Automated cortical auditory response detection strategy

Fabrice Bardy PhD [a,b]

Bram Van Dun PhD [a,c]

Mark Seeto [a,c]

Harvey Dillon PhD[a,d,e]

[a] HEARing Co-operative Research Centre, Australia

[b] University of Auckland, New Zealand

[c] National Acoustic Laboratories, NSW, Australia

[d] Macquarie University, NSW, Australia

[e] University of Manchester, United Kingdom

Contact: Fabrice Bardy

fabrice.bardy@auckland.ac.nz

M&HS BUILDING 507 - Bldg 507

28 PARK AVE

GRAFTON

Auckland 1023, New Zealand

+64 29 123 06 63

Abstract

Objective

This study describes a new automated strategy to determine the detection status of an electrophysiological response.

Design

Response, noise and signal-to-noise ratio of the cortical auditory evoked potential (CAEP) were characterized. Detection rules were defined: when to start testing, when to conduct subsequent statistical tests using residual noise as an objective criterion, and when to stop testing.

Study sample

Simulations were run to determine optimal parameters on a large combined CAEP data set collected in 45 normal-hearing adults and 17 adults with hearing loss.

Results

The proposed strategy to detect CAEPs is fully automated. The first statistical test is conducted when the residual noise level is equal to or smaller than 5.1 µV. The succeeding Hotelling's T^2 statistical tests are conducted using pre-defined residual noise levels criteria ranging from 5.1 to 1.2 µV. A rule was introduced allowing to stop testing before the maximum number of recorded epochs is reached, depending on a minimum p-value criterion.

Conclusion

The proposed framework can be applied to systems which involves detection of electrophysiological responses in biological systems containing background noise. The proposed detection algorithm which optimize sensitivity, specificity, and recording time has the potential to be in clinical setting.

Keywords

- Electrophysiology
- Objective response detection
- Automated algorithm
- Cortical auditory evoked potentials
- Residual noise level criteria

Introduction

A relevant question when recording electrophysiological responses of any kind is to know whether a response is present, absent or if the recording is inconclusive. This can be evaluated by a human tester or by an automated algorithm.

Automated algorithms have the potential to be more efficient than human testers by avoiding tester bias, controlling the false positive rate, and reducing recording time by using sophisticated detection methods. Objective evoked potential (EP) detection can be achieved using different statistical techniques, e.g. cross-correlation, detection theory or parametric approaches like the Hotelling's T^2 (Golding et al, 2009; Valdes Sosa et al, 1987; Hyde et al, 1998). Technique selection is aimed at correctly identifying physiologically present responses (i.e., a high sensitivity) and correctly rejecting non-physiologically present responses (i.e, a high specificity or a low false-positive rate). Given different recordings have different response and noise characteristics, the number of required response averages is recording dependent. The likelihood of detecting a response increases with increasing response size and decreases with increasing background activity of the recorded signal (i.e. electrical noise). Multiple responses need to be acquired in order to lower residual noise (RN; the noise in the averaged response) and reach a signal-to-noise ratio with an acceptable likelihood of response detection (British Society of Audiology, 2016).

When evaluating response presence, there is a need to define time intervals between statistical tests and criteria to stop testing in the case of response absence. To address the question of when to statistically evaluate response detection, classic approaches include testing at regular time intervals or performing one single test at the end of a recording with predefined length. These approaches have some drawbacks. When applying a statistical test only once after the recording is finished, one encounters the risk that if the chosen recording length is on the short side, small responses will not be detected (resulting in a lower than optimal sensitivity). If the predefined recording length is long however, recording times will be unnecessarily extended when response amplitudes are large.

Conversely, testing at regular intervals and repeating this until a response is detected, might be a better option. In this strategy, potentially time can be saved as a response might be detected after only a few statistical tests. On the other hand, because the residual noise in the averaged response is assumed to proportionally decrease with recording length (Elberling & Don, 1984), a fixed interval between two statistical tests will result in smaller and smaller decreases in residual noise with each succeeding test. As a result, statistical tests are increasingly likely to be conducted without there being significant improvements in the signal-to-noise ratio. There is therefore only a small chance that the response detection is more likely than in the previous statistical test. An additional issue with this strategy is the large number of tests. Multiple testing increases the probability of false rejection of the null hypothesis (Type II error or a low specificity). Therefore, the larger the number of tests, the stricter the correction of the p-value needs to be to keep the false positive rate (FPR) at 5%.

To address these drawbacks, another approach will be presented in this paper which takes the residual noise values into consideration. It will be shown that this approach allows a balance between test interval lengths and the number of statistical tests. This method is adaptive in the sense that the interval between two statistical tests will vary depending to the noisiness of the tested subject. The method allows also to control for the number of statistical tests. Appropriate criteria can be derived through simulations on a large sample of real-life data sets (Stürzebecher et al, 2005). Finally, deciding when a response is absent is critical, creating the need for an appropriate stopping criterion. This criterion is generally determined by a maximum number of epochs or by a sufficiently low residual noise to allow a likely detection of a predefined response amplitude. A comprehensive overview with guidelines and suggestions for CAEP testing, detection and absence criteria as used by the British Society of Audiology, a leading body on CAEP testing, can be found in the British Society of Audiology Recommended Procedure for CAEP testing (2016). They highlight the relevance of the residual noise level required for response presence and absence.

In this article, we will be describing a fully automated algorithm for response detection and its optimization procedure. The aim is to have an objective detection of electrophysiological responses with the highest sensitivity and a controlled specificity in the shortest possible recording time. Although the proposed techniques in this paper can be applied to any type of electrophysiological response, the idea is conceptualized through real-life data sets involving the recording of cortical auditory evoked potentials (CAEPs) for objective hearing threshold estimation in adult subjects. Given some people cannot provide reliable behavioral feedback due to medical reasons (dementia and stroke), age (babies and young children), or do not want to because of medico-legal situations (workers compensation), objective measures offer an alternative to behavioral methods. One of these objective measures are CAEPs, which are electric responses from the auditory cortex which have been shown to be a good measure to estimate hearing thresholds in adults (Perl et al, 1953; Beagley & Kellogg, 1969; Pratt & Sohmer, 1978; Coles & Mason, 1984; Ross et al, 1999; Lightfoot & Kennedy, 2006). Therefore, in more specific terms, we will be describing a fully automated hearing threshold estimation algorithm. In the first part of the paper, we characterize the response of interest. In the second part, we define the rules of the algorithm and its optimization procedure using simulations with several real-life data sets.

Materials and Methods

A large combined CAEP data set recorded on adult subjects was used for the simulations. Optimal parameters for response detection of CAEPs in adults are derived using the proposed strategy.

