UNIVERSITY OF SOUTHAMPTON

Objective Detection of Auditory Brainstem Responses Using a Bootstrap Technique

by

Jing Lv

A thesis submitted in partial fulfillment for the degree of Doctor of Philosophy

in the
Signal Processing and Control Group
Institute of Sound and Vibration Research

April 2007

Abstract

Auditory Evoked Potentials (AEPs) measure the responses from the auditory nervous system structures following presentation of an acoustic stimulus (clicks or tone-burst). Usually, the responses are interpreted subjectively, by visual inspection. However, this requires well trained professionals, and is strongly dependent on the experience of the observer. Objective, automated methods for detecting responses are clearly desirable, especially for screening (e.g. neonatal hearing tests) and monitoring (e.g. during surgery). The aim of this work is to investigate new methods to objectively detect the responses.

A novel bootstrap technique was proposed, allowing the statistical significance (p-value) to be estimated for a wide range of different signal parameters, and detect the response in an easy and very flexible manner. The bootstrap method is based on randomly resampling (with replacement) the original data and gives an estimate of the probability that the response obtained is due to random variation in the data rather than a physiological response. Furthermore, the bootstrap technique provides a simple way to compare different methods for response detection using p-values. Even though existing methods have proved to be effective in detecting a response, comparing them is usually a problem because different approaches have different criteria. The proposed method helps to solve that problem.

A modified bootstrap method with three artefact rejection schemes was then proposed and they can efficiently eliminate the effect of stimulus and/or movement artefacts. This modification makes the bootstrap procedures more effective to deal with 'real data' from patients, where artefacts are often present.

The performance of the bootstrap method was evaluated on simulated signals by receiver operating characteristic (ROC) analysis and compared with other methods. On data recorded from normal-hearing volunteers, the techniques provided similar hearing thresholds to those obtained by visual inspection of the auditory brainstem response (ABR). The flexibility of this approach allows the method to be used with a range of parameters, numbers of sweeps, and with user-defined false positive rates.

Acknowledgements

I would like to thank David Simpson, my supervisor, for his many suggestions, constant support and kind encouragement during this research.

Many thanks to Steven Bell for providing the ABR data (Set A), for his guidance in the experiment of collecting additional ABR data (Set B), his help in the calibrations of the equipments, and for his comments on all the publications.

Thanks also to all colleagues in SPCG and HABC of ISVR for their help, and all of the subjects involved with the data collection.

Scholarships awarding to me by the ORS (UK) and Rayleigh (ISVR/University of Southampton) were crucial to the completion of the PhD.

Of course, I am grateful to my parents, my husband and my brother for their patience, love and constant encouragement.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Outline and motivation

Auditory Evoked Potentials (AEPs) (Katz, 2001; Hall, 1992b) measure the responses from the auditory nervous system structures following presentation of an acoustic stimulus, which ranges from clicks (very brief, sharp sounds) to tone-bursts, or more complex sounds, such as speech. The response is an electrical activity in the cochlea, auditory nerve, and various structures in the auditory brainstem, through to the cortex, which are captured by recording the electroencephalogram (EEG). AEPs can be divided into two categories: transient or onset potentials and sustained potentials. Transient potentials represent a single response that results from presentation of a single stimulus. Neural units generating these responses are onset-sensitive, thus responding to the onset of a stimulus. In contrast, sustained potentials are responses that reflect either repeated or continual stimulation. Typical transient potentials are the eighth nerve action potential (AP) seen in electrocochleography (ECochG), the auditory brainstem response (ABR), the middle latency response (MLR),and cortical potentials such as the N1-P2 (vertex) response and the P300 response. Sustained potentials include the cochlear microphonic (CM) and the 40-Hz response, which is referred to as a steady state potential because of its repetitive nature (Hood, 1998).

AEPs are widely and successfully used in clinical practice. ABR is primarily utilized in (a) identification of neurological abnormalities in the eighth cranial nerve and auditory pathways of the brainstem and (b) estimation of hearing sensitivity based on the presence of a response at various intensity levels. A potential application of

the MLR is as indicators of depth of anesthesia in patient undergoing surgery (Beer et al., 1996).

Of several AEPs, the potential generated in the brain stem provides the most accurate information about the integrity of the auditory system. Unlike the cortical evoked potentials, the auditory brainstem response (ABR) is unaffected by a variety of psycho-physiological parameters, e.g. sleep (Amadeo and Shagass, 1973), attention and arousal (Picton and Hillyard, 1976), or anesthetic agents (Bobbin et al., 1979; Stockard et al., 1978). Its low variability has led to wide applications for neurology, otology, and audiology. One of the primary application of ABR is to determine hearing thresholds in patients that are unable or unwilling to cooperate with behavioural testing.

The conventional way to analyze and interpret the ABR is visual inspection by experienced audiologists, who usually identify significant peaks (the most important are denoted with roman numerals I, III and V - see Figure 1.1[1]). However, this identification is subjective, and considerable inconsistency has been found between different experienced professionals in estimating hearing thresholds (Vidler and Parker, 2004; Arnold, 1985) from the ABR. As a result of this, a number of methods and algorithms for automated ABR identification and detection have been described in the literature. Some of these identify the highest amplitudes in latency regions where peaks are expected to occur in the normal ABR (Mason, 1984; Ozdamar et al., 1994; Pool and Finitzo, 1989). Others are based on different statistical properties, either in the time-domain (e.g. Cross-correlation (Weber and Fletcher, 1980; Ozdamar et al., 1990), F_{sp} (Elberling and Don, 1984), $\pm$ difference (Wong and Bickford, 1980), Friedman test (Cebullar et al., 2000), and Cochran's Q-test (Cebullar et al., 2000)), or in the frequency domain (e.g. magnitude-squared coherence (MSC) (Dobie and Wilson, 1989), phase coherence (Jerger et al., 1986), spectral F-test (Zurek, 1992), q-sample uniform scores test (Sturzebecher et al., 1999), Rayleigh test (Cebullar et al., 1996; Lutkenhoner, 1991; Sturzebecher and Cebullar, 1997), Hotelling's T^2 (Picton et al., 1987), modified Hotelling's T^2 (Valds-Sosa et al., 1987), and circular T^2 (Victor and Mast, 1991)). Some of these methods provide an exact statistical criterion (p-value) when a response can be considered to be significant, others do not. The advantage

[1]In common with other publications, the unit microvolt is presented as uV due to the lack of the symbol μV in many graphics software packages.

of the former is that the false-positive-rate provides a clearly defined criterion for detecting responses, whereas for the latter empirically derived threshold criteria are used, so it becomes difficult to compare techniques based on the trade-off between sensitivity and specificity. These difficulties motivated the current work to improve the available techniques and explore new methods.

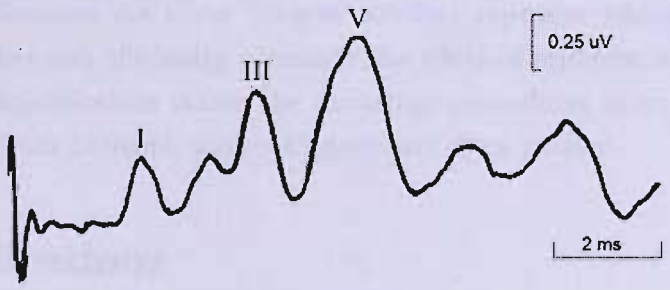


Figure 1.1: ABR waveform labelled by Roman numerals. The figure is drawn based on Figure 2-1 in Hood (1998).

We will describe the proposed methods, based on the statistical bootstrap technique, for which we give some detail on its performance in ABRs. However, the technique proposed could also readily be applied in other modalities of AEPs, as well as visual, somatosensory, and other event-related evoked potentials.

1.2 Original contributions

There are three main contributions in this work:

One is that a novel method, a bootstrap technique to detect the response based on the statistical p-values, is proposed. The bootstrap method is based on randomly resampling (with replacement) the original data and gives an estimate of the probability that the response obtained is due to random variation in the data rather than a physiological response. Furthermore, the bootstrap technique provides a simple way to compare different methods for response detection using p-values. Even though existing methods have proved to be effective in detecting a response, comparing them is

usually a problem because different approaches have different criteria. This proposed method solves the problem.

The second contribution is to evaluate the performance of the bootstrap method for detecting the response in well-controlled simulations and recordings from normal-hearing subjects.

The third contributions are three 'plug-in' artefact rejection schemes for the bootstrap method, that can efficiently eliminate the effect of stimulus or/and movement artefacts. This modification makes the bootstrap procedures more effective to deal with 'real data' from patients, where artefacts are often present.

1.3 Publications

Journal papers:

1. Lv, J, Simpson, D. M, Bell, S. L. A modified bootstrap method for the detection of auditory brainstem responses, submitted to Biomedical Signal Processing and Control on 12/12/2006.

2. Lv, J, Simpson, D. M, Bell, S. L. 'Objective Detection of Evoked Potentials Using a Bootstrap Technique,' *Medical Engineering and Physics*, 29, p.191-198, 2007.

Conference papers:

1. Lv, J, Simpson, D. M, Bell, S. L. A modified bootstrap test for the detection of evoked responses, with artefact rejection. World congress on Medical Physics and Biomedical Engineering (WC2006), Seoul, Korea, 27 August - 1 September, 2006.

2. Lv, J, Simpson, D. M, Bell, S. L. Detection of evoked responses by bootstrap methods: new parameters and ROC analysis. Proceedings of MEDSIP 2006 (Advances in Medical Signal and Information Processing), Glasgow, UK, 17-19, July, 2006.

3. Lv, J, Simpson, D. M, Bell, S. L. A new approach in the automated detection of evoked potentials. Proceedings of Faculty of Medicine, Health and Life Sciences Postgraduate Conference, Southampton, UK, 6-7, June, 2006.

4. Lv, J, Simpson, D. M, Bell, S. L. Application of the bootstrap technique to detect auditory evoked potentials. Proceedings of the 2nd Life Science Interfaces Conference, Southampton, UK, 1, December, 2005.

5. Lv, J, Simpson, D. M, Bell, S. L. A novel statistical test to detect auditory evoked potentials. The 3rd European Medical and Biological Engineering Conference, Prague, Czech Republic, 20-25, November, 2005.

6. Lv, J, Simpson, D. M, Bell, S. L. A Statistical Approach to Measuring Hearing Thresholds from Auditory Brainstem Responses. International Evoked Response Audiometry Study Group (IERASG), Havana, Cuba, 13-16, June, 2005.

7. Lv, J, Simpson, D. M, Bell, S. L. Objective tests for the detection of auditory evoked potentials. Pages: 1-2 Proceedings of PGBIOMED04 (The 3rd IEEE EMBSS UK and RI Postgraduate Conference in Biomedical Engineering and Medical Physics), Southampton, UK, 9-11 August 2004.

8. Lv, J, Simpson, D. M, Bell, S. L. A statistical test for the detection of auditory evoked potentials. Proceedings of the Institute of Physics and Engineering in Medicine (IPEM) Meeting on Signal Processing Applications in Clinical Neurophysiology, York, UK, 10 February 2004.

1.4 Structure of the thesis

Chapter 1 has provided an introduction to outline the problem and motivations. Then the main contributions of this study were briefly described, and the publications listed.

Chapter 2 is a literature review which includes an overview of the auditory brainstem response (ABR), automated assessments of the ABR, overview of the bootstrap technique and its applications, and some of the statistical analysis used in this work.

Chapter 3 is a summary of the signal acquisition carried out, which contains the experimental work such as the equipment configuration, calibration, and the procedures of recording the ABR.

Chapter 4 investigates the properties of the background EEG which is present when recording ABRs. Simulations of the background EEG by an autoregressive model, which is used in testing the bootstrap method, are then described.

Chapter 5 describes the procedures of the bootstrap method in detail. The evaluation of the method is then performed based on Monte-Carlo simulations. Following that, the bootstrap technique is employed on real recordings to estimate the hearing threshold, and investigate the minimal number of sweeps required for detection of a response.

In Chapter 6, three artefact rejection schemes are proposed and investigated to remove the effect of stimulus and movement artefacts. Then these schemes, combined with the bootstrap technique, are applied to both simulations and recordings, in the presence or absence of the artefacts. The results demonstrates the improvement of sensitivity achieved by the artefact rejection schemes, in the presence of artefacts.

Chapter 7 compares the bootstrap technique with other methods which were proposed by other authors, and used in detecting the response.

Finally, Chapter 8 summarizes the findings and suggests future work leading on from the investigations that have been performed.

Chapter 2

Literature Review

2.1 Introduction

In this chapter, we will summarize background knowledge related to the problem mentioned in the previous chapter. Auditory brainstem response (ABR) are the subject of all the remaining chapters and therefore a good understanding of them is important. We provide an overview including how to obtain the ABR, the characteristics of a normal ABR, neural generators of it, and factors that influence it, as well as applications of the ABR. Then we will list the objective approaches for their detection available in the literature, with a brief description of the principles behind them and the algorithms. The first five methods are applied in time domain and another nine are applied in the frequency domain. Following those, we will introduce the bootstrap technique, including its principles and some applications. Finally we will introduce some statistical analysis methods used in this study. Some are well-known but others are more complex and not as often applied, particularly in this field. This background should provide a sound basis for the following investigations.

2.2 Overview of auditory brainstem response

2.2.1 Estimating the ABR

The ABRs are usually impossible to recognize in the background scalp-recorded EEG. Averaging methods help to estimate the ABR and three most popular approaches will be introduced in the following.

Coherent Averaging

Coherent averaging is a method that is conventionally used to improve the signal-to-noise ratio so that the signal of interest can be extracted. An important principle of coherent averaging is that the measured signal (EEG, including both ABR and background EEG) is acquired in epochs (sweeps) that are exactly time-locked to the repeated stimulus. In the recordings, normally hundreds (at high stimulus intensities), and often thousands (at low intensities) epochs are acquired. Under the assumption that: (1) the ABR is deterministic and the same in response to a constant stimulus, (2) the background EEG is random (i.e., it is not correlated with the stimuli) (Robert and Carrie, 2001); the process of coherent averaging is shown in Figure 2.1 and 2.2. This process leads to a signal that remains the same size as the ABR, while the background EEG tends to cancel and become smaller. Thus the SNR increases as the number of averaged epochs increases.

When the background EEG is stationary with zero mean and equal variance, coherent averaging method reduces the standard deviation of the background EEG in the coherently averaged signal by the square root of the number of sweeps, $\sqrt{K}$ (K represents the number of sweeps in the recording), and correspondingly increases the SNR by the same rate.

In order to investigate this relationship from theory point of view, the background EEG signal can be written in a matrix:

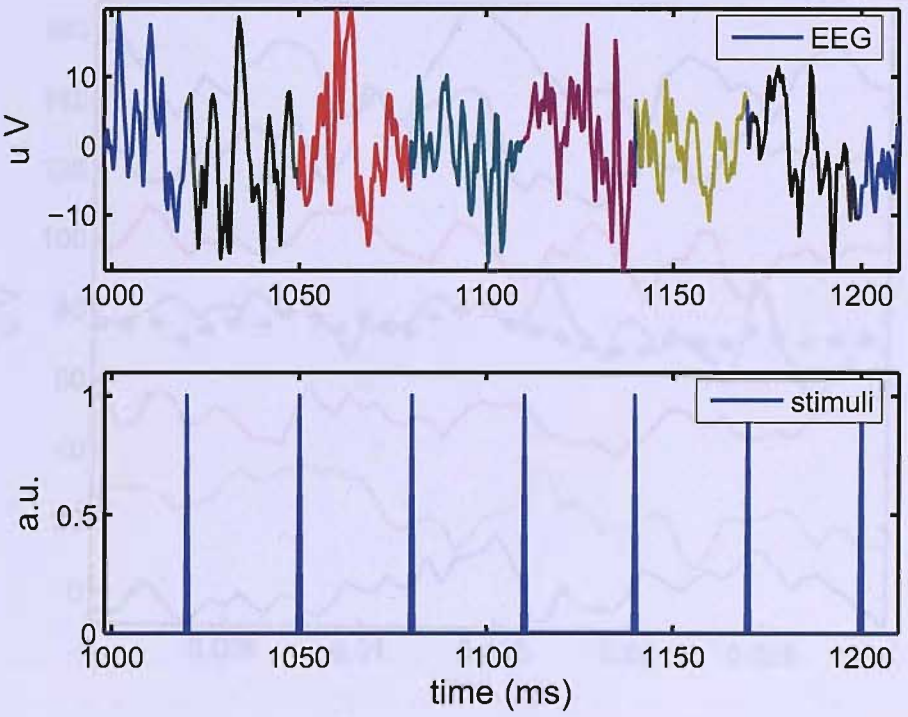


Figure 2.1: In order to show the process of coherent averaging clearly, only a part of a typical raw EEG data is plotted. In the bottom plot, the peaks represent the repeated stimuli. Corresponding to each stimulus, the top figure shows the EEG including both the ABR and the background EEG. The signal is then segmented into many epochs whose starting point is exactly the onset of the stimuli. The ensemble of the different sweeps is shown in the Figure 2.2 (time range is 0-30 ms).

$$X = \begin{pmatrix} x_{11} & \cdots & x_{1M} \\ x_{21} & \cdots & x_{2M} \\ . & . & . \\ x_{n1} & \cdots & x_{nM} \\ x_{K1} & \cdots & x_{KM} \end{pmatrix} \tag{2.2.1}$$

where $x_{k,m}$ represents one sample, $k = 1, 2, ..., K$, where K is the number of the epochs (sweeps), and $m = 1, 2, ..., M$, where M is the number of the samples in each

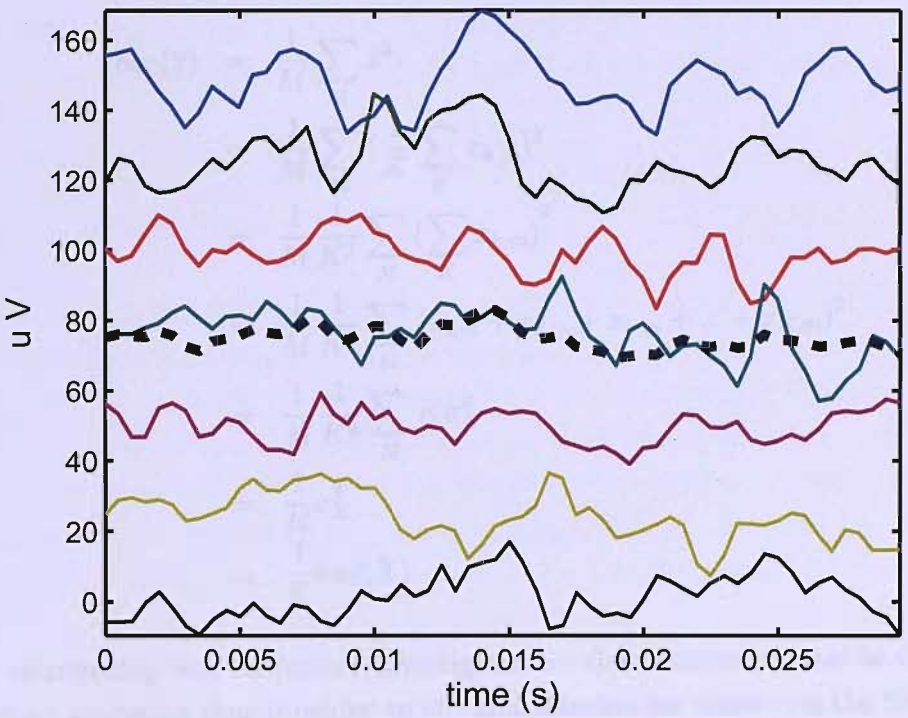


Figure 2.2: Corresponding to the repeated stimuli, the coherent average of the selected synchronized epochs is calculated. The seven solid lines represent the epochs of raw EEG following each stimulus (an offset has been added to show them more clearly), and the dotted line is the result of the averaging of the seven signals. That is usually called the coherently averaged signal.

epoch, and $N = M * K$. The variance of all the components in the matrix is:

$$var(X) = \frac{1}{N} \sum_{M} \sum_{K} x_{k,m}^2 = \sigma_X^2 \tag{2.2.2}$$

And the variance of the coherently averaged signal is calculated from the following equations, with the assumption of signals being zero mean and with equal variance, uncorrelated between sweeps.

$$
\begin{aligned}
var(\overline{x}) &= \frac{1}{M}\sum_{M}\bar{x}^2 \\
&= \frac{1}{M}\sum_{M}(\frac{1}{K}\sum_{K}x_{k,m})^2 \\
&= \frac{1}{M}\frac{1}{K^2}\sum_{M}(\sum_{K}x_{k,m})^2 \\
&= \frac{1}{M}\frac{1}{K^2}\sum_{M}(x_{1,m}+x_{2,m}+x_{k,m}+...+x_{K,m})^2 \\
&= \frac{1}{M}\frac{1}{K^2}\sum_{M}K\sigma_X^2 \\
&= \frac{1}{K}\sigma_X^2 \\
&= \frac{1}{K}var(X) \qquad (2.2.3)
\end{aligned}
$$

This relationship will be further investigated on the recorded signals in Chapter 4. Coherent averaging thus provides an efficient solution for improving the SNR.

Median Averaging

Coherent averaging has some inherent problems that make it a less than desirable tool, although it has many desired characteristics. Coherent averaging is founded on the principle that the signal in a response is constant and phase locked to the stimulus, whereas the noise is stationary and random with no phase locking to the stimulus. In the real world, however, noise is nonstationary. Thus, coherent averaging may produce suboptimal extraction of the signal from noise. Coherent averaging is highly sensitive to nonstationary noise. An alternative method, median averaging (Ozdamar and Kalayci, 1999) was investigated for reducing the deleterious effects of noise. The median may provide a more reliable representation of a group than the mean, especially, when there are extreme values in the group. Borda and Frost (1968) first suggested the use of a median averaging method for reducing the sensitivity of conventional coherent averaging to noise fluctuations in small samples. After experiments, Ozdamar and Kalayci (1999) demonstrated that the median averaging of the

ABR is a feasible and reliable method. The responses obtained with median averaging of 512 sweeps showed somewhat better characteristics in terms of wave detection and SNR than coherent averaging of the same data using visual identification of the ABR waves.

Weighted Averaging

The well-established coherent averaging technique presupposes that the physiological background noise is stationary, but as discussed in the previous section that is not always the case especially when the test subject or the patient changes his or her state of relaxation. The high level background noise introduces larger uncertainty in the ABR estimate than those from low levels. Aware of this problem, Elberling and Wahlgreen (1985) proposed a weighted averaging technique with the only assumption that the physiological background noise has a Gaussian distribution. The data is split into sub-blocks of say 250 sweeps. The average of each block is weighted according to the reciprocal of the variance of the block (an estimate of the noise level of the block) before being included in the overall average. Bayesian statistics were applied to the ABR estimate in the following way. Let S_i and V_i indicate the mean waveform of the ith block, including 250 sweeps, and the estimated variance of the corresponding background noise, respectively. $\widehat{ABR_i}$ denotes the Bayesian estimate of the ABR after the ith block:

After the first block, calculate:

$$\widehat{ABR_1} = (\frac{S_1}{V_1})\ \frac{1}{C_1}\ ;\ C_1 = \frac{1}{V_1} \tag{2.2.4}$$

After the second block,

$$\widehat{ABR_2} = (\frac{S_1}{V_1} + \frac{S_2}{V_2})\ \frac{1}{C_2}\ ;\ C_2 = \frac{1}{V_1} + \frac{1}{V_2} \tag{2.2.5}$$

Likewise, after the nth block, calculate:

$$\widehat{ABR_n} = (\frac{S_1}{V_1} + \frac{S_2}{V_2} + \cdots \frac{S_n}{V_n})\ \frac{1}{C_n}\ ;\ C_n = \frac{1}{V_1} + \frac{1}{V_2} + \cdots \frac{1}{V_n} \tag{2.2.6}$$

Equation 2.2.6 describes how the Bayesian inference is used to produce an ABR estimate. In this study, Elberling and Wahlgreen demonstrated the weighted averaging technique was more efficient in recovering the ABR from the background noise than the classic coherent averaging technique.

2.2.2 The normal ABR waveform and its characteristics

A normal ABR waveform is characterized by five to seven vertex-positive peaks that occur in the time period from 1.4 to 10 ms after the onset of a stimulus (Hood, 1998) as shown in Figure 1.1. The peaks of the ABR represent sums of neural activity from one or more sources at various discrete points in time.

Responses are displayed with positive peaks reflecting activity toward the vertex of the head (vertex-positive), and the peaks are labelled by Roman numerals I through VII (Figure 1.1) following recommendations of Jewett and Williston (Jewett et al., 1970).

In normal individuals, the absolute latency of Wave I usually occurs approximately 1.6 ms after stimulus onset, Wave III at about 3.7 ms and Wave V at about 5.6 ms for click presented at 75 dB above the normal hearing threshold (Hood, 1998). The latency of the ABR is consistent and repeatable in normal individuals, and peak latencies should replicate[1] within 0.1 ms. A normal ABR ranges in amplitude from 0.1 to 1.0 μ V. As the stimulus intensity decreases, the amplitude decreases, and the latency increases.

2.2.3 Neural generators of the ABR in humans

The usefulness of ABR in making otoneurological diagnoses depends upon knowledge of the anatomical origins of the various components of the ABR that can be identified and how different pathologies change these potentials.

In the early 1980s, some of the first published systematic studies of neural generators of ABRs in neurosurgical patients, using simultaneous near-field, and far-field

[1]In visual inspection of the ABR, two recordings in the same conditions are collected and called replicates.

recordings, appeared in the literature. The near-field recording is situated close to the source of an electrical potential. And the far-field occurs when recording electrodes are at a distance from the source of the electrical potential. Moller et al. (1981) and Hashimoto et al. (1981) did a series of studies and identified the origins of different peaks (Figure 1.1):

- Wave I of the ABR occurred with the same latency as N1 of the ECoG potentials, which suggested that peak I was generated by the auditory nerve.
- Wave II is generated by the proximal (brainstem) portion of the eighth nerve.
- Wave III is generated by auditory pathways and structures in the pons (e.g., trapezoid body, superior olivary complex).
- The origin of wave IV is controversial, and it has been difficult to identify specific anatomical structures that generate this peak.
- Wave V is generated by the lateral lemniscus, where it terminates in the contralateral inferior colliculus.

Wave I and II arise on the side of the auditory system ipsilateral to the stimulus, whereas wave III and later components probably receive bilateral contributions. The response presumably reflects synchronous activation primarily of onset-type neurons within the auditory system. Wave I and II are action potentials, whereas later waves may reflect post-synaptic activity in major brainstem auditory regions.

2.2.4 Factors affecting the ABR

A number of patient, stimulus, and recording factors influence the ABR. These factors may affect the amplitude and latency of the waveform. So it is very important to understand these effects and consider them in recording procedures and interpretations.

Patient factors

Patient factors may influence the outcome of the ABR recordings in any subject, even those with a normal peripheral and central auditory system. These include age, gender, medication, attention, body temperature, and muscle activity.

1. **Age**

   The ABR changes as a function of age, particularly during the first 12 to 18 months of life, as the central auditory system continues to mature. However, the effect of age in adults is less clear (Fria and Doyle, 1984).

2. **Gender**

   Females usually have shorter latency and higher amplitude ABRs than males (Allison et al., 1983; Jerger and Hall, 1980; Michalewski et al., 1980). These differences may be related to shorter cochlea response times in females than males (Don et al., 1994).

   When gender and age in adults are considered together, the shortest latencies are obtained from younger females, with latency increasing for older females, then young males, and finally the longest latencies are obtained from older males (Don et al., 1994).

