**Efficient detection of cortical auditory evoked potentials in adults using bootstrapped methods**

**Chesnaye M.A1., Bell S.L1., Harte J.M4., Simonsen L.B4., Visram A2,3, Stone M.A.,2,3, Munro K.J., 2,3  & Simpson D.M1.**

1Institute of Sound and Vibration Research, Faculty of Engineering and the Environment, University of Southampton, UK.

2Manchester Centre for Audiology and Deafness, School of Health Sciences, University of Manchester, UK

3Manchester University Hospitals NHS Foundation Trust, Manchester Academic Health Science Centre, Manchester, UK

4Interacoustics Research Unit, c/o Technical University of Denmark, Denmark. jmha@iru.interacoustics.com

**Keywords**

Cortical Auditory Evoked Potentials, objective detection methods, bootstrap

**Abstract**

Statistical detection methods are useful tools for assisting clinicians with cortical auditory evoked potential (CAEP) detection, and can help improve the overall efficiency and reliability of the test. However, many of these detection methods are parametric, and require various assumptions regarding the electroencephalogram (EEG) data to be satisfied. When these assumptions are violated, then reduced test sensitivities and/or increased or decreased false-positive rates can be expected. As an alternative to the parametric approach, test significance can also be evaluated using a bootstrap, which does not require some of the aforementioned assumptions. Bootstrapping also permits a large amount of freedom when choosing or designing the statistical test for response detection, as the theoretical null distributions underlying the test statistic no longer need to be known in advance.***Objectives****:* The aim of the current study was to evaluate the specificity and sensitivity of various methods for automatic CAEP detection. ***Design*:** The methods included in the assessment were the Hotelling’s T2 test, the Fmp, four modified q-sample statistics, and various template-based detection methods (calculated between the ensemble coherent average and some pre-defined template), including the correlation coefficient, covariance, and dynamic time-warping (DTW). The assessment was carried out using both simulations and a CAEP threshold series collected from 23 adults with normal hearing. ***Results:*** The most sensitive method was DTW, evaluated using the bootstrap, with maximum increases in test sensitivity (relative to the conventional Hotelling’s T2 test) of up to 30%. An important factor underlying the performance of DTW is that the template adopted for the analysis correlates well with the subjects’ CAEP. ***Conclusion*:** When subjects’ CAEP morphology is approximately known prior to the test, then the DTW algorithm provides a highly sensitive method for CAEP detection.

**Introduction**

Cortical auditory evoked potentials (CAEPs) are changes in neural activity at or near cortical brain regions in response to acoustic stimuli (Picton 2011). They are typically recorded non-invasively using the electroencephalogram (EEG), and have main clinical applications in the diagnosis of psychiatric and neurological disorders (Duncan et al. 2009), along with hearing threshold estimation in subjects who cannot or will not provide reliable behavioural responses (Lightfoot and Kennedy 2006). For many of these applications, the challenge is to determine whether a response is present or absent within an acceptable timespan. This is usually achieved by highly trained individuals, who are given the task to visually inspect the acquired waveforms. When doing so, the examiner can be assisted by an objective detection method, which has the goal to reduce dependency on the examiner, and to provide an overall more time-efficient and reliable test outcome. This is particularly important when testing infants and the hearing-impaired where variability in response morphology is considered to be substantial (Picton 2011), i.e. it is not always clear for the examiners what to look for in the CAEP waveforms. For the current study, the overall goal is to improve the reliability and efficiency of CAEP-related applications by improving the performance of objective CAEP detection methods. To do so, simulations and subject-recorded data were used to evaluate and compare the performance of various competing detection methods. The following paragraphs present a brief overview of these methods.

Starting with some promising methods from the literature, the Hotelling's T2 test is frequently used for CAEP detection (Carter et al. 2010; Chang et al. 2012; Van Dun et al. 2012; Van Dun et al. 2015), and has been shown to have a test sensitivity at least comparable to that of visual inspection by examiners (Golding et al. 2009). A potential competitor to the Hotelling's T2 test is the q-sample uniform scores test and its modifications (Stürzebecher et al. 1999; Cebulla et al. 2006), which have been shown to have a good test sensitivity for auditory steady-state response (ASSR) detection, but have not yet been explored for CAEP detection. Some additional test statistics of interest include the frequently used ‘F for multiple points’ (Fmp) statistic (Martin et al. 1994) and the correlation coefficient (CC; Mason et al. 1977; Hoth 1993). A recurring complication with the Fmp and the CC, however, is that their null distributions are often recording-dependent, and are hence typically unknown prior to the test. Evaluating test significance for these methods can therefore be problematic. One solution is to use a bootstrap approach (Efron and Tibshirani 1994), which is a random resampling with replacement procedure that can be used to approximate the recording-dependent null distribution for the EEG feature of interest (Lv et al. 2007). The approximated null distribution can then be used to conduct statistical inference (see the Methods section for more details). Besides providing robust, non-parametric evaluations of test significance, the bootstrap also gives the user a large amount of freedom when choosing which EEG features to use for response detection, as the null distributions no longer needs to be known prior to the test.

In this study, the bootstrap is used to evaluate the test significance of the Fmp, the CC, and the modified q-sample statistics. An additional template-based detection method that is explored (and evaluated using the bootstrap) is covariance (COV), calculated between the ensemble coherent average and some template. COV has a potential advantage over the standard CC as it is sensitive to both the shape of the evoked response (as is the CC), and its amplitude. A shortcoming for both the CC and COV, however, is that the adopted template might not correlate well with the CAEP waveform from some subjects. It might therefore be beneficial to allow some degree of flexibility in the adopted template. The latter can be achieved using the dynamic time-warping (DTW) algorithm (Sakoe and Chiba 1978), which allows the template to be stretched and compressed in time, such that correlation with the subject CAEP is optimised. DTW has previously been used for evoked response classification and for improving SNRs by re-aligning single trial responses (Huang and Jansen 1985: Picton et al. 1988; Gupta et al. 1996; Wang et al. 2001; Casarotto et al. 2005; Assecondi et al. 2009, Goffredo et al. 2016), but not yet been explored as automatic detection method with specificity controlled using the bootstrap from Lv et al. (2007).

To summarise, the aim of this study was to evaluate and compare the specificity and sensitivity of the aforementioned detection methods for adult CAEP detection, with the overall goal of improving the reliability and efficiency of CAEP-related applications. The objective detection methods included in the assessment are (i) the Hotelling's T2 test, (ii) various modified q-sample statistics from Cebulla et al (2006), (iii) the Fmp, and (iv) multiple template-based detection methods, including the CC, COV, and DTW, all calculated between the ensemble coherent average and a pre-defined template. The assessment is carried out using extensive Monte-Carlo simulations, along with a sample of data from 23 adults with normal hearing where CAEPs were used to estimate behavioural hearing thresholds. Results from visual inspection by three experienced audiologists are also considered.