Subjects and stimuli

The data used for the simulation were collected during four studies conducted at the National Acoustic Laboratories. Overall, CAEPs were recorded in 45 normal-hearing adults and 17 adults with hearing loss in response to short pure- and multi-tone auditory stimuli (either 50 or 70 ms) presented monaurally via insert earphones (Etymotic Research ER-3A). Sensation levels of the stimuli were 10, 20 and 40 dB SL and the stimulus onset asynchrony (SOA) was randomized uniformly between 1000 and 3000 ms. All stimuli were acoustically calibrated at 70 dB HL according to the ISO standard 389-2 (ISO 1994) in an HA-2 2-cc coupler, incorporating.a 1-inch 4144 microphone, a 1-to-1/2 inch DB0375 adaptor, and a 4230 sound level meter (all Brüel & Kjær). Non-response epochs were collected using portion of EEG signal selected randomly between 1 and 1.3 s after any stimulus onset in case the SOA was higher than 2s. In total 66442 non-response epochs were collected in the 62 subjects. The false detection rate for the non-response data was 4.9% when conducting a single Hotelling's T2 with a p-value criterion of 0.05. This confirmed that the non-response data have similar characteristics of true non-response data. Table 1 summarizes the relevant details of the four studies. All subjects were in good general health and reported normal neurological status.

CAEP recordings

The EEG recording equipment used in the four studies was a Neuroscan Synamps2 version 4.3 (Compumedics, Charlotte, NC, USA). The EEG was obtained from 3 gold-plated electrodes placed at Cz (active), the mastoid contralateral to the ear of stimulation (reference) and the forehead as the common (ground) channel. Electrode impedance was checked before and after each recording, and kept under 5 kOhms between active and ground, and between reference and ground. During testing, subjects were seated comfortably in a dimmed, sound attenuated booth. Subjects watched a muted close-captioned DVD of their choice and were instructed to ignore the stimulus being presented in their ear.

All EEG channels were amplified by a factor of 1210, sampled at 1 kHz, and band-pass filtered online between 0.1 and 30 Hz. The recording window consisted of a 300 ms pre- and 600 ms post-stimulus interval (900 ms per epoch). Baseline correction was applied to each individual sweep based on the average over 100 ms prior to stimulus onset. Epochs exceeding ±75 µV were excluded. Matlab (MathWorks) and the EEGLAB toolbox (Delorme & Makeig, 2004) were used to process the EEG files.

Results and simulations:

PART 1: Characterization of the electrophysiological response, noise and SNR at detection

In order to determine optimal response detection parameters, the response and noise properties of the signal of interest need to be characterised first. In the case of CAEP, an estimate of the CAEP amplitude as well as the residual noise (RN) amplitude need to be calculated. This allows determination of the signal-to-noise ratio (SNR) at detection.

Residual noise (RN) amplitude

The rms amplitude of the RN is estimated based on the epoch-to-epoch variation at each and every point within the epoch (in a region of interest from 51 to 347 ms after onset). That is, at each point in the epoch, the variance across epochs is calculated. These values are averaged across all such points in the epoch, and the square root of that average is taken (Elberling & Don, 1984). This estimation assumes EEG stationarity. Although this assumption is not completely valid, the accuracy of the RN prediction is sufficient in a practical sense if noise variance is not changing considerably between epochs. Figure 1a shows the median RN rms amplitudes and standard deviations (SDs) across participants after averaging a specific number of epochs for a group of normal-hearing adults (Bardy et al, 2015a). A logarithmic decrease of RN inversely proportional to the square root of the number of epochs can be observed. The mean rms amplitude per epoch, calculated as the RN amplitude after n epochs (i.e. 70 in this case) multiplied by the square root of number of epochs, was 12.5 µV (SD 2.65 µV).

CAEP amplitude estimation

To obtain an estimate of the CAEP amplitude, a correction is required by accounting for the RN. This correction can be applied under the assumption of independence between RN and the true CAEP (Elberling & Don, 1984). An automated estimate of the CAEP amplitude can be calculated by first subtracting the RN power from the CAEP power which is calculated as the average waveform power in a region of interest time interval (i.e. from 51 to 347 ms after onset). Then, the CAEP amplitude estimate is calculated as the square root of this subtraction.

Figure 1b shows the CAEP amplitude distributions at 3 sensation levels: 10, 20 and 40 dB SL for normal-hearing adults (Bardy et al, 2015a). When combining the distributions obtained at 10, 20 and 40 dB SL, only 15.5% of CAEP peak amplitudes were larger than 5.1 µV.

Knowing now both the CAEP amplitude and RN rms amplitude distributions, the only measure which still needs to be characterised is the signal-to-noise ratio (SNR) required for a response likely to be detected. If this specific SNR is known, it is then possible to estimate the maximally allowable RN rms amplitude at which the first statistical test should occur, still guaranteeing a high likelihood to detect a CAEP.

[Insert Fig. 1 here]

Response detection using Hotelling's T^2 & control of the false positive rate

One objective measure for detection of CAEP waveforms is the Hotelling's T^2 statistic, which has been validated in both adults (Golding et al, 2009) and infants (Carter et al, 2010), and which has been shown to be at least as accurate as human examiners. Several steps were taken before the Hotelling's T^2 was applied. First, each epoch was divided into 9 bins, with each bin covering a predefined latency range. The 9 bins covered the range from 51 to 347 ms, with each bin being 33 ms wide. The bin width and number of bins were chosen based on earlier data (Golding et al, 2009). Second, EEG samples in these bins were averaged. Hence, each epoch was reduced to a 9-dimensional binned epoch, and the recorded waveform to a N-by-9 matrix with N the number of collected epochs. Finally, for response detection, a p-value was obtained from a one-sample Hotelling's T^2 test on the N-by-9 matrix, which tests the null hypothesis that the true mean vector equals the zero vector (i.e., whether the true cortical response in every bin is equal to zero).

To guarantee a 5% FPR when multiple statistical tests were conducted sequentially, non-response data were used in simulations to derive a statistical p-value criterion for the Hotelling's T^2 statistic. When only one statistical test is conducted at the end of a recording, it can be shown for real data using simulations that a p-value of 0.05 corresponds to a FPR of approximately 5%. This is assuming that the epochs are independent observations from the same multivariate normal distribution. However, as different – and more complex – algorithms are employed here, simulations need to be conducted to control the FPR as it is difficult to mathematically derive which p-value needs to be applied.

Signal-to-noise ratio at detection

Signal detection depends primarily on the characteristics of signal and noise, both reflected in the SNR measure. In this section, the SNR is defined first. Then, the SNR at which a response is likely to be detected using the Hotelling's T^2 statistic is investigated. These characteristics will allow the derivation of criteria guiding when to conduct statistical tests.