3. **Medication**

   Sanders et al. (1979) and Starr et al. (1977) suggested that the ABR is not affected by sedative, relaxants, barbiturates, or anesthesia. However, abnormal ABRs have been found in conjunction with medications such as phenytoin (anti-convulsants agents to control seizure activity), lidocaine (antiarrhythmic agents to control heart activity), and diazepam (antianxiety drug) (Hood, 1998).

4. **Attention**

   ABRs are not affected by sleep (Jewett and Williston, 1971), do not change by metabolic or toxic coma (Starr et al., 1977), and do not differ as a function of attention (Picton and Hillyard, 1974).

5. **Body Temperature**

   A correlation between rises in body temperature and decreases in latency in cats has been reported by Jones et al. (1980).

6. **Muscle artefact**

   In theory, the ABR is hardly influenced by myogenic potentials (muscle activity), although it is well-known that a quiet patient facilitates detection of the response, particularly at stimulus intensity near threshold.

Stimulus factors

Stimulus properties, such as intensity, rate, polarity, duration and rise time, mode of stimulus presentation (monaural and binaural), exert significant and interrelated effects on the ABR measurement.

1. **Stimulus Intensity**

   All waves of the ABR show the tendency to increase in latency and decrease in amplitude as stimulus intensity decrease from 70 or 80 nHL to the threshold of the normal-hearing subjects (Picton and Hillyard, 1974; Starr and Achor, 1975).

2. **Stimulus Rate**

   The stimulus rate influences both the latency and the amplitude of the ABR. Generally, with a stimulus rate over 30/s, the latency of all waves of the ABR increases and the amplitude of the earlier waveforms decreases (Don et al., 1977; Fowler and Noffsinger, 1983).

3. **Stimulus Polarity**

   Three types of stimulus polarity are available rarefaction, condensation, and alternating between rarefaction and condensation, and these can be selected in recording the ABR. They have different effects on the waveforms of the ABR. For a rarefaction stimulus, latency is slightly shorter and the amplitude is higher for the earlier waves than for condensation stimuli.

4. **Stimulus Duration and Rise Time**

   The standard pulse duration in ABR testing is 100 μs. Because the ABR is an onset response, the stimulus duration should not alter the response.

   The rise time of the stimulus affects the ABR significantly. Slower rise time is related to reduced synchrony in that fewer neurons are firing simultaneously. As rise time increases, latency increases, amplitude decreases and the morphology deteriorates (Hood, 1998).

5. **Monaural versus Binaural stimulus**

   ABRs to binaural stimuli show an average of 60% increase in amplitude over those by monaural stimulation (Blegvad, 1975). Latencies of the ABRs by monaural and binaural stimulation are similar.

Recording factors

An understanding of recording factors, such as electrode placement, filter, time window, one-channel or two-channel recording, number of sweeps, is essential for successful clinical applications of the ABR.

1. **Electrode Placement**

   Placement of electrodes at the forehead (F_z) and the mastoids (A_1 and A_2) (Figure 2.3) is optimum for recording the ABR in most conditions (Martin and Moore, 1977). Wave I, II, III are most prominent in ipsilateral recording, and wave IV and V are better isolated in contralateral recordings (Mizrahi et al., 1983).

2. **Filter Settings**

   The filtering aims to reduce the internal noise (e.g. unrelated muscle potentials) and the external electrical interferences (e.g. mains noise). Changes in frequency band affect the latency and amplitude. Increasing the high-pass filter cut-off frequency (i.e., reducing the low-frequency energy) from 30 Hz to 100 Hz results in decreases in the amplitude and latency of the ABR. Allowing more

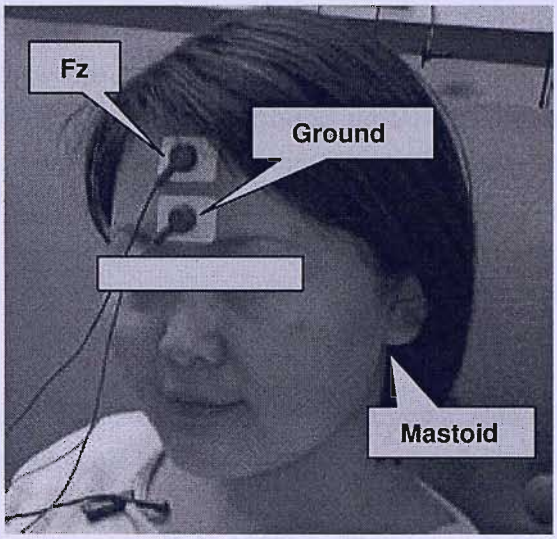


Figure 2.3: The positions of the three electrodes for recording the ABR.

low-frequency information into the average generally leads to increased amplitude, especially for the later components, and slightly longer latencies. Decreasing the cut-off frequency of the low-pass filter (e.g., from 3000 Hz to 1500 Hz) has less effect on amplitude and latency. When the interference from electrical sources and muscle activity (mainly low frequencies) are included, wave V amplitude increases because it is dependent on the amount of low-frequency energy. Very narrow-band filters are not recommended because phase shifting may appear in frequency regions near the cut-off frequencies.

3. **Time Window**

The time window for analysis of the ABR should be set to cover all the waves of the ABR. The length of the window will vary with the age of the patient and the intensity and type of the stimulus. For click stimulus in adult, a time window from stimulus onset to 10-12 ms is adequate for ABR recording because wave V for normal individuals occurs at 5 or 6 ms at high stimulus intensity and 8 or 9 ms for intensities close to the threshold. Insert earphones will delay the response less than 1 ms but 10 ms still can guarantee the inclusion of all the components.

4. **Number of Sweeps**

   The number of sweeps required varies due to the inherent amplitude of the response and the amount of background noise, such as muscle activity, mains noise (50 Hz), spontaneous EEG, and so on. For ABR recording, 1000-2000 are usually sufficient to obtain a clear response in a quiet patient fairly without movement, at high stimulus intensity. At low intensity near threshold more sweeps may be required to get an adequate signal-to-noise ratio.

5. **Artefacts**

   When collecting ABR data, many sources of noise, also called artefacts, mainly classified as physiological and non-physiological noise will interference the quality of the data. Physiological sources here refer to muscle activity, often arising from neck and jaw muscles (Hall, 1992b). These signals have a low frequency and a large amplitude, and can be eliminated by instructing the patient to relax or sleep, or raising the high-pass filter. Non-physiological sources of noise include electrostatic potentials, eletromagnetic interference, internal instrument noise, power line radiation (50 Hz, UK) and stimulus transducer radiation (Hyde, 1985). Among these non-physiological artefacts, the 50 Hz electrical power can be reduced by setting up a notch filter, the stimulus interference (excessively early artefacts) can be eliminated by moving transducer (earphone) away from the electrode, and others can be removed by verifying a good 'ground', electrode impedance and etc. In the Chapter 6, the artefact rejection will be focused on the stimulus and movement artefacts (muscle and any randomly occurring artefacts).

2.2.5 Applications of the ABR

Assessment of hearing sensitivity and hearing screening

The primary audiologic application of the ABR is the assessment of hearing sensitivity, which is the determination of the ABR threshold. This is known to correlate well with behavioral hearing thresholds for mid- and high-frequency stimuli (Gorga et al., 1985). In particular, the ABR is used to predict the hearing sensitivity for difficult-to-test populations, such as infants and children. Nowadays, ABR is often applied in

infant hearing screening (Kileny, 1988; Herrmann et al., 1995; Mason and Herrmann, 1998).

The early detection and treatment of hearing impairment has become an increasingly important component of pediatric care, as their importance in the prevention and treatment of speech and language disorders and ear disease in children is recognized (Ozdamar et al., 1990). For preterm infants, newborns, infants less than 6 months old, and multiply-handicapped people, electrophysiologic methods are used to evaluate auditory function. ABR is used as an infant hearing screening tool for four main reasons: (1) Near-threshold stimuli can be used, allowing for the detection of even mild hearing impairments; (2) Each ear is tested separately; (3) It is a neurophysiological response not influenced by state (attention or sleeping) and anesthetic agents; and (4) The ABR are affected by maturation and neurological status, and provides additional diagnostic information.

The above indicates the importance and justifies the popularity of the ABR application for hearing screening. The question may arise, why not use Otoacoustic Emission (OAE) which is also widely used in hearing screening tasks. One of the main advantages of ABR screening is that it provides information not only on conductive hearing loss and cochlear pathology, as OAE screening does, but also on the more central auditory pathology up to the midbrain. The ABR technology tests the entire hearing pathway from ear to, and including the brainstem (see Figure 2.4). However, OAE technology (Transient Evoked Otoacoustic Emissions (TEOAE) and Distortion Product Otoacoustic Emissions (DPOAE)) only tests a portion of the hearing pathway from the outer ear to the cochlear (inner ear) (van Straaten, 1999). Thus the ABR is accepted as an accurate means of infant hearing screening.

Differential diagnosis

Another application of the ABR is for differential diagnosis of diseases of the eighth nerve and brainstem. The ABR is composed of several voltage deflections (peaks and troughs) which represent far-field synchronous activity produced by onset responses of neural elements and abrupt bends in the neural fiber tracts of the eighth nerve and the auditory brainstem pathway (Stegeman et al., 1987; Deupree and Jewett, 1988). The determination of ABR component latencies, amplitudes and wave shapes are used

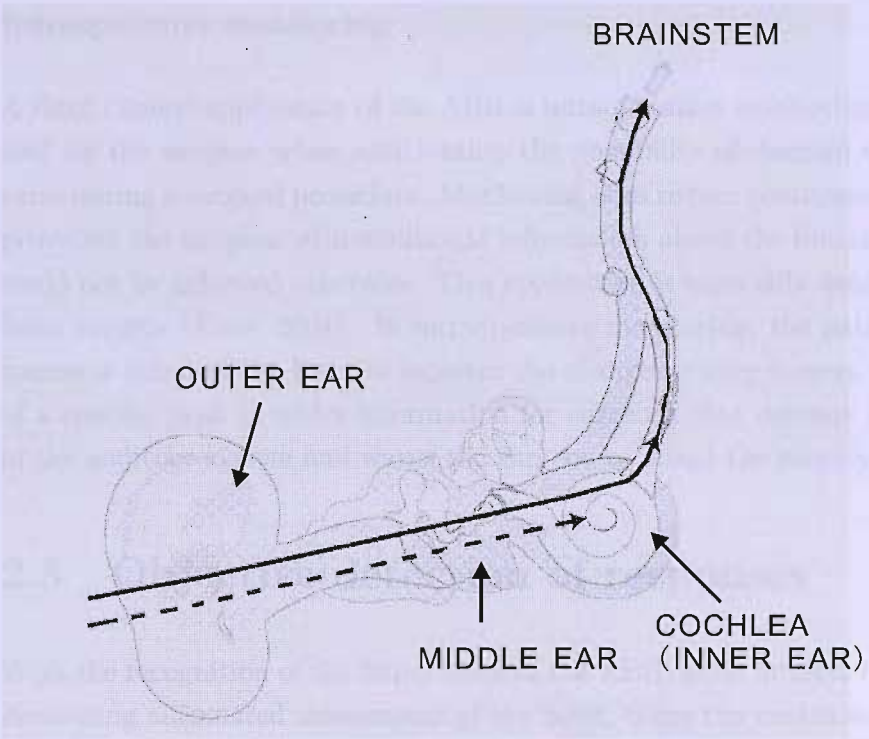


Figure 2.4: The pathway of ABR and OAE. The solid line indicates the entire hearing pathway of ABR from outer ear to brainstem. The dashed line shows the OAE's pathway from outer ear to the cochlea (inner ear). This figure is drawn based on the figure in the product brochure of echo-screen produced by Natus (US).

in indicating types of hearing loss and neurological problems. The abnormal ABR is correlated with structural brain lesions (brainstem tumors, vascular dysfunction, multiple sclerosis, and demyelination) (Starr and Achor, 1975; Starr and Hamilton, 1976; Stockard and Rossiter, 1977). For example, the lesions can result in prolongation of interpeak latencies (e.g., I-V delay) and abnormal peak amplitude ratios (V/I). The procedure is sensitive enough to detect mild loss due to otitis media with effusion (Mendelson et al., 1979). ABR is used to monitor changes in hearing sensitivity that may relate to the recovery from meningitis (Ozdamar et al., 1983; Ozdamar and Kraus, 1983) or hydrocephalus (Kraus et al., 1984).

Intraoperative monitoring

A third clinical application of the ABR is intraoperative monitoring which is a useful tool for the surgeon when anticipating the possibility of damage to auditory structures during a surgical procedure. Monitoring does reduce postoperative morbidity by providing the surgeon with additional information about the functional status which could not be achieved otherwise. This application is especially used during posterior fossa surgery (Katz, 2001). In intraoperative monitoring, the patient's baseline response is taken as the basis to monitor the changes during surgery. The degradation of a specific peak provides information for surgeons that damage may be occurring in the auditory system and warns the surgeon to adapt the surgery procedures.

2.3 Objective detection of responses

With the recognition of the importance of the ABR, more interest has been shown in developing automated assessments of the ABR. Since the traditional ABR screening is based on visual inspection, which is very costly in terms of personnel and administration time, various automated methods have been described in the time domain and the frequency domain, in the literature. In the following,the first five methods are carried out in the time domain and the other nine methods are applied in the frequency domain.

2.3.1 Cross-Correlation

Cross-correlation has been used for the quantitative assessment of the degree of similarity between two waveforms. Its applications to response detection fall into quite distinct classes: *replicate cross-correlation* and *template cross-correlation.*

For replicate cross-correlation, two independent averages are obtained for any particular stimulus condition. The well-known Pearson product-moment correlation coefficient is calculated. Perfect correlation (r=1) means that the two waveforms have identical shape. An r of -1 would be obtained if the two averages were mirror images. Values of r close to zero lead to acceptance of the null hypothesis of no response

whereas values approaching 1 suggest response presence. One of the earliest reports of replicate cross-correlation methods for objective detection of the ABR was by Weber and Fletcher (1980).

A detailed analysis of how to use the template cross-correlation techniques was reported by Ozdamar et al. (1990). The template cross-correlation technique differs from replicate cross-correlation in that one of the waveform is pre-established, usually from prior normative data regarding the expected or desired response waveform. Since the mechanics of calculation are the same as replicate cross-correlation, this method is highly directed toward determining the extent to which the observed data resemble the desired target. Thus the template is very important. If the template is correct, the method will be more powerful than replicate cross-correlation method which makes less use of response information. However, if the template is wrong, great loss of power of this method may occur.

2.3.2 F_{sp}

F_{sp} technique is a well-known response detection algorithm based on statistical principles (Elberling and Don, 1984). This technique quantifies the noise contribution in the recording by specifying a single digitized point (or small number of points) in the response window and calculating the sweep-to-sweep variance of the amplitude measured at that point. Because the contribution from the evoked potentials at any fixed point in time should be the same for each sweep, the only contribution to the sweep-to-sweep variance should be the noise. The variance across successive points in the average is also measured. This value represents the overall energy in the average, which includes signal and noise. In a large response with good resolution of peaks, this variance across the window would be large. The F_{sp} statistic is then defined as:

$$F_{sp} = \frac{var(\overline{ABR})}{var(SP)/K} \tag{2.3.1}$$

where var $(\overline{ABR})$ is the variance within the averaged ABR between say 5 and 15 ms after the onset of the stimulus and $var(SP)$ is the variance of a single point, say 10 ms, after stimulus onset calculated across all the segments recorded and K is the number of segments. As the averaging process reduces the noise, the denominator of

the ratio is reduced more quickly than the numerator, if a response is present. When no response is present, the F_{sp} will be close to 1.0, and the growth curve will be flat because both numerator and denominator contain only noise. In fact, the slope of the growth curve has been shown to be correlated to the number of sweeps and the click level. With the increase of the number of sweeps and click level, the F_{sp} value grows.

2.3.3 ± difference

Wong and Bickford (1980) suggested a technique called ± difference. For this technique, when an ABR is averaged, alternate sweeps are put into different buffers to generate two averages (alternatively the ABR can be acquired twice to produce two averages). Then, noise is estimated and quantified by subtracting one averaged waveform from the other, and the ABR is estimated as the sum of the two averages. ± difference is defined as:

$$\pm\, difference = \frac{var(Sum)}{var(Diff)} \tag{2.3.2}$$

where var is the variance, Sum and Diff are found by addition and subtraction of the two averages respectively. In terms of their experience, Wong and Bickford indicated that if ± difference > 2, an ABR is likely present at or near threshold stimuli.

2.3.4 Friedman test

Cebullar et al. (2000) proposed a q-sample test, based on Friedman's two way analysis of variance (Friedman, 1937; Altman, 1991) to objectively detect the ABR. In order to understand the definition of the test statistic, the data could be considered as a matrix as described in equation 2.2.1.

Where K is the number of sweeps, M is the number of samples in each sweep. Each row in the matrix represents one sweep. In the q-sample test, the M amplitudes of each row of the data matrix are ranked and the amplitudes are replaced by their ranks. Then the test statistic is:

$$\hat{\chi}^2_R = \left(\frac{12}{KM(M+1)} \sum_{i=1}^{M} R_i^2 \right) - 3K(M+1) \tag{2.3.3}$$

where Ri is the sum of the ranks in the ith sample (ith column). The null hypothesis is that the amplitudes in all columns are same, i.e., there is no significant difference between samples. Under the null hypothesis, the statistic has a chi-squared distribution with $M-1$ degrees of freedom. According to the significance level (α), the decision can then be made to accept or reject the null hypothesis.

2.3.5 Cochran's Q-test

Cochran's Q-test was originally proposed in (Cochran, 1950). In the data matrix (equation 2.2.1), for this test the amplitude values are replaced by their signs (+/-) of the amplitude values. The test statistic is:

$$Q = \frac{(M-1)\left[M\sum_{i=1}^{M} T_i^2 - \left(\sum_{i=1}^{M} T_i\right)^2\right]}{M\sum_{n=1}^{K} L_n - \sum_{n=1}^{K} L_n^2} \tag{2.3.4}$$

where L_n is the sum of the number of '+' signs in the nth row of the data matrix and T_i is the sum of the number of '+' signs in the ith column. Under the null-hypothesis H_0 of no signal added (no response), the probability functions is estimated by the histogram of the Q statistic values on Monte-Carlo simulations. Based on the choice of the significance level, the critical value of Q is determined.

Now nine approaches used in the frequency domain will be introduced. Some of the methods only considered the phase, and others are based on both the phase and amplitude. Some statistics are calculated based on polar coordinates, and others on Cartesian coordinates as shown in Figure 2.5.

2.3.6 Magnitude-squared coherence (MSC)

The magnitude-squared coherence function (MSC) is formally defined as a squared, normalized cross-spectral density function relating input and output time series (Dobie and Wilson, 1994). MSC is estimated, for a given frequency, as the ratio between the power in the grand average (power of the mean denoted by PM) and the mean

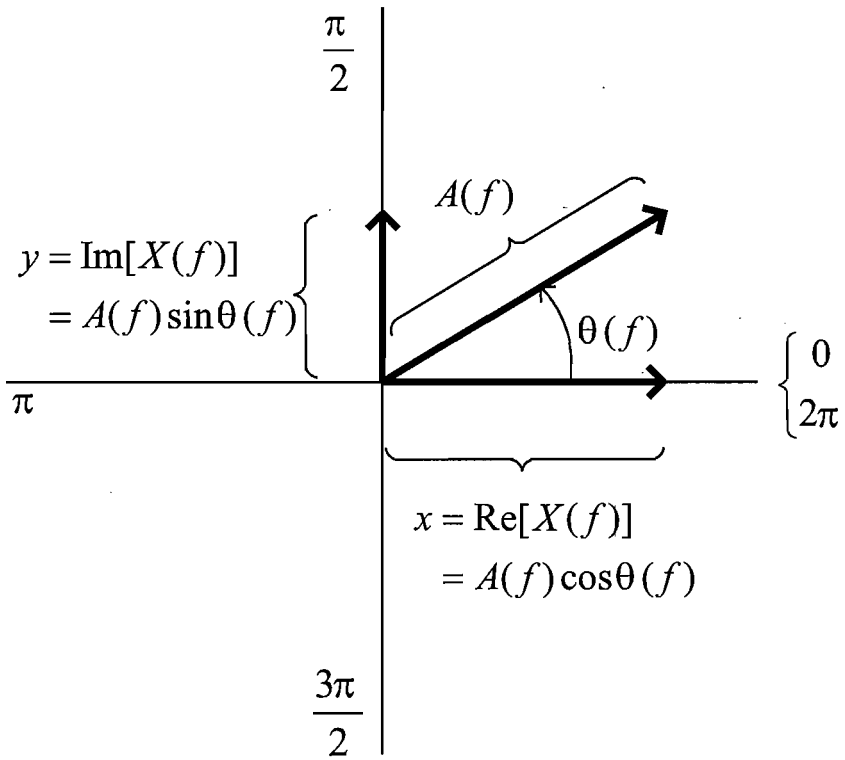


Figure 2.5: The amplitude and phase of a single frequency component (f) of a signal $X(f)$ are shown in polar coordinates (A(f) and $\theta(f)$). This vector can be equivalently represented by its Cartesian coordinates, the real (x) and imaginary (y) parts of $X(f)$.

power in the subaverages (mean power is written as MP). The subaverages are obtained by breaking up a single evoked potential average, e.g., 16 subaverages of 256 responses each instead of a single average of 4096. MSC is defined as (Dobie and Wilson, 1989):

$$MSC = \frac{PM}{MP} = \frac{L \bullet Y_{grand}^2}{\sum_{i=1}^{L} Y_i^2} \tag{2.3.5}$$

where Y_{grand} is the Fourier (complex) representation of the grand average at the frequency of interest, Y_i is the complex representation of *ith* subaverage, and L is the number of the subaverages. The MSC ranges from 0 to 1, and makes use of

both amplitude and phase information. Given L and the desired false positive (α) rate, critical values are available to determine whether or not a response is present at the stimulus frequency (or its harmonics) (Dobie and Wilson, 1996). The MSC identifies those frequencies contributing significantly to an evoked potential. This information can also be useful in specifying analog and/or digital filter parameters and sampling frequencies for improving evoked potential detection in time domain waveforms (Dobie and Wilson, 1989).

2.3.7 Phase coherence (PC)

The phase coherence method also used grand average and subaverage data, and was originally proposed by Jerger et al. (1986) for evaluating the effects of sleep on the auditory steady state evoked potential (SSEP). Similar to MSC, the raw signal was broken into L subaverages and these subaverages were then Fourier transformed. At the frequency of interest, i.e., 40 Hz, the phases of each subaverage was considered. These phase angles were then projected onto a unit circle, and their sines and cosines were separately averaged. For this frequency, PC is calculated from the sine and cosine values (see Figure 2.5) in the subaverages (L is the number of subaverages) (Dobie and Wilson, 1994; Jerger et al., 1986):

$$PC = \left[\left(\frac{1}{L} \sum_{i=1}^{L} \cos\theta_i \right)^2 + \left(\frac{1}{L} \sum_{i=1}^{L} \sin\theta_i \right)^2 \right]^{1/2} \tag{2.3.6}$$

where θ_i is the phase angle of the Fourier component of the ith subaverage. The PC value varies from 0 to 1 and quantifies the degree to which the phases of the frequency of interest are dispersed. Since the sine and cosine of an angle can vary from -1 to 1, randomly dispersed angles of the subaverages will have sines and cosines whose averages approach zero (PC=0, no response present). Conversely, a group of nearly identical angles will lead to average of sines and cosines forming a right triangle with hypotenuse approaching 1 (PC=1, a response present).

For a given number of phase values (subaverages), Mardia (1972) gave critical values for the PC. For sets of 16 phase values, the critical values of PC are 0.429 and 0.525

respectively for $\alpha = 5\%$ and $\alpha = 1\%$. This was validated by Dobie and Wilson (Dobie and Wilson, 1994).

In fact, if the amplitudes of all subaverages are set to a constant prior to analysis, i.e., if the amplitude information is ignored, the relationship between MSC and PC will be:

$$PC = \sqrt{MSC} \tag{2.3.7}$$

Champlin (1992) used MSC and PC to detect 40 Hz auditory steady-state potentials in normal human subjects and MSC was only slightly superior to PC. Similar investigation was performed by Dobie and Wilson (1994), who did not find any difference in the response detection performance between MSC and PC.

2.3.8 Rayleigh test

The Rayleigh test was originally described in Mardia (1972) and used to evaluate the null hypothesis that a sample of angular observations has arisen from the uniform circular distribution. For each frequency of interest, the degree of dispersion or aggregation of the phase is taken as a measure of noise (random) versus genuine response (clustered). The test statistic of the Rayleigh method is written as (the meanings of the variables are the same as those defined in PC method):

$$R = \sqrt{C^2 + S^2} \tag{2.3.8}$$

with $C = \frac{1}{L}\sum_{i=1}^{L}\cos\theta_i$; and $S = \frac{1}{L}\sum_{i=1}^{L}\sin\theta_i$

When substituting C and S into 2.3.8, $R = PC$ is obtained and the Rayleigh test is identical to PC method. Therefore in the literature, these two methods are mentioned together.

2.3.9 Modified Rayleigh test

In addition to the phase information, a modified modified Rayleigh test was proposed by Jervis et al. (1983), taking the spectral amplitude information into account. And

Moore (1980) offered an alternative modification of the Rayleigh test by considering the ranks of the spectral amplitudes and this modified test statistic was:

$$R_m^* = \frac{R_m}{L\sqrt{L}}, R_m = \sqrt{C_m^2 + S_m^2} \tag{2.3.9}$$

with $C_m = \frac{1}{L}\sum_{i=1}^{L} r_i \cos\theta_i$ and $S_m = \frac{1}{L}\sum_{i=1}^{L} r_i \sin\theta_i$.

where r_i=rank of the amplitude A_i (see Figure 2.5), $1 \leq r \leq L$. The critical value of the test statistic was determined from Monte-Carlo simulations. Cebullar et al. (1996) and Sturzebecher and Cebullar (1997) reported that the use of the ranks (Moore, 1980) instead of the spectral amplitudes themselves (Jervis et al., 1983) was more advantageous in that ranking reduces the influence of non-stationary noise and the artefacts as well. But the detection performances of both modifications of the Rayleigh test were similar when non-stationary noise was absent. And both methods were superior to the Rayleigh test (Cebullar et al., 1996; Sturzebecher and Cebullar, 1997), when applying on the near-hearing threshold ABR signals.

2.3.10 Hotelling's T^2

Hotelling's T^2, like the MSC and modified Rayleigh test, considers both amplitude and phase information and is the multivariate analogue of the well-known t test (Hotelling, 1931). It was applied to frequency domain ABR detection by Valds-Sosa et al. (1987) and 40 Hz steady-state evoked potentials by Picton et al. (1987). Similar to the above frequency methods, the subaverages were Fourier transformed. For each frequency of interest, the multiple subaverage spectral estimates (L estimates on L subaverages which were obtained as the way described in the MSC method) can be displayed in the complex plane as a swarm of points, or a group of vectors.

