Objective measures for detecting the Auditory Brainstem Response: comparisons of specificity, sensitivity, and detection time**.

Chesnaye M.A[1], Bell S.L[2], Harte J.M[3], & Simpson D.M[4].

1. Institute of Sound and Vibration Research, Faculty of Engineering and the Environment, University of Southampton, United Kingdom. mac1f14@soton.ac.uk
2. Institute of Sound and Vibration Research, Faculty of Engineering and the Environment, University of Southampton, United Kingdom. s.l.bell@soton.ac.uk
3. Interacoustics Research Unit, c/o Technical University of Denmark, Denmark. jmha@iru.interacoustics.com
4. Institute of Sound and Vibration Research, Faculty of Engineering and the Environment, University of Southampton, United Kingdom. ds@isvr.soton.ac.uk

Abstract

Objective: To evaluate and compare the specificity, sensitivity, and detection time of various time-domain and multi-band frequency domain methods when detecting the auditory brainstem response (ABR).

Design: Simulations and subject recorded data were used to assess and compare the performance of the Hotelling's T^2 test (applied in either time or frequency domain), two versions of the modified q-sample uniform scores test, and both the Fsp and Fmp, which were evaluated using both conventional F-distributions with assumed degrees of freedom and a bootstrap approach.

Study Sample: Data consisted of simulations along with click-evoked ABRs and recordings of EEG background activity from 12 and 17 normal hearing adults respectively.

Results: An overall advantage in sensitivity and detection time was demonstrated for the Hotelling's T^2 test. The false-positive rates (FPRs) of the Fsp and Fmp were also closer to the nominal alpha level when evaluating statistical significance using the bootstrap approach, as opposed to using conventional F-distributions. The FPRs of the remaining methods were slightly higher than expected.

Conclusions: In the current work, Hotelling's T^2 outperformed the alternative methods for automatically detecting ABRs. Its promise as a sensitive and efficient detection method should now be tested in a larger clinical study.

Keywords

Auditory Brainstem Response detection, Hotelling's T^2, Fsp, Fmp, Bootstrapping, modified q-sample uniform scores test

1. Introduction

Transient auditory brainstem responses (ABRs) are defined as short changes in neural activity along the auditory pathway in response to a brief acoustic stimulus, such as a click, chirp, or tone burst. They are typically recorded non-invasively using surface mounted electrodes, and are used primarily for diagnosing abnormalities within the auditory system, such as hearing loss and various neurological disorders. Determining whether an ABR is present (by either inspecting the data visually, or by applying an objective statistical test) is usually the first step for these applications, after which additional analysis can be performed on, for example, the morphology of the response.

The focus for this paper is on objective methods for detecting the ABR. The goal is to compare the performance of various statistical detection methods in terms of (i) true-positive rates (TPRs – in the current paper defined as the fraction of ABR responses that is detected), (ii) false-positive rates (FPRs – defined as the fraction of cases with no response that were incorrectly deemed to have a response), and (iii) detection time, i.e. the number of stimuli (expressed in time) required for detecting a significant response, which can be considered as the three most important properties for ABR detection methods.

Most ABR detection methods are classified as either time or frequency domain techniques. In the frequency domain detection is more challenging due to the spectral content of the ABR being spread across multiple bands (Elberling, 1976; Kevanishvili & Aphonchenko, 1979; Elberling, 1979; Suzuki et al., 1982). As most frequency domain techniques are applied to a single spectral band, they would need to be applied multiple times to cover the bandwidth of a typical ABR. The latter can result in an inflated FPR, and adjusted critical decision boundaries are required in order to preserve the desired alpha level of the test. This process may result in a significantly lower test sensitivity.

The broadband spectral content of the ABR has therefore led the majority of scientific investigations to explore methods for assessing multiple spectral bands within a single test, i.e. multi-band detection methods. A powerful multi-band detection method is the modified q-sample uniform scores test (Stürzebecher et al., 1996; Stürzebecher et al, 1999). The modified q-sample uniform scores test is applied to the ranks of the phases and amplitudes of multiple spectral bands, and has outperformed various alternative methods when detecting the ABR. These include the original q-sample uniform scores test and the q-sample analogue to Watson's U2 statistic (Stürzebecher et al, 1999), along with the F for a Single Point (Fsp), Friedman's test, and Cochran's Q test (Cebulla et al., 2000). Moreover, Cebulla et al (2006) have proposed various additional modifications to the q-sample uniform scores test. These modifications have shown a high performance when detecting auditory steady state responses (ASSRs), but have not yet been compared for ABR detection. The present paper will investigate if these proposed modifications are also suitable for ABR detection.

Another promising multi-band detection method is the Hotelling's T^2 test (Hotelling, 1931), which has outperformed both the Standard Deviation Ratio and the correlation coefficient (between two replicates of the coherent average) when detecting ABRs in subject recorded data (Valdes et al., 1987), along with the F for multiple points statistic (Fmp) when detecting ABRs extracted from quasi ASSRs (Lachowska et al, 2012). The Hotelling's T^2 test has recently also been applied in the time domain for detecting the slow cortical response (Golding et al., 2009; Carter et al., 2010; Chang et al., 2012; Van Dun et al., 2012; Van Dun et al., 2015). These time-domain features (see section 3.1) have not yet been evaluated for ABR detection, and may be a preferable alternative to frequency domain analysis due to the broadband spectral content of the ABR.

Additional time-domain techniques for ABR detection methods that are of interest include the Fsp and the Fmp, both of which can be tested for significance using F-distributions with v_1 and v_2 degrees of freedom. A recurring complication, however, is that the degrees of freedom are typically unknown for EEG data, and hence have to be assumed before statistical inference can be realized. Because false-negatives (i.e. failure to detect an ABR response that is in fact present) are typically less detrimental to the performance of ABR applications (ABR hearing screening tests in particular) than false-positives, Elberling & Don (1984) have recommended a conservative approach (fewer false-positives than the nominal target) by setting v_1 to 5. The drawback is a decrease in test power, which may result in an increased cost of service delivery due to prolonged test times and/or increased false negative rates.

An alternative approach for evaluating the significance of the Fsp (and the Fmp) has been proposed by Lv et al (2007). Lv et al use bootstrapping (Efron & Tibshirani, 1993 - see also section 3.4) to approximate the statistic's underlying null distribution. The approximated null distribution can then be used for statistical inference without needing to explicitly estimate or assume the degrees of freedom of the data. It is therefore hypothesized that evaluating statistical significance with the bootstrap approach, as opposed to using F-distributions with assumed degrees of freedom, would provide more consistent results across datasets with degrees of freedom that may vary between individuals and recordings.

The goal for this paper is to evaluate and compare the performance of various objective ABR detection methods in terms of specificity, sensitivity and detection time by using simulations and a small sample of normal-hearing adults. The methods selected for the analysis include two versions of the modified q-sample uniform scores test, which use either the ranks or the actual values of the phases and amplitudes of the Fourier components of multiple spectral bands (see section 3.2 for details), Hotelling's T^2 test applied in either the time or the frequency domain (section 3.1), and both the Fsp and the Fmp (section 3.3), which were evaluated using either F-distributions with assumed degrees of freedom or with the bootstrap approach.