The signal-to-noise ratio (in dB) of the CAEP at detection is defined as:

$$\mathrm{SNR\ (dB)} = 20 \log 10 \frac{\mathrm{CAEP\ amplitude}}{\mathrm{RN\ amplitude}}, \text{ with} \qquad \text{(Eq. 1)}$$

- SNR: Signal to noise ratio of the CAEP amplitude and the RN amplitudes;
- CAEP amplitude, as defined in section "CAEP amplitude estimation "; and
- RN amplitude, based on the epoch-to-epoch standard deviation (see section Residual noise (RN) amplitude).

We determined the median SNR needed for a CAEP to be detected using data reported in Bardy et al (2015a). Using a sequential test strategy, the p-value was calculated after the collection of nine epochs and subsequently, every additional two epochs. For response detection, a correction for multiple testing of the p-value (to 0.006) was derived using non-response data to keep the FPR at 5%. For every test condition in each subject, the SNR at CAEP detection was collected. Figure 2 shows the distribution of SNRs at which a significant CAEP could just be detected. As can be derived from Figure 2, 50% of CAEPs needed a SNR of 3.3 dB or greater to be detected and 18% of detections occurred at negative SNRs. While the relationship between SNR and Hotelling's T^2 is highly correlated there is variation which depends on the shape of the response and characteristics of the noise. In fact, for a particular shape of response, and a particular distribution of noise rms values (calculated across epochs) along the epoch, if the noise rms value was lower at every point along the epoch by the same proportion, then both Hotelling's T^2 and SNR would increase. Thus for this type of variation, there would be a perfectly monotonic relationship between the two. However, Hotelling's T^2 gives the greatest weight to the time windows that have the best combination of signal amplitude to noise rms. SNR however, weights all time points the same; only the total signal rms and the total noise rms matter. So, it's possible that a change in the signal, or the noise has a different effect on Hotelling's T^2 than it does on the SNR measure. For example an increase of the noise rms in a time window where the signal is zero will have virtually no effect on Hotelling's T^2, but will decrease SNR. A noise Increase in a time window where the signal is at a maximum will have cause a large decrease in Hotelling's T^2, but only a small decrease in SNR. So, it's easy to see that although detection with Hotelling's T^2 generally gets easier as SNR improves (over a wide range of possible SNRs), there will be variations from this relationship. Consequently, it will sometimes be possible to detect signals below some criterion SNR (such as 0 dB) while sometimes not being able to detect them for SNRs above this criterion SNR.

[Insert Fig. 2 here]

PART 2: Detection rules within a single stimulus condition

The number of statistical tests conducted during the recording of a single stimulus condition needs to be limited in order to avoid either an unnecessary increase of the FPR or an excessive decrease of the p criterion used for each test. The purpose of this section is to define the strategy and rules to be used in real-time during data collection for response detection. The rules described are based on the characteristics of the response amplitude, the RN and SNR at detection described in Part 1. The aim is: 1) to determine the criteria to start statistical testing, 2) to define when to perform successive statistical tests, and 3) to define when to stop collecting data. Finally, the validity of the method is demonstrated using real data through simulations.

When to start statistical testing for response detection?

First, the minimum number of epochs to conduct the first statistical test needs to be larger than the number of bins to calculate the Hotelling's T^2. Second, the RN level for the first statistical test is data driven and depends on the RN at which there already is a reasonable chance to detect a CAEP. From the data displayed in Figure 1b, it was calculated that 86% of true CAEP peak amplitudes of the response detected tend to be smaller than 5.1 µV. Hence, testing at RN higher than 5.1 µV would only allow appropriate response detection conditions for a minority of CAEPs. This results in the waste of one (or several) statistical tests at the early stages of the recording. Considering these data, one criterion for the first statistical test to be conducted is a RN amplitude below 5.1 µV.

When to conduct statistical tests later on? Residual noise as an objective criterion.

When the first statistical test has been conducted, the question is when to conduct the remaining statistical tests. There are two approaches which are commonly used when performing statistical tests on a recording consisting of multiple epochs:

1) Apply a statistical test only once, after the collection of a fixed number of epochs; and
2) Multiple tests at fixed intervals (equidistant epochs).

We propose a novel approach that relies on multiple tests at predefined RN amplitudes. This approach guarantees that the SNR improves by a predefined ratio from the time when the previous statistical test was conducted (assuming the CAEP is constant in amplitude), allowing an increased chance of detection. Moreover, it implicitly adapts to the noise condition within each recording (which is highly different depending on the population tested). For example, in cases of increased noise during the recording, statistical testing will be automatically postponed until it reaches the predefined RN criterion. In addition, the number of statistical tests and the space between two statistical tests can be controlled. Figure 3a represents the strategy that has been derived based on the following constraints:

The number of statistical tests and their spacing is a trade-off between test duration, detection sensitivity and clinical applicability. Clinical applicability can be defined as having short test durations tolerable to the patient and a sufficient number of statistical updates for the clinician. A low number of statistical tests results in higher (less strict) p-value criteria (typically with p = 0.05 in the extreme case when there is only 1 statistical test at the end of the recording). However, because of the low number of statistical tests, test duration will be longer as fewer opportunities are available to end a recording early (with the maximum test duration achieved in the extreme case of only one statistical test at the end of the recording). Conversely, a higher number of statistical tests results in lower (stricter) p-value thresholds. Test durations will likely be shorter as more opportunities arise to stop testing earlier. However, the very strict p-values that are the consequence of a large number of tests may also delay detection, or in some cases prevent it from occurring. To summarize, when test duration and detection sensitivity are traded off, an optimal number of statistical tests can be derived, providing clinical applicability is still acceptable.

The statistical tests that are executed need to be distributed over a range. A balance needs to be found between: (1) keeping the spacing between two neighboring tests as small as possible to allow early response detection, and (2) keeping the spacing as large as possible such that a test would be worthwhile, given the adverse effects that additional tests have on either the FDR or the p-value criterion, or both. The RN needs to have dropped by a significant fraction since the previous test making it more likely that a response is detected were there to be one present. The following strategy achieves this balance.

Figure 3 shows the proposed spacing of tests in three inter-related ways. Panel (a) shows the residual noise values at which each successive test is carried out. The first test is carried out when RN equals 5.1 µV. Succeeding tests occur as RN decreases to the values determined by the exponential function shown. This curve applies irrespective of the actual noisiness of any individual person being tested. For a typical person with an rms noise level (per epoch) of 12.5 µV, panel (b) shows the total number of epochs that will have elapsed when each test has been carried out, and panel (c) shows the corresponding number of epochs between immediately adjacent tests. As expected, the number of epochs between adjacent tests increases with increasing test number. Note, however, that the number of epochs between adjacent tests never exceeds 40, which was one of the design goals that helped determine the function shown in panel (a).