Since the T^2 statistic is related the statistic of t distribution, thus description will start with a univariate distribution and its t statistic. If a univariate distribution is sampled L [2] times to give a sample mean of $\overline{x}$ and a sample standard deviation of

[2] L used here is in accordance with the number of the subsaverages spectral estimates.

s, the actual mean of the population occurs within the limits of x described by the inequality:

$$\frac{\sqrt{L}(\overline{x} - x)}{s} \leq t \tag{2.3.10}$$

where t is taken from the two-tailed Student's t distribution with $L - 1$ degrees of freedom. For a multivariate distribution, the confidence region for the mean vector is given by the equation (Anderson, 1984):

$$L(\overline{x} - x)' S^{-1} (\overline{x} - x) \leq T_0^2 \tag{2.3.11}$$

where S^{-1} is the inverted variance-covariance matrix of the sample and T_0^2 is derived from F distribution by:

$$T_0^2 = \frac{(L-1)k}{L-k} F \tag{2.3.12}$$

where k is the dimension of the multivariate vector and the degrees of freedom of F are k and $L - k$. The spectral estimates of the ABR are a two-dimensional vector, the confidence region for the mean of this vector is plotted as an ellipse. If this ellipse does not include zero (origin (0,0)), the ABR recording can be considered significantly different from zero at the probability for which the ellipse is determined.

The great advantage of using Hotelling's T^2 is that it is convenient and straightforward. The Rayleigh test, as also PC, is essentially an amplitude-free version of the Hotelling's T^2 test.

2.3.11 Circular T^2

The Hotelling's T^2 method ignores relationships between the real and imaginary parts of the Fourier components. With recognition of this, Victor and Mast (1991) proposed a new circular T^2 (T^2_{circ}) test which is specifically designed for the analysis of variability of Fourier components, with an assumption of equal variances for the real and imaginary components of the Fourier vectors.

The Fourier components of the L subaverages are written as the cartesian representation $z = x + iy$ as described in Figure 2.5, and the real quantities x and y represent the cosine and sine components of the response. The L estimates of the complex Fourier components are denoted by $z_1, z_2, ..., z_L$, and their mean value is denoted by $\langle x \rangle_{est} = (\sum z_j)/L$ and the population mean [3] is denoted by ζ. If the set of experimental estimates are indeed drawn from a population whose mean is equal to ζ, there are two independent estimates of the population variance V of real and imaginary parts (Victor and Mast, 1991). One is V_{indiv}, derived from the scatter of the individually determined components z_j about their mean. There are $2(L-1)$ degrees of freedom, since the means of the x_js are constrained to be $\langle x \rangle_{est}$ and the means of the y_js are constrained to be $\langle y \rangle_{est}$. Thus, one estimate of the population variance V is:

$$\begin{aligned} V_{indiv} &= \frac{1}{2(L-1)} \sum_{j=1}^{L} [(x_j - \langle x \rangle_{est})^2 + (y_j - \langle y \rangle_{est})^2] \\ &= \frac{1}{2(L-1)} \sum_{j=1}^{L} |z_j - \langle z \rangle_{est}|^2 \end{aligned} \tag{2.3.13}$$

The other estimate of the population variance is based on the assumed population mean ζ. As $\langle x \rangle_{est}$, the sample mean, is the mean of L independent estimates, both its real and imaginary parts have variance V/L about the population mean ζ. Therefore, the second estimate of V:

$$\begin{aligned} V_{group} &= \frac{L}{2}[(\langle x \rangle_{est} - \xi)^2 + (\langle y \rangle_{est} - \eta)^2] \\ &= \frac{L}{2} |\langle z \rangle_{est} - \zeta|^2 \end{aligned} \tag{2.3.14}$$

Under the hypothesis that the experimental data z_j are samples of a population whose mean is ζ, each of V_{indiv} and V_{group} are estimates of the variance V derived from independent quantities. Therefore, the ratio of V_{group}/V_{indiv} is an F distribution (Sokal and Rohlf, 1995), with 2 and $2(L-1)$ degrees of freedom for numerator and

[3] A population is any collection of individuals of interest, where these individuals may be anything, and the number of individuals may be finite or infinite (Bland, 2000). The population mean is the mean from these individuals (population), and here refers to the mean of all possible Fourier components.

denominator, respectively. In order to make a close analogy with the Hotelling's T^2 statistic, T^2_{circ} is defined as the variance ratio V_{group}/V_{indiv} normalized by the number of the samples L (Victor and Mast, 1991):

$$\begin{aligned} T^2_{circ} &= \frac{1}{L}\frac{V_{group}}{V_{indiv}} \\ &= (L-1)\frac{|\langle z\rangle_{est} - \zeta|^2}{\sum_{j=1}^{L} |z_j - \langle z\rangle_{est}|^2} \end{aligned} \tag{2.3.15}$$

Since $L \cdot T^2_{circ}$ is distributed according to $F_{[2,2L-2]}$, under the null hypothesis that no signal is present (assumed population mean $\zeta = 0$), the critical value of T^2_{circ} at the pre-set significance level (α) can be obtained by:

$$T^2_{circ(\alpha)} = \frac{1}{L} F_{(\alpha)[2,2L-2]} \tag{2.3.16}$$

Looking at Figure 2.5, graphically, the T^2 statistic considers the cluster of estimates of Fourier components to form an ellipse whose axes and orientation are unknown. For T^2_{circ}, the assumption of equal variances and zero covariance are made and those correspond to that the cluster of estimates of Fourier components is circularly symmetric. This increases the number of degrees of freedom from $L-2$ (degree of freedom in T^2) to $2L-2$ (that in T^2_{circ}), and the more information of the variances and covariances make circular T^2 a more precise statistical test.

Victor and Mast (1991) applied both T^2 and T^2_{circ} on the steady-state evoked potentials, and demonstrated that circular T^2 performed better than T^2; T^2_{circ} detected signals earlier than the Hotelling's T^2 (Hotelling, 1931) and the confidence regions derived from T^2_{circ} are consistently smaller than those derived from T^2; and the calculation of the critical value by T^2_{circ} is simpler than that for T^2.

Dobie and Wilson (1993) reported that the circular T^2 was a simple algebraic transform of the MSC, with identical statistical power as MSC, and was superior to PC for response detection.

2.3.12 F test for power spectral density

Zurek (1992) pointed out that power estimates at both f_s (known frequency) and neighbouring frequencies are distributed as a chi-square distribution. Thus their ratio can be tested as an F statistic. At each frequency, measured power is the sum of two independent squared variables (real and imaginary components); thus, at f_s, the power estimate is a chi-square variable with two degrees of freedom ($df = 2$). If noise power is estimated by averaging across m neighbouring frequencies, this unbiased estimate ($\hat{P}_n$, power of background noise) is a chi-square variable with $df = 2m$ (Dobie and Wilson, 1996). The power ratio is calculated as:

$$F = \frac{\hat{P}_{(s+n)}}{\hat{P}_n} \tag{2.3.17}$$

where $\hat{P}_{(s+n)}$ is an unbiased estimate of the sum of signal power and noise power at the stimulus frequency or its harmonics. The statistical significance of F can then be tested using standard tables or statistical software programs with $df = 2, 2m$. Furthermore an unbiased estimate of SNR can be calculated:

$$\hat{P}_s = \hat{P}_{(s+n)} - \hat{P}_n \tag{2.3.18}$$

$$S\hat{N}R^2 = \frac{\hat{P}_s}{\hat{P}_n} = \frac{\hat{P}_{(s+n)} - \hat{P}_n}{\hat{P}_n} = F - 1 \tag{2.3.19}$$

$$S\hat{N}R = \sqrt{F - 1} \tag{2.3.20}$$

For a selected significance level (1% or 5%, false positive rate), the critical value for SNR estimates can be obtained. If it is smaller than the SNR, a 'significant' response is detected.

Detection performance for the F test increases rapidly with the number of neighbouring frequencies used to estimate noise power (a range between 3 and 7 was tested in Zurek (1992)). However, using large numbers of neighbouring frequencies adds computational time without improving performance; indeed, performance may even be degraded if the background noise is not white.

The F test was applied in another way by Simpson et al. (2000) who defined the spectral F test (SFT) as the ratio of the power spectra obtained from the EEG during (with stimuli) and before or after stimulation (without stimuli). By comparing the performance of MSC, phase-synchrony measure (PSM), which is related to phase coherence (PC) and SFT, Simpson concluded that the SFT showed much poorer results, requiring a far higher number of data segments (K) or SNR in order to achieve the same detection rate as the MSC or PSM.

2.3.13 q-sample uniform scores test

The q-sample uniform scores test is to check the null hypothesis H_0 that q samples are taken from populations with the same continuous distribution. The q-sample uniform scores test was first proposed for auditory evoked potential detection by Mardia (1972), considering the phase angle in the form of their ranks whereas the amplitude information is neglected.

Let X_{mk}; $1 \leq m \leq FM$, $1 \leq k \leq L$ be a collection of phase angles; L is the number of samples with the sample size FM (in order to differ from the sample size M in the time domain). There are $FN = L \times FM$ phase angle values, which are ranked in a signal sequence. Let r_{mk}, $m = 1, 2, ..., FM$, be the ranks of the phase angles in the kth sample.

The phase angles are then replaced by the uniform scores

$$\beta_{mk} = \frac{2 \cdot \pi \cdot r_{mk}}{FN} \tag{2.3.21}$$

The test statistic used is

$$W = \frac{2}{FM} \sum_{k=1}^{L} (C_k^2 + S_k^2) \tag{2.3.22}$$

with $C_k = \sum_{m=1}^{FM} \cos\beta_{mk}$; and $S_k = \sum_{m=1}^{FM} \sin\beta_{mk}$.

W is distributed as Chi-squared with $2(L-1)$ degrees of freedom. The critical value is obtained from the Chi-squared distribution.

2.3.14 Modified q-sample uniform test

Additional to the phase angles, a modified q-sample uniform test was developed by Sturzebecher et al. (1999), taking the amplitude information into account. Similar to the phase angles, the spectral amplitudes A_{mk} are ranked in a single sequence. a_{mk} is the rank of the amplitude A_{mk} in the kth sample. The phase angles are also replaced by the uniform scores (Equation 2.3.21).

The test statistic for the modified q-sample uniform test is

$$W^* = \frac{2^2}{L^2(L+1)^2}\frac{120}{FM}\sum_{k=1}^{L}(C_k^{*2} + S_k^{*2}) \tag{2.3.23}$$

with $C_k^* = \sum_{m=1}^{FM} a_{mk}\cos\beta_{mk}; S_k^* = \sum_{m=1}^{FM} a_{mk}\sin\beta_{mk}$

and

$$\beta_{mk} = \frac{2\cdot\pi\cdot r_{mk}}{FN}$$

The modification is not derived mathematically (Sturzebecher et al., 1999). The amplitude in the form of the ranks of the spectral amplitudes was introduced analogy to the modification of the Rayleigh test proposed by Moore (1980). The critical values are derived from Monte Carlo simulations.

2.4 Overview of the bootstrap technique and its applications

2.4.1 Introduction

Bootstrap methods are computer-intensive methods of statistical analysis that use simulation to calculate standard errors, confidence intervals and significance tests (Davison and Hinkley, 1997). Bootstrap technique received considerable attention and were initially introduced by Efron (1979a) due to the availability of affordable and powerful computers. Bootstrap is a statistical method for estimating the sampling

distribution of an estimator by sampling with replacement from the original sample. Particularly when conventional methods cannot be applied, the bootstrap technique can be used to derive robust estimates of standard errors and confidence intervals (Efron and Gong, 1983) of a population parameter like a mean, median, proportion, odds ratio, correlation coefficient or regression coefficient, and it can also be used to construct hypothesis tests.

Efron investigated and discussed the connections between the various nonparametric methods, and also the relationship to familiar parametric techniques (Efron, 1979a,b, 1981b,a). The bootstrap method is shown to be successful in many situations. In fact, it is better than some other asymptotic methods, such as the traditional normal approximation.

In this study, the most important contribution is to apply the bootstrap technique in a significance test to detect the auditory evoked potentials. To the best of our knowledge, the bootstrap technique has not been used for this purpose before. Therefore understanding the bootstrap principles and theorem is important and we will briefly describe these in the next section. Since it was introduced in 1979, bootstrap methods have been widely used in statistics, life sciences, medical sciences, social sciences and business (Davison and Hinkley, 1997). As the fields exceed the scope of the current work, the focus will be on previous applications in biomedical signal processing, in the final section of this chapter.

2.4.2 Principles and theorem

Suppose we observe $X_i = x_i$, $i = 1, 2, ..., n$, where X_i is independent and identically distributed (i.i.d.) according to some probability function F (X_1,X_2, ... ,$X_n \sim F$). The sample is studied in order to estimate a certain parameter, maybe the mean, median, correlation, and so on, generally denoted as $\theta(F)$, associated with the distribution F. The estimate is then given by $\hat{\theta} = \theta(\hat{F})$, where $\hat{F}$ is the empirical distribution function putting mass $1/n$ at each observed value x_i.

A way to obtain the distribution of $\hat{\theta}$ or its characteristics is to repeat the experiment a sufficient number of times and approximate the distribution of $\hat{\theta}$ by the empirical distributions. However, in many practical situations, this method is inapplicable for cost reasons or because the experimental conditions are not reproducible.

The bootstrap method suggests that a distribution, i.e. the sample (or empirical) distribution, $\hat{F}$, could be resampled, and the distribution can approximate F as sample size $n \rightarrow \infty$.

There are two situations to distinguish, the parametric and nonparametric. When there is a particular mathematical model, with an adaptive constants or parameter ψ which fully determine F, such a model is called *parametric* and statistical methods based on this model are parametric methods. In this case the parameter of interest θ is a component of or function of ψ. When no such mathematical model is applied, the statistical analysis is *nonparametric*, and uses only the fact that the random variables X_i are i.i.d (Davison and Hinkley, 1997). Even if a parametric model could be applied, a nonparametric analysis is still useful and helpful to assess the robustness of conclusions drawn from a parametric analysis.

Therefore the choice of $\hat{F}$ is not unique, any distribution that can approach F as $n \rightarrow \infty$, can be used. The parametric bootstrap method is a particular case with partial information on F. But in most cases F is unknown, and we will concentrate on the principles of nonparametric bootstrap.

The basic bootstrap principle

The nonparametric bootstrap procedures could be performed as the following steps.

Step 1. Conduct the experiment to obtain the random sample $x = \{X_1, X_2, ..., X_n\}$ and calculate the estimate $\hat{\theta}$ from the sample x.

Step 2. Construct the empirical distribution, $\hat{F}$, which puts equal mass, $1/n$, at each observation, $X_1 = x_1, X_2 = x_2, ..., X_n = x_n$.

Step 3. From the empirical distribution, $\hat{F}$, draw a random sample of size n with replacement. This is a *resample*. Calculate the statistic of interest, $\hat{\theta}$, for this resample, yielding $\hat{\theta}^*$.

Step 4. Repeat step 3 B times, where B is a large number, in order to create B samples. The practical size of B depends on the tests to be run on the data. This will be discussed later.

Step 5. Construct the relative frequency histogram from the B number of $\hat{\theta}^*$ by putting a probability of $1/B$ at each point, $\hat{\theta}_1^*, \hat{\theta}_2^*, ..., \hat{\theta}_B^*$. The distribution obtained is the bootstrapped estimate of the sampling distribution of $\hat{\theta}$. This distribution can now be used to make inferences about the parameter θ, which is to be estimated by $\hat{\theta}$.

Example 1: The bootstrap principle for calculating a confidence interval for the mean

Let $X_1, X_2, ..., X_n$ be n i.i.d. random variables from some unknown distribution, and suppose an estimator and a $(1-\alpha)100\%$ interval for the mean μ are to be found. Traditionally, μ is estimated by the sample mean $\hat{\mu}$

$$\hat{\mu} = \frac{X_1 + ... + X_n}{n} \tag{2.4.1}$$

A confidence interval for μ can be found by determining the distribution of $\hat{\mu}$ by drawing repeated samples of size n from the underlying distribution (if the experiment can be repeated), and calculating the upper and lower limit of $\hat{\mu}$ to meet

$$P(\hat{\mu}_L \leq \mu \leq \hat{\mu}_U) = 1 - \alpha$$

The distribution of $\hat{\mu}$ depends on the distribution of the X_i, i.e. F, which is unknown. When n is large enough, the distribution of $\hat{\mu}$ could be approximated by the normal distribution, but in many cases n is small and this approximation is not valid.

The nonparametric bootstrap provides a way to estimate the confidence interval. The principles of that is described as follows:

Step 1. *Experiment.* Perform the experiment. The size of random variables is n and the estimated mean $(\hat{\mu})$ is calculated by Equation 2.4.1.

Step 2. *Resampling.* Using a pseudo-random number generator with a uniform distribution, draw a random sample of n values, with replacement. Thus some of the original sample values appear more than once, and some not at all. Denote this as x^*.

Step 3. *Calculation of the bootstrap estimate.* Calculate the mean of all values in x^*.

Step 4. *Repetition.* Repeat step 2 and 3 B times and obtain B bootstrap estimates of the mean $\hat{\mu}_1^*, \hat{\mu}_2^*, ..., \hat{\mu}_B^*$.

Step 5. Approximation of the distribution of $\hat{\mu}$. Sort the bootstrap estimates into increasing order to obtain $\hat{\mu}_{(1)}^* \leq \hat{\mu}_{(2)}^* \leq ... \leq \hat{\mu}_{(B)}^*$, where $\hat{\mu}_{(k)}^*$ is the kth smallest of $\hat{\mu}_1^*, \hat{\mu}_2^*, ..., \hat{\mu}_B^*$.

Step 6. Confidence interval. The desired $(1-\alpha)100\%$ bootstrap confidence interval is $(\hat{\mu}_{(q_1)}^*, \hat{\mu}_{(q_2)}^*)$, where $q_1 = B\alpha/2$, and $q_2 = B - q_1 + 1$. Note q_1, q_2 must be integers and the selection of B should consider this point.

Example 2: The bootstrap principle for estimating standard errors

The sample mean $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i$ is an estimate of μ and has expectation μ and variance σ^2/n. The standard error $(se(\hat{\mu}))$ of the sample mean $\hat{\mu}$ is the square root of its variance,

$$se(\hat{\mu}) = \sqrt{var(\hat{\mu})} = \sigma/\sqrt{n} \tag{2.4.2}$$

Standard error is a general term for the standard deviation of a summary statistic and is the most common way of indicating statistical accuracy (Efron, 1993).

The bootstrap estimate of $se(\hat{\theta})$ is the standard error of $\hat{\theta}$ for data sets of size n randomly sampled from $\hat{F}$, which is defined by $se_{\hat{F}}(\hat{\theta}^*)$. The bootstrap procedures for estimating standard error are very similar as those for calculating the confidence interval and are therefore described by five steps: (1) Experiment; (2) Resampling; (3) Calculation of the bootstrap estimate; (4) Repetition; and finally (5) Estimating the standard error as the sample standard deviation of the B replications,

$$\hat{se}_B = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}[\hat{\theta}_b^* - \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}_b^*]^2} \tag{2.4.3}$$

Example 3: The bootstrap principle for testing a hypothesis

The use of bootstrap technique for hypothesis testing is the key application in the current work. This work is based on the application of bootstrap technique for testing a specific hypothesis, i.e. that there is no significant response present and rejecting the null-hypothesis at a known significance level (the false positive rate). The parameters representing particular features of the signal are calculated and tested against the null-hypothesis. For this, the sample distribution of the parameter is required, but it may be difficult to derive from theory or conventional experiment procedures (e.g. repeated experiment on the same condition for many times), because of the large inter and intra-subject variability, and time variation.

For detecting the auditory brainstem responses, the parameters (θ) are chosen to represent the strength of the response, such as the power of the signal or the dynamic range. First of all, the parameter θ is obtained from the coherent average (described in the previous section), it is then tested against the null-hypothesis that there is no response present, using the bootstrap technique. If no response is present in the coherent average, one could expect the parameter θ and estimated parameter (denoted by θ^*) from bootstrap resampling to be of compatible value. However, if θ is greater than any of the θ^*, there should be something special about the response following the onset of the stimuli, compared to the data randomly selected within bootstrap process. The detailed procedures of the bootstrap method based on ABR will be introduced in Chapter 5.

2.4.3 Applications

The bootstrap method is an attractive tool for assessing the accuracy of estimators and testing hypothesis for parameters in small data-sample situations. From its introduction by Efron (1979a), it has been developed intensively and also applied in many fields: biomedical engineering , radar signal processing, geophysics, control, vibration analysis, and artificial neural networks. A review of the application in some areas will be given in the following, and the focus will be placed on the applications in biomedical engineering, particularly for the signals from the brain, as there are of most concern in the current work.

Biomedical engineering

The bootstrap methods have been used to compared evoked responses in psychophysiological experiments, such as to compare the difference of the cerebral response (even-related potentials, ERP) to 'old' and 'new' words (Nocera and Ferlazzo, 2000). In this case, the mean value of responses is found to be different in the latency range of 400-800 ms after presenting the words. In order to test whether this difference is significant, the bootstrap significance test was applied under the null-hypothesis of 'no difference present'. The procedures will be further described in Chapter 7 and will then be employed on the ABR data. Similar methods were used to compare measures other than the mean of the coherent averages, such as the maximum (Neelon et al., 2006a,b). The maximum is less amenable to traditional statistical analysis than the mean.

Event-related changes of energy in different EEG frequency bands are an important indicator of the underlying brain processes. Sensory processing and motor behavior are connected with the localized decrease of power in certain frequency bands and this phenomenon is called event-related desynchronization (ERD). On the other hand, the increase in the power is named event-related synchronization (ERS) (Graimann et al., 2002). These effects relate to movement planning, and the ERD/ERS is defined as the change in power in particular frequency range. Identifying significant such changes in multi-channel EEG recordings has provided a further application of bootstrap methods (Graimann et al., 2002; Zygierewicz et al., 2005). They used the t-percentile method in which the data was converted into a 't-statistic' by subtracting the mean and dividing by the standard deviation, and then determined the bootstrap distribution. A similar approach was utilized on spectral estimators (parameters) based on Matching Pursuit (Durka et al., 2004). This parameter produced highly non-normal distributions, which can not be assessed by conventional parametric statistical analysis.

The applications above considered single channel of EEG signals. A number of problems in the brain research is localizing the source of cerebral activity within the skull, on non-invasive multi-channel electroencephalography and magnetoencephalography (EEG and MEG) signals. Bootstrap methods have been proposed to assess the uncertainty of the localization (Gross et al., 2003; Darvas et al., 2004; Rodriguez-Rivera

et al., 2006) by computing the confidence interval for example of local maxima of activation in Gross et al. (2003). The assignment of the local maxima of activation to specific anatomical structures can be used to test the differences in source localization in different experimental situations. The estimation of the location area from surface recordings is of course prone to estimation errors (Darvas et al., 2004). In order to estimate these errors, a process similar to that described in Graimann et al. (2002) can be adopted: the 'bootstrap' resamples are obtained by selecting the data epoch by epoch with replacement, and then average these to get new coherent averages. These then provide new estimates of the source localization, and their scatter of an estimate of estimation errors. The main benefit of the bootstrap method is that it takes the characteristics of the source localization algorithm and of the signals into account. The assumption for this is that each individual stimulus-response is 'typical', and any random combination of these responses 'may have occurred'.

The correlation and coherence between the EEG signals recorded from different locations have been extensively used to investigate the functional interactions between different brain regions, and bootstrap methods have been employed in Menon et al. (1996). Spatio-temporal correlations were calculated in the 20-50 Hz range of the signals recorded from the surface of the cerebral cortex. Bootstrap methods were used to test for significance and to investigate the spatial distance over which the recorded signals were significantly correlated. Whitcher et al. (2005) proposed the bootstrap method for testing the significance of the time-varying coherence between the local field potential picked up from the subthalamic nucleus and the EEG recorded over motor areas of the cerebral cortex.

The analysis of relationships between recording sites in intracranial EEG is essential to reveal active abnormal couplings and to detect possible causal relationships between signals. Here bootstrapping was also proposed (Chávez et al., 2003) in order to test the null-hypothesis of non-causality between two time series. One of the signals was randomly resampled and the other was left unchanged, and thus in accordance with the null-hypothesis, any temporal relationship between them was destroyed. They used the 'stationary bootstrap' (Politis and Romano, 1994) by resampling in blocks and block lengths follow a random geometric distribution. The stationary bootstrap was applied to recordings made within different regions of the brain and to identify the direction of casual origin between different cerebral regions during particular periods

in the seizures.

The relationship between slow eye movement (SEM) and EEG was assessed by means of product-moment correlations which has been analyzed by a bootstrap significance test (Gennaro et al., 2000). Conventional measures of the wake-sleep transition (i.e. slow eye movement) present some weaknesses. These include their variable rate in different subjects, discontinuity of the sleep onset process and the subject's position on the wakefulness/sleep continuum. Therefore in order to reduce the effect of the variability of the subjects, the estimate of the correlation coefficients between SEM percentages and power of EEG was performed for each subject separately. The small number of not-statistically independent signals for each subject may make conventional estimates unreliable. The bootstrap method provides an efficient solution to solve the problem.

Kannurpatti and Biswal (2005) introduced an application of bootstrap resampling in conjunction with cross-correlation to estimate the confidence intervals of activation-induced blood flow recorded over an area of the cortex. Bootstrap analysis can take the variability of noise between pixels in the blood flow image into account. Another application on blood flow was to assess the inter-relationship between blood flow (using Doppler Ultrasound) and arterial blood pressure in order to investigate the blood flow control system in the brain (Simpson et al., 2004). Constrained system identification, selecting one from a set of ten possible impulse responses (models) was used and bootstrapping applied for determining the empirical sampling distribution for the selected models.

Other applications

Nagaoka and Amai (1991) discuss a bootstrap application in which the distribution of the estimated 'close approach probability' is derived to be used as an index of collision risk in air traffic control.

Fisher and Hall applied the bootstrap in Geophysics to the problem of deciding whether or not paleomagnetic specimens sampled from a folded rock surface were magnetized before of after folding occurred (Fisher and Hall, 1989, 1990, 1991) . They conclude that the bootstrap method provides the only feasible approach in this

common paleomagnetic problem (Zoubir and Boashash, 1998). Another application in paleomagnetism has been reported in (Tauxe et al., 1991).

Kukreja et al. (2004) developed a bootstrap structure detection algorithm as a means of determining the structure of highly over-parameterized models of the system. It provides accurate estimates of parameter statistics without depending on assumptions made by traditional procedures and yields a parsimonious description of the system.

Zoubir and Bohme (1995) use bootstrap technique to construct multiple hypothesis tests for finding optimal sensor locations for knock detection in spark ignition engines.

Recently, bootstrap techniques were also applied in artificial neural networks. Tibshirani (1996) discussed a number of approaches for estimating the standard error of predicted values from a multi-layered perception. He found the bootstrap performed best, to some extent because they capture variability due to the choice of starting weights.

From a range of applications outlined above, the main benefits of the bootstrap method can be highlighted. First, bootstrap methods are very flexible, and can be easily applied to the study of unusual signal parameters. Second, bootstrap methods can be applied on small data sets. Finally, bootstrap methods make minimal assumptions about the data.

2.5 Statistical Analysis

Some of the statistical analysis methods using in this work will now be outlined.