**Methods**

This section describes the CAEP data used for the assessment, along with the objective CAEP detection methods. The specificity and sensitivity assessment using simulations and subject-recorded CAEP data are also described.

**CAEP** **data**

CAEPs were recorded from 23 adults with normal hearing (19 female and 4 male, aged 19-30 years, mean age 24.8 years), i.e. ≤ 20 dB hearing level (HL) for audiometric frequencies 250-8000 Hz. All subjects underwent standard otoscopy and tympanometry to screen for outer and middle ear dysfunction. The stimulus for evoking the CAEP was a 70 ms mid-frequency (1120 – 1820 Hz) synthetic speech token (2/3 octave wide harmonic complex, harmonics 8-13 of 140 Hz fundamental; Stone et al., submitted 2019). The stimulus was presented unilaterally to the subjects at a rate of 0.9 Hz through an ER3A insert earphone at a range of sensation levels (SL), i.e. relative to the behavioural hearing thresholds. The test ear was by default the right ear unless excessive wax made placement of the insert phone troublesome, in which case the left ear was chosen. The behavioural hearing thresholds were first obtained using a standard audiometric procedure where the intensity of the stimulus was increased in steps of 5 dB for every missed response, and decreased in steps of 10 dB for every correct response. The stimulus was then presented to the subjects at 10, 20, 30 and 50 dB SL, along with a no-stimulus condition. The order of the test conditions was also randomised, per subject, along with the additional constraint that test orders were not repeated across subjects. For each stimulus sensation level, there were two runs of data collection. The criterion for stopping each run was either (i) 60 stimuli had been presented to the subject, or (ii) 40 artefact free post-stimulus windows (each with a duration of 1100 ms) had been accepted with online artefact rejection set at +/- 50 uV. For the no-stimulus condition, the stopping criteria were the same, except that the maximum number of stimuli was increased to 160, or that 80 artefact-free 1100 ms windows had been recorded. For every run of stimulus data collected, a run of no-stimulus data was also collected. During data collection, subjects were seated upright in a comfortable chair (with support being offered for the head and neck) whilst watching a muted DVD with subtitles. Data were then recorded using the Interacoustics Eclipse EP25 system with electrodes (impedances < 5 kΩ) placed at the high forehead (active electrode), the right mastoid (reference electrode), and the left mastoid (ground electrode). All recordings were downsampled to 1 kHz and band-pass filtered using a 3rd order Butterworth filter (implemented as a forward-backwards filter to preserve phase) from 1-15 Hz, after which they were stored for offline analysis.

*Offline artefact rejection*

Visually inspecting the acquired EEG recordings showed a large number of blink artefacts in the data. Preliminary analysis showed that these blinks caused notable degradation of CAEP estimates and needed to be removed or attenuated prior to further processing of the recordings. Blinks were therefore removed (offline) using a two-stage procedure. In the first stage, subject-specific ‘blink templates’ were constructed as follows: First, a 450 ms window that contained a stereotypical blink was manually selected from within a recording from the subject in question. This window functioned as an initial blink template, which was used to locate additional blinks. The latter was achieved by shifting the blink template through all EEG recordings from the subject in question, and finding all 450 ms segments that (i) contained both a maximum value larger than 50 uV and a minimum value smaller than -50 uV, and (ii) had a correlation with the blink template of at least 0.9 (calculated using Pearson’s correlation coefficient). For each identified blink, the match with the template was optimised by shifting the template across a 100 ms interval relative to the initial matching location, and finding the highest correlation with the template. Then, once all additional blinks had been located (and their locations optimised), the mean was taken across all blinks, after which the ‘mean blink’ was used as the new updated blink template. This procedure was repeated (now using the updated blink template) until the final blink template had stabilised, i.e. no new blinks were identified. In the second stage of the blink removal procedure, all identified blinks were subtracted from the recording using the optimised blink template. To do so, for each blink, the blink template was first re-scaled, such that its peak-to-peak amplitude matched the peak-to-peak amplitude of the blink-to-be-removed (rescaling was necessary to account for variation in the amplitudes of the blinks), after which the template was subtracted from the EEG segment in question. The full procedure was repeated on a subject-by-subject basis. After subtracting the blinks, the EEG recordings were again low-pass filtered at 15 Hz (using a 3rd-order Butterworth filter) to remove discontinuities in the data that were introduced by the blink removal procedure. The recordings were then structured into 1100 ms windows, consisting of an initial 500 ms pre-stimulus window, followed by a 600 ms post-stimulus window. For each dB SL condition, there were a minimum of 100 artefact free 1100 ms windows available, per subject. For the no-stimulus condition, there was approximately 6 hours of artefact free data available (roughly 15 minutes per subject).

**Objective detection methods**

This section describes the objective detection methods that were included in the specificity and sensitivity assessments. All detection methods were applied to the 50-300 ms windows (giving 250 ms epochs) following stimulus onset. When a CAEP is indeed present, then the 50-300 ms windows span a typical P1-N1-P2 complex in adults (Picton 2011).

*The one-sample Hotelling's T2 test*

The one-sample Hotelling's T2 test (Hotelling 1931) is the multivariate equivalent to Student’s one-sample *t-*test, and can be used to test the null hypothesis H0 that the mean values of Q features are equal to Q hypothesized values. When used for CAEP detection, the features consist of mean voltages, taken across short time-intervals within epochs (Golding et al. 2009), henceforth ‘voltage-means’. Each epoch is thus ‘compressed’ into Q voltage-means, giving an N x Q-dimensional feature matrix **V**:

**V** =

where N is the total number of epochs and is the jth voltage-mean extracted from the ith epoch. As suggested by Golding et al (2009), each voltage-mean is calculated across a 50 ms interval. The 250 ms epochs in this study are hence compressed into five voltage-means (**V** is N x 5-dimensional). The hypothesized values to test against are furthermore all zero, as the mean of the recording is shifted to baseline due to the high-pass filter. The T2 statistic itself is given by (Rencher 2001):

*Equation 1*

where is a vector containing the Q mean feature values (found by taking the means down the columns of **V**), is avector containing the Q hypothesized values to test against (all zeros here), **S**-1 is the inverse of the covariance matrix of **V**, and H superscript denotes Hermitian (complex conjugate) transpose. The T2 statistic is then transformed into an F statistic using:

*Equation 2*

Which is F-distributed with Q and N-Q degrees of freedom under H0. Note that will be singular for N<=Q. The number of epochs N should therefore be larger than the number of features Q, else T2 cannot be calculated.