Figure 3a shows the RN amplitudes at which to conduct the sequential statistical tests. It is independent from the subject being tested: some subjects might be less noisy, therefore needing a lower number of epochs to reach the RN criteria.

Figure 3b presents the number of epochs needed to reach the RN amplitudes shown in Figure 3a, when the RN per epoch is equal to 12.5 µV, which corresponds to the mean RN per epoch in normal hearing adults. The exponential curves indicate that increasingly more epochs need to be collected after each sequential statistical test to reach the next RN criterion.

Figure 3c shows the differential number of epochs between 2 tests before reaching the next RN criterion and is based on Figure b. It is clear that with increasing statistical test index, the spacing distance (expressed in number of epochs) between tests increases as well. To keep this spacing under control, it was opted to limit the distance to about 40 epochs for practical reasons. This constraint in turn controlled the maximum slope of the RN graph and the minimum number of statistical tests in Figure 3a.

To recapitulate, the exponential function displayed in Figure 3a dictates when to perform each statistical test. It is determined by its slope, the number of statistical tests and the minimum RN amplitude that needs to be reached before the first test can be conducted.

[Insert Fig. 3 here]

When to stop averaging?

Stopping criteria are determined by CAEP amplitude distributions at various stimulus levels, RN estimates, and the objective detection algorithm with FPR. Five stopping criteria can be identified:

1. ***When a response is detected*** - using an objective detection technique like e.g. the Hotelling's T^2 with a predefined detection criterion that has been determined a priori using non-response data.
2. ***When the maximum number of epochs* in a recording has been acquired.** The maximum number of epochs is a trade-off between the maximum acceptable test duration and the required sensitivity for response detection at low sensation levels. Sensitivity in turn depends on the RN level of the averaged waveform when the maximum number of epochs has been collected.
3. ***When the subject appears too noisy*** to continue the process. This can be identified at any stage in the recording.
4. ***When it is clear the required objective criterion will not be reached.*** If the p-value is still above a specific value after a certain number of epochs, the recording can be stopped immediately. Simulations will be conducted to determine these p-values.
5. **When reaching a predefined minimum RN criterion and no response has been detected.** We advise against the use of this criterion as it was noticed during live recordings that this approach can lead to inappropriate results in some people with genuinely low RN (and CAEP) amplitudes. A statistical failure to detect a response may be due to insufficient averaging to achieve the required SNR, rather than the absence of a response.

The concept of adaptive stopping criteria has been introduced in previous research (Kelley et al, 2018; Botella et al, 2006). Moreover, diagnostic evoked potential devices, including most commercially available auditory brainstem response (ABR) and auditory steady-state response (ASSR) instruments, utilise automatic stopping rules. The theoretical literature describes several approaches to stop testing earlier than anticipated. For example, in sequential estimation the sample size to use is not specified at the start, and instead outcomes are employed to evaluate a predefined stopping rule if sampling should continue or stop (Kelley et al, 2018; Botella et al, 2006) . In addition, Bayesian statistics have been used to allow premature stopping of behavioural experiments or clinical trials while keeping the false positive rate constant (Psioda et al, 2018; Komaki & Biswas, 2018; Alcalá-Quintana & García-PÉREZ, 2005)

PART 3: Simulations and validation

According to the rules defined in Part 2, the first simulations were conducted to adjust the statistical detection criterion (p-value) for multiple testing. Second, parameters were defined for p-values that do not reach a certain minimum value after a number of collected epochs (allowing early stopping).

<u>Simulation 1: p-value correction for multiple testing to control the FPR</u>

The general aim of Simulation 1 is to find the p-value which keeps the FPR at 5%. A strict and conservative estimate can be derived using the Bonferroni-correction, which divides the p-value by the number of statistical tests. A better approximation can be obtained through Monte Carlo simulations on EEG data containing no CAEPs, calculating the FPR for a range of p-values while adhering to the rules defined in Part 2. Simulations were conducted using EEG data collected during 4 different studies described in the Materials and Methods section.

[Insert Table 1 here]

A total number of 66442 epochs formed 552 simulated recordings of 120 epochs each. The following procedure was followed for the Monte Carlo simulation to calculate the p-value criterion needed to achieve a FPR of 0.05.

- **for** p varied from 0.0001 **to** 0.05 **in steps of** 0.0001
  - **for** each simulated recording out of 552
    - **for** each epoch ranging from 20 to 120
      - **add** the epoch to the grand average
      - **if** a predetermined RN amplitude is reached

        - conduct a statistical test, providing a p-value P
        - **if** P <= p, stop and FP = FP + 1
    - **calculate** FPR = FP / 552.
    - **if** FPR < 0.05, stop

The p-value criterion adopted was the highest p-value tested with FPR < 0.05. Each simulated recording allowed between 3-9 statistical tests (mean = 6.2) (depending on the RN amplitudes that have been reached), while satisfying an FPR of 5%. For different maximum numbers of epochs, the following p-value criteria were determined (in brackets): 120 (0.0077), 110 (0.0119), 100 (0.0129).

<u>Rule to stop testing if the p-value is still above a certain value after n epochs</u>

When we perform a sequence of statistical tests with each test using all epochs in the run up to that time, the sets of epochs used for different tests overlap, so the tests are not independent. Because of this non-independence, knowledge of the p-value for a particular test allows us to be sure that after a specified number of additional epochs, the new p-value will be in a certain range. It follows that if we have set a maximum number of epochs to be recorded, then it will sometimes be possible to know before reaching the maximum number of epochs that the p-values from later tests will not be less than the critical p-value for detection. In other words, it is sometimes possible to know in advance that subsequent tests in a run will not detect a response, in which case the run can be stopped early to save time.

To state the condition for early stopping we require some notation. Consider a test after n epochs and a later test after N epochs (so $N > n$), and denote the respective p-values by p_n and p_N. Let k be the number of bins, so in our framework we have $k = 9$. Let Ψ_{ν_1,ν_2} be the cumulative distribution function of an F random variable with degrees of freedom ν_1 and ν_2.

The general result is that if $0 < q < 1$ and

$$p_n > 1 - \Psi_{k,n-k}\left[\left(\frac{n-k}{N}\right)\left(\left(\frac{n}{N-k}\right)\Psi^{-1}_{k,N-k}(1-q) - \frac{N-n}{k}\right)\right],$$

then $p_N > q$. An outline of the proof of this result is given in the appendix.