2. ABR and no-stimulus EEG data

The data used throughout this study consists of (i) a small sample of normal-hearing adults where physiological hearing thresholds were estimated using click-evoked ABRs of various intensity levels, thus yielding a wide range of ABR waveform morphologies, and (ii) a relatively large database of no-stimulus EEG recordings. The database of no-stimulus EEG recordings was initially used to assess the specificities of the methods (section 3.5), after which it was used in combination with the subject recorded ABR data in simulations to assess sensitivity (section 3.6). The simulated data provides a test-bed in which large amounts of well controlled data are generated to assess performance when signal characteristics are repeatable. The next step was to assess sensitivities and detection times of the methods using just the subject recorded ABR data (section 3.7), which reflect real-world features of routine recordings. Extended clinical studies including data from participants with a range of hearing impairments are beyond the scope of the current work, but should follow in progressing this research further.

2.1 Subject recorded ABR data

The subject recorded ABR data, previously described in Lv et al (2007), was collected from 12 subjects (6 female and 6 males) ranging from 18 to 30 years of age. The stimulus was a rectangular 100 μs click delivered at a stimulus rate of 33.3 Hz through ER-2 insert phones (Etymotic, USA). The click intensities ranged from 0 to 50 dB SL (sensation level, i.e. relative to individual hearing thresholds) in steps of 10 dB. The behavioural thresholds were estimated using a simple 'up-down' approach where the click intensity was reduced in steps of 10 dB for every correct response, and increased in steps of 5 dB for every missed response. ABRs were recorded with the active electrode placed at vertex, a reference electrode at the nape of the neck, and a ground electrode placed at mid-forehead. Measurements were obtained at a sampling rate of 10 kHz using a Cambridge Electronic Design (CED) micro 1401 data acquisition unit along with a CED 1902 amplifier. Electrode impedances remained below 5 kΩ throughout the recording. The recordings were band-pass filtered offline from 30 to 1500 Hz with a 3rd-order Butterworth filter. Each recording was furthermore downsampled to 5 kHz, and an artefact rejection method was applied by discarding 15% of the noisiest epochs, as determined by their mean square values. Approximately 3600 clicks were delivered per subject and per stimulus condition, resulting in a minimum of 3000 epochs after artefact rejection. The 30.03 ms intervals following the onset of each stimulus (henceforth referred to as epochs) were saved for offline analysis.

2.2 No-stimulus EEG recordings

Recordings of spontaneous EEG background activity (no stimulus was used) were previously collected by Madsen et al (2017) and Madsen (2010) from 17 subjects (12 male and 5 female) under four conditions. The conditions were (i) *asleep*, where the subjects were asked to try and fall asleep, though sleep was not confirmed, (ii) *still*, where the subjects were instructed to lie still with their eyes closed, but not to fall asleep, (iii) *blink*, where the subjects were instructed to blink every 1 to 3 seconds as a circle appeared on a screen in front of them, and (iv) *move*, where the subjects were asked to move according to a random animation, also shown on a screen in front of them. Measurements were then obtained using a Compumedics Neuroscan II EEG amplifier at a sampling rate of 20 kHz with three silver-silver chloride (Ag/AgCl) electrodes placed on the left mastoid, the right cheek (ground), and the upper forehead (reference). The electrode impedances remained below 1 kΩ throughout the recording for all subjects.

In the present study, the background EEG recordings were band-pass filtered with a 3rd-order Butterworth filter from 30 to 1500 Hz, after which they were downsampled to 5 kHz. Each recording was then structured into 30.03 ms epochs, and artefact rejection was applied by discarding 15% of the noisiest epochs, as determined by their mean square values. A total of 149 continuous EEG recordings were available, with an average of 6800 pre-processed epochs per recording, resulting in approximately 8 hours' of EEG.

3. Methods

This section first provides a description of the ABR detection methods and the bootstrap approach, after which the adopted methodologies for evaluating the specificity, sensitivity, and detection time of the methods are described.

The data to which the methods are applied consists of ensembles of epochs (for details on how these ensembles were pre-processed and constructed, the reader is referred to sections 2.1, 2.2, 3.5 and 3.6). Each ensemble is structured according to matrix **D**:

Equation 1.

$$\mathbf{D} = \begin{bmatrix} \mathrm{d}_{11} & \cdots & \mathrm{d}_{1K} \\ \vdots & \ddots & \vdots \\ \mathrm{d}_{N1} & \cdots & \mathrm{d}_{NK} \end{bmatrix}$$

where N is the ensemble size, K is the number of samples per epoch, and d_{ij} is the j^{th} sample of the i^{th} epoch. The mean epoch $\overline{E}$ (also known as the coherent average) is found by taking the K averages across the columns. The frequency domain representation of **D** is furthermore obtained by taking the Fast Fourier Transform (FFT) of each row. Features can then be extracted from either the time or frequency domain representations of the data. Extracting L features from each epoch results in the NxL-dimensional feature matrix **V**:

Equation 2.

$$\mathbf{V} = \begin{bmatrix} \mathrm{v}_{11} & \cdots & \mathrm{v}_{1L} \\ \vdots & \ddots & \vdots \\ \mathrm{v}_{N1} & \cdots & \mathrm{v}_{NL} \end{bmatrix}$$

where v_{ij} is the j^{th} feature extracted from the i^{th} epoch.

3.1 The one-sample Hotelling's T^2 test

The one-sample Hotelling's T^2 test is the multivariate extension to Students t-test, and can be used to test whether the means of L features are significantly different from L hypothesized values. In the present work it is assumed that the expected values of the features are zero. The statistic itself is a weighted sum of the L feature means where the weights are determined by the variances and covariances of the features. These weights have the convenient property of normalizing the L means, which allows features with different scales and units to be combined appropriately. The T^2 statistic is given by (Rencher, 2001):

Equation 3.

$$\mathrm{T}^2 = \mathrm{N}\,(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)\mathbf{S}^{-1}(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)^{\mathrm{H}}$$

where $\bar{\mathbf{x}}$ is the L-dimensional vector of means (found by taking the means down the L columns of **V)**, $\boldsymbol{\mu}_0$ is the L-dimensional vector of hypothesized values to test against, $\mathbf{S}^{-1}$ is the inverse of the covariance matrix of the NxL-dimensional feature matrix **V**, and the $^{\mathrm{H}}$ superscript denotes the Hermitian transpose. The T^2 statistic can then be transformed into an F statistic with v_1 and v_2 degrees of freedom using:

Equation 4.

$$\mathrm{F} = \frac{\mathrm{N}-\mathrm{L}}{\mathrm{L}(\mathrm{N}-1)}\mathrm{T}^2 \quad \sim_{\mathrm{F}_{\mathrm{L,N-L}}}$$

where v_1 = L and v_2 = N-L. The significance of F can be determined with an F-distribution look-up table, or by finding the area under an F-distribution with L and N-L degrees of freedom on the interval 0 to F. Note that in order to calculate $\mathbf{S}^{-1}$, the number of epochs N should be larger than the number of features L. Note also that when the features are highly correlated, that $\mathbf{S}^{-1}$ can be close to singular, in which case rounding errors might occur. A solution would then be to use the pseudoinverse (e.g. the Moore-Penrose pseudoinverse; Moore, 1920; Penrose, 1955) instead of the regular inverse.

Time domain features

When applied in the time domain, the features for the one-sample Hotelling's T^2 test consist of `time-voltage means' (TVMs), which are defined as mean voltages, calculated across short time-intervals within each epoch (see e.g. Golding et al., 2009; Carter et al., 2010; Chang et al., 2012; Van Dun et al., 2012; Van Dun et al., 2015). As an example, when NxL TVMs are extracted, then each epoch is divided into L segments of approximately equal duration, and the mean is taken across each segment, resulting in the NxL-dimensional feature matrix **V**. The length of each segment requires a compromise such that the segments are neither too long, thus covering both peaks and troughs (with an average value of approximately zero) nor too short, thus leading to poor statistical robustness and reduced sensitivity. Because the direct current component is removed from the EEG recordings with a high-pass filter, the expected values for the TVMs will be zero. The hypothesized values to test against (defined above as $\boldsymbol{\mu}_0$) are therefore given as an L-dimensional vector of zeros.

