All the recordings were made using rectangular click stimuli with duration of 0.1ms. At each stimulus level, two recordings were collected for every subject, starting at 50 dB sensation level (SL), decreasing in 10 dB steps to 0 dB SL. The click rate was 33.3 Hz. The number of sweeps contributing to each coherent averaged response was 2000. The raw data were band-pass filtered between 30 and 2100 Hz, and sampled at 5 kHz. The ABR was then obtained by coherently averaging the ensemble of data segments following the onset of each stimulus. The raw recorded signal, containing spontaneous background cerebral activity, and noise as well as the ABR, will be referred to as the 'EEG'.

Bootstrap test

In order to test for the presence of a response, we first estimate a parameter from the coherently averaged data (ABR), and then test its statistical significance, using the bootstrap method. The following four parameters were used in this work: *diff* [5], the difference between the maximum and minimum value of the AEP in the range 5-15 ms; *power*, the mean power of the AEP; F_{sp} [6], an estimate of the signal-to-noise ratio of the AEP at 10 ms; ±*difference* [7], an alternative estimate of the signal-to-noise ratio.

Following the conventional procedure for coherent averaging, the EEG signal is first broken into segments beginning at the instant of each stimulus. The resulting ensemble of signal segments (with a length corresponding to the time-interval between stimuli) is then averaged to obtain the evoked response. In the bootstrap method, a similar ensemble is built up, but segments are chosen from random locations within the EEG signal. By averaging the signals in this ensemble, the 'incoherent' average is found. From this 'incoherent' average the parameters (see above) are again calculated (to be denoted by *). This process is repeated 499 times, providing the 'bootstrap distribution' of each of the parameters. This gives an estimate of the sampling distribution of the parameter, under the null-hypothesis of 'no response to the stimuli' – as illustrated by the cumulative distribution shown in Fig. 1 for one parameter (*diff*) and one subject at different stimulus intensities. The bootstrap distributions at each stimulus intensity are shown by curves and the symbols x give the corresponding values for the parameter from the coherent average. From this the statistical significance (p-value) of the parameter estimated can be found: the proportion of bootstrap values (*diff**) that are larger than *diff* gives the p-value for that parameter. If $p \leq 5\%$ (horizontal line in Fig. 1), we consider there to be a significant response present. In Fig. 1, there is a significant response for 20, 30, 40 and 50 dB, but not for 0 dB (p≈85%) or 10 dB (p≈65%).

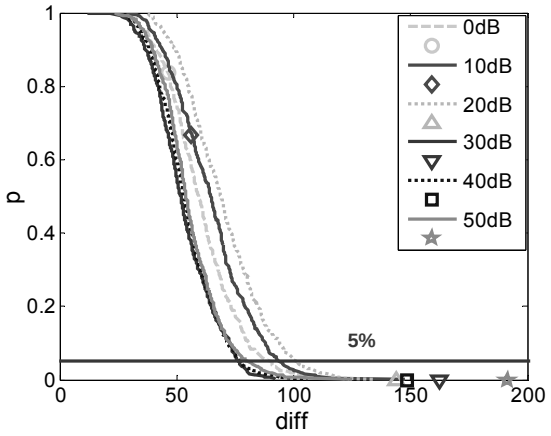

Figure 1: The distribution of diff* (solid lines), for stimulus intensities between 0 and 50 dB (sensation level - SL) for one subject. The symbols x show the corresponding *diff* for the coherent average.

Simulated data

In order to test the proposed bootstrap test, we carried out a Monte-Carlo study, simulating signals without stimulus responses. The aim of this was to estimate whether the pre-defined false positive rate was actually obtained, when no response was present (*coverage error*). We used an autoregressive model (AR) to simulate EEG signals. First we select one EEG signal, whose spectrum may be considered to be most 'typical' for the set of recordings. This was defined as the ones whose spectrum was closest to the median spectrum of the 12 recordings at 0 dB SL. We chose the model order of 16 according to the Final Prediction Error (FPE) (Fig. 2.) to simulate 500 'EEG signals' without a stimulus response.