If we take N to be the maximum number of epochs and q to be the p-value cutoff for detection then $p_N > q$ means that even after the maximum number of epochs, a response is not detected, so the testing can be stopped after epoch n instead of waiting until epoch N.

Table 2 shows the critical p-values for early stopping, assuming a maximum of 120 epochs and a p-value detection criterion of $p < 0.01$, obtained from the expression above with $k = 9$, $q = 0.01$, $N = 120$ and n from 102 to 119. The interpretation is, for example, that if the p-value after 110 epochs is greater than 0.298 then the p-value after 120 epochs cannot be less than 0.01, so the testing can be stopped after 110 epochs.

[Insert Table 2 here]

Discussion

This paper presented a general framework to optimize the detection of time-locked evoked potentials. The optimization allows the determination of when to conduct each statistical test and how long to test for. It aimed to optimize the inevitable tradeoffs between detection sensitivity, FPR (i.e., the specificity) and recording time. This framework can be applied to detection of auditory evoked potentials for threshold estimation.

Comparisons with other automated objective response detection paradigms

While guidelines have been developed with recommended (clinical) criteria for response presence and absence for CAEP testing (British Society of Audiology, 2016), only a handful studies have described objective statistical techniques to detect CAEPs when determining hearing thresholds, with an overview provided by Van Dun et al (2015). They do not provide guidelines however on when to automatically proceed to the next stimulus level, and on how to maximize sensitivity and minimize data collection time. The only study that we are aware of that provides elements towards an automatic objective threshold searching paradigm, is (Elberling & Don, 1984) for auditory brainstem responses (ABRs). They presented practical guidelines related to ABR and residual noise amplitudes, statistical measures for detection (using the Fsp), sensitivities and false positive rates. Based on these parameters, they derived when (or when not) to stop data collection and proceed to the next stimulus level. Overall, the paradigm presented in this study is the first one of its kind, and it allows automation of cortical threshold searching. A practical strategy is provided below.

Practical strategy for an automated response detection of CAEPs in adult populations

- The first statistical test is conducted when the RN level is equal to or smaller than 5.1 µV.
- Succeeding statistical tests are conducted when specific RN levels are reached, as defined in Figure 3a.

- The p-value detection criterion adopted throughout the strategy is equal to $p = 0.01$. This criterion guarantees a FPR of about 5%.

Averaging is stopped when:

- a response is detected using Hotelling's T^2 with a detection criterion of $p = 0.01$; or
- the maximum number of 120 epochs has been collected; or
- a RN level of 5.1 µV will likely not be reached before the maximum number of epochs has been collected; or
- the p-value at a specific number of accepted epochs (≥102) is higher than the critical p-value presented in Table 2.

Possible limitations of the method

A possible limitation of this study arise from the specific functions shown in Figure 3. Although these are reasonable choices, we are not aware of any technique that could be used to derive them such that the combination of sensitivity, specificity and recording time could be mathematically optimized. A better set of functions might therefore exist.

A second limitation is that for the technique to be applied to other populations (such as infants), the true CAEP peak and RN amplitude distributions and resulting SNRs (Figures 1 and 2) need to be determined in those populations. Nevertheless, we believe that the general approach will be valid. However, the maximum number of epochs might need to be adjusted for each population, impacting the spacing of the statistical tests.

A third limitation is the absence of an inconclusive result in case the residual noise levels are too high after the maximum number of epochs has been reached. Currently the system will indicate that the CAEP is absent. A new criterion could be added for this specific case. The test could be categorised as inconclusive instead of absent if the subject's mean noise-per-epoch value is higher than an established age-appropriate threshold when the maximum number of epochs is reached. This option might need to be implemented in the algorithm in the future.

An optimal detection method for CAEPs in a clinical environment

The proposed approach will make CAEP testing more accessible for clinicians, who generally have to rely on their own judgement how and when to interpret the cortical waveforms, and when to stop collecting data. This decision process takes time, which is in short supply in a clinical environment. Moreover, as clinicians all have their own approaches, it is next to impossible to derive a false positive rate for each clinician. The general technique described is intended for response detection in any age group, more specifically those who cannot provide reliable feedback like e.g. infants, young children, malingerers, those with multiple disabilities, who have suffered a stroke or are diagnosed with dementia. The real-time implementation of the proposed algorithm for clinical use has completed on electrophysiological hardware system called 'HEARLab' developed at the National Acoustic Laboratories.

Conclusion

This paper determined how often and when to conduct a statistical test and how long to test for the detection of time-locked evoked responses. When sufficient data are available to run the required simulations for determining specific parameters, the proposed framework can be applied to any system which involves detection of time-locked electrophysiological responses in biological systems containing background noise. Applications of this approach can be found in auditory, visual or motor threshold estimation, or basically any automated system which can converge to a predetermined criterion.

Acknowledgements:

We thank Professor Robert Cowan and the two reviewers for helpful comments on the manuscript. The experimental data used for this research were collected while the authors worked at the National Acoustic Laboratories. The authors acknowledge the financial support of the Australian Department of Health, the NSW Health and the HEARing CRC, established under the Australian

Government's Cooperative Research Centres (CRC) Program. The CRC Program supports industry-led collaborations between industry, researchers and the community. Harvey Dillon acknowledges the support of the NIHR Manchester Biomedical Research Centre and Macquarie University.

Declaration of interest statement

The authors have been involved with the development of HEARLab and its modules, but do not receive financial benefits from its sale.

References

Alcalá-Quintana, R., García-PÉREZ, M.A., 2005. Stopping rules in Bayesian adaptive threshold estimation. *Spat. Vis.*

Bardy, F., Van Dun, B., Dillon, H., 2015a. Bigger Is Better: Increasing Cortical Auditory Response Amplitude Via Stimulus Spectral Complexity. *Ear Hear*, 36(6), p.677–687. Available at: http://www.ncbi.nlm.nih.gov/pubmed/26039014.

Bardy, F., Sjahalam-King, J., Van Dun, B., Dillon, H., 2015b. Cortical auditory evoked potentials (CAEPs) in response to multi-tone (MT) stimuli in hearing-impaired adults. *Am. J. Audiol.*, (27), p.406–415.

Beagley, H.A., Kellogg, S.E., 1969. A comparison of evoked response and subjective auditory thresholds. *Int. J. Audiol.*, 8, p.345–353.

Botella, J., Ximénez, C., Revuelta, J., Suero, M., 2006. Optimization of sample size in controlled experiments: The CLAST rule. *Behav. Res. Methods*.

British Society of Audiology, 2016. *Recommended Procedure for CAEP testing*, Available at: https://www.thebsa.org.uk/wp-content/uploads/2016/05/Cortical-ERA.pdf.

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(5), p.347–356.