*The modified q-sample uniform scores test*

The original q-sample uniform scores test is a non-parametric method for testing whether the phases of W spectral bands share the same underlying distribution (Mardia 1972). The modifications proposed by Stürzebecher et al (1999) and Cebulla et al (2006) consider the amplitudes, in addition to the phases. The modifications are furthermore applied either to the actual values, or to the ranks of the phases and amplitudes, where the ranking is performed across all spectral bands and epochs, i.e. rank values, for either the phases or the amplitudes, range from 1 to N⋅W. The four variations explored in this study include (i) QMod V1, applied to phase ranks and amplitude values, (ii) QMod V2, applied to phase ranks and amplitude ranks, (iii) QMod V3, applied to phase values and amplitude ranks, and (iv) QMod V4, applied to phase values and amplitude values. The spectral bands for the analysis include the 4-, 8-, and 12-Hz bands. The test statistic, say W\*, is then given by (Cebulla et al. 2006):

*Equation 3*

where is either the rank or the actual value of the amplitude of the jth spectral band from the ith epoch, and where is either the rank or the actual value of the phase of the jth spectral band from the ith epoch. The significance of is evaluated with the bootstrap.

*The Fmp*

The Fmp is a ratio between the estimated variance of the evoked response and the estimated variance of the EEG background activity. The variance of the evoked response is estimated with the variance of the coherently averaged epoch, whereas the variance of the EEG background activity is estimated with the mean of multiple ‘single point’ variances. A ‘single point variance’ is found by taking a single value (at a fixed point in time) from each epoch (giving a total of N values), and calculating the variance of the resulting sample. The Fmp is defined as (Martin et al. 1994):

*Equation 4*

where var denotes variance and denotes the ith set of single point values. The significance of the Fmp is evaluated using the bootstrap.

*The correlation coefficient*

The correlation coefficient (CC) gives the linear correlation between two variables (Pearson 1895). It takes values ranging from -1 to 1, with -1 representing perfect negative correlation, 1 representing perfect positive correlation, and 0 representing no correlation at all. The CC calculated between the ensemble coherent average and some template is given by:

*Equation 5*

where is the ith sample of the ensemble coherent average, is the mean of the ensemble coherent average, is the ith sample of the template, is the mean of the template, and L is the number of samples within and The significance of the CC is evaluated using the bootstrap.

*Covariance*

COV is the non-normalised CC, given by the numerator in Equation 5:

*Equation 6*

The significance of COV is evaluated using the bootstrap.

*Dynamic time warping*

DTW is a technique for stretching and compressing two signals in time, such that their alignment is optimised, albeit under some set of constraints. A detailed description, including illustrations and proofs, can be found in (Müller 2007). Here, a summary with application to CAEP detection will be provided. In this study, DTW was used to realign the coherently averaged epoch with some template. To do so, ‘cost matrix’ C is first constructed, with elements given by for (matrix C is L x L-dimensional). There are then many ‘paths’ through C, starting at element and ending at element. Each ‘step’ along a given path requires either the row index or the column index (or both) of C to be increased by no more, and no less, than 1 (the ‘step-wise condition’; Müller, 2007). Any given path then represents a conceivable re-alignment of and, i.e. the row indexes of C (associated with a given path) give the sample indexes of, whereas the column indexes of C give the sample indexes of (or vice versa). For this study, an additional constraint is imposed using a Sakoe-Chiba band (Sakoe and Chiba 1978), which can be used to restrict the realignment of and . In particular, the Sakoe-Chiba band requires the absolute difference between indexes i and j to remain below some upper boundary λ. Ideally, λ should be sufficiently large, such that an optimal re-alignment can be found between the adopted template and all subject CAEP waveforms. However, further increasing λ will result in DTW overfitting the template to the background activity, potentially resulting in a loss of test sensitivity. For this study, pilot simulations (details not presented) suggest that λ should be kept relatively low at just ~5-25 ms. For all analyses that follow, λ=10 ms was used.

After constructing cost matrix C, the ‘cumulative cost matrix’ D can be constructed, with elements defined as:

*Equation 7*

The path for the optimal re-alignment of and is then found by traversing D, now from elementto, and where each step through D is chosen based on the ‘smallest cumulative cost’, i.e. each step through D always takes the element with the smallest possible value. Finally, the actual test statistic for evoked response detection is again given by COV, now calculated between the warped ensemble coherent average and the warped template. This statistic will henceforth be referred to as ‘DTW COV’. The significance of DTW COV is evaluated with the bootstrap.

**The bootstrap**

The bootstrap is a random resampling with replacement procedure (Efron and Tibshirani 1993), previously used by Lv et al (2007) to approximate the null distribution for some EEG feature of interest. The bootstrap proceeds by generating many additional ensembles (1000 ensembles were used for this study) of `bootstrapped epochs’, where each bootstrapped epoch is found by randomly resampling (with replacement) an EEG segment from within the continuous recording. The key concept behind the bootstrap is that time-locking between the stimuli and the resampled epochs is disrupted, i.e. resampling proceeds with no regards to the timing of the stimuli. It is then assumed that, even if the original ensemble of epochs contain a response, this will cancel out in the bootstrapped ensembles. To facilitate the latter, every other randomly resampled epoch is inverted. The EEG feature of interest is then calculated from all bootstrapped ensembles, giving many bootstrapped feature values, which are used to approximate the null distribution. Finally, the approximated null distribution can be used to generate critical decision boundaries (for rejecting H0), or to generate a *p* value. The *p* value is given by the percentile of the approximated null distribution associated with the EEG feature value calculated from the original ensemble of epochs.

**Response detection by audiologists**

CAEP data were analysed by three audiologists with substantial experience in interpreting CAEPs. The main goals for the visual inspection were (i) to obtain a benchmark for evaluating the specificity and sensitivity of the objective detection methods, (ii) to obtain clear CAEP waveforms for emulating a response in the simulations described below, and (iii) for constructing templates for the correlation-based detection methods. Data were presented to the audiologists as two replicates of the coherent average, obtained by averaging across the first and second half of the epochs within the ensemble in question. For the stimulus condition, all available epochs (a minimum of 100) were used to construct the coherent average replicates, whereas for the no-stimulus condition, the ensemble size was fixed at 60 epochs so that more ensembles could be constructed (giving a total of 256 ensembles), which allowed a FPR to be estimated. The coherent average replicates were presented to the audiologists using an in-house MatlabTM interface from -150 ms to +500 ms, relative to stimulus onset, with the y-axis fixed from -15 to +15 uV. Data were furthermore presented randomly, i.e. with no specific subject or dB SL ordering. This emulates the challenge of the automatic detection methods, which decide on response present/absent without knowledge of dB SL. For each pair of coherent average replicates, the examiners were forced to choose between (1) CAEP present, (2) CAEP absent, or (3) ambiguous. As criteria for response detection, the examiners used amplitude and repeatability of the CAEP replicates, but were ultimately left free to decide whether a response was present, absent, or ambiguous.