Figure 2: The relationship between order and Final Prediction Error (FPE) using a recording with stimulation at 0 dB SL.

Muscle artefact rejection

The traditional way to remove muscle activity is by setting a maximum amplitude level, and any stimulus-response that exceeds this limit is not included in the coherent average. Based on observations of our recording and common practice in this field [8], we chose to exclude data whose amplitudes are outside the

ISSN: 1727-1983

range ±20 μV. This procedure was applied in both the coherent and incoherent (bootstrap) averaging.

Minimal number of sweeps required for detecting a response

In order to find the minimum number of sweeps required for detecting a response, we applied the bootstrap method with increasing numbers of sweeps (100, 200, …, 2000), using non-overlapping segments of the original recordings. We then determined, for each stimulus intensity, the percentage of segments (from all subjects), in which a statistically significant AEP-response is found, for each of the four parameters. The above artefact rejection scheme was not used at this stage.

Results

Simulated data

The percentages of false positive tests (out of 500) for each of the four parameters are shown in Table 1, with and without using the muscle artefact rejection scheme described above. As expected, these values were all close to $\alpha = 5\%$, and within the acceptable range of 3.2%-6.8%, given by the binomial distribution for 500 trials with probability of 'success' equal to 5% (95% confidence limits). Note that the four parameters were all calculated from the same set of simulated signals.

Table 1: The percentages of false positives with and without muscle artefact rejection in 500 simulated signals, using the bootstrap test with α=5%.

Parameter	diff	power	F_{sp}	±
With muscle rejection	4.0%	3.6%	3.4%	4.2%
Without muscle rejection	4.2%	3.6%	3.4%	3.8%

ABR recorded data

Subjective inspections: Three experienced audiologists (A, B, and C) visually inspected the responses by comparing two replicates of coherent averages at each stimulus intensity, in order to subjectively determine the hearing thresholds.

The hearing threshold was also estimated by bootstrap approach ($\alpha = 5\%$). The minimum stimulus intensity, at which a significant response is found (and for which $p < 5\%$ also at all higher stimulus intensities), is considered to be the hearing threshold.

In order to compare the difference of hearing threshold between subjective inspections and objective bootstrap approach based on the four parameters, we calculated the average hearing threshold (AHT) (Table 2) of 12 subjects. The parameter *power* appears to be the most sensitive in detecting a response.

Table 2: Average hearing threshold by subjective inspection and objective bootstrap technique.

dB SL	Subjective			Objective			
	A	B	C	diff	power	F_{sp}	±
AHT	20*	25*	15	13.8	10.8	15.8	17.3*

* Significantly different to the threshold found with parameter power (sign-test, $p < 0.05$).

The median value of the three (A, B, C) subjectively evaluated hearing thresholds (MHT) of each of the 12 subjects were calculated and compared to the hearing thresholds for each of the four parameters (HT) obtained by the bootstrap method ($p<5\%$). Results are shown in Fig. 3. For *power* and *diff*, these are lower or equal to MHT in 11 of the 12 subjects; for F_{sp} and ± *difference*, this is the case in 10 subjects.

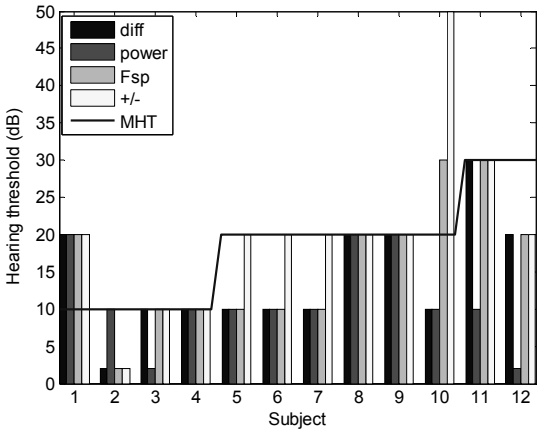

Figure 3: Comparison between median hearing threshold (MHT - median of A, B and C, solid line), and hearing thresholds from the four parameters (the histograms from left to right correspond to the parameters *diff*, *power*, F_{sp}, ± *difference*). In most cases, the latter are smaller than, or equal to the corresponding MHT.