Frequency domain features

When using the frequency domain approach, the features are the real and imaginary parts of the Fourier components of Q spectral bands (resulting in an Nx2Q-dimensional feature matrix **V**). Because the phases of each spectral band are expected to be uniformly distributed between 0 and 2π when no response is present, the expected values for the real and imaginary parts of each spectral band are again 0. The hypothesized values to test against are therefore given as a 2Q-dimensional vector of zeros.

3.2 The modified q-sample uniform scores test

The original q-sample uniform scores test (Mardia, 1972) is a non-parametric test that uses the ranks of the phases of the Fourier components of Q spectral bands to test whether the phases share the same distribution. The modification proposed by Stürzebecher et al (1999) uses the ranks of the amplitudes in addition to the ranks of the phases, and is given by:

Equation 5.

$$W^* = C\sum_{j=1}^{G}\left[\left(\sum_{i=1}^{N} r_{ij}\cos\beta_{ij}\right)^2 + \left(\sum_{i=1}^{N} r_{ij}\sin\beta_{ij}\right)^2\right]$$

where r^{ij} is the rank of the amplitude of the ith Fourier component (obtained from the ith epoch) of the jth spectral band. C is furthermore an additional scaling factor given by

Equation 6.

$$C = \frac{4}{Q^2(Q+1)^2}\frac{2}{N}$$

and β_{ij} is given by:

Equation 7.

$$\beta_{ij} = \frac{a_{ij}2\pi}{NQ}$$

where a_{ij} is the rank of the phase of the ith Fourier component (obtained from the ith epoch) of the jth spectral band. This modification will henceforth be referred to as ‘Modified q-sample (ranks)’ (using the same notation as Cebulla et al, 2006).

In addition to the Modified q-sample V2 test, the ‘Modified q-sample V4’ test (Cebulla et al., 2006) is also included in the analysis. The latter uses the actual values of the phases and amplitudes as opposed to their ranks, in which case r_{ij} in equation 5 refers to the amplitude of the ith Fourier component of the jth spectral band, and β_{ij} to the (untransformed) phase value of the ith Fourier component of the jth spectral band. The significance of these statistics can furthermore be evaluated with pre-determined critical values based on simulations (Stürzebecher et al, 1999; Cebulla et al, 2000; Cebulla et al, 2006). Deviating from the literature, the significance of the Modified q-sample V2 and V4 statistics in this study are evaluated using the bootstrap, as opposed to using pre-determined thresholds generated from no-stimulus data. How the critical decision thresholds might differ between the two approaches is further considered in the discussion.

3.3 The Fsp and the Fmp

The Fsp and the Fmp are defined as the ratio between the variance of the mean epoch $\overline{E}$ (found by taking the K averages across the columns of data matrix **D**) and the estimated variance of the EEG background noise. For the Fsp, the variance of the EEG background noise is estimated by the 'single point' (SP) variance, which is defined as the variance down a single column of data matrix **D**. The Fsp is given by (Elberling & Don, 1984):

Equation 8.

$$\mathrm{Fsp} = \mathrm{N}\frac{\mathrm{VAR}(\bar{E})}{\mathrm{VAR}(\mathrm{SP})}$$

where VAR denotes sample variance and SP refers to the values along an arbitrarily chosen column of **D**. For the Fmp, the variance of the EEG background noise is estimated by taking the average of multiple 'SP variances' (the average of the variances of multiple columns of **D**). The Fmp is given by (Martin et al., 1994):

Equation 9.

$$\mathrm{Fmp} = \mathrm{N}\frac{\mathrm{VAR}(\bar{E})}{\frac{1}{\mathrm{M}}\sum_{\mathrm{j}=1}^{\mathrm{M}}\mathrm{VAR}(\mathrm{SP_j})}$$

where VAR($\mathrm{SP_j}$) is the variance of the j$^{\mathrm{th}}$ included column of **D**, and M is the number of columns (of **D**) to include.

Under the null hypothesis of no response present, it is assumed that the Fsp and the Fmp follow F-distributions with v_1 and N-1 degrees of freedom. The latter is justified by assuming that epochs are sufficiently distant in time to be uncorrelated (and thus independent for normally distributed data).

When the spectrum of the coherent average $\overline{E}$ is white, then the K samples within the coherent average can also be considered independent and v_1 will equal K. The finite frequency content of EEG background activity, however, introduces correlations between the samples, which makes the true degrees of freedom difficult to estimate. A conservative recommendation, i.e. a FPR smaller than the nominal alpha level of the test, for v_1 is 5 (Elberling and Don, 1984). As an alternative, the Fsp and Fmp can be evaluated with the bootstrap approach.

3.4 Bootstrapping

Bootstrapping is a resampling with replacement procedure that can be used to construct a reference distribution so that statistical inference can be performed (Efron & Tibshirani, 1993). For evoked response detection, the goal is to construct the null distribution of some statistic by repeatedly drawing ensembles of epochs from the continuous EEG record, and calculating the statistic of interest on each new ensemble. Each bootstrapped ensemble should therefore represent the no-response condition, achieved by randomly selecting N segments of K samples from within the continuous EEG without regard to where the stimuli occur (Lv et al., 2007). Note that the selected segments may overlap, in accordance with the principles of bootstrapping where samples are picked at random with replacement, i.e. without removing that data from what can be picked later. The null distribution is then approximated by calculating the statistic in question from many bootstrapped ensembles (1000 ensembles were used for this study). Finally, the statistic is also calculated from the original ensemble of epochs, and its significance is evaluated by finding its location (percentile) along the bootstrapped null distribution.

3.5 Specificity assessment

A methods FPR, or 1-specificity, is defined in the current work as the percentage of significant test outcomes when no response is present (note again that this definition differs from studies that aim to detect a clinical disorder, e.g. hearing loss, where sensitivity commonly refers to the detection of the disorder). The FPRs of the methods were evaluated for different ensemble sizes using the pre-processed recordings of EEG background noise (no stimulus was used) described in section 2.2. The ensemble sizes selected for the analysis were 50, 100, 175, 275, 375, 500, 650, and 800 epochs, which were chosen based on results from the sensitivity assessment (see section 3.6). For each ensemble size, the EEG recordings were decomposed into ensembles of (consecutive) 30.03 ms epochs, resulting in a total of 20197, 10060, 5717, 3606, 2640, 1967, 1500, and 1187 ensembles with ensemble sizes of 50, 100, 175, 275, 375, 500, 650, and 800 respectively. Note that no further distinction was made between EEG recordings obtained under different noise conditions. The latter keeps the results concise, and is justified by realizing that all four conditions occur in clinical practice, and that, ideally, the methods should perform adequately under each of them. The detection methods were then applied to the initial 15 ms windows of the 30.03 ms epochs within each ensemble. For the frequency domain methods, the spectral resolution was first increased to 40 Hz by extending each 15 ms segment to 25 ms with zero-padding. The frequency domain methods were then applied to all spectral bands between 80 and 600 Hz. For the Modified q-sample V2 and V4 tests, averaging was used (prior to calculating the FFT) to compress each ensemble into blocks of sub-averages, as recommended by Cebulla et al (2000) (As noted by one of the reviewers, it is worth emphasizing that these recommendations were formulated for ABR detection, and that in later publications on ASSR detection, averaging is not advocated, see Cebulla et al, 2006). In this study, averaging was performed across blocks of 25 epochs so that no epochs were excluded from the analysis (each ensemble size is a multiple of 25), which hence compressed each ensemble into $\frac{N}{25}$ sub-averages. With respect to the time domain methods, a total of 25 TVMs were used for the Hotelling's T^2 test, which were spread equally across the 15 ms analysis window. The choice for 25 TVMs was based on additional simulations, which showed a robust performance for the Hotelling's T^2 test when using anything between ~20 and ~40 TVMs. These simulations were similar to the ones described in section 3.6 below, but used an alternative set of ABR templates for simulating a response (obtained from the coherent averages of the subject data described in Elberling et al, 2010). The column index (of data matrix **D**) for calculating the single point variance for the Fsp was furthermore arbitrarily set to 30, and the number of columns to include in the Fmp was set to 75 (corresponding to the full analysis window, or 15 ms). The significance of the Fsp and Fmp was evaluated using either F-distributions with 5 and N-1 degrees of freedom (denoted by 'Fsp 5 dof' and 'Fmp 5 dof' respectively) or with the bootstrap approach (denoted by 'Fsp bootstrapped' and 'Fmp bootstrapped' respectively).