*CAEP* *waveforms and templates*

The CAEP waveforms for emulating a response in the simulations were given by the ensemble coherent averages, under the condition that these contained a ‘clear response’. The criteria for a ‘clear response’ was that all three audiologists classified the ensemble coherent average as ‘CAEP present’. This resulted in 16, 14, 11, and 8 CAEP waveforms from the 50, 30, 20, and 10 dB SL conditions, respectively. The templates for the correlation-based detection methods were then given by the mean of the CAEP waveforms, per dB condition, after removing the CAEP waveform from the subject in question (if present). The latter was necessary to avoid introducing an unfair bias, i.e. the adopted templates should be constructed independently from the subjects’ true CAEP waveform, as this information is typically unavailable prior to the test.

**Specificity assessment**

The FPRs of the detection methods were first evaluated using no-stimulus CAEP recordings. A more powerful specificity assessment was then obtained using realistic surrogate data, constructed by randomising the phases of the no-stimulus recordings (further described below).

*Subject recorded no-stimulus CAEP* *data*

Each pre-processed recording of no-stimulus CAEP data was structured into ensembles of N 1100 ms artefact-free epochs, where N ranged from 10 to 100 epochs, in steps of 10 epochs. There was sufficient data for constructing 1976, 937, 588, 422, 298, 250, 211, 161, 117, and 80 ensembles for N=10, 20, …, 100, respectively. The aforementioned objective detection methods were then applied to the 50-300 ms windows of the ensembles.

*Simulations*

In order to conduct a more powerful specificity assessment, many no-stimulus surrogate recordings were generated. To do so, continuous segments of EEG were first randomly resampled from within a randomly selected recording of no-stimulus CAEP data. The resampled segments were then structured into N 1100 ms epochs, after which the epochs were transformed to the frequency domain using the Fast Fourier Transform (FFT). In the frequency domain, phase randomisation was applied to all Fourier components up to the Nyquist frequency, i.e. each phase was assigned a new random value from the [-pi, pi] interval. Finally, the phase-randomised epochs were transformed back to the time domain using the inverse FFT, after which they were analysed with the detection methods. A total of 50,000 phase-randomised surrogate recordings were thus generated, per ensemble size N. It is worth noting this phase-randomisation procedure preserves some degree of non-stationarity, but disrupts serial correlation between epochs.

**Sensitivity assessment**

Sensitivity was first evaluated using the previously described subject-recorded CAEP threshold series. A more powerful assessment of sensitivity was then obtained using simulations.

*Subject-recorded CAEP* *data*

Each recording was structured into ensembles of N artefact-free 1100 ms epochs, where N took values of either 20, 40, or 100+ epochs. There were sufficient data for ~ 95 (N=20), ~40 (N=40), and 23 (N=100+) ensembles, per dB SL condition. The 50-300 ms windows of the epochs where analysed with the detection methods.

*Simulations*

Recordings of no-stimulus surrogate data were first constructed using phase randomisation, as described in the specificity assessment above. The ensemble size N took values ranging from 10 to 100 epochs, in steps of 10 epochs. For each ensemble, a CAEP was simulated by adding a re-scaled CAEP waveform to all epochs, under the condition that the CAEP waveform in question contained a ‘clear response’, as described in see section on visual inspection above. The scaling factor for the CAEP waveforms was furthermore chosen such that a specific SNR was obtained, calculated using:

*Equation 8*

where is the mean square of the re-scaled CAEP waveform, and is the mean square of the ensemble of epochs (prior to adding the response) when treated as a continuous recording. The SNRs for the simulated response were given by the estimated SNRs within the subject CAEP data. The latter were similarly calculated using Eq. 8, where was now the mean square of the ensemble coherent average and the mean square of the full ensemble of epochs and when treated as a continuous recording, and after subtracting the ensemble coherent average from each epoch. The estimated SNRs ranged from -15.4 dB to -8.5 dB, with a mean of -12.25 dB. A total of 10,000 recordings were then simulated, per ensemble size N, and the 50-300 ms windows were analysed with the detection methods.

**Results**

This section presents the results from the specificity and sensitivity assessments. To keep this section concise, some results from the modified q-sample statistics have been moved to the supplemental Appendix. In particular, this section presents results from just the best performing q-sample modification, which was QMod V2 (applied to phase ranks and amplitude ranks), although performance of all four q-sample modifications was quite similar (see supplemental Appendix, Figure 1).

**Specificity**

The FPRs (using α = 0.01) from both the subject-recorded no-stimulus data and the simulations are shown in Table 1. For the subject-recorded data, the 99% confidence intervals for the expected 0.01 FPR are also presented, per ensemble size N. The latter were determined using a Binomial distribution, constructed from M Bernoulli trials, where M denotes the number of Bernoulli trials (here the number of tests performed), and where the probability of a single ‘successful’ Bernoulli trial (here a false-positive) was equal to α. Note that confidence intervals could not be constructed for the simulations, as M is unknown, i.e. although 50,000 tests were performed, the total number of independent tests is still unknown. Results show that, for the subject recorded no-stimulus data, a single liberal test performance was observed for ‘DTW COV’ at N=90 epochs, whereas all remaining FPRs fell within the expected 99% CIs. The FPRs from the surrogate data were all very close to the expected FPR of 0.01, which suggests that the FPRs were controlled well for these methods.