2.5.1 Binomial distribution

Binomial distribution is the distribution followed by the number of successes in n independent trials when the probability of any single trial being a success is p (see Figure 2.6). The probability of r successes is

$$PROB(r \text{ successes}) = \frac{n!}{r!\,(n-r)!} p^r (1-p)^{(n-r)} \tag{2.5.1}$$

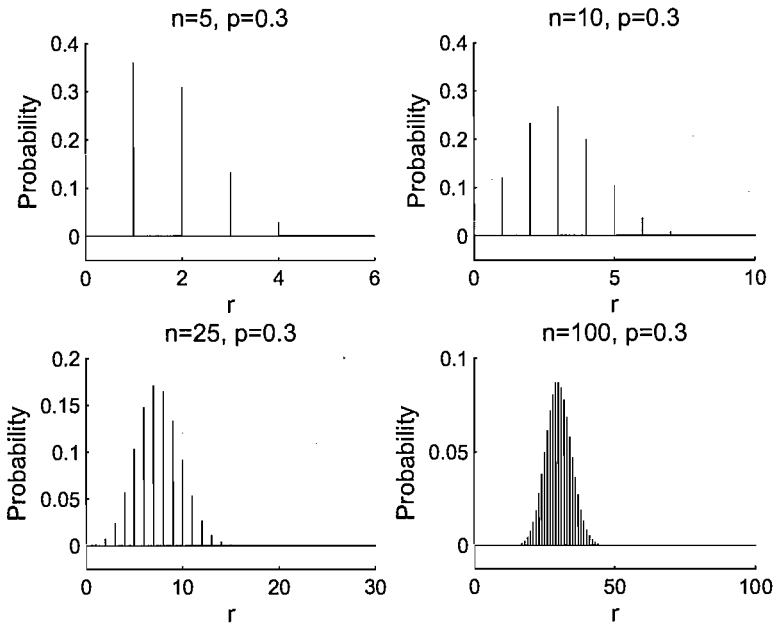


Figure 2.6: Binomial distributions with different number of independent trials ($n = 5, 10, 25, 100$), $p = 0.3$. The probability is obtained by setting different success rates r.

The binomial distribution is used in this study to estimate the acceptable range of false positive cases in 500 independent trials, with a single trial having a success rate of 0.05. Figure 2.7 shows the binomial distribution based on the above success rate and the number of independent trials. With a chosen confidence interval as containing 95% of this probability distribution, the success number could be in the range between 16 and 34, which represents the percentage of 3.2% and 6.8% respectively, as seen in the Figure 2.7. This means in 500 simulations (as used later in this work), the false positive rate in the range between 3.2% and 6.8% can be accepted as a reasonable rate.

2.5.2 Normal distribution

The reason we mention the normal distribution is that some of the following work is based on the normal distribution, which often provides a convenient approximation.

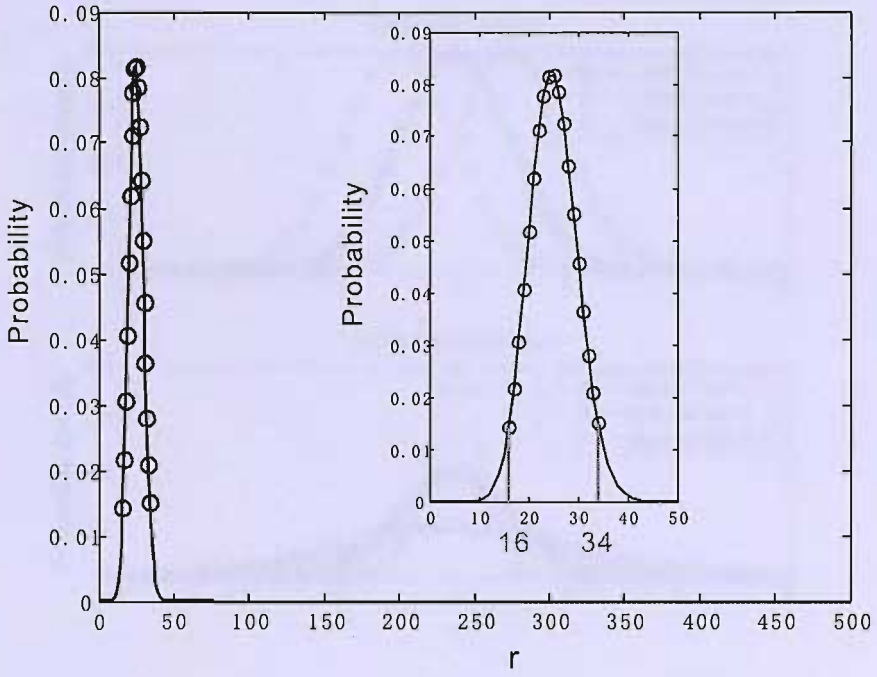


Figure 2.7: Binomial distribution with n=500 and p=0.05. The embedded figure is partly enlarged to emphasize the acceptable range (95% confidence intervals).

The **Normal distribution**, also known as the **Gaussian distribution**, is a family of distributions of the same general form, differing in their location and scale parameters: the mean ('average') and standard deviation ('variability'), respectively. The **standard normal distribution** is the normal distribution with a mean of zero and a standard deviation of one (the solid curves in the plots of Figure 2.8). It is often called the bell curve because the graph of its probability density resembles a bell.

Probability density function (PDF) of the normal distribution with mean μ and variance σ^2 (equivalently, standard deviation σ) is given by,

$$\begin{aligned} f(x;\, \mu,\, \sigma) &= \frac{1}{\sigma\sqrt{2\pi}} \exp(-\frac{(x-\mu)^2}{\sigma^2}) \\ &= \frac{1}{\sigma}\varphi(\frac{x-\mu}{\sigma}) \end{aligned} \tag{2.5.2}$$

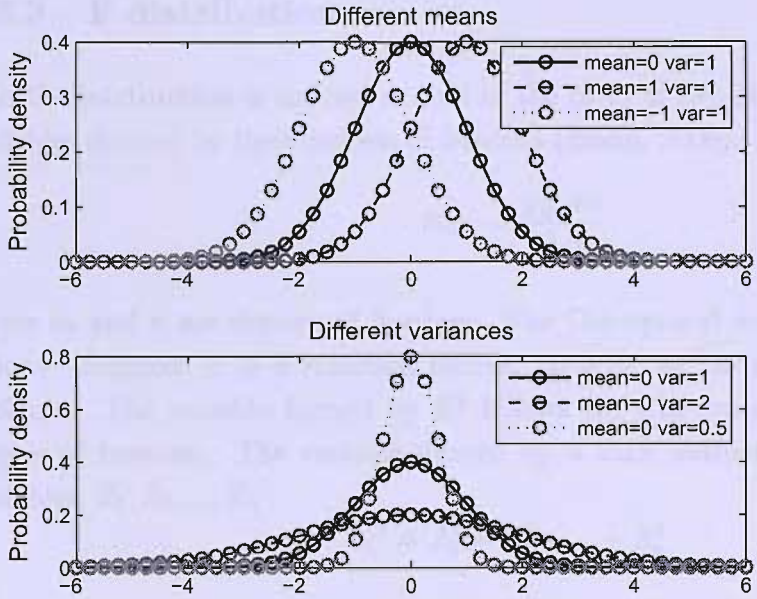


Figure 2.8: Normal distributions with different means and different variances.

where

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

is the PDF of standard normal distribution with $\mu = 0$ and $\sigma = 1$. This PDF has a number of general properties. As with all PDFs, the total area under the curve must be one, since this is the total probability of all possible events. The mean will be in the middle of the curve and most of the area under the curve will be between the mean minus two standard deviations and the mean plus two standard deviations.

The binomial distribution with parameters n and p may be approximated by the normal distribution when both np and $n(1-p)$ exceed 5.

In this study, the background EEG (BEEG) is assumed to follow a Gaussian distribution and its corresponding statistical properties are used for further analysis. Moreover, many probability distributions can be derived for functions of Normal variables, for example, Chi-squared, t, and F distributions. And those will be described in the following and their applications in this study will be mentioned as well.

2.5.3 F distribution

The **F distribution** is defined as that of the ratio of two independent Chi-squared variables divided by their degrees of freedom (Bland, 2000):

$$F_{m,n} = \frac{\chi_m^2/m}{\chi_n^2/n} \tag{2.5.3}$$

where m and n are degrees of freedom. The Chi-squared distribution is defined as follows. Suppose Z is a standard normal variable, so having zero mean and unit variance. The variable formed by Z^2 follows the Chi-squared distribution with 1 degree of freedom. The variable formed by n such independent standard normal variables, $Z_1, Z_2, ..., Z_n$:

$$\chi^2 = Z_1^2 + Z_2^2 + ... + Z_n^2$$

is defined to be the Chi-squared distribution with n degrees of freedom.

This distribution is used for comparing variances. If we have two independent estimates of the same variance calculated from Normal data, the variance ratio will follow the F distribution.

In this study, we will mention two widely used objective methods (F_{sp} and $\pm$ *difference*) for detecting the response. Those are based on estimates of variances and the F distribution.

2.5.4 Sign-test

The **Sign test** is used to test the null hypothesis that two methods have the same effect on the event (no difference). This hypothesis implies that given a random pair of methods (x, y), then both x and y are equally likely to be larger than the other. Suppose that $r+$ is positive difference and $r-$ is negative difference. Under the null hypothesis, the number of $r+$ and $r-$ follow a binomial distribution with $p = 0.5$ and $n =$ equal to the total number of $r+$ and $r-$. The test is performed by finding the maximum number (NM) of $r+$ and $r-$, and then using the table of binomial distribution to find the probability of observing this value of NM. Depending on that value, the null-hypothesis is accepted or rejected.

The sign test will be used to estimate whether different bootstrap techniques have the same ability to detect the response.

2.5.5 The kappa statistic

The **Kappa statistic** (written κ) is regarded as a measure of inter-observer agreement (reliability). The calculation is based on the difference between how much agreement is actually present ('observed' agreement) compared to how much agreement would be expected to be present by chance alone ('expected' agreement) (Viera and Garrett, 2005). Table 2.1 is used to explain the definitions and the procedures of calculation.

		Observer1 - Results		
		Yes	No	Total
Observer2 - Results	Yes	a	b	m1
	No	c	d	m0
	Total	n1	n0	n

Table 2.1: Definition of components for calculating the Kappa value.

The observed agreement is

$$A_o = \frac{a+d}{n}$$

and expected agreement is

$$A_e = [\frac{n1}{n} \bullet \frac{m1}{n}] + [\frac{n0}{n} \bullet \frac{m0}{n}]$$

Therefore Kappa is defined as:

$$\kappa = \frac{A_o - A_e}{1 - A_e} \tag{2.5.4}$$

It has a maximum of 1 when agreement is perfect, a value of zero indicates no agreement better than chance, and negative values show worse than chance agreement, which is unlikely in this context (Altman, 1991). In order to interpret a value between 0 and 1, Landis and Koch (Landis and Koch, 1977) gave a rough guideline as:

Value of κ	Strength of agreement
< 0.20	Poor
0.21-0.40	Fair
0.41-0.60	Moderate
0.61-0.80	Good
0.81-1.00	Perfect

The Kappa statistic is used in the current work to estimate the agreement between subjective inspections (inter-observer) in determining hearing thresholds. Further details for applying Kappa will be given in section 5.4.2.

2.5.6 ROC curve and its area

To compare the response detection performance of different methods investigated, the sensitivity of the tests is calculated. The sensitivity characterizes the performance of a test to detect a response, when present. It is defined as the number of correct positive decisions (CPD) divided by the sum of CPD and the number of false negative decisions (FND) (Cebullar et al., 1996):

$$sensitivity = \frac{CPD}{CPD + FND}$$

Specificity is defined as the number of correct negative decisions (CND) divided by the sum of CND and the number of false positive decisions (FPD).

$$specificity = \frac{CND}{CND + FPD}$$

The sensitivity and false positive rate (1-specificity) are the main issues in the detection task and can be obtained at any cutoff threshold. When comparing two or more tests, these two characteristics at one cutoff threshold are not enough to determine the performance of the tests, since it is possible that at one cutoff threshold the sensitivity of test A is greater than that of test B, but at another cutoff threshold, the sensitivity of test A is smaller than test B. Therefore, a **receiver operating characteristic (ROC)**, defined as a graphical plot of the sensitivity vs. (1 - specificity)

(see an example in Figure 2.9) as its discrimination threshold is varied, is desirable to investigate the sensitivity and false positive rate at different cutoff thresholds.

A point in the ROC curve only represents one cutoff. In order to estimate the overall performance of one or more tests, all the points in the ROC curve should be taken into account. The accuracy of the test is defined as how well the test separates the group being tested into those with and without the responses in question, and it is measured by the area under the ROC curve (Massof and Emmel, 1987). The area is calculated as the sum of the area of many small trapezoids which are obtained by dividing the area according to the points on the curve (an example is shown in Figure 2.9).

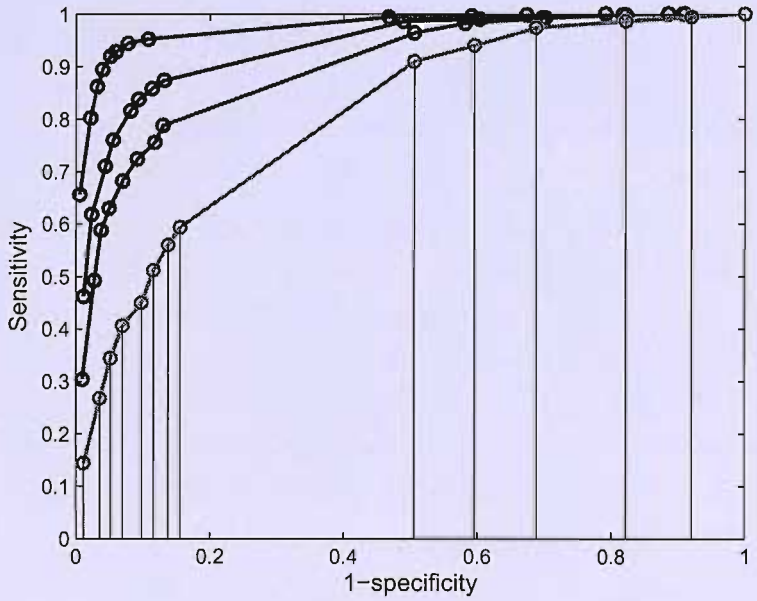


Figure 2.9: An example of ROC curve for four different detection methods and approximation of the area under the ROC curve by dividing the area into many small trapezoids.

An area of 1 represents a perfect test, that is when the curve follows the left and upper axes (see Figure 2.9), such that sensitivity is one for all possible cutoff thresholds. An area of 0.5 means no discrimination exists, and the curve lies along the major

diagonal, where the sensitivity and false positive rate are always equal. A rough guide for classifying the accuracy of a diagnostic test is given below (Egan, 1975):

Value of area	Power of test
0.50-0.60	Fail
0.60-0.70	Poor
0.70-0.80	Fair
0.80-0.90	Good
0.90-1.00	Excellent

ROC and its area are used here is to compare the performance of different ABR detection methods and will be applied in Chapter 6.

Chapter 3

Signal Acquisition

3.1 Introduction

Now the experimental work, calibration of equipment and recording procedures are introduced in the following three sections.

In this study, two sets of recordings of ABR (defined as Set A and Set B) were used for testing the proposed methods. Set A was collected by colleagues (Bell, 2003; Cane, 2002) in 2002 from 12 normal-hearing adults stimulated at 0 to 50 dB sensation level (SL) in steps of 10 dB. Set B was recorded by myself in 2004 in order to obtain signals both with and without stimulation under the same experimental conditions. The equipment settings between data Set A and B were similar with differences shown in Table 3.1. This chapter will focus on the procedures and parameter settings for set B.

3.2 Experimental work

3.2.1 Equipment configuration

For this experiment, click stimuli were generated and EEG signals recorded using a computer controlled Cambridge Electronic Design (CED, UK) micro1401 laboratory interface (consisting of Analogue-to-Digital (AD) and Digital-to-Analogue (DA) converters) and CED 1902 isolated biological amplifier. This equipment was controlled

using an ABR program able to get and individually store (for later analysis) different number of epochs of EEG signal including ABR, which was developed by Prof. Mark Lutman in the ISVR (University of Southampton) and run on a PC. The equipment configuration for this project is shown in Figure 3.1.

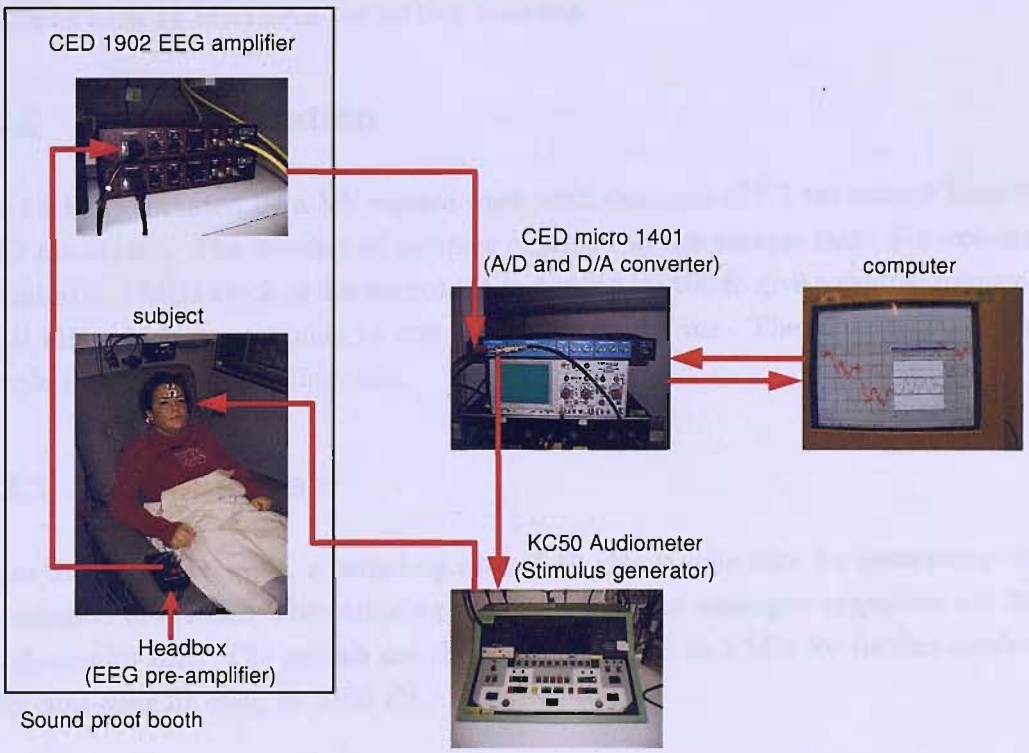


Figure 3.1: Equipment configuration.

The computer (PC) loads an appropriate sound stimulus waveform into the memory of the micro1401, which reads out the buffer from the digital to analogue (DA) converter to generate a repeating stimulus such as a click. This analogue signal is sent to a Kamplex KC50 audiometer, and the input level of the stimulus can be amplified and adjusted in accordance with specific requirements. Finally it is sent to the subject via insert earphones (Etymotic, USA).

The ABR is picked up by electrodes place on the subject's scalp and is sent to the CED 1902 biological amplifier via an isolated EEG 'headbox'. The CED 1902 then amplifies and filters the signal. After that the signal is sent to the micro1401 analogue to digital (AD) converter and then is in digital format, to be transmitted to the computer. The

signal is averaged by the ABR program, after applying artefact rejection. This allows on-line verification of signal quality. As the raw EEG (including ABR, spontanous EEG, and possibly artefacts) is required for the various statistical analysis, it is also exported to disk by the ABR software. The raw EEG can then be imported to other programs such as MATLAB for further analysis.

3.2.2 Click generation

The click is generated by a 5 V square wave with duration of 0.1 ms output from the CED micro1401. The number of samples depends on the sample rate. For example, the internal 1 MHz clock of the micro1401 is divided by 100 to give a sample frequency of 10 kHz which corresponds to one sample every 0.1 ms. Therefore a click is one sample long at this sampling rate.

3.2.3 Sampling rate

From the above example, a sampling rate of 10 kHz is adequate for generating click stimulation of 0.1 ms. The sampling rate, at which the analogue responses are digitized, was 20 kHz. The signals are then downsampled to 5 kHz for further analysis, after anti-alias filtering at 2100 Hz.

3.2.4 Transducer-headphone selection

When choosing the transducers to generate the auditory evoked potentials, two factors should be considered: the frequency response of the transducer and the amount of stimulus artefact the headphones generate. E-A-RTONE 5A (E-A-R Auditory Systems, USA) insert earphones are used here to deliver the sound signal. The major benefit of this earphone is the reduction of background noise that might interact with the presented sound and influence threshold determination. The stimulus artefact also is greatly reduced in two ways: (1) The earphone box can be placed far away from the electrode leads; the greater the separation, the less likely the artefact will be. (2) The tube produces a time delay between stimulus and response. Even if the stimulus artefact is present, the delay actually eliminates the interference with the

early components (e.g. wave I). The insert earphone has the advantage of reducing the stimulus artefact, compared to supraaural earphones (e.g. TDH-39 or TDH-49).

3.2.5 Filters and mains noise

Choosing an appropriate filter is very important to record a clear ABR waveform. There is a compromise in the choice of high and low pass filters, so as to exclude as much background EEG as possible and including the maximum frequency content of the ABR. The CED 1902 amplifier contains a limited number of 12 dB/octave filter settings. For ABR studies, Hall (1992b) recommended the filter settings as 30-3000 Hz with a notch filter at 50 Hz, to remove the influence of mains noise (50 Hz in UK). For recording data Set B, the recommended filter settings were applied.

3.2.6 Electrodes

Electrode placement refers to the 10-20 International system (Jasper, 1958). The conventional locations of the electrodes in recording the ABR are a noninverting electrode on the high forehead (close to Fz) and an inverting electrode on the mastoid. The low forehead is the position of the ground electrode. The specific locations of the electrodes are shown in Figure 2.3.

Inter-electrode electrical impedance is another important factor for data quality. Low and balanced electrode impedances contribute to higher quality ABR recordings by (1) limiting the internal noise of the amplifiers, (2) reducing the effects of external electrical interference (noise) and (3) maintaining higher common mode rejection ratios. The convention for maximum inter-electrode impedance is 5 kΩ (Hall, 1992a). In this experiment, before recording the signals, inter-electrode impedance was tested in each subject. If the requirement was satisfied with less than 5 kΩ, the following procedure was carried out. Otherwise, the problems were found and solved, e.g. by changing the electrode, cleaning the skin, etc., until impedance reached the requirement.

The above parameter settings for data Set B are summarized in Table 3.1. For comparison, those for Set A are placed in the Table as well.

Parameter	**Set A**	**Set B**
Transducer	ER-2A	ER-5A
Type	Click	Click
Duration	0.1 ms	0.1 ms
Polarity	Rarefaction	Rarefaction
Rate	33.3/s	33.3/s
Intensity	0 to 50 dB SL	0 to 60 dB nHL
Without stimulation	No	Yes
Electrode arrays		
Channel 1	Vertex (Cz)	Forehead (Fz)
Channel 2	Nape of neck	Mastoid (Right)
Ground	High forehead	Lower forehead
Bandpass filter	30-3000Hz	30-3000 Hz
Notch filter	Yes (50 Hz)	Yes (50 Hz)
Amplification	30,000	10,000
Analysis window	5-15 ms	5-15 ms
Sweeps	2000	2000

Table 3.1: Acquisition parameters for two data sets.

3.3 Calibration of equipment

There were three calibration procedures carried out. The amplification of the CED biological amplifier was checked, and the click stimulus was calibrated according to peak-equivalent SPL. Normal hearing level (nHL) was calibrated in order to determine 0 dB nHL, relative to which stimuli were then applied.

3.3.1 Calibration of the CED 1902 biological amplifier

It is necessary to determine the gain of the CED 1902 amplifier in order to calculate the magnitude of the ABR. The gain may shift over time, so the calibration was carried out for each experiment.

The relationship between the voltage input to the micro 1401 A-D converter and the scale in MLS-MLR software was calculated. A 5 V peak-to-peak sine-wave burst was generated by MLS-MLR and this was delivered from the analogue output of the micro 1401 back to the input. The level of the sine burst was displayed on an oscilloscope for verifying the actual value. The sine burst was averaged by the MLS-MLR over 100 sweeps and the peak-to-peak amplitude was read. For the condition where the ABR was collected, the peak-to-peak amplitude was 32767 scale units, which means that 1 V corresponds to 6553 MLS-MLR scale units, or 1 scale unit corresponds to 0.153 mV at the input to the micro 1401. The recorded ABR from a subject is normally less than $\pm 1\mu V$, therefore before feeding into the micro 1401, the CED 1902 amplifies the ABR amplitude with a gain of 10,000. Actually 1 μV of the ABR from the subject was $10^{-2}V$ on the input of the micro 1401, and this corresponds to 65.53 scale units shown on the PC.

3.3.2 Input level of the audiometer

The stimulus signals from the micro 1401 went through the input of the audiometer and then were attenuated by the audiometer before reaching the earphone connected to the subject. A change of the gain of the input of the audiometer leads to a change of stimulus levels. So the gain of the input of the audiometer was set before any measurements were made on the subject and calibrations of the equipment.

A 5 V peak-to-peak sine wave generated by the micro 1401 was used to set the input of the audiometer. Before any measurements, the setting of the input was checked using this calibration signal and adjusted so that the LCD display for the input level on the audiometer read 0 dB on the dial.

3.3.3 Calibration of clicks in dB pe SPL

The insert earphone was connected to a Brüel and Kjær 2112 spectrometer and an oscilloscope via a IEC 126 '2cc' coupler to a 1 inch microphone as specified by International Standard ISO 389. The scale of the spectrometer was set up using a reference calibration piston. The piston generates a fixed SPL and the input gain of the spectrometer was adjusted until the spectrometer dial gave the same reference level. The output of the spectrometer was fed back to the AD input of the micro 1401. A clear click response was then recorded by averaging the spectrometer output over 100 clicks using the MLS-MLR software. The peak-to-peak amplitude and the half period of the first positive going cycle of the click were estimated.

For calibration, the frequency of the click is assumed to be 1 over twice the half period measured. A sine wave is adjusted until its peak-to-peak amplitude is equal to the peak-to-peak value of the click estimated above. The value then read on the spectrometer is the peak-equivalent SPL (pe SPL) of the click. That was 88 dB pe SPL here.

3.3.4 Calibration of normal hearing level

Normal hearing level (dB nHL) is the most common reference for describing stimulus intensity for a short duration stimulus. Behavioral hearing tests are performed for a relatively small group of normal-hearing adult subjects, the lowest sound level heard of each subject is then defined as his/her behavioral threshold level. The average behavioral hearing thresholds from 10 subjects (both ears) is calculated and this is defined to be 0 dB nHL. In this case, 35 dB p.e.SPL on the KC50 audiometer (see Figure 3.2) was estimated as 0 dB nHL.

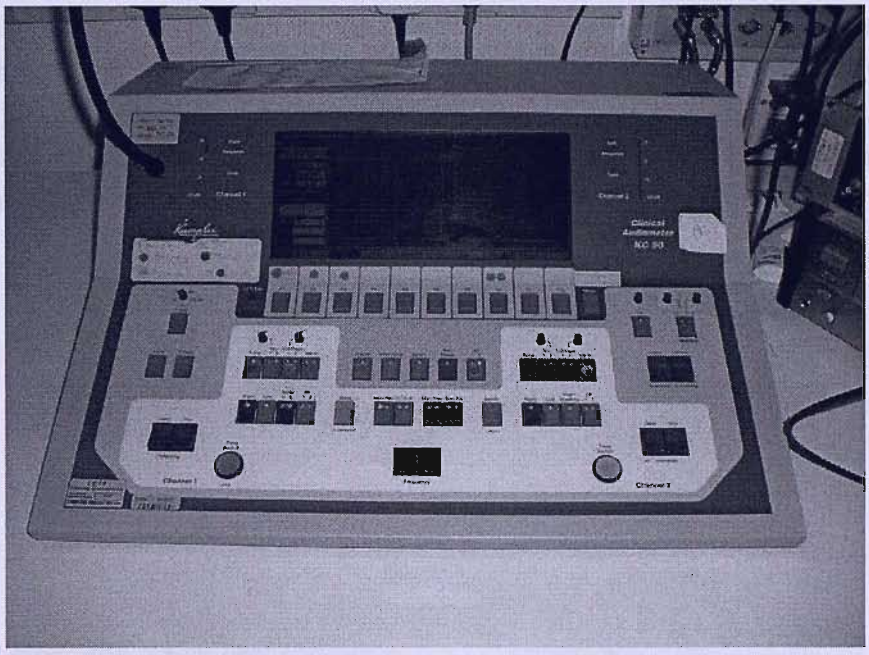


Figure 3.2: Audiometer (KC50).