Coles, R.R.A., Mason, S.M., 1984. The results of cortical electric response audiometry in medico-legal investigations. *Br. J. Audiol.*, 18(2), p.71–78.

Delorme, A., Makeig, S., 2004. EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *J. Neurosci. Methods*, 134(1), p.9–21.

Van Dun, B., Dillon, H., Seeto, M., 2015. Estimating hearing thresholds in hearing-impaired adults through automatic detection of cortical auditory evoked potentials. *J. Am. Acad. Audiol.*, 26(4), p.1–14.

Elberling, C., Don, M., 1984. Quality estimation of averaged auditory brainstem responses. *Scand. Audiol.*, 13(3), p.187–197.

Golding, M., Dillon, 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(12), p.833–842.

Hyde, M., Sininger, Y.D., Don, M., 1998. Objective detection and analysis of auditory brainstem response: an historical perspective. *Semin. Hear.*, 19(1), p.97–117.

Kelley, K., Darku, F.B., Chattopadhyay, B., 2018. Accuracy in parameter estimation for a general class of effect sizes: A sequential approach. *Psychol. Methods*.

Komaki, F., Biswas, A., 2018. Bayesian optimal response-adaptive design for binary responses using stopping rule. *Stat. Methods Med. Res.*

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

Perl, E.R., Galambos, R., Glorig, A., 1953. The estimation of hearing threshold by electroencephalography. *Electroencephalogr. Clin. Neurophysiol.*, 5(4), p.501–512.

Pratt, H., Sohmer, H., 1978. Comparison of hearing threshold determined by auditory pathway electric responses and by behavioural responses. *Int. J. Audiol.*, 17(4), p.285–292.

Psioda, M.A., Soukup, M., Ibrahim, J.G., 2018. A practical Bayesian adaptive design incorporating data from historical controls. *Stat. Med.*

Ross, B., Lutkenhoner, B., Pantev, C., Hoke, M., 1999. Frequency-specific threshold determination with the CERAgram method: basic principle and retrospective evaluation of data. *Audiol. {&}*

Neuro-Otology, 4(1), p.12–27. Available at: http://www.ncbi.nlm.nih.gov/pubmed/9873149.

Stürzebecher, E., Cebulla, M., Elberling, C., 2005. Automated auditory response detection: Statistical problems with repeated testing Evaluación repetida en la detección de respuestas auditivas. *Int. J. Audiol.*, 44(2), p.110–117. Available at: http://www.tandfonline.com/doi/full/10.1080/14992020400029228.

Valdes Sosa, M.J., Bobes, M.A., Perez Abalo, M.C., Perera, M., Carballo, J.A., et al, 1987. Comparison of auditory-evoked potential detection methods using signal detection theory. *Audiology*, 26(3), p.166–178. Available at: file:////C:/Bibliografia sobre Neurociencias/pubneuropdf003/comparison.pdf.

Tables

Table 1: Summary of the EEG data sets used for Simulation 1. NH: normal-hearers, HI: hearing-impaired.

Data source	Adult population	N	Average number of non-response epochs per subject	Total number of non-response epochs
Bardy et al. 2015a	NH	15	928	15776
Bardy et al. (2015b)	HI	17	917	15589
NAL data set 1	NH	17	845	14366
NAL data set 2	NH	13	1218	20711
	Total	**62**	**3908**	**66442**

Table 2: Critical p-values for early stopping, assuming a maximum of 120 epochs and a p-value detection criterion of $p < 0.01$. Testing can be stopped after the number of epochs shown in the table if the p-value is greater than the corresponding p-value in the table.

Epoch	**102**	**103**	**104**	**105**	**106**	**107**	**108**	**109**	**110**
p-value	0.979	0.938	0.872	0.784	0.683	0.578	0.475	0.381	0.298
Epoch	**111**	**112**	**113**	**114**	**115**	**116**	**117**	**118**	**119**
p-value	0.229	0.172	0.127	0.092	0.066	0.047	0.033	0.023	0.015