|  |
| --- |
| **No-stimulus ALR data** |
| **Ensemble size -->** | **10** | **20** | **30** | **40** | **50** | **60** | **70** | **80** | **90** | **100** |
| **Fmp**  | 0.0086 | 0.0139 | 0.0187 | 0.0142 | 0.0134 | 0.012 | 0.019 | 0.0186 | 0.0256 | 0.0125 |
| **HT2** | 0.0081 | 0.0096 | 0.0136 | 0.0118 | 0.0101 | 0.004 | 0.0047 | 0.0062 | 0 | 0.0125 |
| **QMod V2** | 0.0071 | 0.0096 | 0.0136 | 0.0095 | 0.0201 | 0.012 | 0.0047 | 0.0124 | 0.0256 | 0 |
| **CC** | 0.0131 | 0.0085 | 0.0221 | 0.0095 | 0.0067 | 0.012 | 0.0142 | 0.0248 | 0.0171 | 0.0125 |
| **COV** | 0.0101 | 0.0085 | 0.0221 | 0.0071 | 0.0101 | 0.008 | 0.019 | 0.0124 | 0.0427 | 0.0125 |
| **DTW COV** | 0.0106 | 0.0128 | 0.0221 | 0.0071 | 0.0134 | 0.012 | 0.019 | 0.0062 | 0.0513 | 0.0125 |
| *Upper 99% CI* | 0.0165 | 0.0199 | 0.023 | 0.0257 | 0.0295 | 0.0316 | 0.034 | 0.0385 | 0.0453 | 0.0562 |
| *Lower 99% CI* | 0.005 | 0.0032 | 0.0019 | 0.0008 | 0.0003 | 0.0002 | 0.0002 | 0.0002 | ~0 | ~0 |
| **Simulations (phase randomisation)** |
| **Ensemble size -->** | **10** | **20** | **30** | **40** | **50** | **60** | **70** | **80** | **90** | **100** |
| **Fmp**  | 0.0113 | 0.0103 | 0.0112 | 0.0099 | 0.0112 | 0.0116 | 0.0104 | 0.0112 | 0.0124 | 0.0111 |
| **HT2** | 0.0114 | 0.0096 | 0.0098 | 0.0094 | 0.012 | 0.0104 | 0.0106 | 0.0112 | 0.01 | 0.0096 |
| **QMod V2** | 0.0114 | 0.01 | 0.0111 | 0.0108 | 0.0107 | 0.0102 | 0.0113 | 0.0117 | 0.0107 | 0.0095 |
| **CC** | 0.0113 | 0.012 | 0.0119 | 0.0116 | 0.0111 | 0.0112 | 0.0116 | 0.0144 | 0.0122 | 0.0112 |
| **COV** | 0.0097 | 0.012 | 0.0112 | 0.012 | 0.011 | 0.0113 | 0.0103 | 0.0128 | 0.0092 | 0.0108 |
| **DTW COV** | 0.0099 | 0.0123 | 0.0109 | 0.0116 | 0.0109 | 0.0115 | 0.01 | 0.0113 | 0.0095 | 0.0099 |

**Table 1.**

The FPRs (α = 0.01) of the detection methods for different ensemble sizes N when applied to both no-stimulus ALR data (top) and when applied to surrogate data (bottom). For the no-stimulus ALR data, the upper and lower 99% confidence intervals (CIs) are also shown, per ensemble size N. Significant (*p <* 0.01) deviations from the expected 0.01 FPR are indicated by a grey cell. For the surrogate data, each FPR was generated using 50,000 tests, per ensemble size N.

**Sensitivity**

*Simulations*

Results from the simulations are presented in Figure 1, which shows the detection rates (using α = 0.01) of the methods, as a function of the ensemble size N. Results demonstrate a relatively poor performance for the CC, particularly so for large N, which can likely be attributed to the template not correlating well with the CAEP for some subjects. Results also show that removing the normalisation factor from the CC (giving the COV statistic), greatly improves test sensitivity, and that further benefit can be gained by allowing some degree of flexibility in the choice for template, achieved here with the DTW algorithm. For the non-template-based detection methods, the highest test sensitivity was observed for the modified q-sample statistic applied to phase values and amplitude values (QMod V2), followed by the bootstrapped Fmp, and lastly by the Hotelling’s T2 test.



**Figure 1.**

The detection rates (α = 0.01) from the simulations, per method, as a function of the ensemble size N. Data for the simulations consists of phase randomised recordings of no-stimulus ALR data, along with coherently averaged ALR waveforms for simulating a response.

*Subject-recorded CAEP data*

The detection rates (α = 0.01) of the objective methods when applied to the subject recorded CAEP threshold series are presented in Figure 2 for different ensemble sizes N, per dB SL condition. Overall, results show a similar trend as those seen in the simulations (see Figure 1), i.e. the best performing method was DTW COV, followed by COV and QMod V2. The largest differences in test sensitivities can be seen for smaller ensemble sizes (N=20 and N=40). For the largest ensemble size (N=100+), test sensitivities were similar across methods, which might be attributed to ceiling effects.



**Figure 2.**

The detection rates (α = 0.01) of the objective methods when applied to the subject recorded ALR data. Results are presented per dB SL condition, for different ensemble sizes N.

**Response detection by examiners**

The true-positive rates (TPRs) and FPRs of the examiners are presented in Figure 3. Results first demonstrate a good specificity for all three examiners, particularly so for examiner 3 who obtained a FPR of 0. In terms of test sensitivity, TPRs were similar across examiners, although consensus between examiners was not always reached.

****

**Figure 3.**

The observed TPRs and FPRs for the examiners when using visual inspection for ALR detection. Note that the *y*-axis for the FPR (right) has a different scale as the *y*-axis for the detection rate (left).

**Post-hoc statistical analysis**

The permutation test (Fisher 1935; Efron and Tibshirani 1993) was used to test compare the detection rates from the subject data. In order to reduce the number of comparisons, test sensitivity was considered across dB SL conditions and ensemble sizes, i.e. a single detection rate (using α = 0.01) was constructed, per method. This resulted in the following detection rates: 0.40, 0.30, 0.43, 0.25, 0.53, and 0.54 for the Fmp, HT2, QMod V2, CC, COV, and DTW COV, respectively. The permutation test then proceeds by pooling the *p* values from the two detection methods in question (note that this pooled sample space now contains *p* values from two detection methods, and from all dB SL conditions and all ensemble sizes), and then randomly re-assigning the pooled sample space to two new samples. The two new samples are used to construct two new detection rates, and the difference between these detection rates is taken. This is repeated many times to construct a distribution of ‘detection rate differences’. The distribution can be seen as the expected distribution of detection rate differences under the null hypothesis of no difference (when considered across all dB SL conditions and ensemble sizes), which was used to evaluate the significance of the original detection rate discrepancy. The resulting *p* values from all pair-wise comparisons between methods are presented in Table 2. Results confirm that the many comparisons were significant, with the best performing methods being COV and DTW COV.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Fmp**  | **HT2** | **QMod V2** | **CC** | **COV** | **DTW COV** |
| **Fmp**  |  | <0.001 | 0.167 | <0.001 | <0.001 | <0.001 |
| **HT2** |  |  | <0.001 | 0.0171 | <0.001 | <0.001 |
| **QMod V2** |  |  |  | <0.001 | <0.001 | <0.001 |
| **CC** |  |  |  |  | <0.001 | <0.001 |
| **COV** |  |  |  |  |  | 0.787 |
| **DTW COV** |  |  |  |  |  |  |

**Table 2.**

The *p* values generated by the permutation test when drawing pair-wise comparisons between the detection rates of the methods. Further details are presented in the text.