3.4 Pre-assessment and recording

Since data are collected from human beings, strict adherence to the protocol is required, in order to guarantee the safety of the experiment[1]. Before setting up the equipment, the procedure was explained and a consent form (Appendix B) signed by each subject. Then each subject completed a questionnaire (Appendix A), in order to identify any ear problems or previous treatment. The following complementary tests were carried out before recording.

3.4.1 Otoscopy

Otoscopy is an examination that involves looking into the ear with an instrument called an otoscope (Figure 3.3). This is performed in order to examine the 'external ear canal' - the tunnel that leads from the outer ear (pinna) to the eardrum. Inspection

[1]This experiment is approved by the school's safety and ethics committee and adheres to the ISO60601 standard.

of the eardrum can give much information on the condition of the middle ear. This test is carried out to make sure the external ear canal and middle ear structure is normal. In this experiment, almost all the subjects were identified as normal at this stage. One subject had a lot of wax in the ear canal and was asked to clean this, before carrying on with the following test.

Figure 3.3: Otoscope

3.4.2 Tympanometry

Tympanometry is an audiological procedure for measuring the acoustic admittance of the middle ear, allowing abnormal conditions of the eardrum and middle ear to be identified. A tympanometer (Figure 3.4) provides a tympanogram, which shows acoustic compliance as a function of air pressure in the external ear. Figure 3.5 gives an example from a 19 year old female. The shape of an inverted-U indicates normality. Furthermore, the middle ear pressure of $15daPa$, compliance of 0.5 ml eqV, and ear canal volume of 0.9 ml eqV are all in the normal ranges (-100 to +50 daPa for pressure

Figure 3.4: Tympanometer: measuring the compliance of the eardrum.

and 0.3 to 1.5 ml eqV for compliance) recommended in the literature. On the other hand, an abnormal tympanogram may reveal any of the following problems:

- Fluid in the middle ear;
- Perforated ear drum;
- Impacted ear wax;
- Scarring of the tympanic membrane;
- Lack of contact between the conduction bones of the middle ear;
- A tumor in the middle ear.

3.4.3 Pure-tone audiometry

Pure-tone audiometry (PTA) is a manual subjective audiometry technique, which aims to establish an individual's sensitivity to single-frequency tones. PTA is applied

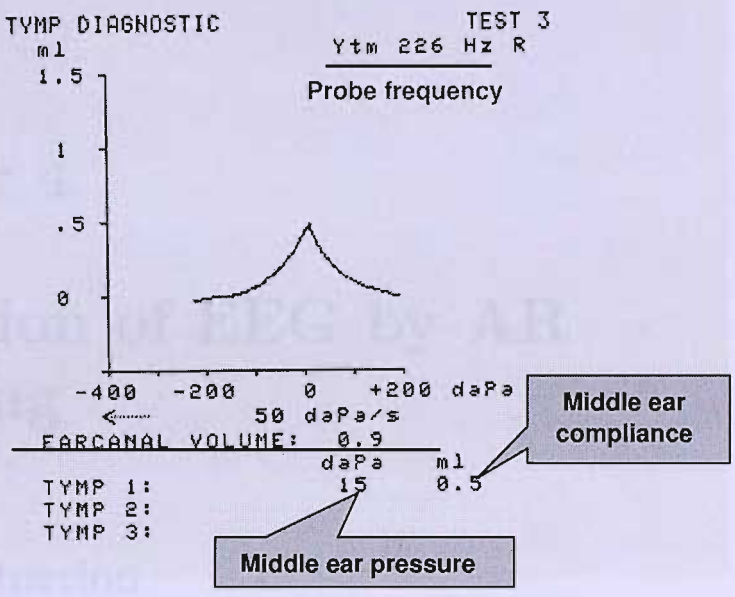


Figure 3.5: An example of a tympanogram.

here to determine first if the individual is within the normal-hearing range or not. Hearing threshold level (HTL) of 20 dB HL or less is defined as normal hearing. That more than 20 dB HL is defined as hearing impairment: (1) mild loss: 20-40 dB HL; (2) moderate loss: 40-60 dB HL; (3) severe loss: 60-90 dB HL; (4) profound loss: over 90 dB HL.

3.4.4 EEG recordings

There were 27 subjects involved in the experiments. One was identified as having moderate hearing loss, and not suitable for ABR recording. Ten cases were required for calibrating 0 dB nHL. Another 16 subjects thus provided ABR recordings. For each subject, 2 replicates at each stimulus level (0, 10, 20, 30, 40, 50, 60 dB nHL) are carried out with 2000 sweeps (stimulus rate of 30/sec). Under the condition of no stimulus, four replicates were acquired, each recording with 4000 sweeps (2 minutes). The raw EEG data were saved on disk for further analysis.

Chapter 4

Simulation of EEG by AR modelling

4.1 Introduction

When collecting ABR data, many sources of noises may interfere with the quality of the data (Chapter 2). If non-physiological sources of noise are eliminated, the residual artefacts will be background EEG and movement. An artefact rejection scheme (MAR) has been developed and will be introduced in Chapter 6, particularly to remove the movement artefact. The background EEG (BEEG) thus becomes the main source of noise affecting ABR recordings.

In this chapter, the common assumptions for BEEG in most measurements for evoked potentials, i.e., BEEG is a stationary and ergodic process, will be introduced. Then characteristics of the BEEG will be analysed, in order to assess how accurate the assumptions are for recorded BEEG. Finally, we will present the method for carrying out the Monte Carlo simulations.

4.2 Assumptions for EEG and their statistical properties

At present, almost all methods of EEG analysis are based on certain implicit assumptions regarding the statistical properties of the underlying random process, particularly with respect to the extent of stationarity and Gaussianity of the process (Elberling and Don, 1984; Wong and Bickford, 1980; Cebullar et al., 2000; Sturzebecher et al., 1999). In this work, it is necessary to recognize and understand the statistical properties of assumptions in order to investigate if these are satisfied in practice. In this section, we will concentrate on the statistical properties of stationarity and ergodicity signals. These properties relate to the specific definitions of the characteristics of random signals. We thus begin this section with an introduction of random signals and then refer to the specific properties.

4.2.1 Random signals

A random signal is a signal from a random phenomenon (Bendat and Piersol, 1986). The essential feature of random signals is that one cannot precisely predict the signals in advance, since each 'realization' of the experiment may result in a different signal. Therefore in order to describe such signals, we can use probability theory and statistical descriptors (mean, variance/standard deviation, and autocorrelation).

The statistical properties are defined from the ensemble of recordings (see Figure 4.1). The **mean** of the ensemble of signals $x(n)$ is given by

$$\widehat{\mu(k)} = \frac{1}{M}\sum_{j=1}^{M} x_j(k) \tag{4.2.1}$$

where M is the number of signals $x_j(k)$ in the ensemble. The mean at a time-instant k is calculated as the mean of all the signals at this sample. This is not necessarily the same as the mean of each signal. It is also clear that the mean $\widehat{\mu(k)}$ may vary over time.

Similarly, the **variance** of the process can be estimated at a time-instant k down the ensemble as shown in Figure 4.1.

$$\widehat{\sigma^2(k)} = \frac{1}{M}\sum_{j=1}^{M}(x_j(k) - \widehat{\mu(k)})^2 \tag{4.2.2}$$

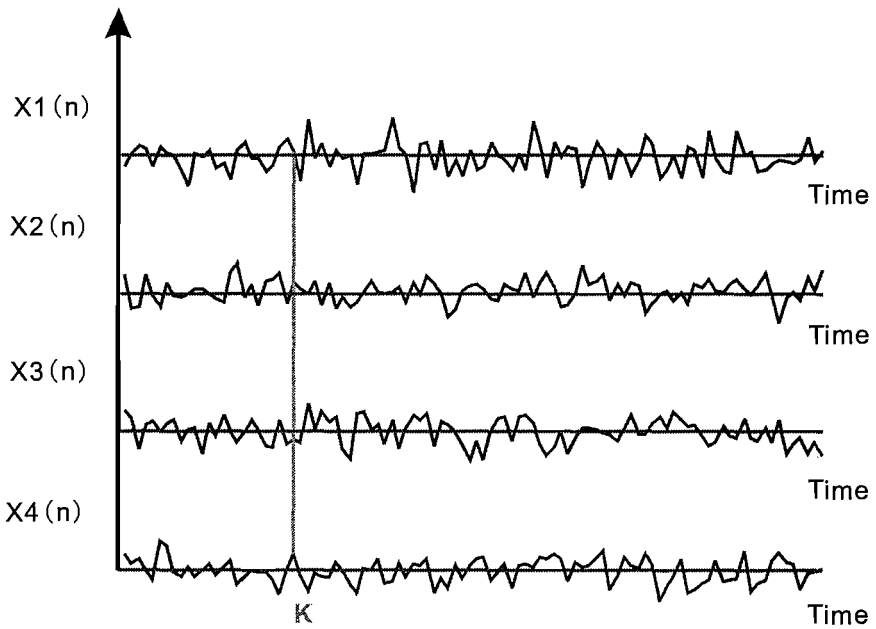


Figure 4.1: An ensemble of random signals.

Another property commonly used to describe the random signals is the **autocorrelation**. This is related to the correlation coefficient between different samples of the same signal, and can be estimated as follows:

$$\widehat{R_{xx}}(k,\, k+\tau) = \frac{1}{M}\sum_{j=1}^{M}[x_j(k)x_j(k+\tau)] \tag{4.2.3}$$

where τ is the time-delay between samples $x_j(k)$ and $x_j(k+\tau)$.

4.2.2 Stationary and ergodic signals

Random signals can be divided into stationary and non-stationary signals (Bendat and Piersol, 1986) depending on whether or not these statistical properties change

with time. Since all biomedical signals change as time evolves, the assumption of stationarity can only be an approximation, but is often useful for further analysis.

In **stationary** signals, the mean, variance and autocorrelation are constant over time.

$$\widehat{\mu(k)} = \mu \tag{4.2.4}$$

$$\widehat{\sigma^2(k)} = \sigma^2 \tag{4.2.5}$$

$$\widehat{R_{xx}}(k,\, k+\tau) = R_{xx}(\tau) \tag{4.2.6}$$

A stationary random signal is considered to be ergodic if it is possible to compute its statistics (mean, variance, etc) using the time average over any single signal instead of an average down the ensemble.

$$\widehat{\mu(k)} = \frac{1}{M}\sum_{j=1}^{M} x_j(k) = \widehat{\mu(x_j)} = \frac{1}{N}\sum_{k=1}^{N} x_j(k) \tag{4.2.7}$$

where M is the number of signals in the ensemble and N is the number of samples in each signal.

In summary, the properties of an ergodic signal are:

$$\widehat{\mu(k)} = \mu = \widehat{\mu(x_j)} \tag{4.2.8}$$

$$\widehat{\sigma^2(k)} = \sigma^2 = \widehat{\sigma^2(x_j)} \tag{4.2.9}$$

$\forall k$ (instants) and $\forall x_j$(sample functions).

4.3 Assessing characteristics of the EEG

With the knowledge of the statistical properties of stationary and ergodic signals from a theory point of view, the assessment of those on the recorded background EEG follows. EEG in this chapter specifically refers to background EEG (BEEG).

4.3.1 EEG components and properties

In order to remove the influence of background EEG from evoked potentials, e.g. the ABR, it is very important to have some knowledge of the frequency distribution of the BEEG. The frequency content of BEEG lies primarily between 0 and 100 Hz, although most typical waves are in the frequency between 1 to 20 Hz. Four major types of BEEG activity are recognized (alpha, beta, delta and theta).

Delta (δ) wave is the frequency range up to 4 Hz and is often associated with the young and certain encephalopathies and underlying lesions. It is seen in stage 3 and 4 of sleep. The frequency range of theta (θ) waves is from 4 Hz to 8 Hz and it is associated with drowsiness, childhood, adolescence and young adulthood. Alpha (α) is the frequency range from 8 Hz to 13 Hz. It is characteristic of a relaxed, alert state of consciousness. Alpha rhythms can be best detected when the eyes are closed. Beta (β) is the frequency range between 13 to 22 Hz. Low amplitude β waves with multiple and varying frequencies are often associated with active, busy or anxious thinking and active concentration.

In this study, a band-pass filter of 30-3000 Hz was applied, in order to analyse dominant wave V of the ABR, and thus the above low frequency waves were filtered out. There is an example of a BEEG during 3 seconds in Figure 4.2.

Furthermore, an estimate of power spectral density (PSD) was performed by the Welch method (Hayes, 1996). The PSD in Figure 4.3 was obtained by averaging the PSDs of 8 raw background EEG segments, each of which had 15 second duration. A strong trough at 50 Hz, derives from the notch filter. The magnitude of the PSD at high frequencies over 500 Hz are very small at about $10^{-3}[\mu V]^2/Hz$. This indicates that the spectrum of the ABR waves would not be affected by the background EEG. Particularly for our interest in wave V, its energy is mainly at about 500 Hz (Boston, 1981).

4.3.2 The effect of averaging

Averaging (in this work coherent averaging) is the most commonly used technique to make the evoked potentials visible. In principle, the standard deviation (SD) of

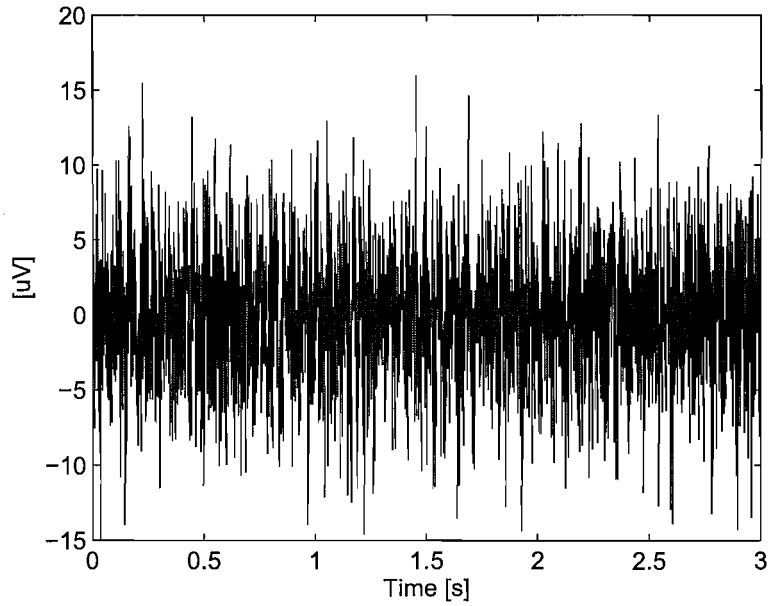


Figure 4.2: A typical raw background EEG data, recorded without stimulation.

the average decreases by the square root of the number of sweeps in the average (Schimmel et al., 1974). This was investigated from mathematical point of view in Chapter 2, and is shown by the equations 2.2.2 and 2.2.3.

The effect of averaging is demonstrated for both white noise with a normal distribution and mean zero, variance one and a typical background EEG from a subject without stimulation. Figure 4.4 shows the effect on the SD of averaging for white noise (solid line) and that of the real recording as a dashed line. Both are normalized by the SD of the first sweep. A plot based on the theory that SD decreases as the square root of the number of sweeps is also shown in the figure as a dotted line. By visual inspection, the SD for both white noise and recording are very similar to the 'theory' line.

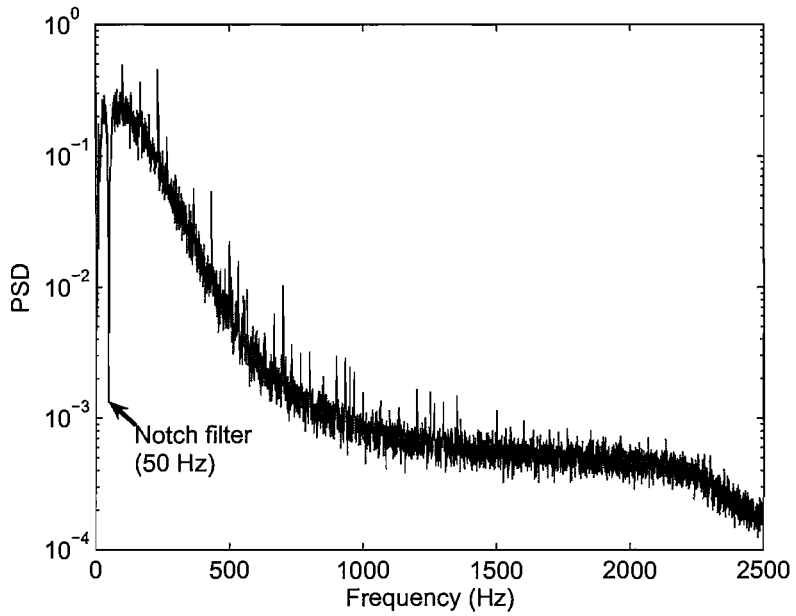


Figure 4.3: Power spectral density of background EEG via Welch method. At 50 Hz, a strong trough appears which derives from the notch filter.

4.4 Autoregressive model for EEG simulation

The background EEG is viewed as a stationary random process and usually is modelled as the output of a linear system, e.g. Autoregressive Moving Average (ARMA), driven by white noise (Isaksson et al., 1981; Cerutti et al., 1985; Liberati et al., 1992; Rossi et al., 2007).

A general form of the ARMA model is that an input signal $x(n)$ is filtered by a system to give an output $y(n)$, where n is the sample index. The response of the system to any input is described by the parameters a and b. The number of a and b parameters (p and q) is defined as the order of the filter. The relationship between input and output is (Marple, 1987):

$$y(n) = \sum_{k=1}^{p} a_k y(n-k) + \sum_{k=0}^{q} b_k x(n-k) \tag{4.4.1}$$

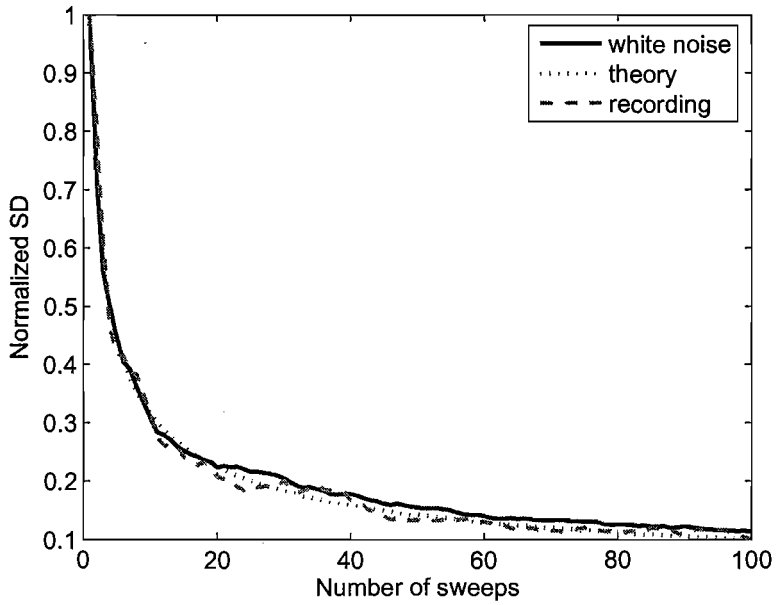


Figure 4.4: The effect of coherent averaging on the standard deviation of white noise and raw EEG data.

The current output $y(n)$ is a a weighted sum of previous outputs, $y(n-k)$ and a sum of filtered previous inputs $x(n-k)$. If all the $b[k]$ coefficients except $b[0]$ are set to zero, this process is strictly an autoregressive (AR) process of order p. The AR model is used here to simulate the background EEG signals following the procedures in Figure 4.5. The two important factors, namely a reference signal for estimating the AR parameters, and a suitable order should be determined, before employing the AR model.

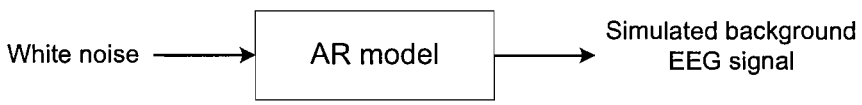


Figure 4.5: The procedures of simulating EEG signals through an AR model. White noise is acting as input and simulated background EEG signal is the output.

4.4.1 Selection of a reference signal

An appropriate reference signal is very important in the AR process and used for estimating the AR parameters. In order to simulate the background EEG signals, a 'typical' signal should be chosen as a reference signal and usually a signal without stimuli is more desirable than that with stimuli, because of the aim for simulating background EEG. As mentioned before, two data sets (A and B) were collected under the different conditions in the current work, and the selection of a reference signal will be described separately.

For data Set A, the reference signal was selected from 12 recordings with stimulation at 0 dB SL, since there were no signals without stimuli. The power spectral density (PSD) of each recording was estimated by the Welch method and then a median PSD among the 12 spectra was obtained. The absolute values of the difference between the median PSD and each of the 12 recordings at each frequency were summed and the minimal summation value was found and its corresponding recording was taken as the reference signal. The reason behind that is that the properties of the signals from different subjects can vary considerably and some randomly occurring artefacts are possibly present. The above procedure aims to find the most 'typical' recording. The simulations in Chapter 5 were all based on this reference signal (denoted as RSA).

With the development of the artefact rejection schemes which will be introduced in Chapter 6, more simulations were required and at that moment we had obtained the second set of recordings (Set B), under conditions with and without the stimulation. Therefore, a second reference signal (here denoted by RSB) was chosen from recordings without stimulation, using a procedure equivalent to that described above. All simulations of background EEG in Chapter 6 and 7 were based on RSB.

4.4.2 Determination of AR model order

The order selection is another issue for the AR model. Too low values for model order will result in a highly smoothed spectral estimates, whereas too high values will increase the resolution and introduce spurious detail into the spectrum. There are three standard methods to estimate the order (Haykin, 1986): Final Prediction Error (FPE), Akaike information criterion (AIC) and criterion autoregressive transfer

(CAT) function. In this work, the FPE was chosen as as the selection criterion of the order. FPE is defined by the following equation:

$$FPE[p] = \hat{\rho}_p \left\{ \frac{N + (p+1)}{N - (p-1)} \right\} \tag{4.4.2}$$

where N is the number of data samples, p is the order, and $\hat{\rho}_p$ is the estimated white noise variance. The term in parentheses increases as the order increases, reflecting the increase in the uncertainty of the estimated $\hat{\rho}_p$ of the prediction error variance. The order p to be selected is the one for which the FPE is minimum. The exact value of the order differs for different reference signals, thus this will be given in Chapter 5 and Chapter 6, respectively.

4.4.3 Autocorrelation (Yule-Walker) method for AR parameter estimation

The Yule-Walker method is also called the autocorrelation method to fit a *pth* order autoregressive (AR) model to the input signal, by minimizing the forward prediction error in the least-squares sense. This formulation leads to the Yule-Walker equations, which are solved by Levinson-Durbin recursion (Akay, 1994) and then the AR parameters (prediction coefficients) can be obtained. In the Levinson-Durbin recursion algorithm, the prediction coefficients of the AR model at the current stage can be obtained recursively from those calculated at the previous stage. The relationship between the new prediction-error filter and the old prediction-error filter is known as the equation of Levinson-Durbin recursion. The Yule-Walker method was used both to estimate the FPE, and hence to determine the model order, and then with this method, to estimate the model parameters from the reference signal (see Figure 4.6).

4.4.4 Summary of EEG simulation

Following the procedures showing in Figure 4.6, the AR parameters could be obtained. Then white noise was input to the AR model and the output of AR model is a simulated background EEG signal.

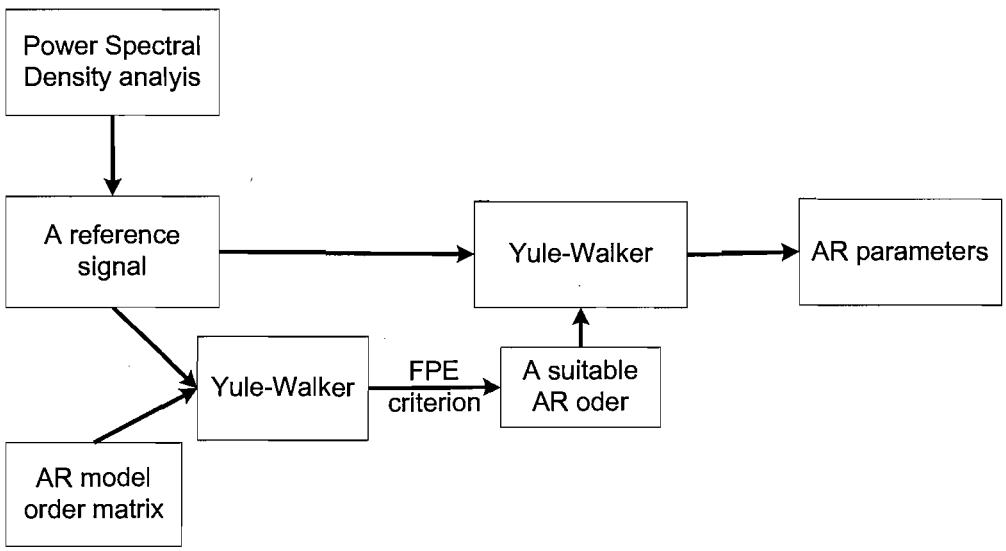
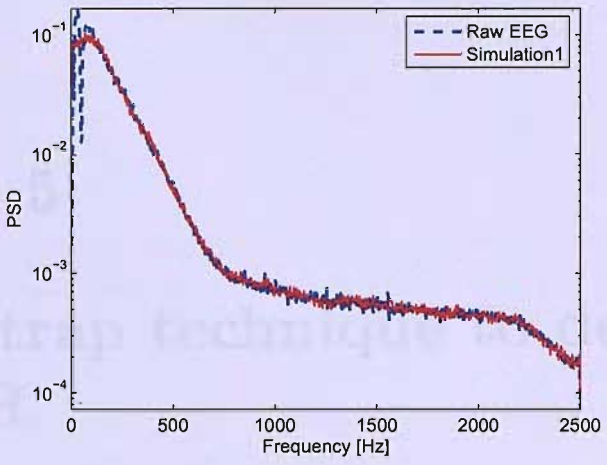
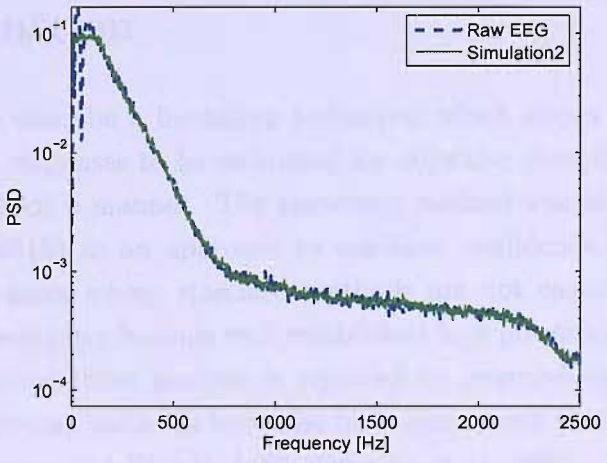


Figure 4.6: The procedures for simulating the background EEG. The power spectral density analysis is aimed at choosing a reference signal. AR model order matrix refers a number of values of the order. FPE is to determine the order of the AR model.