Fig. 4 shows the hearing thresholds obtained with $\alpha = 1\%$ rather than $\alpha = 5\%$ used in the previous results. For 12 subjects, the hearing threshold remains the same in most cases, and increases by 10 or 20 dB in three cases for *diff*, two cases for *power*, one case for F_{sp}, and four cases for ± *difference*.

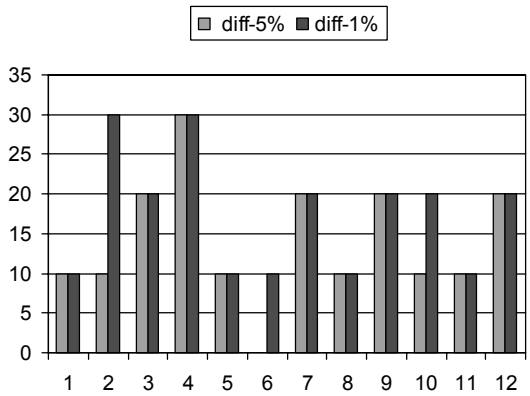

Figure 4: Comparison the hearing threshold for *diff* with $\alpha = 5\%$ and $\alpha = 1\%$.

The artefact rejection scheme aims to eliminate poor data segments, and thus improve the ability to detect the ABR. Table 3 shows the fraction of the 72 recordings in which a response was detected with, and without the artefact rejection scheme. Overall, the number of cases

ISSN: 1727-1983

in which a response was detected however did not change greatly. The p-values obtained from the bootstrap test tended to be smaller without the artefact rejection scheme than with it (the difference was statistically significant for *diff*, F_{sp} and ± *difference;* sign-test, $p < 0.05$).

Table 3: Dectection percentages ($p < 0.05$) with and without muscle artefact rejection.

Parameter	diff	power	F_{sp}	±
With muscle artefact rejection	75.0%	81.9%	76.4%	68.1%
Without muscle artefact rejection	79.2%	83.3%	76.4%	68.1%

In accordance with the methods described above, we then applied the bootstrap tests to progressively increasing numbers of stimuli. The aim is to determine the minimum number of stimuli required in order to detect the ABR. Fig. 5 illustrates the results for the parameter *diff* and F_{sp}. As expected, the fraction of cases in which the ABR is detected increases with increasing stimulus intensity and also with the number of sweeps.