3.6 Sensitivity assessment using simulations

A methods TPR, or sensitivity, is defined as the percentage of significant test outcomes (ABR responses detected) when a response is present, which should of course be as high as possible for some set FPR. In this study, sensitivity was assessed using both simulations and subject recorded ABR data. Simulations were included as these allow a large number of tests to be performed, which allows powerful comparisons to be drawn amongst the methods. This is important for the present study as the analysis regarding the subject recorded ABR data (section 3.7) was based on just 12 subjects.

The data used for the simulations consists of (i) the pre-processed recordings of EEG background noise (see section 2.2), along with (ii) the coherent averages from the subject recorded ABR data that contained a clear response. The latter was determined through visual inspection by an experienced audiologist. As guidance for determining the presence of a clear response, the audiologist inspected the repeatability of the waveform by comparing two replicates of the coherent average (obtained by averaging across epochs 1 to 1500, and again across epochs 1501 to 3000). The audiologist also used the 3 to 1 signal to noise criterion as additional guidance (see Sutton et al, 2003), but was ultimately left free to decide whether a response was present or not. This process resulted in a total of 34 ABR templates with a clear response: 4, 7, 8, 7, and 8 from the 10, 20, 30, 40, and 50 dB SL conditions respectively. Data was then assembled by randomly selecting N consecutive epochs from within a randomly selected recording of EEG background noise, and adding a randomly selected and rescaled ABR template to all epochs within the ensemble. The ensemble size N took values of 50, 100, 175, 275, 375, 500, 650 and 800 epochs, which were chosen based on the results from a pilot simulation that showed a good coverage of TPRs across methods when using these values. The scaling factor was furthermore chosen so that the signal to noise ratio (SNR) was -23 dB, which was calculated according to:

Equation 10.

$$\mathrm{SNR} = 10\log_{10}\left(\frac{\mathrm{P_{Signal}}}{\mathrm{P_{Noise}}}\right)$$

where $\mathrm{P_{Signal}}$ is the mean square of the scaled ABR waveform, and $\mathrm{P_{Noise}}$ the mean square of the ensemble of N epochs (prior to adding the ABR waveform, and treated as a continuous recording). The SNR of -23 dB was based on a brief analysis of the subject recorded ABR data, which showed that the responses were in the proximity of -23 dB. The latter was similarly calculated with equation 10, with $\mathrm{P_{Signal}}$ now being the mean square of the coherently averaged ABR (calculated across all 3000 epochs from the subject and dB SL condition in question), and $\mathrm{P_{Noise}}$ the mean square of the epochs when treated as a continuous recording. A total of 10 000 tests were performed for each ensemble size using the same detection methods and features as those described in section 3.5.

3.7 Sensitivity and detection time assessment using subject recorded ABR data

The sensitivities and detection times of the methods were further evaluated using just the subject recorded ABR data. The methods were applied to the initial 1-16 ms segments of the epochs (the first ms was excluded to avoid potential contaminations from a stimulus artefact), which was repeated for each subject and each stimulus condition. The methods were applied to the data sequentially, every 50 epochs, from 50 epochs onwards. To clarify - a test was first performed using an ensemble size of 50, then again using an ensemble size of 100, etc., until all 3000 epochs had been analysed (a total of 60 tests, per subject, and per dB SL condition). The detection methods and features selected for the analysis were the same as those described in section 3.5.

4. Results

This section presents the results regarding the specificity assessment (section 3.5), the sensitivity assessment using simulations (section 3.6), and the sensitivity and detection time assessment using the subject recorded ABR data (section 3.7).

4.1 Specificity assessment

The FPRs of the methods (using an alpha of 0.01) for the no-stimulus condition are presented in Table 1 for different ensemble sizes N. The upper and lower boundaries for significant deviations (**p<0.05**) from the expected 1% FPRs were found using the binomial distribution (see appendix). Results show that the FPRs of \`Fsp 5 dof' and \`Fmp 5 dof' were significantly (**p<0.05**) lower than 1%, as predicted by Elberling & Don (1984). The remaining methods appear to show a slight tendency towards a higher than expected FPR, which was significant (**p<0.05**) for: 'T2 Time' (for N=100 and N=375) , 'T2 Freq' (for N=375 and N=650), 'Fsp bootstrapped' (N=50, N=100, N=175, and N=375), 'Fmp bootstrapped' (N=100 and N=175), and 'Modified q-sample V2' (N=375 and N=500). Although these deviations are relatively small (all remained below 2%), a higher than expected FPR can be worrisome for some ABR applications. Additional simulations were therefore performed to further test and explore why the methods appear to show a higher than expected FPR. The data for these simulations consisted of a large amount (50 000 recordings) of Gaussian-distributed coloured noise with similar spectral content to real EEG background noise, and with stationary variance and a true mean of zero. The resulting data was pre-processed and analysed as described in sections 2.2 and 3.5, i.e. the settings were the same as those used for evaluating the real EEG background activity. The resulting FPRs were in the range of 1.15% to 1.2% for an expected FPR of 1%. Although very small, the deviations were significant (**P<0.01)** as the statistical power was high (50 000 tests were performed). Note that all underlying statistical assumptions for these simulations were met, except potentially the independence assumption between epochs, which suggests that the slightly higher than expected FPR can be attributed to a violation of this assumption. The latter is further addressed in the discussion.

- INSERT TABLE 1 -

4.2 Sensitivity assessment using simulations

The detection rates of the methods (using an alpha of 0.01) are presented in Figure 1 as a function of the ensemble size N. The best performances (highest TPR) is noted consistently for the HT2 tests, followed by the modified Q-sample tests, the bootstrapped Fmp and Fsp, and lastly by the Fsp and Fmp evaluated with F-distributions with assumed degrees of freedom. As the latter have a lower FPR also (see Table 1), a reduced TPR also might be expected. Moreover, the Fmp and Fsp use only the SNR (i.e. average power values) and it is thus not surprising that they are less sensitive. Note that a potential danger of using detection rates for comparisons in sensitivity is that methods with higher FPRs are given an unfair advantage over those that are more conservative and have a lower FPR. As shown by Table 1, the FPRs were all close to the expected 1% (with the exception of ‘Fsp 5 dof’ and ‘Fmp 5 dof ‘), which suggests that the comparison in sensitivity was fair. The latter was verified by finding the critical alpha values under which the methods obtained a FPR of exactly 1%, and replotting the resulting detection rates. Results (not presented) were almost identical to those presented in Figure 1 (again, with the exception of ‘Fsp 5 dof’ and ‘Fmp 5 dof ‘). With respect to ‘Fsp 5 dof’ and ‘Fmp 5 dof’, their sensitivity was greatly improved (to values similar to those seen with the bootstrapped method) by using the adjusted critical alpha values. It also might be noted that the differences in FPR shown in Table 1 do not consistently explain the differences in TPR between the methods.