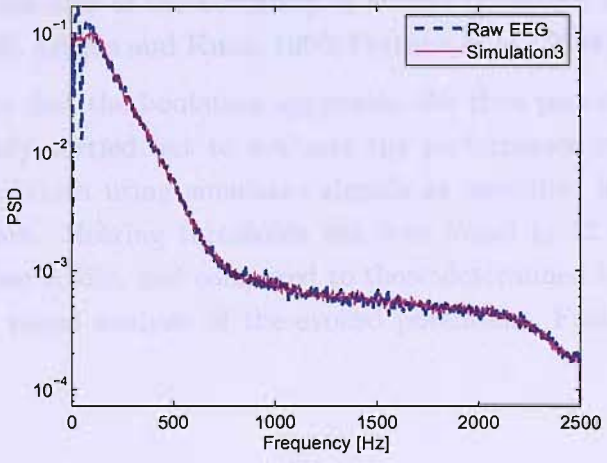
In order to show that the simulations looks like the raw EEG (a reference signal), an example for the PSDs of three simulated signals are shown in Figure 4.7. The PSDs for the simulations are similar to those from the raw EEG (in the same figure as a dashed line).



(a)



(b)



(c)

Figure 4.7: The PSDs for an EEG signal and three different simulations.

Chapter 5

A bootstrap technique to detect the ABR

5.1 Introduction

In this chapter, we describe a bootstrap technique, which allows the statistical significance of evoked responses to be estimated for objective detection, and does so in an easy and very flexible manner. The bootstrap method was introduced by Efron (Efron, 1979a,b, 1981b) as an approach to calculate confidence intervals for parameters in circumstances where standard methods are not easily be applied. The bootstrap has subsequently become well established as a powerful statistical tool, in which complex mathematical analysis is replaced by intensive computational load. In recent years bootstrap methods have also been extensively used in biomedical signal processing (Haynor and Woods, 1989; Simpson et al., 2004). To the best of our knowledge this technique has not previously been used in the detection of evoked potentials, though other uses of the bootstrap in evoked potentials have been reported (Darvas et al., 2005; Adams and Kunz, 1996; Fortune et al., 2004).

In the following, we first the bootstrap approach. We then provide the results from a Monte-Carlo study carried out to evaluate the performance of the technique in well-controlled conditions using simulated signals as described in Chapter 4, with no stimulus response. Hearing thresholds are then found in 12 subjects using the bootstrap method on ABRs, and compared to those determined by experienced professionals through visual analysis of the evoked potentials. Finally, we discuss the

results, and other potential applications of the proposed method, and some of its limitations.

5.2 Data

ABRs were recorded from 12 normal-hearing adults subjects (6 males and 6 females), who were aged between 18 and 30 years following the procedure outlined in Chapter 3. The ABR was recorded between the vertex and the nape of neck, with a frontal electrode serving as ground. The auditory stimulation was a rectangular click stimulus with a duration of 100μs delivered by ER-2 insert phones (Etymotic, USA), at a click rate of 33.3 Hz. Stimulation started at 50 dB sensation level (SL), decreasing in 10 dB steps to 0 dB SL. Here, 'dB SL' refers to the stimulus level above the auditory threshold level of the subject, as determined from conventional audiometry. At each stimulus level, two replicates were collected for every subject. The insert phones and associated cables were screened to minimize electromagnetic artefacts. The number of stimuli contributing to each coherent averaged response was K$\approx$2000. Two recordings were made at each stimulus intensity, in each subject. The acquired raw signals were band-pass filtered between 30 and 2100 Hz in order to emphasize wave V - which is the most important feature of ABRs (Figure 5.1). In addition a notch filter (50 Hz) was applied to remove mains noise. The signal was sampled at 5 kHz. The ABR was then obtained by coherently averaging the ensemble of data segments following each stimulus. The bootstrap method then uses both the averaged waveforms and the raw recorded signal, prior to averaging. The latter, containing spontaneous background cerebral activity, and noise as well as the ABR, will be referred to as the 'EEG'.

5.3 Bootstrap technique for detecting the ABR

5.3.1 Overview of algorithm

The bootstrap technique is widely used in assessing the confidence interval, estimating standard errors and testing hypotheses. The bootstrap method proposed here for detecting the response is actually a significance test of the null-hypothesis that 'no

response is present', and applied to some parameters representing the main features of the signal. First we will introduce the parameters chosen here. Then the detailed bootstrap process will be described.

5.3.2 Parameters used in detecting ABRs

For click stimuli in adults, a time window of 10 ms or 15 ms is usually sufficient to record the ABR, because wave V occurs in normal individuals within 5 to 6 ms of the stimulus at high intensities and within 8 to 9 ms for intensities near the auditory threshold (Hood, 1998). We kept the analysis window from 5 to 15 ms, which should in all cases include wave V. The four parameters described below were then calculated from the ABRs. Each of these provides a measure of the strength of the stimulus response, and is calculated over the time-interval 5 -15 ms.

- *diff* (Lv et al., 2004a,b), is the difference between the maximum and minimum value of the ABR, as shown in Figure 5.1;

- *power* is the mean power of the ABR:

$$power = \frac{1}{M}\sum_{i=1}^{M} x[i]^2 \tag{5.3.1}$$

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

- F_{sp} is an estimate of the signal-to-noise ratio of the evoked potential, which has been used extensively in detecting ABRs (Elberling and Don, 1984; Don et al., 1984)(see section 3.3.2):

$$F_{sp} = \frac{var(\overline{ABR})}{var(SP)/K} \tag{5.3.2}$$

  where $var(\overline{ABR})$ is the variance of the coherently averaged signal between 5 and 15 ms after the onset of the stimulus, and $var(SP)$ is the variance of the ensemble of K (2000 in our application) stimulus-responses at a single point (10

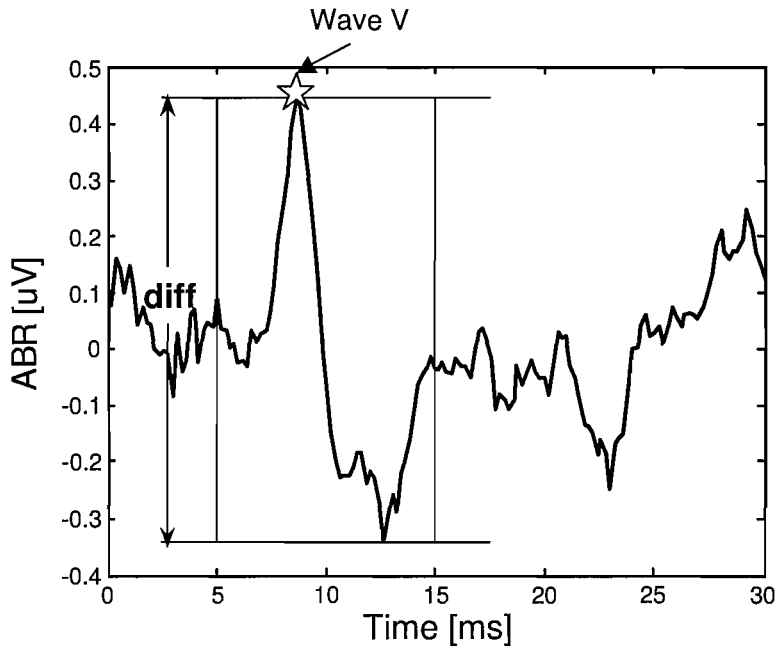


Figure 5.1: The ABR for one subject (click stimulation at 30 dB SL). The vertical lines at 5 and 15 ms show the region of the response that was used in analysis. The parameter *diff* gives the range of the ABR within this interval. The symbol ★ indicates wave V.

ms after stimulus onset was chosen in this work). Thus $var(SP)$ is obtained from the ensemble of signals before averaging, and represents the power of the noise (background activity), and $var(\overline{ABR})$ is found from the coherent average and corresponds to the power of the ABR.

- $\pm$*difference* (see 2.3.3) is an alternative estimate of the signal-to-noise ratio (Wong and Bickford, 1980) and is found by first allocating the even-numbered stimulus responses to one ensemble, and the odd-numbered ones to another. The coherent average of each of these two ensembles is then found. Hence,

$$\pm difference = \frac{var(Sum)}{var(Diff)} \tag{5.3.3}$$

where the numerator refers to the variance of the sum of the two averages, calculated over the time-window from 5 -15 ms following the stimuli, and the

denominator to the variance of the difference of the two averages. Clearly, if there is a strong stimulus-response, the sum of the averages will be much larger than their difference (where stimulus responses are cancelled), leading to relatively large $\pm$*differences*.

Following the calculation of these parameters, the statistical significance of each is tested against the null-hypothesis (H_0) of 'no stimulus-response'.

5.3.3 Bootstrap test

The bootstrap method (Zoubir and Boashash, 1998; Efron and Gong, 1983; Efron, 1979a,b, 1981b) is based on repeatedly drawing random samples (with replacement) from the original data. The parameter of interest is then calculated from these 're-samples', building up an estimate of the sampling distribution of the estimated parameter (we use symbol θ to denote any of the four parameters described above). The bootstrap method allows confidence limits of the estimate to be determined, or the statistical significance (with respect to the null hypothesis) to be tested - as in the current application.

First the coherent average of the EEG is calculated by averaging the K stimulus-responses, from which θ is found. We then apply the bootstrap test, by selecting K random points (the bottom plot of Figure 5.2) anywhere throughout the recorded raw signal, and use these as starting points in obtaining an ensemble of K randomly selected segments. Thus, at this step we ignore the actual timing of the stimuli and use random 'trigger points'. A uniform distribution of starting points covering the entire length of the recorded data is used. The new ensemble is averaged to form an 'incoherent average' (because it is not synchronized with the stimulus-timing- see Figure 5.3), for which the parameter θ is again calculated. The parameter, from the 'bootstrap' resample, will be denoted as $\theta^\star$. The bootstrap resampling process is then repeated $L = 499$ times, and a 'bootstrap distribution' of $\theta^\star$ is obtained. This provides an estimate of the sampling distribution of the parameter $\theta^\star$ as would be expected if there is no stimulus responses present (H_0). By comparing θ with the distribution of $\theta^\star$ (see Figure 5.4), the fraction of $\theta^\star$ that are larger than θ is found: this is the estimated p-value. If this is smaller than some chosen significance level α (say $\alpha = 5\%$), we reject the null-hypothesis of no response (Figure 5.4, right) and

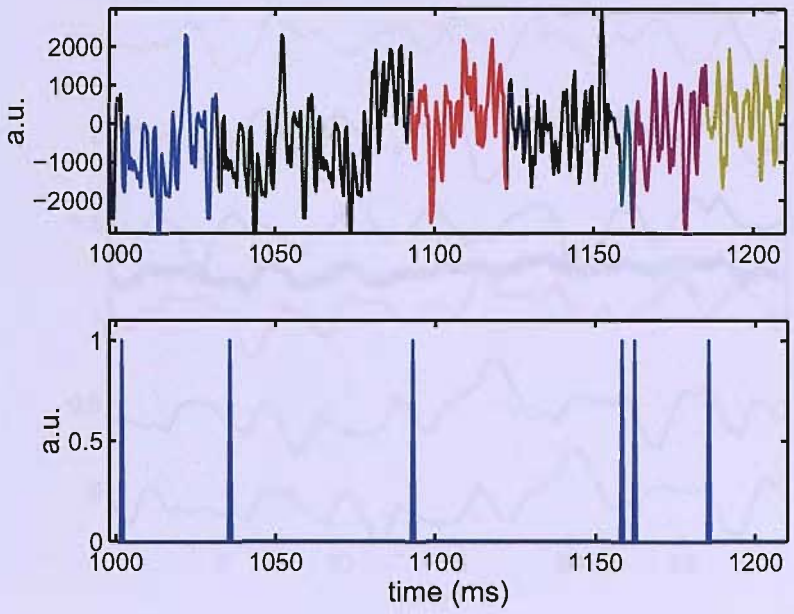


Figure 5.2: Similar to Figure 2.1, this figure shows the process of randomly selecting the segments of the signal. With the random starting points indicated in the lower part, we obtain the random segments of the AEP. The segments correspond to the random starting points and are not time-locked to the stimuli.

consider the value of θ to be statistically significant, i.e. a response has been detected. If all $\theta < \theta^*$, we say $p < 1/L$ (i.e. $p < 0.002$ in our case of $L = 499$). If $p > \alpha$ and θ is towards the left of the distribution of θ^* (Figure 5.4, left plot) we accept the null hypothesis of no response.

5.4 Evaluation of the bootstrap method

The bootstrap method was evaluated by applying the technique to different simulated signals, e.g. background EEG (noise) only and background EEG plus ABR, and recorded signals. Simulations only containing background EEG were used to estimate the false positive rates, and then determine whether the rates were close to the expected $\alpha = 5\%$. Simulations with both background EEG and ABR were employed for testing the power of the bootstrap technique, as a function of the signal-to-noise

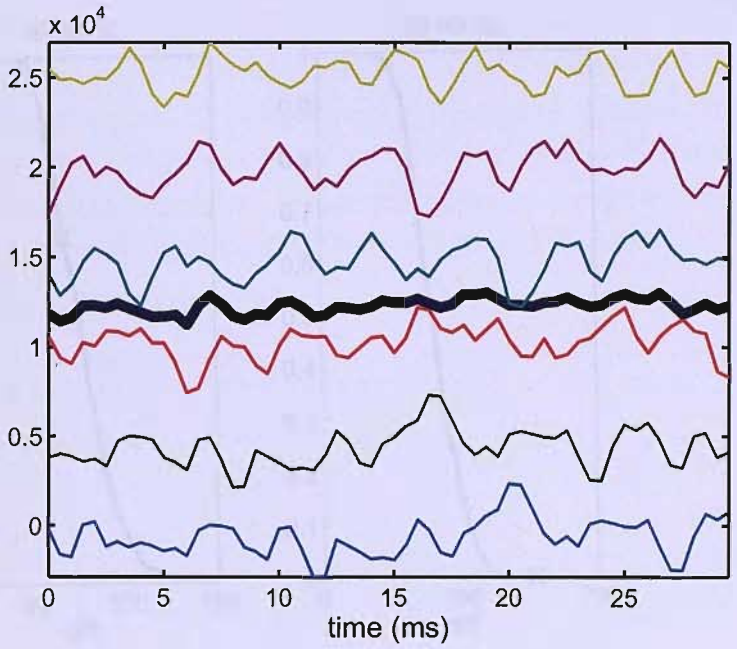


Figure 5.3: Similar to Figure 2.2, the randomly selected segments (6 thin lines) are averaged at each time point. The result of the averaging is shown by the thick line. For the ABR, usually 2000 responses are included.

(SNR) ratios. Finally, the bootstrap technique was also applied to the recordings to estimate the hearing threshold and investigate minimal number of sweeps required for detection.

5.4.1 Monte-Carlo simulations

In order to test the proposed methods, first a Monte-Carlo study on simulated signals with no stimulus response, was carried out. The aim was to determine whether the selected false positive rate ($\alpha = 5\%$) is actually obtained, when no response is present. The method to simulate these signals was described in section 4.4, and the specific order here was found to be 16 (Figure 5.5), according to the FPE. It was found that the FPE did not give a minimum but showed an initial sharp decrease, and after a 'knee' an almost flat section, where higher orders would lead to minimal improvements in FPE. The order chosen corresponds to the point just after the 'knee'.

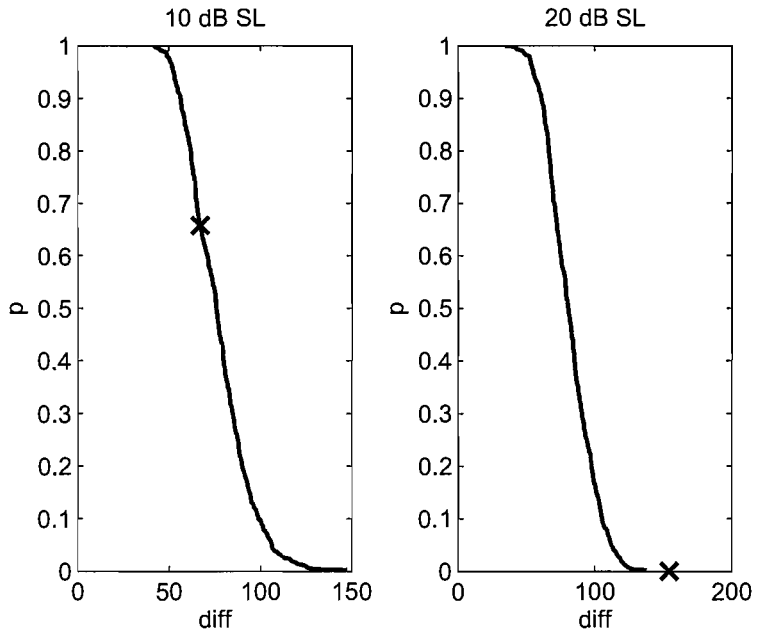


Figure 5.4: Bootstrap distribution of diff* from one subject at two different stimulus intensities. The p-value gives the fraction of cases (out of L=499) which were larger than a given value of diff. The x marks the value of diff obtained from the original data, and the corresponding p-value gives the statistical significance of that value. The example on the left did not give a statistically significant response (p=0.65), but for the one on the right, a response is detected ($p < 0.002$).

Then 500 simulated EEG signals were generated. All four parameters (*diff*, *power*, F_{sp}, $\pm$*difference*) were calculated from the coherent average of these signals (with trigger points at 30 ms intervals, and analysing the time-interval from 5 - 15 ms following each stimulus) and tested the significance (with $\alpha = 5\%$) using the bootstrap method. Since this signal does not contain a stimulus response, false-positive detection of a stimulus response is expected in approximately 5% of cases.

Then the power of the proposed method to detect responses when present was investigated. To this end, simulated ABR data was generated by adding a 'response' to a random background EEG signal. The stimulus response used corresponds to the coherent average from one of the signals recorded in a normal subject at 40 dB SL, which was then multiplied by a gain factor to obtain the desired SNR. This process

was carried out for nine different signal-to-noise ratios (SNR= -20 dB to 20 dB in the steps of 5 dB, calculated on the averaged signals, corresponding to -53 to -13 dB in the raw data). The background EEG signals were obtained by the same AR process used above. At each SNR, 500 simulated ABR data were generated. As before, the four parameters were calculated and their significance tested using the bootstrap method. The fraction of these 500 signals, at which $p < 0.05$ was obtained, indicate the power of the method.

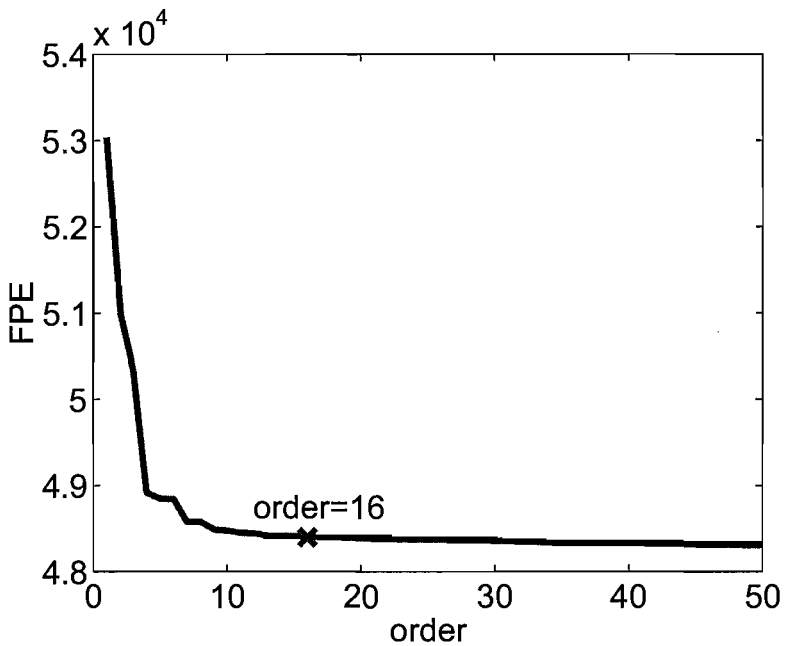


Figure 5.5: The relationship between order and Final Prediction Error (FPE) using a recording with stimulation at 0 dB SL.

5.4.2 Application to recorded signals

The bootstrap tests were then applied to the data recorded from the normal subjects, and hearing thresholds were found for each of the four parameters. The threshold was defined as the minimum stimulus intensity at which $p < 0.05$ (with $p < 0.05$ for all higher stimulus intensities also). We also show the change in hearing threshold when

$p < 0.01$ is used. These thresholds were compared to those determined by three experienced audiologists, who independently inspected the ABRs visually. Furthermore, inter-observer reliability for the visual inspections in this experiment was measured by Cohen's Kappa statistic (Altman, 1991). Kappa is defined as the 'proportion of observed agreement after correction for chance agreement'. Its value is between 0 and 1, which accounts for the range from poor to excellent reliability.

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

5.5 Results

5.5.1 Simulation

False positive rate tested by Monte-Carlo simulation

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

Sensitivity in presence of response

The results of the simulation with added responses are shown in Figure 5.6. As expected, the percentage of detected responses consistently increases with the increase

of the SNR levels for all four parameters. For all parameters (*diff*, *power*, F_{sp} and $\pm$ *difference*) results converge to 100% detection at high SNR, and to the expected $\alpha = 5\%$ at low SNR. At mid-range SNRs (from -38 dB to -23 dB), there is no significant difference between results for F_{sp} and *power* (t-test, $p > 0.05$), but *diff* and $\pm$ *difference* are better and worse, respectively (t-test, $p < 0.05$).

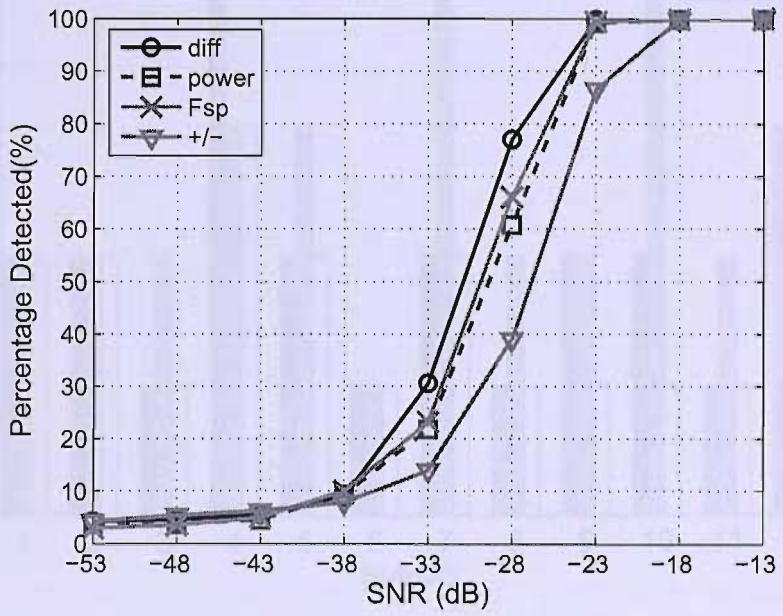


Figure 5.6: Percentage of responses detected as a function of signal-to-noise ratio (SNR) of the raw data. Results correspond to K=2000 averages (SNR=-20 to 20 dB in the coherent average).

5.5.2 Recorded data

Hearing threshold by subjective inspection

Three experienced audiologists determined the hearing thresholds by comparing the two replicate coherent averages of ABR data at the same stimulus intensity, and then finding the minimal stimulus level at which a consistent response was obtained. The results are given in Figure 5.7, showing quite large variations between raters, consistent with the observations in (Mason, 1984), and underlining the need for objective

methods for response detection.

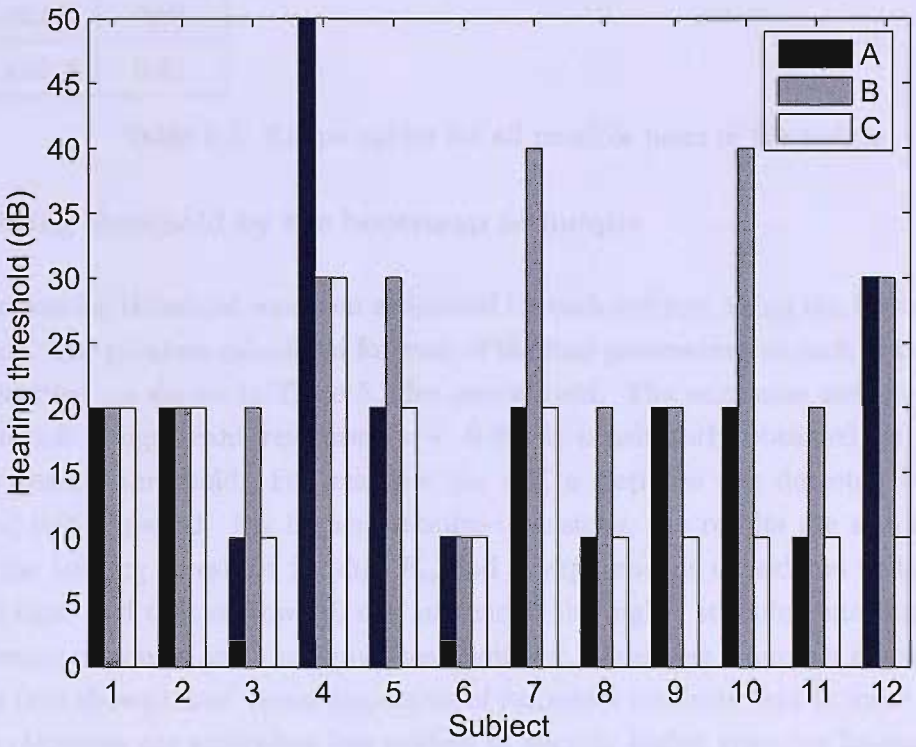


Figure 5.7: Hearing thresholds for 12 normal hearing subjects, as determined from the ABR by three experienced audiologists (A, B, and C) through visual inspection. For each subject, the three bars represent the hearing threshold estimate of A, B and C respectively.

Inter-observer reliability was measured by the Kappa statistic (Altman, 1991). The values of Kappa for all three pairs of judges are shown in Table 5.1. The common interpretation of the reliability is that Kappa should be no less than 0.90 to be regarded as high (Arnold, 1985), i.e. for there to be a good agreement between judges. Clearly this is not the case in Table 5.1.

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

Table 5.1: Kappa values for all possible pairs of the judges

Hearing threshold by the bootstrap technique

The hearing threshold was then estimated for each subject, using the bootstrap technique. The p-values calculated for each of the four parameters, at each of the stimulus intensities are shown in Table 5.2 for one subject. The minimum stimulus intensity at which a significant response ($p < 0.05$) is consistently obtained, is considered the hearing threshold. For example, for *diff*, a response was detected from 10 dB ($p < 0.05$) upward. For higher stimulus-intensities, the results are also significant. So the hearing threshold for *diff*, F_{sp} and $\pm$ *difference* is considered to be 10 dB in this case, and that of power 0 dB. In general the higher stimulus intensities provide stronger responses and lower p-values. However, there were a number of exceptions to this (not shown), and visual inspection of responses confirms that in some recordings the responses are somewhat less evident at slightly higher stimulus intensities. Note that for these cases we define hearing threshold to be the lowest stimulus intensity at which $p < \alpha$ and for which all higher stimulus intensities also showed a significant response.