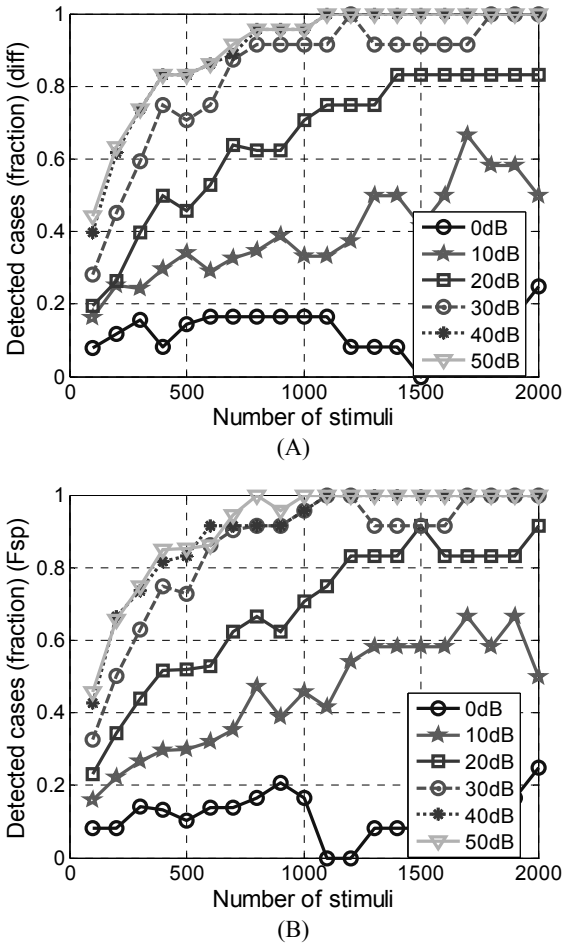

Figure 5: The detection percentages correspond to different number of stimuli (sweeps) at various stimulus intensities (0 dB to 50 dB in steps of 10 dB SL) for parameters *diff* (A) and F_{sp} (B).

At 40 and 50 dB SL, 800 stimuli were enough to detect the response in all of the 12 subjects with the parameter *power*; 1100 stimuli were required for *diff* and F_{sp}.

Discussion and Conclusion

The commonly used way to interpret the auditory evoked potentials is visual inspection, which greatly depends on the experience of the practitioners and can lead to large disagreement between different observers [9]. A number of methods [10-15] have been proposed in the literature to objectively detect the presence of a response, but they do not all rely on statistical criteria, nor could conventional statistical tests be readily applied in detecting a response. The bootstrap method proposed here provides a simple and flexible means for statistically detecting the presence of a response using a wide range of parameters extracted from the signal.

The only parameter to be chosen with the bootstrap method is the significance level (α). Clearly, the choice of this parameter can affect the detected hearing threshold, as illustrated in Fig. 4: lower values of α may lead to higher thresholds. The choice of a suitable α depends on the aim to the specific applications. For instance, in screening tests for hearing loss, a false positive response may lead to missing a hearing impairment, and a low false-positive rate is therefore desirable. As emphasized in [16] , even with α =0.1% in screening tests one out of 1000 newborns with hearing impairment would be missed, with potentially dramatic consequences for the baby. However, in monitoring the depth of anaesthesia using middle latency responses (MLR), the presence of a response may indicate that the patient is 'waking up', and missing this response (a false negative) may have serious consequences.

The proposed artefact rejection scheme did not provide the expected improvement in detecting the hearing threshold; in fact, p-values increased, reducing out ability to detect the response. Visual inspection of the signals indicated that there were not, in fact, large artefacts present in the data, and this may explain why the benefit of the artefact rejection scheme was not evident here. It may also be that the threshold of ± 20 μV selected was not the most appropriate for this data. Too large rejection level will lead to poor data being included in the analysis. On the other hand, the major disadvantage of using low rejection levels is that this may greatly increase the time required to acquire acceptable quality recordings. Artefact rejection schemes such as the one we used are routinely used in clinical work, and the approach we presented allows this to be taken into account in calculating the bootstrap p-values. The proposed artefact rejection schemes lead to small increases of the p-values, and consequently a small decrease in the fraction of recordings in which the response was detected. This could suggest that the artefact rejection scheme has lead to decreased sensitivity in this data. An alternative explanation is that by eliminating muscle artefact we have reduced the

ISSN: 1727-1983

number of false-positives, i.e. some of the artefact (at low stimulus intensities) may have been falsely interpreted as an evoked response when not using the artefact rejection scheme.

The bootstrap technique can deal with varying numbers of stimuli, while maintaining the pre-defined false-positive rate. In Figure 5, it is evident that at 40 and 50 dB SL, approximately 1000 stimuli were enough to detect the response, which is rather less than the 2000 usually recommended in the literature. Thus in normal hearing subjects, at these levels of stimulus the duration of the test could be considerably reduced [17]. Clearly it would be ideal to have a procedure that allows the test to terminate as soon as the response has been detected, without needing to continue to the pre-defined number of stimuli (e.g. 2000). This would require 'sequential statistical tests', which is not a simple matter, as pointed out in a recent investigation [18].

The bootstrap method clearly provides a very simple and flexible means of testing for the statistical significance of stimulus responses in auditory evoked potentials. While we have shown its use in ABRs, it could readily be adopted to other stimulus modalities, and other parameters that may be tuned to the specific features of those signals.