- INSERT FIGURE 1 –

4.3 Sensitivity and detection time assessment using subject recorded ABR data

The detection rates for an ensemble size of 3000 are presented in Figure 2 for the 0, 10, 20, and 30 dB SL conditions (the 40 and 50 dB SL conditions are excluded as all methods obtained a 100% detection rate here). The required time for detecting a response was then found by finding the number of stimuli (expressed in seconds) required for the p-value to drop and remain below the 0.01 threshold for the remainder of the test. The additional requirement that the p-value remains below the 0.01 threshold ensures that the FPR is not inflated due to multiple tests being performed. If a test did not drop below the 0.01 significance threshold, then the full ~90 seconds test time was used (corresponding to 3000 epochs), which may have resulted in an underestimation of the required test time in the case of a false negative. The mean of the resulting detection times (taken across subjects) are presented in Figure 3 as bar graphs for each method, and dB SL condition. HT2 consistently showed the best performance.

- INSERT FIGURE 2 -

- INSERT FIGURE 3 -

Visually inspecting the distributions of the detection rates and detection times showed that both were strongly non-Gaussian, which was confirmed with the Kolmogorov-Smirnov goodness of fit test (**p<0.01** for all distributions). Non-parametric statistical analysis was therefore used to test whether the discrepancy amongst the methods in terms of detection rates and detection times was significant. With respect to detection rates, Cochran's Q test was first used to test for equivalence in performance across all 8 methods for each dB SL condition. Results indicate a significant difference in performance for the 10 (**p < 0.01**) and 20 (**p < 0.05**) dB SL conditions. As a follow-up, Fishers exact test was used to draw pairwise comparisons amongst the methods for the 10 and 20 dB SL conditions.

Results show that the performance of T2 Time and T2 Freq both differed significantly (**p<0.05**) from 'Fmp 5 dof' and 'Modified q-sample V4' for the 10 dB SL condition. The latter is presented in Figure 2 using asterisks and crosses, where an asterisk denotes a significant discrepancy with T2 Time, and a cross a significant discrepancy with T2 Freq. Comparisons between the remaining methods were not significant. Similarly, with respect to detection times (Figure 3), non-parametric statistical analysis was first used to test for equivalence in performance across all 8 methods (now using Friedman's test), per dB SL condition. Results indicate a significant difference in performance for the 10, 20, 30, 40, and 50 dB SL conditions (all **p < 0.001**). The Wilcoxon rank sum test was then used to draw pairwise comparisons between all methods, for the 10, 20, 30, 40, and 50 dB SL conditions. The majority of the comparisons were again not significant, with the exception of the Hotelling's T2 test. Significant (**p<0.05**) discrepancies between T2 Time, T2 Freq, and the remaining methods are again represented by asterisks and crosses in Figure 3, with asterisks denoting a significant discrepancy with T2 Time and crosses with T2 Freq.

5. Discussion

This study used simulations and subject recorded data to compare the specificity, sensitivity and detection time of various objective ABR detection methods. With respect to specificity, although the FPRs mostly fell within the lower and upper 95% confidence intervals of the expected FPR, a slightly higher than expected FPR was observed. The Fsp and Fmp evaluated with theoretical F-distributions were the exception, both of which showed consistently lower than expected FPRs. In terms of sensitivity and detection time, the Hotelling's T^2 test came out on top in both the simulations and the subject recorded data. The results regarding these properties (specificity, sensitivity, and detection time) are now discussed in more detail.

5.1 Specificity

The results from the specificity assessment (Table 1) suggest a slightly higher than expected FPR for most methods (excluding 'Fsp 5 dof' and 'Fmp 5 dof'). Although the deviations were small, this is a concern for many ABR applications where a higher than expected FPR can potentially have severe repercussions. In ABR hearing screening applications, for example, a higher than expected FPR can result in additional cases of undetected hearing loss (ABR responses are falsely detected), which (when left untreated) have been associated with an impaired language development in children (Ramkalawan & Davis, 1991), along with various other more obvious handicaps, such as discrimination, less effective education, a reduced life expectancy, and higher unemployment rates (Miziara, 2012), to name a few. Supplementary simulations (see section 4.1) were therefore performed to explore why the FPRs appear to be higher than expected. Results suggest that the increased FPR can be attributed to a violation of the independence assumption between epochs. For recorded EEG data, which is known to be dominated by low frequencies with an approximate $\frac{1}{f^{\alpha}}$ spectrum (with $\alpha \approx 1$, see Pritchard, 1992), a similar violation can be expected, with the extent of the violation depending primarily on the cut-off frequency of the high-pass filter and the inter-epoch interval (i.e. the stimulus rate). It can thus be expected that increasing the high-pass cut-off frequency or decreasing the stimulus rate would reduce the long term correlations between epochs, and increase their independence. Results from an additional post-hoc analysis indeed show no significant (**p<0.05)** deviations from the expected 1% FPR when repeating the specificity assessment (section 3.5) with an adjusted high-pass cut-off frequency of 100 Hz. An alternative solution to increased FPRs may be to adjust the significance level (the alpha value) of the test. Post hoc analysis showed that FPRs of exactly 1% (across all ensemble sizes in Table 1) could be obtained when using the following alpha values: 0.0087 (T2 Time), 0.0088 (T2 Freq), 0.021 (Fsp 5 dof), 0.0321 (Fmp 5 dof), 0.0071 (Fsp bootstrapped), 0.0074 (Fmp bootstrapped), 0.0088 (Modified q-sample V2), and 0.009 (Modified q-sample V4). Although these adjusted alpha values would result in a small loss of sensitivity, they may be a safer option when detecting ABRs in practice.

It should furthermore be emphasized that the adjusted alpha values for 'Fsp 5 dof' and 'Fmp 5 dof' are expected to be more susceptible to the high-pass cut-off frequency than the remaining methods due to an additional violation of the independence assumption amongst samples within epochs. In particular, increasing the high-pass cut-off frequency results in fewer low frequency components, which reduces the correlations amongst the samples within the epochs, thus increasing the degrees of freedom of the data and removing it farther from the assumed degrees of freedom v_1=5. It can therefore be expected that the performance of 'Fsp 5 dof' and 'Fmp 5 dof' would be even more conservative when the high-pass cut-off frequency is increased, which was confirmed when repeating the specificity assessment (section 3.5) with an adjusted high-pass cut-off frequency of 100 Hz. In particular, the degrees of freedom v_1 for the no-stimulus EEG recordings (section 2.2) ranged from 3 to 15 (with mean 8.4 and standard deviation 2.6) when using a high-pass cut-off frequency of 30 Hz, and from 3 to 20 (with mean 11.3 and standard deviation 4.9) when using a high-pass cut-off frequency of 100 Hz (the latter was achieved by fitting F-distributions to each bootstrapped null distribution, and finding the best fitting function). It might be noted that the conservative estimate of 5 degrees of freedom was originally intended for the higher cut-off frequency of 100 Hz (Elberling & Don, 1984). Note also that the Hotelling's T^2 test and the bootstrapped statistics are immune to independence violations amongst samples within epochs (but not between epochs). In particular, the Hotelling's T^2 test accounts for correlated samples within epochs by scaling the features by their covariance matrix (see methods section), whereas the bootstrapped statistics account for correlated samples by resampling on an epoch to epoch basis, which preserves the correlations between samples within epochs. This further allows the bootstrapped confidence intervals to more accurately reflect test-dependent factors, such as the EEG background noise, the electrode impedances, and ultimately the degrees of freedom of the data. The latter is important for many ABR applications where the objective detection methods are expected to perform adequately across EEG recordings with varying degrees of freedom. It is hence hypothesized that the Hotelling's T^2 test and the bootstrapped statistics would provide more consistent results relative to 'Fsp 5 dof' and 'Fmp 5 dof' across a wider range of test conditions. A similar argument might be made in favour of the bootstrap approach over the use of pre-determined thresholds generated from no-stimulus data (see e.g. Stürzebecher et al, 1999; Cebulla et al, 2000; Cebulla et al, 2006), i.e. pre-determined thresholds may not generalize well across data sets and test conditions, whereas the bootstrap approach estimates confidence intervals specifically for the recording in question.