Comparisons between different methods

In order to compare the difference of hearing threshold between subjective inspections and objective bootstrap approach based on the four parameters, we calculated the average hearing threshold (AHT) (Table 5.3) of 12 subjects. The parameter power appears to be the most sensitive in detecting a response.

The median value of the three (A, B, C) subjectively evaluated hearing thresholds (MHT) of each of the 12 subjects were calculated and compared to the hearing thresholds for each of the four parameters (HT) obtained by the bootstrap method ($p < 5\%$).

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

Table 5.2: Examples of p-values for the four parameters at different stimulus intensities, for one subject. p-values are obtained from the bootstrap test using roughly 2000 stimuli. The p-values marked in bold indicate the hearing threshold.

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

Table 5.3: Average hearing threshold by subjective inspection and objective bootstrap technique. * Significantly different to the threshold found with parameter power (sign-test, $p < 0.05$).

Results are shown in Figure 5.8. For power and diff, these are lower or equal to MHT in 11 of the 12 subjects; for F_{sp} and $\pm difference$, this is the case in 10 subjects.

Significance level

Figure 5.9 shows the hearing thresholds obtained with $\alpha = 1\%$ rather than $\alpha = 5\%$ used in the previous results. For the 12 subjects, the hearing threshold remains the same in most cases, and increases by 10 or 20 dB in three cases for *diff*, two cases for *power*, one case for F_{sp}, and four cases for ± *difference*.

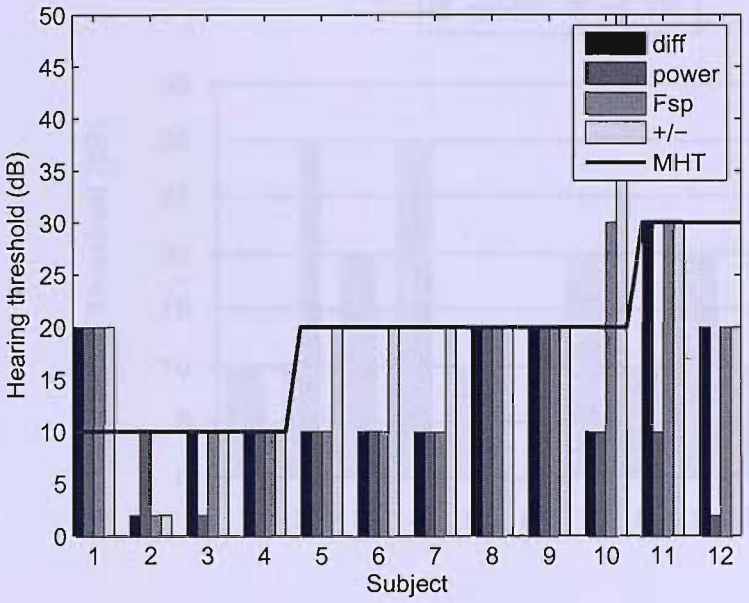


Figure 5.8: Comparison between median hearing threshold (MHT - median of A, B and C, solid line), and hearing thresholds from the four parameters (the bars from left to right correspond to the parameters *diff*, *power*, F_{sp}, $\pm$ *difference*). In most cases, the latter are smaller than, or equal to the corresponding MHT.

Minimal number of stimuli for detection

We also investigated the effect of the number of epochs (stimulus responses) recorded, on the ability to detect a response using the bootstrap approach. We therefore applied the bootstrap tests to progressively increasing numbers of stimuli. Figure 5.10 illustrates the results for the parameter *power*. As expected, the fraction of cases in which the ABR is detected increases with increasing stimulus intensity and also with the number of sweeps. At 40 and 50 dB SL, 800 stimuli were enough to detect the response in all of the 12 subjects with the parameter *power* (see Figure 5.10); 1100 stimuli were required for *diff* and F_{sp}. For $\pm$ *difference*, 2000 stimuli at 50 dB were required to achieve 100% detection. As the other three figures are very similar to Figure 5.10, they are not shown here.

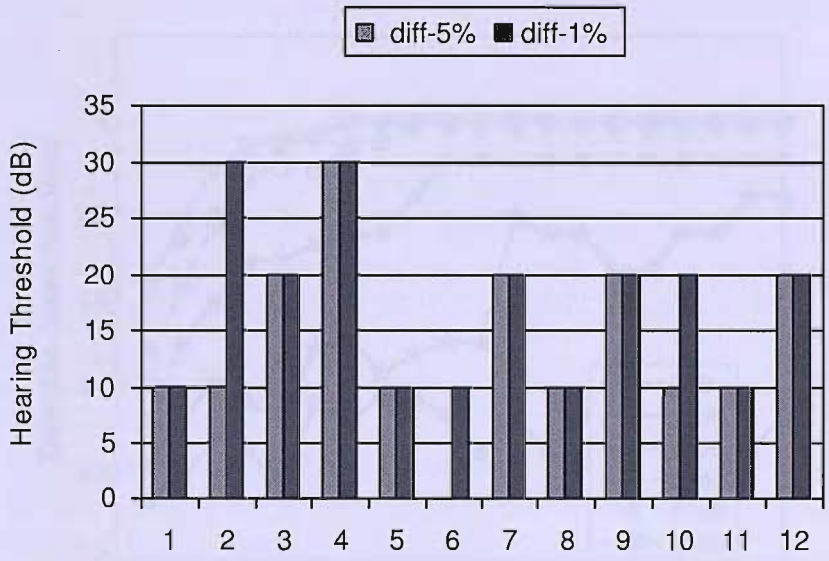


Figure 5.9: Comparison of the hearing threshold of $\alpha = 5\%$ with that of $\alpha = 1\%$. Here the hearing thresholds were estimated by parameter *diff* (other parameter showed a similar pattern of results.

5.6 Discussion

The need for objective methods to detect evoked responses was clearly illustrated by the example of the ABR presented here. There was considerable disagreement between the subjectively selected hearing thresholds given by the three experienced audiologists (A, B, C) and this was reflected in the relatively low values of Kappa. Techniques for the automated detection of evoked responses usually involve the calculation of a parameter, for which a threshold is then selected, above which the response is deemed to have occurred. The selection of this threshold may be based on experience and experimental work e.g.(Ozdamar et al., 1990). The bootstrap technique presents a very attractive and flexible alternative, by providing a simple means of estimating the statistical significance (p-value) of a parameter. It does so by comparing the parameter-value to that expected under the null-hypothesis of no stimulus response.

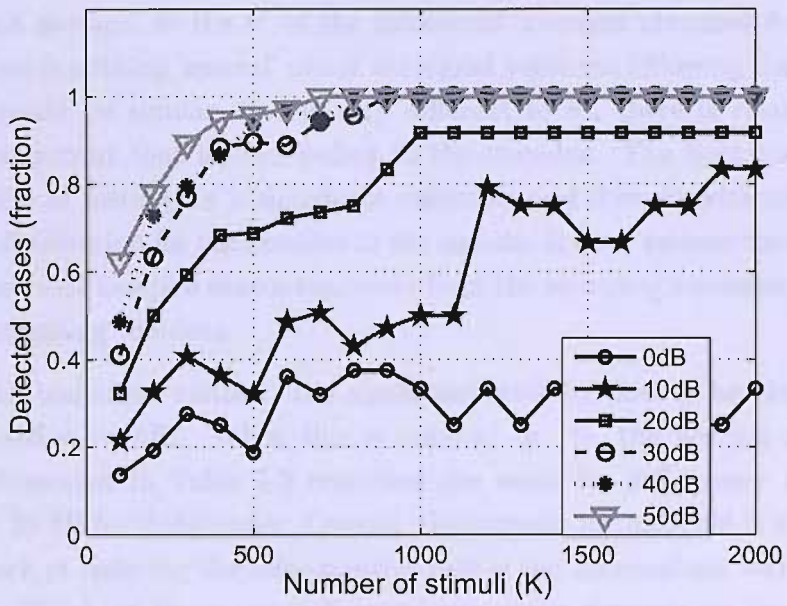


Figure 5.10: The fraction of cases in which the ABR was detected is shown as a function of number of epochs (stimuli) and the parameter *power*. The bootstrap method ($p < 0.05$) was applied with increasing numbers of stimuli, and stimulus intensities between 0 and 50 dB SL. Note that for this result the signals were broken down into non-overlapping blocks of K stimuli, such that for example at K=100 each of the 12 subjects provided 20 blocks, but at K=2000, only a single block.

The use of the bootstrap method circumvents potentially intractable statistical analysis, which would otherwise have to be carried out, in order to obtain a closed-form solution for each case. Such analyses would also usually involve assumptions regarding signal statistics, which it may be difficult to justify or test, for each recording. Conventional statistical analysis is also complicated in this work by the autocorrelation of the signals, such that successive samples are not independent. This, for example is the reason why the F_{sp} does not correspond to the F-statistic, with the degrees of freedom corresponding to the number of samples analysed (Elberling and Don, 1984).

The bootstrap method makes few assumptions about the data, which is one of its main benefits. A 'significant' response to stimulation may be considered to be one in which the parameter θ of the coherent average has 'surprisingly' large values.

The bootstrap method allows this to be tested directly, by comparing the θ from the coherent average, to the θ^* of the incoherent averages obtained from the same data. If there is nothing 'special' about the signal segments following the stimulation, θ and θ^* would be similar; if θ is very different to θ^*, there is clear evidence of a signal component that is time-locked to the stimulus. The bootstrap method is thus intuitive in testing for a significant response, and does so without assuming a statistical distribution for the samples in the signals. It does assume that the signal is ergodic, such that samples drawn randomly from the recording represent the 'random process' generating the data.

In using the bootstrap method, the significant level (α) has to be chosen. In this work we used $\alpha = 5\%$. When this is reduced to 1%, the hearing threshold for the case illustrated in Table 5.2 remained the same for *diff*, *power* and F_{sp}, but increase to 20 dB for $\pm$*difference*. Overall, the increase in threshold is small. Clearly the drawback of reducing the false-positive rate is the concomitant increase in false-negatives. Which of these errors is more important depends on the application: for example, in monitoring depth of anaesthesia (Aceto et al., 2003) a significant mid-latency auditory evoked responses may indicate that the patient is awakening, which might require prompt intervention by the anaesthetist. Thus high sensitivity to the presence of a response (and hence higher α) is desirable. On the other hand, in screening tests for hearing loss, a false positive response may lead to missing a hearing impairment, and a lower false-positive rate is desirable.

The bootstrap technique can deal with varying numbers of stimuli, while maintaining pre-defined false-positive rates. In Figure 5.10, it is evident that at 40 and 50 dB SL, 800 stimuli were enough to detect the response, which is rather less than the 2000 recommended in the literature. Thus in normal hearing subjects, at these levels of stimulus the duration of the test could be considerably reduced, as already indicated by Don et al. (Don et al., 1984).

Although we obtained the encouraging results, some limitations are evident. Inevitably, movement artefact and stimulus artefact are present with the raw recorded EEG. Therefore, a scheme to remove them is desirable. In Chapter 6, we will provide more details about the methods of artefact rejection.

Chapter 6

Artefact Rejection

6.1 Introduction

The bootstrap method was proposed and tested on the recordings and simulated signals in the last chapter. Reasonable false positive rates and detection fractions were obtained, as expected. While very encouraging results were obtained, it became evident that a clinically useful method would need to be able to deal adequately with artefacts (mainly those due to movement and stimuli) in the signals. It is the purpose of this chapter to address this issue, and propose three artefact rejection schemes and evaluate them. They are: movement artefact rejection (called MAR-bootstrap, sometime MAR in brief), stimulus artefact rejection (SAR), and a combination of MAR and SAR named as SMAR-bootstrap. The previous bootstrap technique without any artefact rejection scheme will be called Basic-bootstrap, in order to distinguish this from the three modified bootstrap methods.

In this chapter, first the three artefact rejection schemes will be introduced. The simulated components to generate the simulated data, and the acquisition of the EEG recordings (data Set B) are described. In addition to the parameters described in the Chapter 5, two alternative parameters are investigated and will be employed in the Basic and the modified bootstrap methods. Following that, MAR, SAR and SMAR bootstrap methods will be applied to the different types of simulated data and on data Set B, and their results and evaluations will be provided. Finally, a discussion of some points on the artefact rejection scheme will be given.

6.2 Methodology

6.2.1 Movement artefact rejection

Movement artefacts (MA) here refer to the muscle artefacts mainly originated from neck and jaw muscles, and any sudden or randomly occurring artefacts. As the muscle artefact is in the same frequency range as the ABR, the detection of the ABR by filtering technique becomes difficult (Pantev and Khvoles, 1984). Random artefacts may appear in any time of the recording. A rejection scheme is desirable to deal with this data. In the literature, a widely used rejection method called amplitude threshold (or limit) excludes any sweep with samples exceeding a certain pre-set level (Cebullar et al., 2000; Elberling and Don, 1984; Wong and Bickford, 1980; Ozdamar and Kalayci, 1999; James and Lowe, 2003). However, for different sets of the recordings on the different equipments and depending on the experience of the experts, the thresholds chosen differ in the range between $\pm 10\mu V$ and $\pm 30\mu V$. In the current work, $\pm 30\mu V$ was chosen, based on the suggestion of an audiologist who has more experience on auditory evoked potential and are familiar with the recording system.

In general, this rejection threshold is simply applied to the raw EEG signals and the unwanted sweeps are then eliminated. However, for the bootstrap method, the process of randomly resampling is included, and requires a continuous signal. Removing entire sweeps in the raw data, prior to bootstrap resampling would generate additional artefacts because the removal of some undesirable sweeps breaks the signal for resampling into separate blocks (having one or more sweeps). The rejection scheme presented here is called movement artefact rejection (MAR), and overcomes this problem.

Before coherently averaging the raw EEG signal, the threshold is applied to all the sweeps and the unwanted sweeps were removed. Then the remaining sweeps ('good') are used for coherent averaging, and the number of the 'good' sweeps is recorded. The signal parameter based on the coherent average is estimated. With the MAR scheme, similar to the Basic bootstrap, the resampling is on the raw signal and bootstrap resamples exceeding the threshold are again rejected. The resampling is repeated until the number of 'good' sweeps matches that of the coherent average. These 'good' sweeps of a 'bootstrap' signal are then used to calculate the 'incoherent' average. The

p-values are then found in the same manner as for the Basic bootstrap.

6.2.2 Stimulus artefact rejection

The sources of stimulus artefacts (SA) have been identified due to capacitive and inductive coupling of the transducer and its electrode leads (Cooper and Parker, 1981). In the current work, the insert earphone was used and can reduce SA greatly, compared to the supraaural earphones. However, there is a possibility that SA is present in the recording, and for other earphones the SA could be a major problem. In order to make the bootstrap method be able to used in many conditions, the stimulus artefact rejection scheme should be developed.

For the bootstrap method, in estimating the coherent average, the stimulus artefact (SA) can be largely avoided, by an appropriate selection of the time-window. However this is not possible in bootstrap resampling because in the resampling procedures all the samples have the same chance of being selected. In the presence of strong stimulus artefacts, the bootstrap method may overestimate the p-value for the following reason. The randomly selected segments will sometimes include the stimulus artefact, and this may provide larger than expected values of θ^*. Consequently (see Figure 5.4), increased values of p may be obtained, thus reducing the probability of rejecting the null hypothesis, decreasing the sensitivity of the test. An adaptation of the bootstrap method is thus required, in order to overcome this problem. An approach akin to that employed in the $\pm$ difference is proposed, as illustrated in the two blue boxes of Figure 6.1.

Firstly the signal is split in the middle, to obtain two segments, each corresponding to $K/2$ stimulus-responses. Then these are added to give a new signal (x_+) right branch in Figure 6.1 from which the coherent average and θ are obtained. Next, the two halves are subtract from each other (sample by sample), giving the signal x_- (left branch in Figure 6.1). Since the stimulus artefact and stimulus-response is cancelled in x_-, this signal conforms to the null-hypothesis of no response (ABR or artefact) to the stimulus. This signal is then resampled to give the bootstrap distribution of θ^*.

6.2.3 Stimulus and movement artefact rejection

SA and MA can both contaminate the ABR recordings simultaneously and therefore it is desirable to develop a scheme to remove both, within a bootstrap method. Such a method, combining both SAR and MAR is therefore proposed as SMAR-bootstrap, and its complete procedures are shown in Figure 6.1.

There are two considerations behind the idea of combining the two schemes. One is the order, i.e. SAR before MAR or the opposite, and the other is the value of the threshold of MAR. As regards the first, it is necessary to apply SAR before MAR. The reason is that the MAR procedures will remove samples and make the data discontinuous, as mentioned earlier (section 6.2.1). That will generate errors in the calculation of the parameters, and further provide poor results in detection. The second issue is the threshold value for MAR. The procedures of addition and subtraction within the SAR scheme results in the increase of the variance of the random component of the signal. When the signal is split into two halves, these are regarded as two random signals assumed to have the same mean and variance. If the two variables are independent, the mean of their sum is the sum of the means, and the variance of the sum is the sum of their variances; when we subtract one random variable from another, the mean of the difference is again the difference of the means, and the variance of the difference is the sum of their variances. Therefore, following the assumption that the two random variables have the same variance, the standard deviation of x_- is $\sqrt{2}$ times that of the raw signals. The rejection threshold of SMAR is then increased by the same amount, in our case, $\pm 30\mu V \times \sqrt{2} = \pm 42.43\mu V$.

6.2.4 Evaluation by area under the ROC curve

The receiver operating characteristics (ROC) curve is an effective method of evaluating the performance of diagnostic tests (DeLong et al., 1988; Park et al., 2004). The curve is constructed by varying the cut-off points for detection of a response (here the significance level α), and plotting the sensitivity against 1-specificity. Good performance of the detector would be indicated by high sensitivity and specificity, and thus a curve running close to the top-left corner of the graph. The area under the ROC curve (AROC) is often used to assess discrimination or accuracy of detection methods

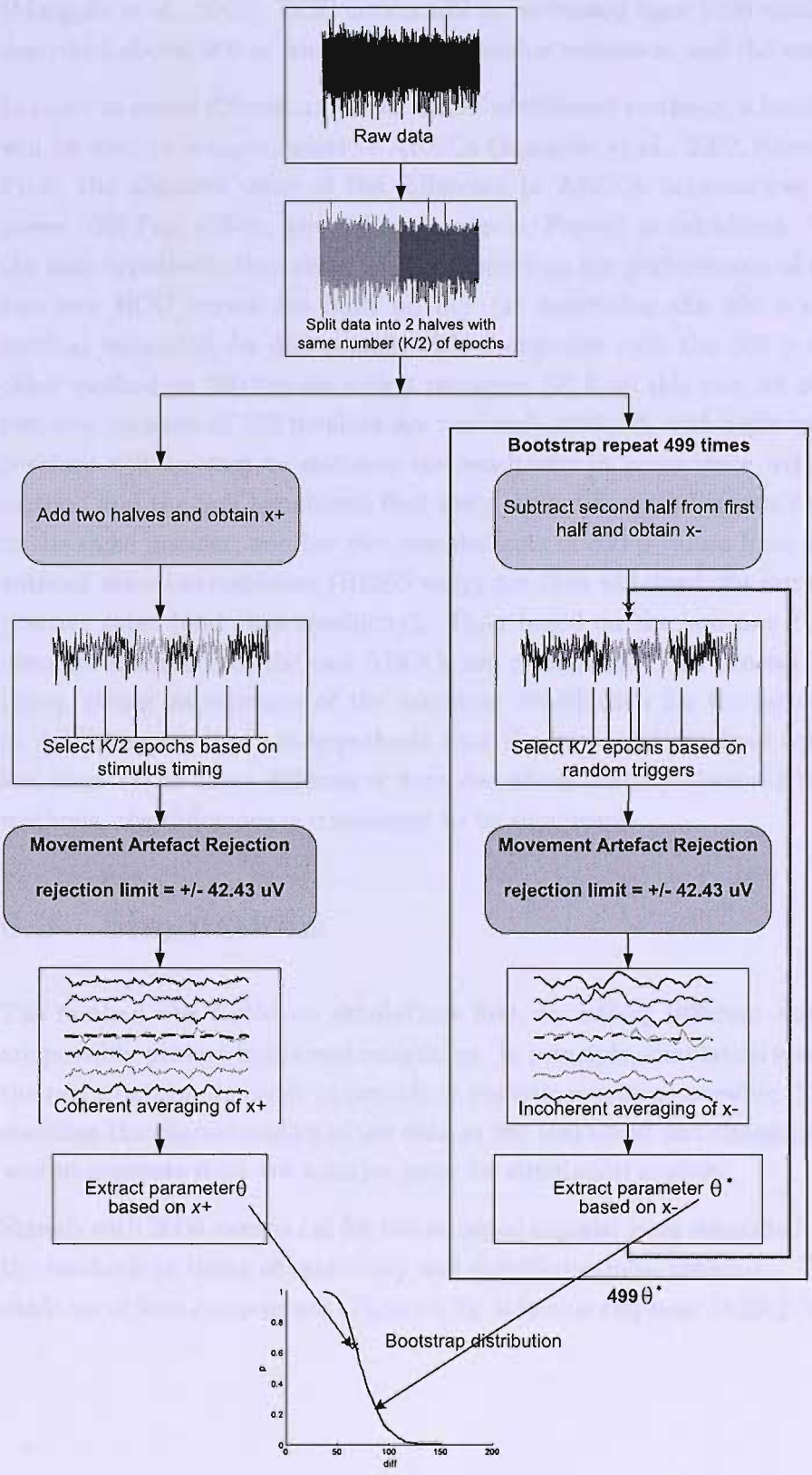


Figure 6.1: Procedures of SMAR-bootstrap.

(Margolis et al., 2002). ROC curves will be estimated from 1000 simulated signals as described above, 500 of which contain stimulus responses, and the remainder do not.

In order to assess differences in the AROC of different methods, a bootstrap technique will be used to compare pairs of AROCs (Margolis et al., 2002; Ramos et al., 2006). First, the absolute value of the difference in AROCs between two methods (diff-power, diff-Fsp, diff-cc, power-Fsp, power-cc, Fsp-cc) is calculated. Then, based on the null-hypothesis that there is no difference in the performance of two techniques, two new ROC curves are built up by: (1) combining the 500 p-values from one method estimated for 500 signals with a response with the 500 p-values from the other method on 500 signals with a response; (2) from this new set of 1000 p-values, two new datasets of 500 p-values are randomly selected, with replacement, and these p-values will be used to estimate the sensitivity in accordance with the bootstrap method and the null hypothesis that there is no difference between two methods; (3) in the same manner, another two new datasets of 500 p-values from simulated signal without stimulus-responses (BEEG only) are then obtained, for estimating the false positive rates (and thus specificity). Then based on the two new ROC curves, the absolute difference of the two AROCs are calculated. This process is repeated 499 times, giving an estimate of the sampling distribution for the estimated difference in AROC under the null-hypothesis that the two methods have equal AROCs. If less than 5% of these differences exceeded those initially observed between the two methods, the difference is considered to be significant.

6.3 Simulations

The method was tested on simulations first, by adding different components which are possibly present in the real recordings. In principle, simulation work should mimic the real situation, in order to provide as realistic results as possible. Therefore understanding the characteristics of the data in the real world and choosing an appropriate way to generate data are a major issue for simulation studies.

Signals with 2000 sweeps (as for the recorded signals) were simulated, in order to test the methods in terms of sensitivity and specificity (false-positives). The signals were made up of four components (Figure 6.2): stimulus response (ABR) (upper-left plot),

stationary random noise simulating background (spontaneous) EEG (BEEG - upper-right plot)(James and Lowe, 2003), movement artefact (MA) (bottom-left plot), and stimulus artefact (SA) (bottom-right plot).

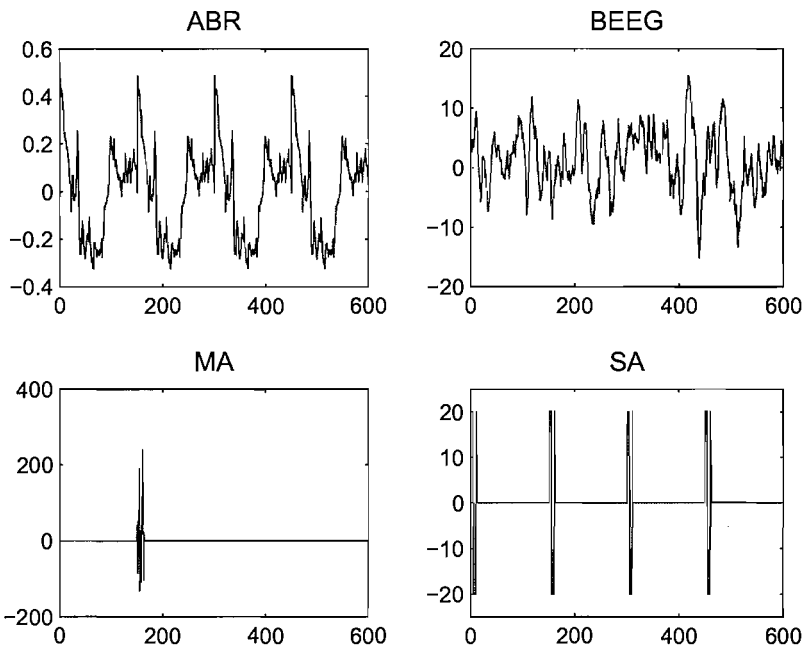


Figure 6.2: The four components of the simulated signals. The x-axis is sample number and y-axis corresponds to magnitude of the signals in μV.

6.3.1 ABR

The ABR included in the simulations is to test the sensitivity, i.e., in what percentage of cases does the X-bootstrap ('X' represents any one of the four methods) detect the response. Simulations without ABR are used to investigate false positive rates.

In order to approximate the real ABR recorded from the human brain more closely, ABR simulation was carried out in accordance with an assumption of a 40 dB SL or higher stimulus level. In this case, $K = 2000$ sweeps are usually sufficient to recover the ABR by coherent averaging. Thus, first the coherent average for recordings from

all normal hearing subjects stimulated at 40 and 50 dB SL, were calculated. Then by visual inspection of all coherent averages, one with a typical wave-shape , and in particular a clear wave V, was chosen. This provided the template ABR, and was periodically repeated K times to represent the ABR component of the simulated signal. The amplitude of this 'ABR' was then adjusted to provide the desired SNR level for different investigations.

6.3.2 Background EEG

The procedures for generating BEEG have been described in Chapter 4. In this chapter, the recording without a stimulus in Set B was used, and the Yule-Walker method employed to estimate the AR parameters. Five hundred Monte-Carlo simulations of BEEG were carried out to synthesize different simulated signals.

6.3.3 Movement artefacts

Movement artefacts are random and possibly appear any time in the recordings with large amplitudes. For simulated signals, the amount of the movement artefacts is of concern. In practice, if too many MA contaminate the ABR, this signal will be discarded. Bell (2003) proposed an algorithm of artefact rejection with a rejection level which would exclude about 10% of the sweeps from the data, and this was considered a realistic level. Therefore signals of K sweeps with 10% (in our case, $2000 \times 10\% = 200$ sweeps) having MA were generated.

Another issue is how to select the shape and amplitude of the MA. The random timing of the MA refers to two aspects: one is the location of MA in one sweep and the other is which sweep of the recording might be contaminated with the MA. Independent uniformly distributed random numbers were used for both and the standard deviation of the MA was set to $\pm 100\mu$ V (see Figure 6.2). Considering MA are usually not present for long periods of time, each MA lasted 3 ms (15 samples of one sweep).