Figure 1

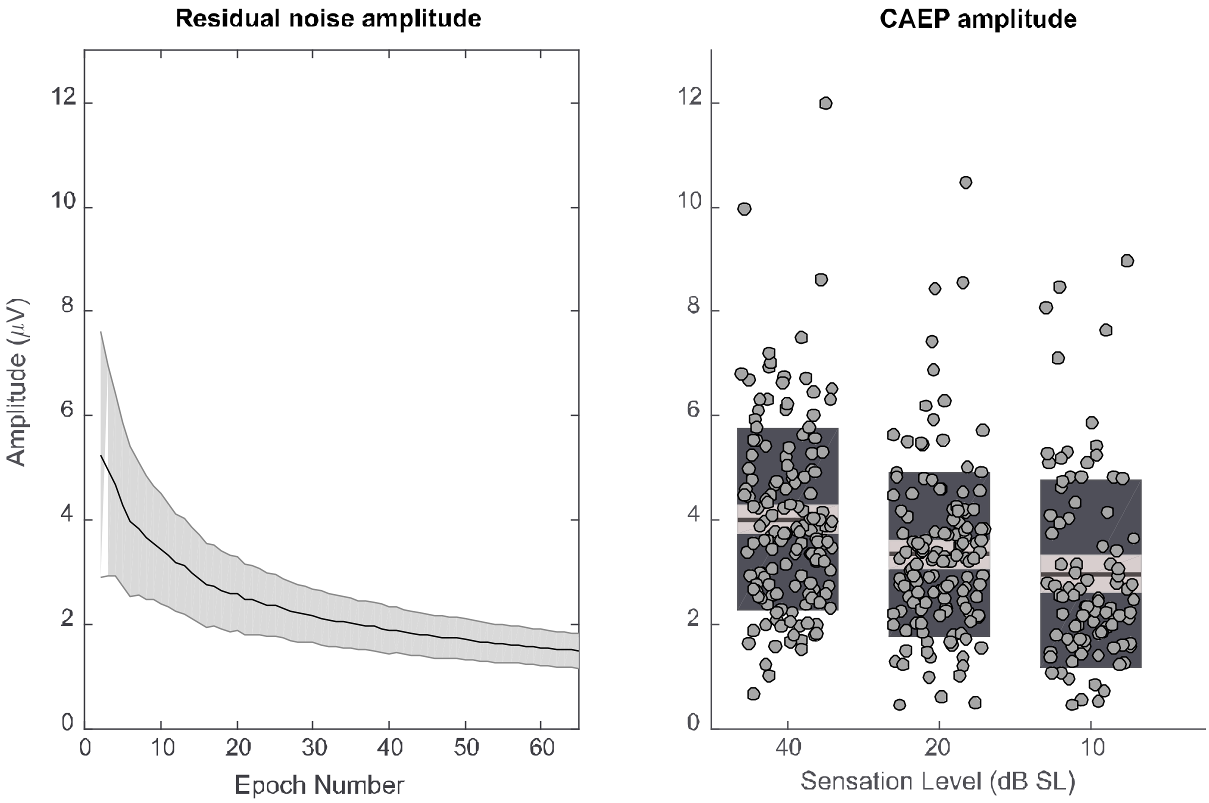

Figure 1. a) Represents the residual noise present in the EEG recording as a function of the number of epochs collected in a normal-hearing adult population. The grey shaded area represents the median and the epoch-to-epoch standard deviation of the RN rms amplitudes,. b) CAEP rms amplitude in normal-hearing adults plotted as a function of sensentation levels (i.e. 40, 20 and 10 dB SL). For the signal and noise amplitudes to be comparable, residual noise was expressed as its rms amplitude, while the signal amplitude was expressed as as the square root of the difference between the average waveform power in the time window from 51 to 347 ms post-stimulus onset and the estimated residual noise power.

Figure 2

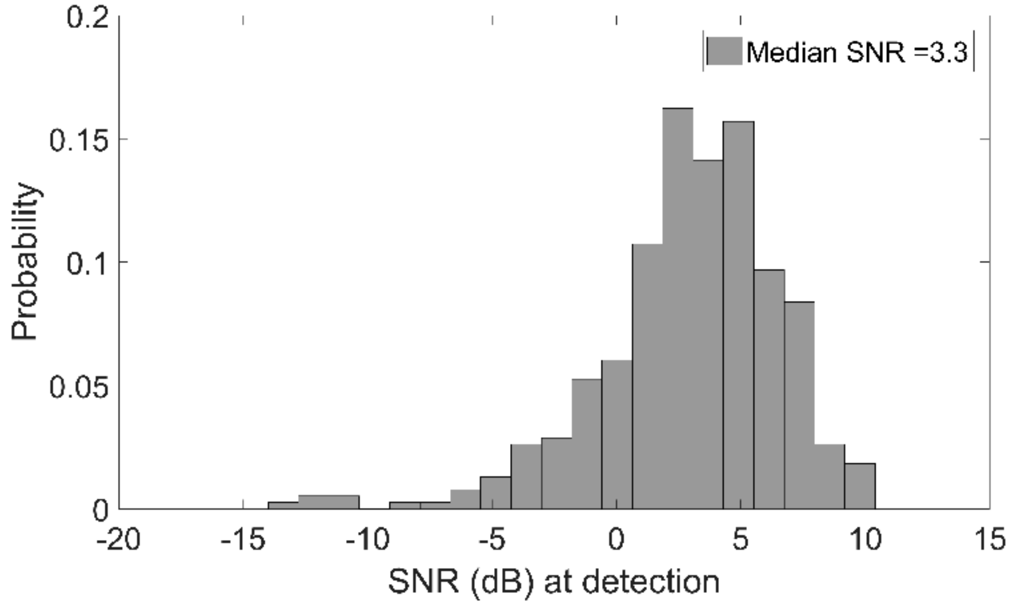

Figure 2: Distribution of SNRs when a CAEP is first detected using the Hotelling's T2 statistic constrained by a 5% FPR (normal-hearing adults).

Figure 3

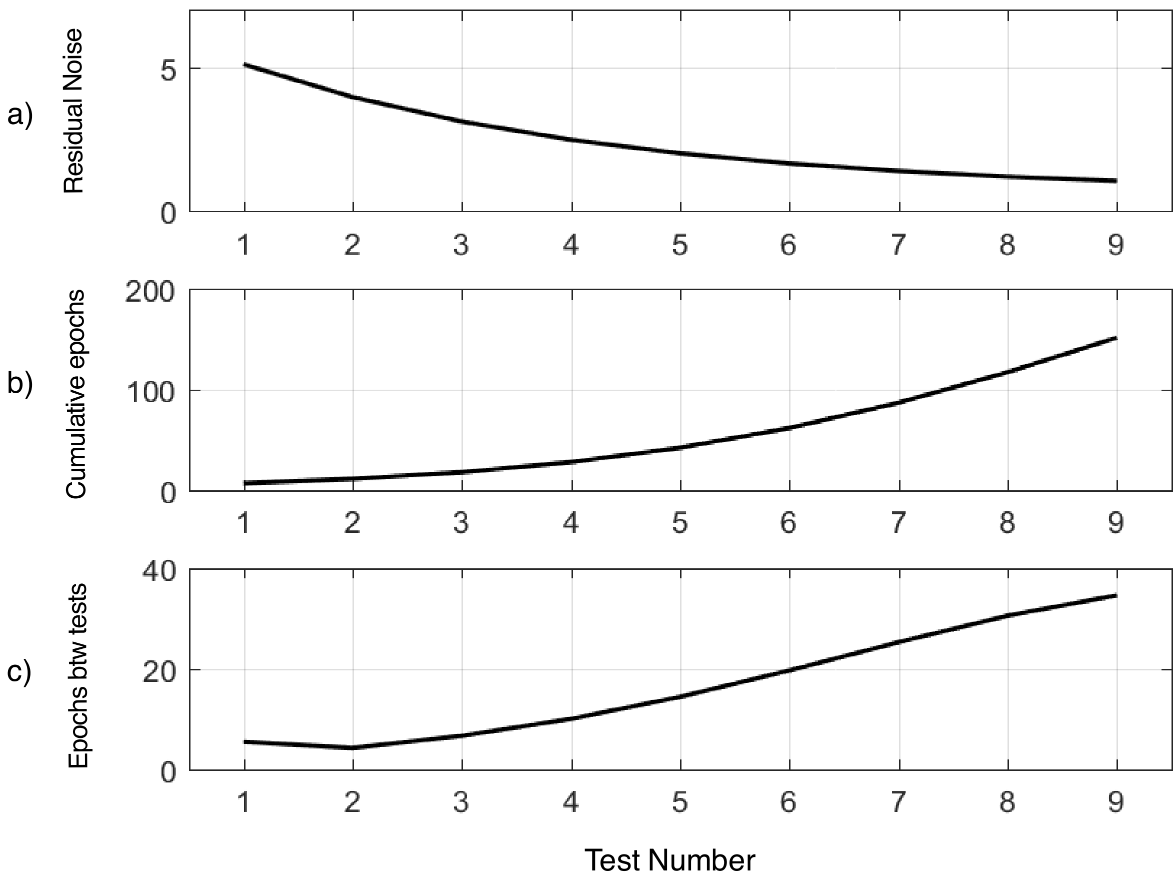

Figure 3: a) Representates the RN amplitudes at which Hotelling's T2 statistical tests are conducted. The equation of the RN criterion is 6*exp(-X/3.4)+0.63. b) Number of epochs necessary to reach the RN criteria displayed above for a subject having a residual noise level per epoch of 12.5 µV. c) Representation of the number of epochs between 2 tests before reaching the next noise criterion.