5.2 Sensitivity and detection time

With respect to sensitivity (the proportion of correctly identified responses) and detection time, these should ideally be as high and low as possible respectively for some set FPR. In ABR audiometry, for example, one would expect thresholds to decrease as the sensitivity of the detection method is increased, which may lead to greater convergence between behavioural and estimated hearing thresholds. In terms of reduced test time, one would expect an increased sensitivity to result in (i) a decreased cost of service delivery, (ii) reduced patient discomfort, and (iii) a smaller time window within which noise artefacts can be introduced to the data. Reduced detection times would be particularly beneficial in patients who cannot cooperate, such as infants or some with dementia.

In this study, sensitivity was evaluated using detection rates, which have the desirable properties of being intuitive and simple. However, as noted in results section 4.2, a potential risk of using detection rates is that methods with higher FPRs receive an unfair advantage over those with lower FPRs (the latter is most notably the case for ‘Fsp 5 dof’ and ‘Fmp 5 dof’, which were indeed designed to have lower FPRs). The problem can be resolved by adjusting the nominal alpha values so that the FPRs are equal across methods. Note however that although this allows for a more fair comparison, it is not necessarily a realistic one as adjustment of the FPR may need to be carried out on an individual basis using prior knowledge that is not always available. Results from the simulations nevertheless suggest an advantage for the Hotelling's T^2 test when using both the adjusted and unadjusted critical alpha values (results section 4.2). This is further supported by FPRs in Table 1: consistent differences in detection rates (Figure 1) cannot be readily explained from relatively inconsistent FPRs. Results from the subject recorded ABR data similarly suggest an overall advantage for the Hotelling’s T^2 test

(Figures 2 and 3). The relative discrepancy in performance amongst the remaining methods, however, is less clear, which can likely be attributed to a small sample size of just 12 subjects.

With respect to the frequency domain features for the Hotelling's T^2 test, it is worth noting that these are essentially the same as those used by the Modified q-sample V4 test (the Hotelling's T^2 test is applied to the real and imaginary parts of the Fourier components, whereas the Modified q-sample V4 test is applied to the phases and amplitudes), and yet a relatively large discrepancy in performance was still observed. This can likely be attributed to the way in which features are weighted and combined into a single statistic. In particular, the Hotelling's T^2 test weights the features according to their variance and covariance, whereas the Modified q-sample V4 test does not. The latter results in an L-dimensional hyper-ellipsoid (centred at features means $\bar{\mathbf{x}}$) as H0 rejection region for the Hotelling's T^2 statistic, where the shape of the ellipsoid is determined by the variance and covariance of the features. Having an ellipsoid as rejection region means that the null hypothesis is more easily rejected in some directions relative to others, meaning it has the potential of providing a more powerful test relative to, for example, a spherical rejection region.

Based on the preceding paragraph, an identical performance between the Modified q-sample V4 test and the Hotelling's T^2 test might be expected when applied to uncorrelated features with equal variance, which was tested with additional simulations. In particular, simulations described in Stürzebecher et al (1999) and Cebulla et al (2000) were implemented, which used Gaussian zero mean white noise with stationary variance to represent the EEG background noise, along with a sinewave multiplied with a Gaussian window for representing a response. The detection methods included for these simulations were (i) the original q-sample uniform scores test (Mardia, 1972), (ii) both the Modified q-sample V2 and V4 tests (Stürzebecher et al, 1999; Cebulla et al, 2006), and (iii) the Hotelling's T^2 test using the frequency domain approach. As predicted, the Hotelling's T^2 test and the Modified q-sample V4 test both came out on top in terms of sensitivity (with very similar performances), followed by the Modified q-sample V2 test (using ranks rather than measured values), and lastly by the original q-sample uniform scores test (which only uses phase ranks).

Study limitations

In this study, the investigators strived to present a comprehensive and fair comparison between various time and frequency domain ABR detection methods in terms of their sensitivity, specificity, and detection time. Whenever possible, feature selection and pre-processing parameters were based on recommendations or findings from the literature. That said, it remains to be seen whether the results presented in this study generalize across alternative feature and data sets. Various additional parameters worth investigating include the time window selected for the analysis, how many and which spectral bands to include for the frequency domain methods, and the selection of TVMs for T2 Time. With respect to the latter, a total of 25 TVMs spread equally across a 15 ms analysis window were used for this study. The choice for 25 TVMs was based on results from additional simulations described in section 3.5, which showed a good performance for the Hotelling's T^2 test when using anything between ~20 and ~40 TVMs. It is however worth noting that these simulations did not distinguish between stimuli of different intensities. Hence, although 25 TVMs may be a relatively robust set of features for ABR detection, it is not necessarily optimal, and it may be beneficial to use more specific arrangements of TVMs depending on the type of stimulus and/or stimulus parameters being used. A general rule of thumb is that the optimal number of TVMs will tend to increase along with the number of peaks in the ABR, since consecutive time-domain peaks within the ABR would cancel each other out when the number of TVMs is too low (Golding, 2009). Hence, when the stimulus intensity is decreased, and ABR waves I and III begin to disappear (Hall, 2006), it may be beneficial to use fewer TVMs.

5. Conclusion

Comparisons were drawn between various objective ABR detection methods in terms of specificity, sensitivity, and detection time. Results from the specificity assessment suggest a tendency towards slightly higher than expected FPR across methods, which likely can be attributed to a violation of the independence assumption between epochs. With respect to sensitivity and detection time, the Hotelling's T^2 test came out on top, which was primarily attributed to a more suitable weighting of the features. Finally, bootstrapping was shown to improve the reliability of the Fsp and the Fmp, as opposed to when test significance was evaluated using F-distributions with the recommended assumption of 5 degrees of freedom.

Acknowledgements

The authors would like to thank Debbie Cane for collecting the subject recorded ABR data and Sara M.K. Madsen for collecting the recordings of EEG background noise. This work was supported by the Oticon Foundation and the Engineering and Physical Sciences Research Council (EPSRC, grant No. EP/M026728/1). The data used throughout this study is openly available at the University of Southampton repository at http://doi.org/10.5258/SOTON/D0168.

Declarations of interest

The authors report no conflicts of interest.