**5. Discussion**

This paper evaluated the performance of various new and existing objective methods for adult CAEP detection. Results firstly demonstrate a good performance for some template-based methods, i.e. increases in test sensitivity of up to ~30% were observed for DTW COV over the more commonly used Hotelling’s T2 test (primarily for small ensemble sizes of N=20 and N=40). A caveat for DTW COV, and for template-based detection methods in general, is that test sensitivity is dependent on the choice of template, and in particular on how well the template matches with the subject-specific CAEP waveforms. This may limit their application to patient groups with relatively robust and predictable CAEP waveforms, e.g. normal-hearing adults. Indeed, when testing infants or the hearing-impaired, template-based detection methods might be less successful, as it is well known that hearing-impairment can result in delayed CAEP peak and trough latencies (Oates et al. 2002), and that the CAEP undergoes dramatic change during the first two years of life (Ponton and Eggermont 2007). Choosing a suitable template for these patient groups might therefore be challenging.

Additional factors that should be considered when choosing the template include electrode placement, filter settings, and both the SL and spectral content of the stimulus, both of which can affect CAEP peak and trough latencies (e.g. Adler and Adler 1989; Picton et al. 1976). Hence, even when testing normal-hearing adults, care should be taken to use suitable, stimulus-specific templates for the analysis. In this study, the relatively poor performance of the CC suggests that the adopted templates may not have matched well with some subjects’ CAEP (Figures 1 and 2). It is also worth pointing out that, when using the CC, additional data collection cannot always compensate for an inadequate template, as moderate correlations between the EEG background activity and the template remain feasible under H0, even when the residual background activity is very small (due to e.g. large ensemble sizes). The latter can be attributed to the normalisation factor in equation 5 (the denominator), which is absent when using COV. As a result, COV will tend to be sensitive to the choice of template. Robustness to sub-optimal templates can be further increased by using the DTW algorithm, which was found to be beneficial in this work, under the condition that DTW did not receive too much freedom when re-aligning the template. In this study, the freedom of the DTW algorithm was restricted using a Sakoe-Chiba band of 10 ms. The choice for 10 ms was based on simulation results, which suggest that a 0-25 ms Sakoe-Chiba band was a good choice for adult CAEP detection. For subject groups where response variability between individuals is known to be more substantial (e.g. infants and the hearing-impaired), it may be beneficial to use a wider Sakoe-Chiba band. Various additional DTW modifications that might be considered can be found in Müller (2007).

For the non-template-based detection methods, the best performing test was modified q-sample V2 (applied to phase ranks and amplitude ranks), followed by the Fmp, and lastly by the Hotellings T2 test. This was surprising, as the Hotellings T2 test has previously shown a good test sensitivity for both ABR detection (Chesnaye et al. 2018) and envelope frequency following response detection (Vanheusden et al., 2018), and has recently received much attention for CAEP detection (Golding et al. 2009; Carter et al. 2010; Chang et al. 2012; Van Dun et al. 2012; Van Dun et al. 2015). This hence first raises the question about the choice for features (the voltage-means), and if test sensitivity might be improved by optimising the feature set. For this study, the voltage-means were calculated across 50 ms windows, as recommended by Golding et al (2009). Additional simulations (presented in the supplementary Appendix) indeed suggest that these 50 ms intervals were close to optimal (albeit for the current data set), although a small increase in test sensitivity could be gained by using slightly smaller intervals of 40-45 ms (see supplemental Appendix, Figure 2). This increase was, however, insufficient to close the gap in performance between the aforementioned detection methods. A second reason for the underperformance of the Hotelling’s T2 test might be related to uncertainty in the estimation of the feature covariance matrix. In particular, when the sample size is small, uncertainty in the variance and covariance estimations will be large. This is accounted for in Hotelling’s T2 by ‘flattening out’ the theoretical null distributions (given by F-distributions), i.e. they will cover a wider range of values to account for this uncertainty. As a result, the critical thresholds for rejecting H0 are increased, and test sensitivity is reduced.

**Hearing threshold estimation**

For most subjects in this study, the behavioural hearing thresholds were approximated using visual inspection to within 10-30 dB, although for some subjects this was increased to 50 dB. These results therefore fall within the upper range of what is usually observed in the literature, i.e. most studies find that the CAEP can approximate behavioural thresholds to within 10 dB (for references, see Martin et al. 2007; Picton 2011). The relatively large deviations observed in the current study might be due to a large number of blink artefacts, which greatly increased test time for some subjects (many trials were rejected). Consequently, subject alertness may have waned, giving less robust CAEPs. It is also worth noting that the test protocol for visual inspection did not follow a typical clinical protocol whereby the clinician would hone in on the threshold in an efficient manner by stopping recordings in cases where highly repeatable responses were seen, and repeating recordings in cases where responses were inconclusive. Instead, audiologists were presented with a pre-defined number of epochs, had no knowledge of stimulus level, and received no assistance from objective detection methods. For the objective detection methods, the approximations were slightly better than those obtained through visual inspection. When using the most sensitive methods (COV and DTW COV), behavioural hearing thresholds were approximated to within 20 dB for ~90% of the subjects, and to within 30 dB for 100% of the subjects.

**The bootstrap and CAEP detection**

In the past, most objective methods for evoked response detection have been designed around statistics with *a priori* assumed theoretical null distributions, e.g. the Fsp, the Fmp, and correlation coefficients. This can be problematic, as the true null distributions underlying these methods are often recording-dependent, and are affected by auto-correlation (which greatly impacts on the degrees of freedom), non-stationarity, and non-Gaussianity. Hence, in order to prevent higher than expected FPRs, the null distributions are often chosen conservatively (e.g. Elberling and Don, 1984). The disadvantage with a conservative test, of course, is that test sensitivity is reduced.