Appendix

We give an outline of the proof of the early stopping result. In the following, that result is called Corollary 2.

We use the same notation as in the body of the article: k is the number of bins, n is the "current" number of epochs, N is a number of epochs larger than n, and Ψ_{ν_1,ν_2} is the cumulative distribution function of an F random variable with degrees of freedom ν_1 and ν_2.

The situation we consider is that we have observed n epochs $x_1, \dots, x_n$, with each x_i being a k-dimensional vector of numbers representing the ith epoch after binning. We then ask, if an additional $N - n$ epochs are added to the original sample to get an extended sample $x_1, \dots x_n, x_{n+1}, \dots, x_N$ of size N, what values can the extended sample's Hotelling p-value take?

Let T_r^2 and p_r be the Hotelling T^2 statistic and the associated p-value, respectively, of the sample $x_1, \dots, x_r$, for $r = n$ and $r = N$.

We consider the original sample $x_1, \dots, x_n$ to be fixed, so the values of T_n^2 and p_n are fixed. We assume $T_n^2 > 0$, which is equivalent to assuming $p_n < 1$, and we make the usual assumption that covariance matrices of the data are positive definite.

Under these assumptions, we have the following results.

Theorem 1. The maximum possible value of T_N^2 is

$$\left(\frac{N-1}{n}\right)\left(\left(\frac{N}{n-1}\right)T_n^2 + N - n\right).$$

Corollary 1. The minimum possible value of p_N is

$$1 - \Psi_{k,N-k}\left[\left(\frac{N-k}{n}\right)\left(\left(\frac{N}{n-k}\right)\Psi_{k,n-k}^{-1}(1-p_n) + \frac{N-n}{k}\right)\right].$$

Corollary 2. If $0 < q < 1$ and

$$p_n > 1 - \Psi_{k,n-k}\left[\left(\frac{n-k}{N}\right)\left(\left(\frac{n}{N-k}\right)\Psi^{-1}_{k,N-k}(1-q) - \frac{N-n}{k}\right)\right],$$

then $p_N > q$.

As an aside, it is easy to show that the minimum possible value of T_N^2 is 0 and the maximum possible value of p_N is 1.

To prove Theorem 1, we start by considering the simplest case of $N = n + 1$ and $k = 1$, that is, one additional epoch and one-dimensional data. Note that with one-dimensional data, the Hotelling T^2 statistic is the square of the one-sample t-test statistic.

For each r, let $\bar{x}_r$, s_r^2 and t_r^2 be the sample mean, sample variance and squared t statistic, respectively, of the first r epochs. Note that the assumption that t_n^2 is non-zero means that $\bar{x}_n$ is non-zero.

We want an expression for t_{n+1}^2 in which the only variable quantity is x_{n+1}, keeping in mind that the first n epochs are assumed to be fixed. By definition we have

$$t_{n+1}^2 = \frac{(n+1)\bar{x}_{n+1}^2}{s_{n+1}^2}, \qquad (A1)$$

so we want to express $\bar{x}_{n+1}$ and s_{n+1}^2 in terms of fixed quantities and x_{n+1}. The required expressions are

$$\bar{x}_{n+1} = \frac{n\bar{x}_n + x_{n+1}}{n+1}$$

and

$$s_{n+1}^2 = \frac{(n+1)(n-1)s_n^2 + n(\bar{x}_n - x_{n+1})^2}{n(n+1)}.$$

Substituting these into (A1) gives

$$t_{n+1}^2 = \frac{n(n\bar{x}_n + x_{n+1})^2}{(n+1)(n-1)s_n^2 + n(\bar{x}_n - x_{n+1})^2}, \qquad (A2)$$

and we define $g(x_{n+1})$ to be the right-hand side of (A2), so $t_{n+1}^2 = g(x_{n+1})$.

We want to maximise the function g, so we look for points at which its derivative is zero. The derivative of g can be written as

$$g'(x_{n+1}) = \frac{2n(n+1)(n\bar{x}_n + x_{n+1})((n-1)s_n^2 + n\bar{x}_n^2 - n\bar{x}_n x_{n+1})}{[(n+1)(n-1)s_n^2 + n(\bar{x}_n - x_{n+1})^2]^2},$$

so the solutions of $g'(x_{n+1}) = 0$ are $x_{n+1} = -n\bar{x}_n$ and

$$x_{n+1} = \bar{x}_n + \frac{(n-1)s_n^2}{n\bar{x}_n}. \qquad (A3)$$

At the first of these solutions g is 0, and since g can't be negative, this must be a minimum. It can be verified that at the second solution, the second derivative of g is negative, so that is where g is maximum.

The maximum possible value of t_{n+1}^2 therefore occurs when x_{n+1} has the value in (A3). Substituting this value into (A2), and denoting the maximum possible value of t_{n+1}^2 by $\max t_{n+1}^2$, gives

$$\max t_{n+1}^2 = \left(\frac{n+1}{n-1}\right) t_n^2 + 1, \qquad (A4)$$

which is Theorem 1 for $N = n + 1$ and $k = 1$.

We now consider the case with $N = n + 1$ and $k > 1$. Using the notation $t^2(v_1, \dots, v_r)$ to mean the squared one-sample t statistic of the univariate sample $v_1, \dots, v_r$, it is well-known (e.g., Johnson & Wichern, 2007) that

$$T_r^2 = \max_{a \neq 0} t^2(a'x_1, \dots, a'x_r),$$

with a and the x_i being viewed as $k \times 1$ matrices and a' denoting the transpose of a.

Using this fact together with (A4) gives

$$\max T_{n+1}^2 = \left(\frac{n+1}{n-1}\right) T_n^2 + 1,$$

which is Theorem 1 for $N = n + 1$. From this, the result for general $N > n$ can be proved by induction.

Corollary 1 can be obtained by using the fundamental relation

$$p_r = 1 - \Psi_{k,r-k}\left(\frac{r-k}{k(r-1)}T_r^2\right)$$

(Johnson & Wichern, 2007) for $r = n$ and $r = N$ together with Theorem 1, and by noting that the p-value is minimised when the corresponding T^2 is maximised.

Corollary 2 can be obtained by rearranging Corollary 1.

Reference:

Johnson, R. A., & Wichern, D. W. (2007). Applied multivariate statistical analysis (6th ed.). Upper Saddle River, NJ: Pearson Prentice Hall.