Appendix

The binomial distribution

A Bernoulli trial is a random experiment with exactly two possible outcomes, typically interpreted as ‘success’ and ‘failure’. When X Bernoulli trials are performed and the probability of a successful trial is P, then the binomial distribution gives the probability densities of observing x successful trials. The distribution is given by:

$$\mathrm{B(x|X,P)} = \frac{\mathrm{X!}}{\mathrm{x!\,(X-x)!}}\mathrm{P^{x}\,(1-P)^{X-x}}$$

For the specificity assessment, the number of Bernoulli trials X was set to the number of ensembles tested, and the probability of a successful trial P to the expected probability of observing a false positive (P=0.01). The resulting distribution was used to find the 95% confidence intervals for the expected number of false positives.

References

Carter L., Golding M., Dillon H., Seymour J. 2010. The Detection of Infant Cortical Auditory Evoked Potentials (CAEPs) Using Statistical and Visual Detection Techniques. *J Am Acad Audiol,* 21, 347-356.

Cebulla M., Stürzebecher E., Wernecke K.D. 2000. Objective detection of auditory brainstem potentials - Comparison of statistical tests in the time and frequency domain. *Scand Audiol*, 29, 44-51.

Cebulla M., Stürzebecher E., Elberling C. 2006. Objective detection of Auditory Steady State Responses: Comparison of One-Sample and q-Sample Tests. *J Am Acad Audiol*, 17, 93-103.

Chang H.W., Dillon H., Carter L., van Dun B., Young S.T. 2012. The relationship between cortical auditory evoked potential (CAEP) detection and estimated audibility in infants with sensorineural hearing loss. *Int J Audiol*, 51, 663-670.

Efron B., Tibshirani R.J. 1993. An Introduction to the Bootstrap. Chapman & Hall/CRC. Editors: Cox D.R., Hinkley D.V., Reid N., Rubin D.B., Silverman B.W. Boca Raton, London, New York, Washington D.C.

Elberling C. 1976. Action Potentials Recorded from the Promontory and the Surface, Compared with Recordings from the Ear Canal in Man. *Scand Audiol,* 5, 69-78

Elberling C. 1979. Auditory electrophysiology: spectral analysis of cochlear and brain stem evoked potentials. A comment on: Kevanishvili and Aponchenko: “Frequency composition of brain stem auditory evoked potentials”. *Scand Audiol*, 8, 57-64.

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

Elberling C., Callø J., & Don M. 2010. Evaluating auditory brainstem responses to different chirp stimuli at three levels of stimulation**.** J. Acoust. Soc. Am., 128 (1) 215-223.

Golding M., Dilon H., Seymour J., Carter L. 2009. The detection of adult cortical auditory evoked potentials (CAEPs) using an automated statistic and visual detection. *Int J Audiol*, 48, 833-842.

Hall J.W., 2006. New Handbook of Auditory Evoked Responses. Pearson, 1st edition.

Hotelling H. 1931. The Generalization of Student's Ratio. *Ann Math Statist*, 2, 360-378.

Kevanishvili Z., Aphonchenko V. 1979. Frequency composition of the brain-stem auditory evoked potential. *Scand Audiol*, 8, 51-55.

Lachowska M., Bohórquez J., Ozdamar O. 2012. Simultaneous acquisition of 80 Hz ASSRs and ABRs from quasi ASSRs for threshold estimation. *Ear Hear*, 33, 660-671.

Lv J., Simpson D.M., Bell S.L. 2007. Objective detection of evoked potentials using a bootstrap technique. *Med Eng Phys*, 29, 191-198.

Madsen S.M.K. 2010. Accuracy of averaged auditory evoked potential amplitude and latency estimates. Msc thesis, Dept. Elec. Eng., Tech. Uni of Denmark.

Madsen S.M.K., Harte J.M., Elberling C., Dau T. 2017. Accuracy of averaged auditory brainstem response amplitude and latency estimates. *Int. Jnl. Audiol,* 1-9.

Mardia K.V. 1972. Statistics of directional data. 1972. London: Academic Press.

Martin W.H., Schwegler J.W., Gleeson A. L., Shi Y.B. 1994. New techniques of hearing assessment. *Otolaryngol Clin North Am*, 27, 487-510.

Miziara I.D., Miziara C.S., Tsuji R.K., Bento R.F. 2012. Bioethics and medical/legal considerations on cochlear implants in children. *Braz. J. Otorhinolaryngol*, 78: 70-79.

Moore E.H. 1920. On the reciprocal of the general algebraic matrix. *Bulletin of the American Mathematical Society,* 26, 394–395.

Penrose R. 1955. A generalized inverse for matrices, *Proceedings of the Cambridge Philosophical Society,* 51, 406–413

Pritchard, W. (1992). The brain in fractal time: 1/f-like power spectrum scaling of the human electroencephalogram. *Intern. J. Neurosci*, 66, 119–129.

Ramkalawan T.W., Davis A.C. 1992. The effects of hearing loss and age of intervention on some language metrics in young hearing-impaired children. *Br. J. Audiol*. 26: 97-107.

Rencher A.C. 2001. Methods of Multivariate Analysis. Second Edition. John Wiley & Sons, Inc. Chapter 5, page 118.

Stürzebecher E., Cebulla M., Baag M., Thie R. 1996. Verfahren zur automatischen Hörschwellenbestimmung, insbesondere bei Neugeborenen und Kleinkindern. Patent application PCT/DE96/02453.

Stürzebecher E., Cebulla M., Wernecke K. 1999. Objective response detection in the frequency domain: comparison of several q-sample tests. *Audiol Neurootol*, 4, 2-11.

Sutton G., Lightfoot G., Stevens J., Booth R., Brennan S., Feirn R., Meredith R. Guidance for Auditory Brainstem Response testing in babies. March 2013, Version 2.1.

Suzuki T., Sakabe N., Miyashita Y. 1982. Power Spectral Analysis of Auditory Brain Stem Responses to Pure Tone Stimuli. *Scand Audiol*, 11, 25-30.

Valdes-Sosa M.J., Bobes M.A., Perez-Abalo M.C., Perera M., Carballo J.A., Valdes-Sosa P. 1987. Comparison of Auditory-Evoked Potential Detection Methods Using Signal Detection Theory. *Audiology*, 26: 166-178.

Van Dun B., Carter L., Dillon H. 2012. Sensitivity of cortical auditory evoked potential (CAEP) detection for hearing-impaired infants in response to short speech sounds. *Audiol Res*, 2e13, 65-76.

Van Dun B., Dillon H., Seeto M. 2015. Estimating Hearing Thresholds in Hearing-Impaired Adults through Objective Detection of Cortical Auditory Evoked Potentials. *J Am Acad Audiol*, 26, 370-383.

Tables

Table 1. The percentage of significant (**p<0.01**) tests for the no-response condition, per method and per ensemble size. Significant deviations (**p<0.05**) from the expected 1% FPR are indicated by a red asterisk.

Figures

Figure 1. The detection rates of the methods (using an alpha of 0.01) as a function of the ensemble size when detecting a simulated -23 dB response (see section 3.6 for details).

Figure 2. The percentage of detected responses (**p<0.01**) for each method, presented as bar graphs for each dB SL condition. Non-parametric statistical analysis was used to test whether the discrepancy between methods was significant (see section 4.3 for details). The majority of the comparisons were not significant, with the exception of the Hotellings T^2 test. Significant discrepancies (**p<0.05**) with T2 Time are indicated by an asterisk (placed above the bar of the corresponding method), whereas significant discrepancies with T2 Freq are indicated by a cross.