Besides a reduced test sensitivity, designing the statistical analysis around EEG features with theoretical null distributions means that the distributions should remain mathematically tractable throughout the examination, which greatly limits the choice and design of the statistical detection method. Non-parametric methods such as the bootstrap (Lv et al. 2007) and the permutation test (Maris and Oostenveld 2007), on the other hand, allow the underlying distributions to be approximated, as opposed to assumed, which gives the user a great deal of freedom when designing the test. That said, it was initially hypothesized that the bootstrap would be biased for CAEP detection due to small sample sizes and sampling errors (Efron and Tibshirani 1993). An additional concern was that the resampling with replacement procedure would result in overlap between some of the resampled epochs, thus introducing independence violations to the analysis. Results from the simulations and subject data nevertheless suggest that violations originating from the bootstrap were likely negligible in this study (Table 1). Indeed, even when analysing 50,000 surrogate recordings, deviations from the nominal α-levels were all relatively small. Randomly inverting the resampled epochs might have played a role here, as this essentially expands the number of possible resampled data sets, i.e. it helps to reduce sampling errors. A caveat with randomly inverting resampled epochs is that the true mean of the resampled recordings will be zero. It is therefore important that the original EEG recordings are de-meaned prior to the statistical analysis, else a bias might be introduced.

**Conclusion**

This study evaluated and compared the specificity and sensitivity of various objective methods for adult CAEP detection. Results show that the most sensitive method was the DTW algorithm, evaluated using the bootstrap. An important factor affecting the performance of the DTW algorithm is that the adopted template matches well with the subjects’ CAEP, which may restrict its application to patient groups with relatively robust and predictable CAEPs, e.g. adults with normal hearing. When the subjects’ CAEP is difficult to predict (e.g. for hearing-impaired infants), it may be preferable to use non-template-based detection methods, in which case results from this study suggest that the modified q-sample test may be a good choice. Finally, in the current study, test significance for all detection methods (except for the Hotellings T2 test) was evaluated using the bootstrap approach. Results suggest a good control over the FPR for the bootstrapped test statistics. Besides providing a robust, non-parametric assessment of test significance, the bootstrap also gives the user a large amount of freedom when designing the statistical detection method, which might be exploited in future studies when further improving CAEP detection methods.

**Acknowledgements**

This work was supported by the NIHR RfPB (**Grant number to be added**) and Demant A/S (**Grant number to be added**). The authors would also like to thank Jo Brooks and Sara Al-Hanbali for data collection.

**References**

Adler, G., & Adler, J. (1989). Influence of stimulus intensity on AEP components in the 80-200- millisecond latency range. *Audiology*, 28(6), 316-324.

Assecondi, S., Bianchi, A.M., Hallez, H., et al. (2009). Automated identification of ERP peaks through dynamic time warping: An application to developmental dyslexia. *Clinical Neurophysiology*, 120(10), 1819-1827.

doi: 10.1016/j.clinph.2009.06.023

Carter, L., Golding, M., Dillon, H., et al. (2010). The detection of infant cortical auditory evoked potentials (CAEPs) using statistical and visual detection techniques. *J. Am. Acad. Audiol*, 21(5), 347-356.

doi:[10.3766/jaaa.21.5.6](https://doi.org/10.3766/jaaa.21.5.6%22%20%5Ct%20%22_blank)

Casarotto, S., Bianchi, A., Cerutti, S., et al. (2005). Dynamic time-warping in the analysis of event-related potentials. *EMB Mag,* 24(1), 68-77.

doi: [10.1109/MEMB.2005.1384103](https://doi.org/10.1109/MEMB.2005.1384103%22%20%5Ct%20%22_blank)

Cebulla, M., St**ü**rzebecher, E., & Elberling, C. (2006). Objective detection of auditory steady state responses: Comparison of one-sample and q-sample tests. J. Am. Acad. Audiol., 17(2), 93-103.

Chang, H.W., Dillon, H., Carter, L., et al. (2012). The relationship between cortical auditory evoked potential (CAEP) detection and estimated audibility in infants with sensorineural hearing loss*. International Journal of Audiology*, 51(9), 663-670.

doi:[10.3109/14992027.2012.690076](https://doi.org/10.3109/14992027.2012.690076%22%20%5Ct%20%22_blank)

Chesnaye, M.A., Bell, S.L., Harte, J.M., et al. (2018). Objective measures for the auditory brainstem response: Comparisons in specificity, sensitivity and detection time. *[Int J Audiol.](https://www.ncbi.nlm.nih.gov/pubmed/29537327%22%20%5Co%20%22International%20journal%20of%20audiology.)* 57(6), 468-478.

doi:10.1080/14992027.2018.1447697

Duncan, C.C., Barry, R.J., Connolly, J.F., et al. (2009). Event-related potentials in clinical research: Guidelines for eliciting, recording, and quantifying mismatch negativity, P300, and N400. *Clinical Neurophysiolog*, 120(11), 1883-1908.

doi:[10.1016/j.clinph.2009.07.045](https://doi.org/10.1016/j.clinph.2009.07.045%22%20%5Ct%20%22_blank)

Efron, B., & Tibshirani, R.J. (1993). An Introduction to the bootstrap. Chapman & Hall/CRC.

Elberling, C., & Don, M. (1984). Quality Estimation of Averaged Auditory Brainstem Responses. *Scandinavian Audiology,* 13(3), 187–197.

doi:10.3109/01050398409043059.

Fisher, R.A. (1935). The design of experiments. Hafner, New York.

Goffredo, M., Nardi, C. Di., Giannantonio, A.S., et al. (2016). A DTW-based method for detecting electrically evoked auditory brainstem responses. GNB, June 20th-22nd, Naples, Italy

Golding, M., Dillon, H., Seymour, J., et al. (2009). The detection of adult cortical auditory evoked potentials (CAEPs) using an automated statistic and visual detection. *Int. J. Audiol*, 48 (12), 833-842.

doi:10.3109/14992020903140928.

Gupta, L., Molfese, D.L., Tammana, R., et al. (1996). Nonlinear alignment and averaging for estimating the evoked potential. *IEEE Transactions on Biomedical Engineering*, 43(4), 348-356.

doi:[10.1109/10.486255](https://doi.org/10.1109/10.486255%22%20%5Ct%20%22_blank)

Hotelling, H. (1931). The Generalization of student's ratio. *Ann. Math. Statist*., 2(3), 360-378.