References

[1] CERUTTI S., CHIARENZA G, LIBERATI D., MASCELLANI P., AND PAVESI G. (1988): 'A Parametric Method of Identification of Single-Trial Event Related Potentials in the Brain,' *IEEE. Trans. Biomed. Eng.,* **35,** pp. 701-711

[2] VIDLER M., AND PARKER D. (2004) 'Auditory brainstem response threshold estimation: subjective threshold estimation by experienced clinicians in a computer simulation of the clinical test,' *Int. J. Audiol.*, **43**, pp. 417-429

[3] LV J., SIMPSON D. M., AND BELL S. L. (2004): 'Objective tests for the detection of auditory evoked potentials,' Proc. of the 3rd IEEE EMBSS UK and RI Postgraduate Conf. in Biomed. Eng. and Med. Phys., Southampton, UK, 2004, pp. 1-2.

[4] EFRON B. (1979): 'Bootstrap Methods. Another Look at the Jackknife,' *Ann. Stat.*, **7**, pp. 1-26

[5] LV J., BELL S. L., AND SIMPSON D. M. (2004): 'A statistical test for the detection of auditory evoked potentials,' IPEM meeting: Signal. Proces. Applicat. in Clin. Neurophysiol., 2004.

[6] ELBERLING C., AND DON M. (1984): 'Quality estimation of averaged auditory brainstem responses,' *Scand. Audiol.*, **13**, pp. 187-197

[7] WONG P. K. H., AND BICKFORD R. G. (1980): 'Brain stem auditory evoked potentials: the use of noise estimate,' *Electroencephalogr. Clin. Neurophysiol*, **50**, pp. 25-34

[8] STEVENS J., ELLIOTT C., LIGHTFOOT G., MASON S., PARKER D., STAPELLS D., SUTTON G., AND VIDLER M. (1999): 'Click auditory brainstem response testing in babies A recommended test protocol,' *Universal Neonatal Hearing Screening*

[9] ARNOLD S. A. (1985): 'Objective versus Visual detection of the Auditory Brain Stem Response,' *Ear. Hear.*, **6**, pp. 144-150

[10] DOBIE R. A., AND WILSON M. J. (1989): 'Analysis of auditory evoked potentials by magnitude-squared coherence,' *Ear. Hear.*, **10**, pp. 2-13

[11] JERGER J., CHMIEL R., FROST J. D., AND COKER N. (1986): 'Effect of sleep on the auditory steady state evoked potential,' *Ear. Hear.*, **7**, pp. 240-245

[12] MASON S. M. (1984): 'On-line computer scoring of the auditory brainstem response for estimation of hearing threshold,' *Audiol.*, **23**, pp. 277-296

[13] OZDAMAR O., DELGADO R. E., EILERS R. E., AND URBANO R. C. (1994): 'Automated electro-physiologic hearing testing using a threshold-seeking algorithm,' *J. Am. Acad. Audiol.*, **5**, pp. 77-88

[14] POOL K. D., AND FINITZO T. (1989): 'Evaluation of a computer-automated program for clinical assessment of the auditory brainstem response,' *Ear. Hear.*, **10**, pp. 304-310

[15] ZUREK P. M. (1992): 'Detectability of transient and sinusoidal otoacoustic emissions,' *Ear. Hear.*, **13**, pp. 307-310

[16] STURZEBECHER E., CEBULLA M., AND WERNECKE K. D. (2001): 'Objective detection of transiently evoked otoacoustic emissions,' *Scand. Audiol.*, **30**, pp. 78-88

[17] DON M., ELBERLING C., AND WARING M. (1984): 'Objective Detection of Averaged Auditory Brainstem Responses,' *Scand. Audiol.*, **13**, pp. 219-228

[18] STURZEBECHER E., CEBULLAR M., AND ELBERLING C. (2005): 'Automated auditory response detection: Statistical problems with repeated testing,' *Int. J. Audiol.*, **44**, pp. 110-117

ISSN: 1727-1983