Figure 3. The mean of the detection times of the methods (calculated across subjects), presented as bar graphs for each per dB SL condition. Non-parametric statistical analysis was used to test whether the discrepancy between methods was significant (details presented in section 4.3). The majority of the comparisons were not significant, with the exception of the Hotellings T^2 test. Significant discrepancies (**p<0.05**) with T2 Time are indicated by an asterisk, placed above the bar of the corresponding method. Significant discrepancies with T2 Freq are indicated by a cross, similarly placed above the bar of the corresponding method.

Figure 1

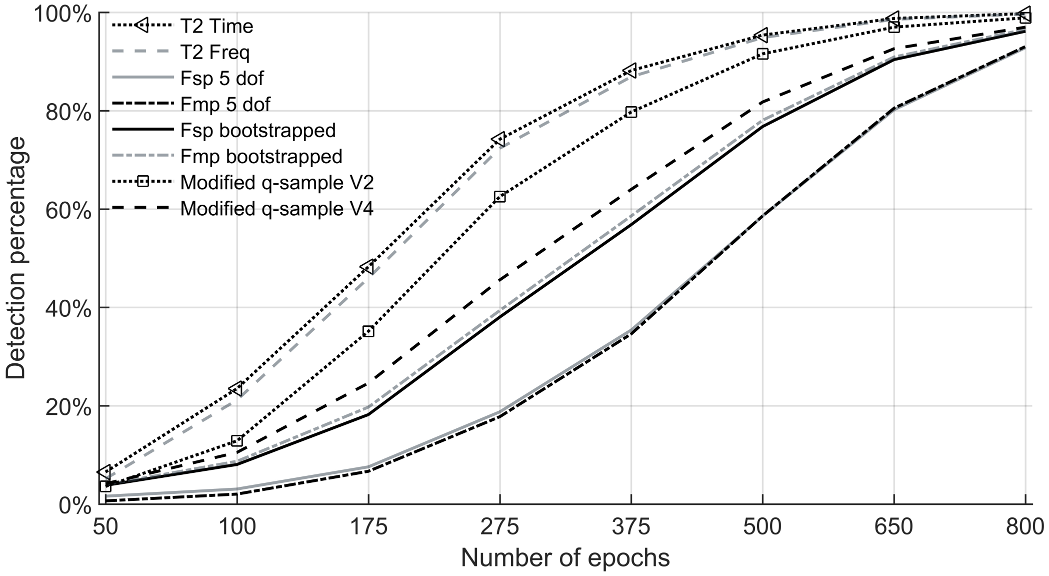
T2 Time
T2 Freq
Fsp 5 dof
Fmp 5 dof
Fsp bootstrapped
Fmp bootstrapped
Modified q-sample V2
Modified q-sample V4
100%
80%
60%
40%
20%
0%
Detection percentage
50
100
175
275
375
500
650
800
Number of epochs

Figure 2

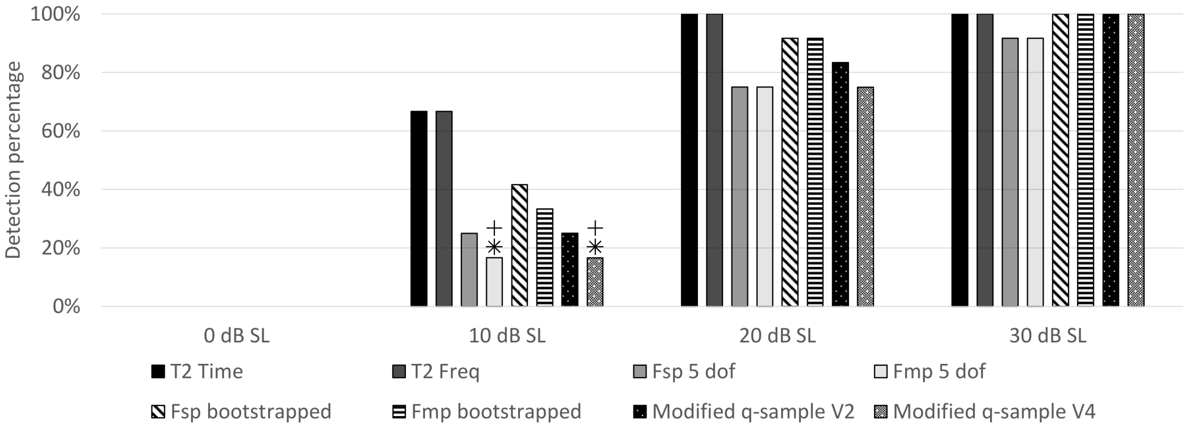
100%
80%
60%
40%
20%
0%
Detection percentage
0 dB SL
10 dB SL
20 dB SL
30 dB SL
T2 Time
T2 Freq
Fsp 5 dof
Fmp 5 dof
Fsp bootstrapped
Fmp bootstrapped
Modified q-sample V2
Modified q-sample V4
+
*
+
*

Figure 3

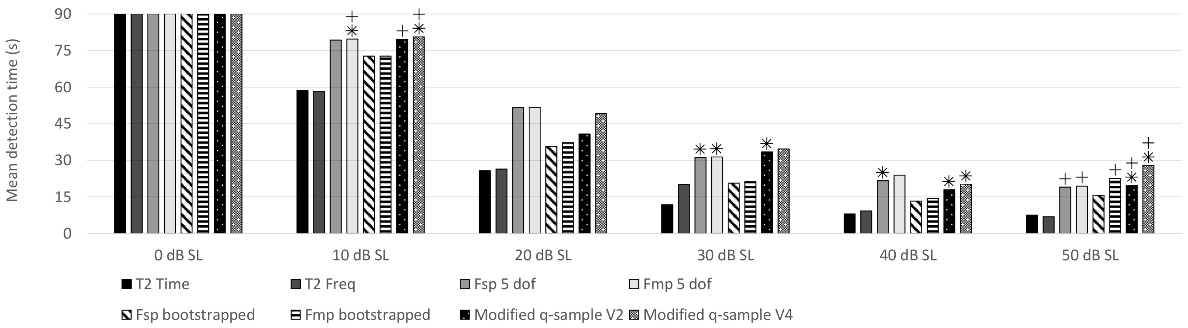
Mean detection time (s)
90
75
60
45
30
15
0
0 dB SL
10 dB SL
20 dB SL
30 dB SL
40 dB SL
50 dB SL
T2 Time
T2 Freq
Fsp 5 dof
Fmp 5 dof
Fsp bootstrapped
Fmp bootstrapped
Modified q-sample V2
Modified q-sample V4
+
✳

Table 1

Number of epochs per ensemble ->	50	100	175	275	375	500	650	800
T2 Time	1.08%	1.25%*	0.98%	1.33%	1.48%*	0.92%	1.13%	1.26%
T2 Freq	1.09%	1.08%	1.14%	1.19%	1.59%*	1.32%	1.73%*	1.26%
Fsp 5 dof	0.54%*	0.53%*	0.51%*	0.5%*	0.80%	0.56%*	0.4%*	0.51%*
Fmp 5 dof	0.23%*	0.36%*	0.37%*	0.44%*	0.61%*	0.56%*	0.33%*	0.42%*
Fsp bootstrapped	1.15%*	1.27%*	1.4%*	0.94%	1.44%*	1.17%	1.27%	1.52%
Fmp bootstrapped	1.12%	1.24%*	1.24%	0.97%	1.48%*	1.02%	1.20%	1.43%
Modified q-sample V2	0.94%	0.96%	1.17%	1.25%	1.44%*	1.88%*	0.87%	1.52%
Modified q-sample V4	0.86%	1.05%	1.15%	0.89%	1.14%	0.97%	1.13%	1.43%