Hoth, S. (1993). Computer-aided hearing threshold determination from cortical auditory evoked potentials. *Scandinavian Audiology*, 22(3), 165-177.

doi: 10.3109/01050399309047463

Huang, H.C., & Jansen, B.H. (1985). EEG waveform analysis by means of dynamic time-warping. *[Int J Biomed Comput](https://www.ncbi.nlm.nih.gov/pubmed/4055119%22%20%5Co%20%22International%20journal%20of%20bio-medical%20computing.)*[.](https://www.ncbi.nlm.nih.gov/pubmed/4055119%22%20%5Co%20%22International%20journal%20of%20bio-medical%20computing.) 17(2), 135-144.

doi: 10.1016/0020-7101(85)90084-4

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

doi:[10.1097/01.aud.0000233902.53432.48](https://doi.org/10.1097/01.aud.0000233902.53432.48%22%20%5Ct%20%22_blank)

Lv, J., Simpson, D.M., & Bell, S.L. (2007). Objective detection of evoked potentials using a bootstrap technique. *Medical Engineering & Physics*, 29(2), 191-198.

doi: [10.1016/j.medengphy.2006.03.001](https://doi.org/10.1016/j.medengphy.2006.03.001%22%20%5Ct%20%22_blank)

Mardia, K.V. (1972). Statistics of directional data. Academic Press, London and New York.

Maris, E. & Oostenveld, R. (2007). Nonparametric statistical testing of EEG- and MEG-data. *Journal of Neuroscience Methods*, 164(1), 177-190.

doi:[10.1016/j.jneumeth.2007.03.024](https://doi.org/10.1016/j.jneumeth.2007.03.024%22%20%5Ct%20%22_blank)

Martin, W.H., Schwegler, J.W., Gleeson, A.L, et al. (1994). New techniques of hearing assessment. *Otolaryngol. Clin. North Am*., 27(3), 487-510.

Mason, S.M., Su, A.P., & Hayes, R.A. (1977). Simple online detector of auditory evoked cortical potentials. Med Biol Eng Comput, 15(6), 641–647.

doi: [10.1007/BF02457922](http://doi.cnki.net/doi/Resolution/Handler?doi=%2010.1007/BF02457922" \t "_blank)

Martin, B.A., Tremblay, K.L., & Stapells, D.R. (2007). Auditory evoked potentials. Basic principles and clinical application, Chapter 23. Wolters Kluwer, Lippincott, Williams & Williams.

Müller, M. (2007). Information retrieval for music and motion, Chapter 4. Springer-Verlag Berlin Heidelberg.

Oates, P.A., Kurtzberg, D., & Stapells, D.R. (2002). Effects of sensorineural hearing loss on cortical event-related potential and behavioural measures of speech-sound processing. Ear Hear, 23(5), 399-415

doi:[10.1097/01.AUD.0000034777.12562.31](https://doi.org/10.1097/01.AUD.0000034777.12562.31%22%20%5Ct%20%22_blank)

Pearson, K. (1895). Notes on regression and inheritance in the case of two parents. Proceedings of the Royal Society of London, 58([347-352](https://royalsocietypublishing.org/toc/rspl/58/347-352)), pp. 240-242.

doi: [https://doi.org/10.1098/rspl.1895.0041](https://doi.org/10.1098/rspl.1895.0041%22%20%5Co%20%22doi%20link)

Picton, T.W. (2011). Human auditory evoked potentials. Chapters 11 & 12. Plural Publishing, San Diego, Oxford, Brisbane.

Picton, T.W., Hunt, M., Mowrey, R., et al. (1988). Evaluation of brain-stem auditory evoked potentials using dynamic time warping. *Electroencephalography and clinical Neurophysiology*, 71(3), 212-225.

doi:10.1016/0168-5597(88)90006-8

Picton, T.W., Woods, D.L., Baribeau-Braun, J., et al. (1976). Evoked potential audiometry. *J. Otolaryngol*, 6(2), 90-117.

Ponton, C.W., & Eggermont, J.J. (2007). Auditory evoked potentials. Basic principles and clinical application, Chapter 18. Wolters Kluwer, Lippincott, Williams & Williams.

Rencher, A.C. (2001). Methods of multivariate analysis. 2nd Ed., John Wiley & Sons.

Sakoe, H., & Chiba, S. (1978). Dynamic programming algorithm optimization for spoken word recognition. IEEE Trans Acoust Speech Signal Process; ASSP, 26(1), 43-49.

doi:10.1109/TASSP.1978.1163055

Stone, M.A., Visram, A., Harte, J.M. & Munro K.J. (in revision 2019). A set of time-and-frequency-localised short-duration speech-like stimuli for assessing hearing-aid performance via Cortical Auditory-Evoked Potentials. Trends Hear

Stürzebecher, E., Cebulla, M., & Wernecke, K.D. (1999). Objective response detection in the frequency domain: Comparison of several q-sample tests. Audiology & neurootology, 4(1), 2-11.

doi:[10.1159/000013815](https://doi.org/10.1159/000013815%22%20%5Ct%20%22_blank)

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. Audiology Research, 2(1), e13.

doi:[10.4081/audiores.2012.e13](https://doi.org/10.4081/audiores.2012.e13%22%20%5Ct%20%22_blank)

Van Dun, B., Dillon, H., & Seeto, M. (2015). Estimating hearing thresholds in hearing-impaired adults through objective detection of cortical auditory evoked potentials. Journal of the American Academy of Audiology, 26(4), 370-383.

doi:10.3766/jaaa.26.4.5

Wang, K., Begleiter, H., & Porjesz, B. (2001). Warp-averaging event-related potentials. Clinical Neurophysiology, 112(10), 1917-1924.

[doi.org/10.1016/S1388-2457(01)00640-X](https://doi.org/10.1016/S1388-2457%2801%2900640-X%22%20%5Ct%20%22_blank%22%20%5Co%20%22Persistent%20link%20using%20digital%20object%20identifier)

**Appendix**

Simulation results from all four modified q-sample statistics are presented in Figure A.1. The highest test sensitivity was observed for QMod V2 (applied to phase ranks and amplitude ranks), followed by QMod V1 (phase ranks, amplitude values), QMod V4 (phase values and amplitude values), and lastly by QMod V3 (phase values and amplitude ranks), although performance for all four modifications was quite similar. Additional simulation results from the optimisation procedure for the Hotelling’s T2 test are shown in Figure A.2. These results show the detection rates, as a function of the number of voltage-means used for the analysis for ensemble sizes N=20, 40, 60, and 100. The highest detection rates were observed when using 5, 6, or 7 voltage-means, i.e. voltage-means were calculated across 50, ~42, and ~36 ms segments, respectively. Data for these simulations were the same as that described in the main text (see “Sensitivity assessment” in methodssection)



**Figure A.1**

The detection rates from all four modified q-sample statistics as a function of the ensemble size N when detecting simulated CAEPs in no-stimulus CAEP surrogate data.



**Figure A.2**

Detection rates for the Hotelling’s T2 test when detecting simulated CAEPs in no-stimulus CAEP surrogate data. The detection rates are plotted as a function of the number of voltage-means, for ensemble sizes N=20, 40, 60, and 100 epochs.