Aim to test	No.	Type of simulations	Basic	MAR	SAR	SMAR
False Positive	1	BEEG	√	√	√	√
	2	BEEG+MA	√	√		
	3	BEEG+SA	√		√	
	4	BEEG+MA+SA	√			√
Sensitivity	5	BEEG+ABR	√	√	√	√
	6	BEEG+ABR+MA	√	√		
	7	BEEG+ABR+SA	√		√	
	8	BEEG+ABR+MA+SA	√			√

Table 6.1: Eight types of simulations containing different components were applied on four bootstrap methods for testing false positives or sensitivity. '√' indicates the X-bootstrap method carried out in that simulation.

6.3.4 Stimulus artefacts

Stimulus artefacts (SA) usually occur at an early latency (Hood, 1998) and can show a larger amplitude than the stimulus response. In the current study, the SA is represented by a square wave located in the first 2 ms of each sweep. As mentioned before, the rejection threshold for the MAR is $\pm$ 30 μV, and in order not to be removed by the MAR, the amplitude of the square wave was set to $\pm$ 20 μV. Therefore the MAR and SAR only have an effect on the corresponding artefacts.

6.3.5 Types of simulations

The simulations were mainly classified into two groups in terms of the aim for testing. The signals without ABR are BEEG, BEEG with MA, BEEG with SA, and BEEG plus MA and SA. These were used to test the false positives. Those with ABR are BEEG plus ABR, BEEG plus ABR and MA, BEEG plus ABR and SA, and BEEG plus ABR plus MA and SA. These are used to calculate the sensitivity. Then the four bootstrap methods Basic, MAR, SAR and SMAR were applied to the simulations, as given in Table 6.1. Each type of simulations included 500 signals.

6.4 Recorded ABR-Set B

The artefact rejection schemes (MAR, SAR and SMAR) and the Basic-bootstrap were tested on recorded ABR as well as the simulations. In this section, the data Set B will be described.

Sixteen normal-hearing adults (11 Males and 5 females), aged from 18 to 34 years old (mean 25.6 years), participated in this experiment, following ethical approval by the appropriate Institutional Committee. No subjects had a history of ear disease or unsuitable noise exposure, as confirmed by questionnaire. The hearing thresholds (pure-tone audiogram) were all better than 20 dB throughout the frequency range of 250-8000 Hz in both ears, all had normal-shaped tympanograms, and were further checked by otoscopy. The subjects lay comfortably on an examination couch in a sound-proof and electrically shielded booth throughout the tests. The stimuli settings and recording parameters have described in Chapter 3. The signals containing both the background EEG and stimulus-responses will be denoted as 'EEG', those containing only background EEG (under the no-stimulus condition) as 'BEEG'.

6.5 Additional parameters

In addition to the four parameters introduced in the Chapter 5, two more parameters are investigated and will be tested by the four bootstrap methods on all simulated signals and the real recordings (Set B). They will be evaluated in the coherently averaged ABR over the time-window from 5-15 ms after the stimulus.

6.5.1 Parameter abs

abs is the mean of the absolute value of the amplitude over the time-window, and obtained similarly to *power*, by replacing $x[i]^2$ with $|x[i]|$. The reason for choosing this is that when a sample with a large amplitude is present in the signal, the *power* is very sensitive, but *abs* is able to reduce the effect of this sample.

6.5.2 Parameter cc

The correlation coefficient (*cc*) is calculated between two replicates which are formed from the coherent averages of the even and odd numbered sweeps in the recording.

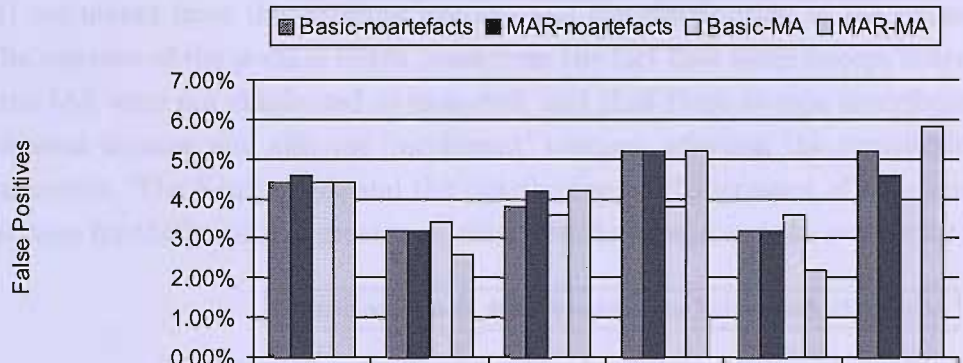
6.6 Application of movement artefact rejection scheme

The MAR will be tested on both simulations and recordings. Simulations without ABR components will give the false positive rates and those with ABR will provide the sensitivity. Set B of the recordings will be used for testing the sensitivity in hearing threshold estimation, and the false positive rates will be evaluated on 'real data' based on the recordings under the no-stimulus condition.

6.6.1 False positives for MAR in simulations

False positive rates were tested based on the well-controlled simulations which included background EEG and background EEG plus movement artefacts (MA), and results are shown in Figure 6.3. For the background EEG simulations, the MAR scheme provided similar false positive rates as the Basic bootstrap method, as expected. When no artefacts were present, MAR did not affect the results. When MA was added to the background EEG, the false positive rates of the Basic bootstrap still stayed in the range of 3.2%-6.8% according to the binomial distribution, however, those of the MAR scheme with parameter *power* and *abs* decreased beyond the range. The lower false positive rate is of less concern, compared to greater rate (beyond 6.8%), because in hearing screening, with the latter more than the expected 5% of subjects with impairment may be missed. For the MAR scheme with the other four parameters, the false positive rates were in the expected range.

The results suggest that the false positive rate is not greatly affected by the MAR scheme. In the next section, the sensitivity will be considered.



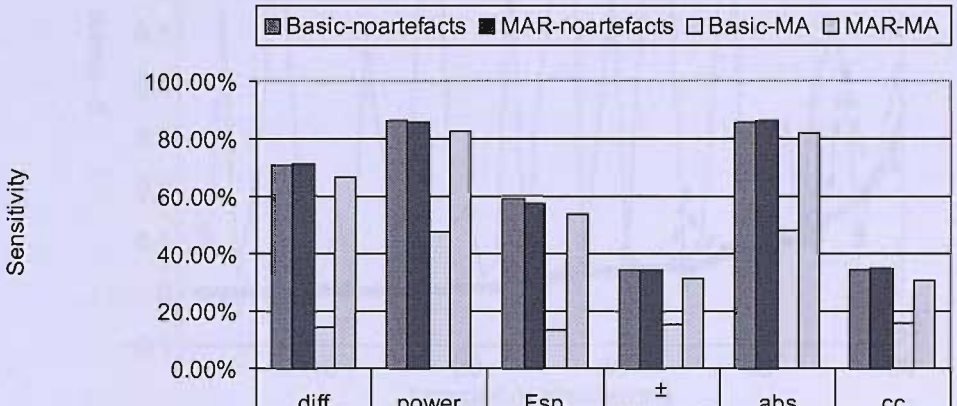
	diff	power	Fsp	± difference	abs	cc
Basic-noartefacts	4.40%	3.20%	3.80%	5.20%	3.20%	5.20%
MAR-noartefacts	4.60%	3.20%	4.20%	5.20%	3.20%	4.60%
Basic-MA	4.40%	3.40%	3.60%	3.80%	3.60%	4.00%
MAR-MA	4.40%	2.60%	4.20%	5.20%	2.20%	5.80%

Figure 6.3: False positive rates for simulations with and without MA measured by Basic and MAR algorithms, respectively. The $\alpha = 5\%$ significance level was used. Noartefacts refer to simulations without artefacts, and MA to those with movement artefact.

6.6.2 Sensitivity for MAR in simulations

Here two types of simulations differing by adding movement artefact or not, were employed on MAR and Basic bootstrap methods, respectively. The results in Figure 6.4 show that MAR and Basic bootstrap methods provide similar sensitivities (left two bars in Figure 6.4) when movement artefacts are not added to background EEG and ABR. This was in accordance with our expectations and meant that MAR scheme did not influence (in particular, did not reduce) the sensitivity when MA was absent. However, a statistically significant difference ($p < 0.05$, t-test) between these two methods (MAR and Basic) was present when artefacts were added (right two bars in Figure 6.4). The sensitivity of the Basic bootstrap method dramatically decreased, and that of MAR was much closer to that calculated from simulations without MA. The difference of sensitivity between 'MAR-noartefact' and 'MAR-MA' identified in the Figure, resulted from the increase of p-values related to the values of the parameter

(θ) calculated from the coherent average and the distribution of the parameter θ^*. The increase of the p-value might come from the fact that some sweeps contaminated with MA were not eliminated as expected, and thus these sweeps contributed to the coherent average and also the 'incoherent' average, affecting the distribution of the parameter. The θ increased, and the distribution of θ^* increased, if the degree of the increase for the latter was greater, p-value would increase and the sensitivity increase.

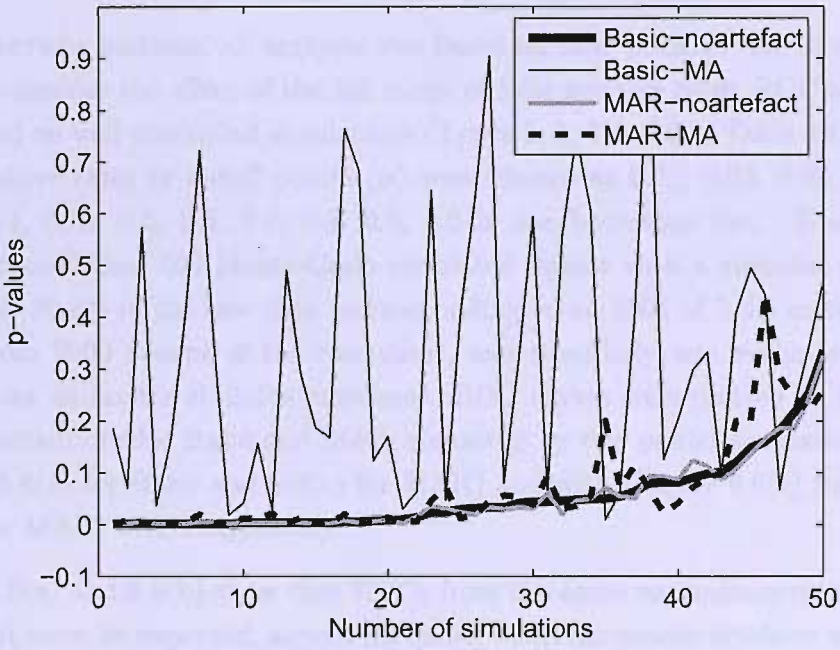


	diff	power	Fsp	± difference	abs	cc
Basic-noartefacts	71.00%	86.20%	58.80%	34.40%	85.60%	34.40%
MAR-noartefacts	71.20%	85.40%	57.40%	34.40%	86.40%	34.80%
Basic-MA	14.40%	47.80%	13.40%	15.60%	48.20%	16.00%
MAR-MA	66.60%	82.80%	54.00%	31.40%	82.00%	30.80%

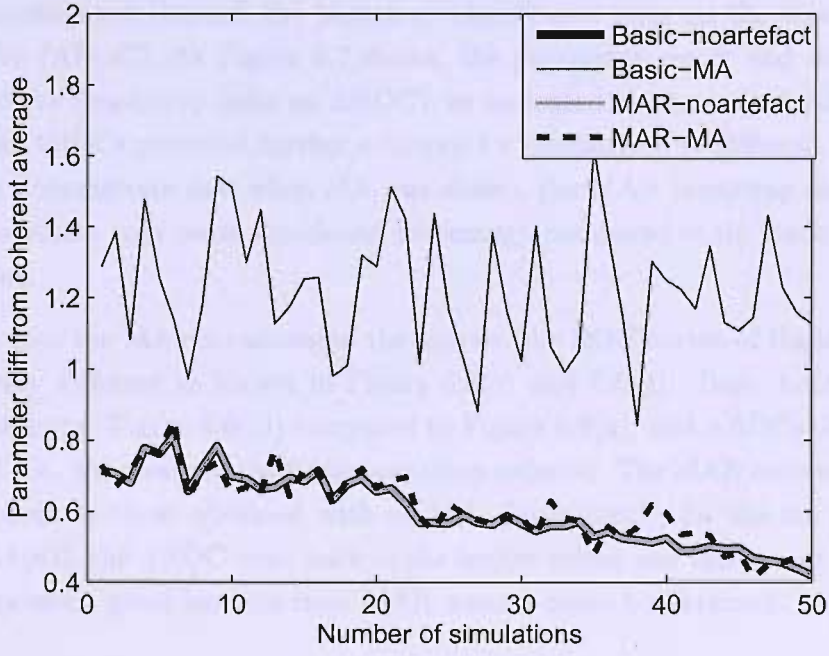
Figure 6.4: Sensitivity for simulations with and without MA measured using Basic and MAR algorithms, respectively.

In order to compare the results for these four conditions, an example for *diff* and its p-values are shown in Figure 6.5(a) and 6.5(b) which were obtained from 50 simulations with and without MA. When MA was present, the values of *diff* from the Basic bootstrap were always greater than those from MAR and their p-values from the Basic were greater as well.

When MA was present, the MAR scheme helped to recover the sensitivity lost with Basic bootstrap. These results therefore suggested that the MAR scheme was effective for signals with MA and furthermore that in the absence of MA, it did not reduce the sensitivity.



(a) p-values.



(b) Parameter *diff* from coherent average.

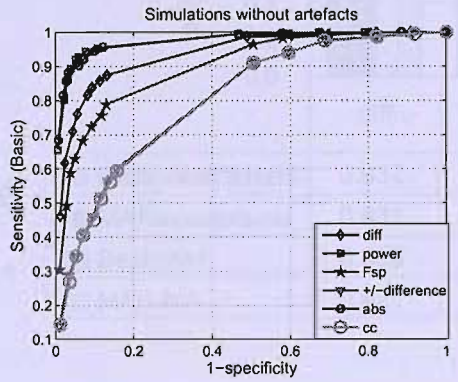
Figure 6.5: An example for parameter *diff* and its p-values on 50 simulations with a stimulus response and with and without MA using the Basic and MAR bootstrap methods. The simulations were sorted in order of increasing p-values (for Basic-noartefact).

6.6.3 ROC analysis for MAR

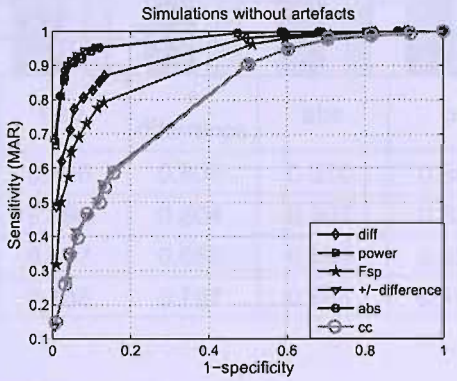
In the previous sections, all analysis was based on false positive rate of $\alpha = 5\%$. In order to consider the effect of the full range of false positive rates, ROC analysis was performed on well-controlled simulations (Types 1, 2, 5 and 6 in Table 6.1). Fourteen false positive rates or cutoff points (α) were chosen as 0.01, 0.03, 0.05, 0.07, 0.09, 0.11, 0.13, 0.15, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0 in the bootstrap test. The sensitivity was estimated from 500 Monte-Carlo simulated signals with a stimulus response at a SNR of -30 dB in the raw data (corresponding to an SNR of 3 dB in the averaged signal from 2000 sweeps of the raw data), and specificity was evaluated from 500 simulations without a stimulus responses. ROC curves were plotted as 1-specificity against sensitivity for Basic and MAR bootstrap by two paired simulations without (Figure 6.6(a) for Basic and 6.6(b) for MAR) and with (Figure 6.6(c) for Basic and 6.6(d) for MAR) MA, respectively.

Figure 6.6(a) and 6.6(b) show that ROCs from the same parameters using the Basic and MAR were, as expected, almost the same, when movement artefacts were absent. These two figures also demonstrate that the parameter *power* and *abs*, as well as $\pm$ *difference* and *cc* gave almost identical curves, i.e., the same performance. This was further investigated through the bootstrap significance tests on the area under the ROC curve (AROC). As Figure 6.7 shows, the parameter *power* and *abs* were the most sensitive (bootstrap tests on AROC), as indicated by the largest AROC. ROC curves and AROCs provided further evidence for comparison of different parameters and again demonstrate that when MA was absent, the MAR bootstrap did not make any improvement (nor cause significant worsening), compared to the Basic bootstrap, as expected.

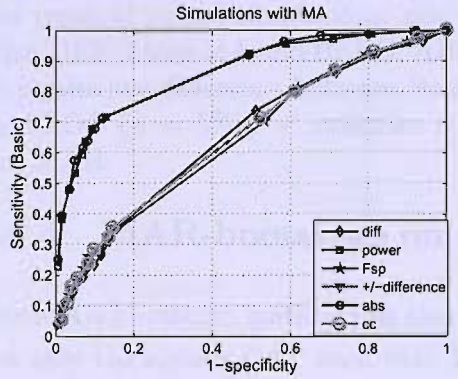
However, once the MA contaminated the signals, the ROC curves of Basic and MAR became very different as shown in Figure 6.6(c) and 6.6(d). Basic bootstrap gave lower sensitivity (Figure 6.6(c)) compared to Figure 6.6(a), and AROCs dramatically decreased, i.e., the power of the Basic bootstrap reduced. The MAR however provided similar curves as those obtained with no MA. Consistently, for the six parameters (Figure 6.6(d)), the AROC went back to the higher values and this meant that, when MA was present, great benefits from MAR scheme could be obtained.



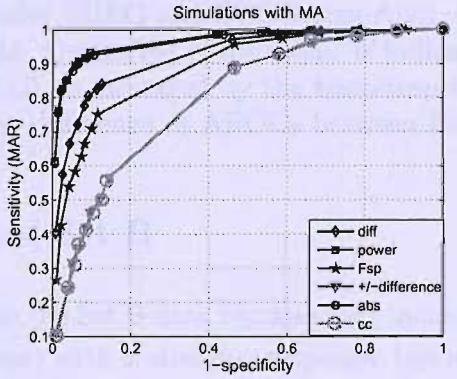
(a) ROC curves for Basic bootstrap method on 500 simulations without artefacts.



(b) ROC curves for MAR bootstrap method on 500 simulations without artefacts.

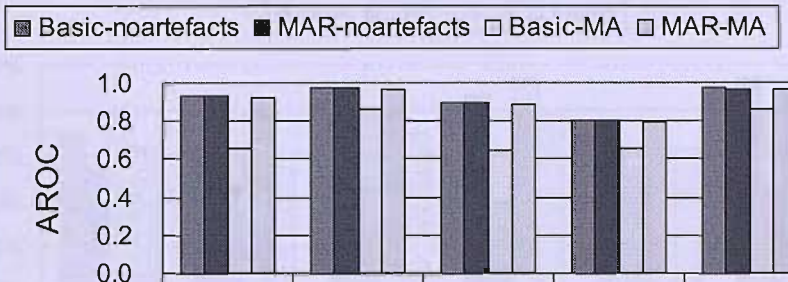


(c) ROC curves for Basic bootstrap method on 500 simulations with movement artefacts.



(d) ROC curves for MAR bootstrap method on 500 simulations with movement artefacts.

Figure 6.6: ROC curves estimated by Basic and MAR bootstrap methods on simulations with and without movement artefacts (MA) for the six parameters introduced in this study.



		diff	power	Fsp	± difference	abs	cc
o	Basic-noartefacts	0.932	0.972	0.896	0.805	0.970	0.805
o	MAR-noartefacts	0.931	0.971	0.896	0.804	0.967	0.804
*	Basic-MA	0.655	0.863	0.647	0.652	0.866	0.654
*	MAR-MA	0.921	0.966	0.888	0.797	0.964	0.796

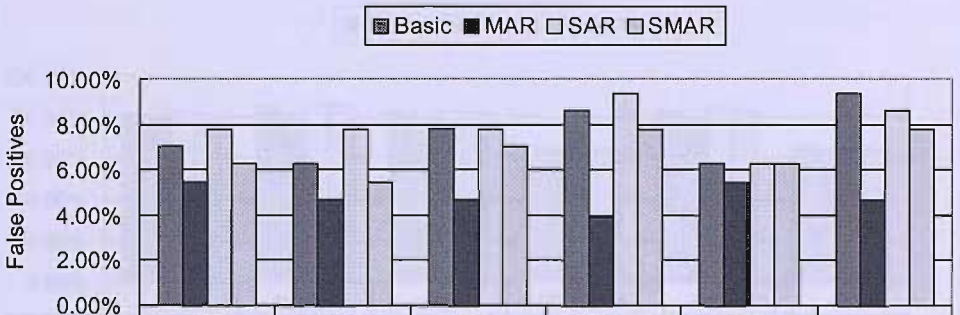
Figure 6.7: AROCs (bars and table) for the Basic method and MAR estimated from two types of paired simulations: one pair includes BEEG and BEEG plus ABR, the other BEEG plus MA and BEEG, ABR, plus MA. On the left of the table, '*o*' indicates no significant difference between Basic and MAR as indicated by the bootstrap test on AROC ($\alpha = 5\%$); '$*$' indicates a significant difference of AROCs between Basic and MAR.

6.6.4 MAR-bootstrap on recorded Set B

The MAR bootstrap method was also applied on the Set B data because they included not only the signals (112, each with 2000 sweeps) with a stimulus response, but also signals (128, each with 2000 sweeps) without stimulations, i.e., background EEG only. This was used to estimate the false positive rates based on real data.

The false positive rates for the six parameters by the MAR in Figure 6.8 were consistently within the range of 1.56%-8.59% as is given by the binomial distribution for 128 trials with a probability of 'success' of 5%. The values observed with MAR were slightly lower than those from the Basic bootstrap.

The sensitivities with MAR (assuming that there is a response present in signals) are shown in Figure 6.9 and do not change greatly compared to those from the Basic bootstrap method. A slight drop observed may be because some sweeps with large



	diff	power	Fsp	±difference	abs	cc
Basic	7.03%	6.25%	7.81%	8.59%	6.25%	9.38%
MAR	5.47%	4.69%	4.69%	3.91%	5.47%	4.69%
SAR	7.81%	7.81%	7.81%	9.38%	6.25%	8.59%
SMAR	6.25%	5.47%	7.03%	7.81%	6.25%	7.81%

Figure 6.8: False positive rates estimated from Set B using Basic, MAR, SAR and SMAR-bootstrap methods. A range of 1.56%-8.59% is given by binomial distribution for 128 trials with a probability of 'success' of 5%.

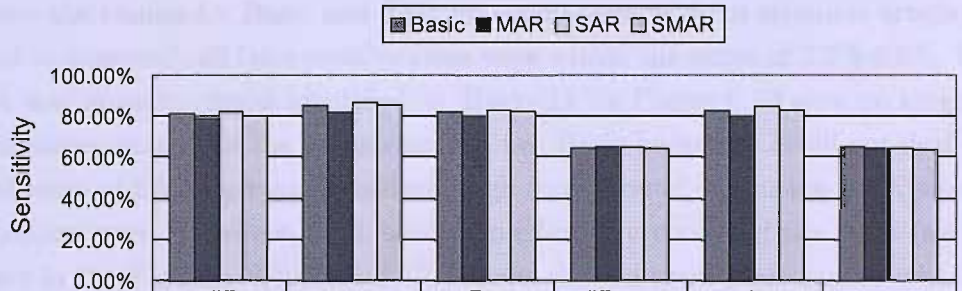
amplitudes are wrongly regarded as a stimulus response by the Basic bootstrap, but removed when employing the MAR. Overall, no great effect with MAR is obtained.

6.6.5 Summary of MAR

The simulation studies for false positive rates and sensitivity indicated that when artefacts were present, MAR scheme should be applied in order to improve sensitivity.

The MAR scheme did not provide the expected improvement for the recorded ABRs of Set B; in fact, p-values increased slightly, reducing our ability to detect the response. Visual inspection of the signals indicated that there were not, in fact, large movement artefacts present in the data, and this may explain why the benefit of the MAR scheme was not evident here in the recorded data.

The rejection threshold can also be used for checking the quality of the recordings when collecting data. If a great number of the sweeps (e.g. 50%) exceeded the threshold, this recording would not be used and would need to be re-collected again because of the unreliable parts of the artefacts.



	diff	power	Fsp	±difference	abs	cc
Basic	81.25%	85.71%	82.14%	64.29%	83.04%	65.18%
MAR	78.57%	82.14%	80.36%	65.18%	80.36%	64.29%
SAR	82.14%	86.61%	83.04%	65.18%	84.82%	64.29%
SMAR	80.36%	85.71%	82.14%	64.29%	83.04%	63.39%

Figure 6.9: Sensitivity estimated from Set B using Basic, MAR, SAR and SMAR-bootstrap methods.

Another aspect should be considered is that for different recordings, the rejection level of $\pm$ 30 μV may need to be changed, because of the differences in amplitudes of the signals recorded from different individuals. Too large rejection level will lead to poor data being included in the analysis. On the other hand, if the number of the sweeps was fixed, too low rejection levels will increase the acquisition time because many sweeps would be excluded by the threshold.

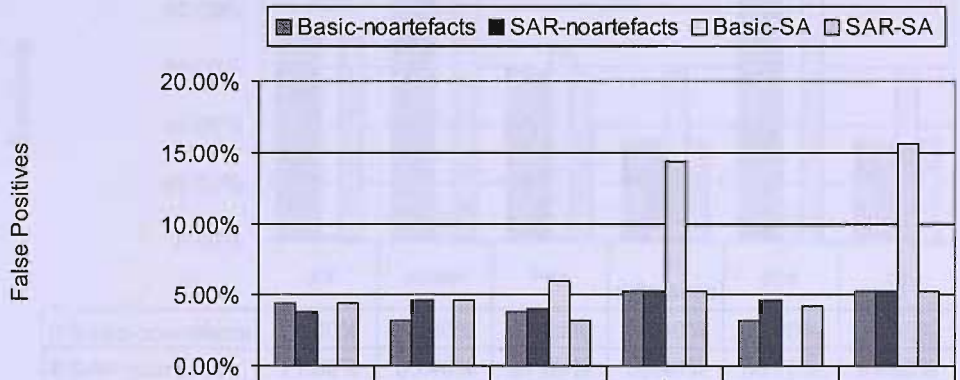
6.7 Application of stimulus artefact rejection scheme

Similarly to the work presented on the MAR bootstrap, stimulus artefact rejection (SAR) was applied to both simulations and recorded signals of Set B.

6.7.1 False positives for SAR in simulations

First, the false positive rates of SAR on two types of simulations (No. 1 and 3 in Table 6.1) were investigated. For comparison, the results from the Basic bootstrap for these simulations are also shown. The two left bars for each parameter in Figure 6.10

show the results for Basic and SAR on simulations without stimulus artefacts (SA), and as expected, all false positive rates were within the range of 3.2%-6.8%. But once SA was present, results identified as 'Basic-SA' in Figure 6.10 were no longer within the range, except for the parameter F_{sp}, i.e., Basic bootstrap could not deal with the influence of SA by giving excessively high (*±difference, cc*) or low (*diff, power*) false positive rates. However, SAR bootstrap efficiently corrected the rates (see the last bars in the Figure). Thus the SAR scheme showed good performance when SA was present.



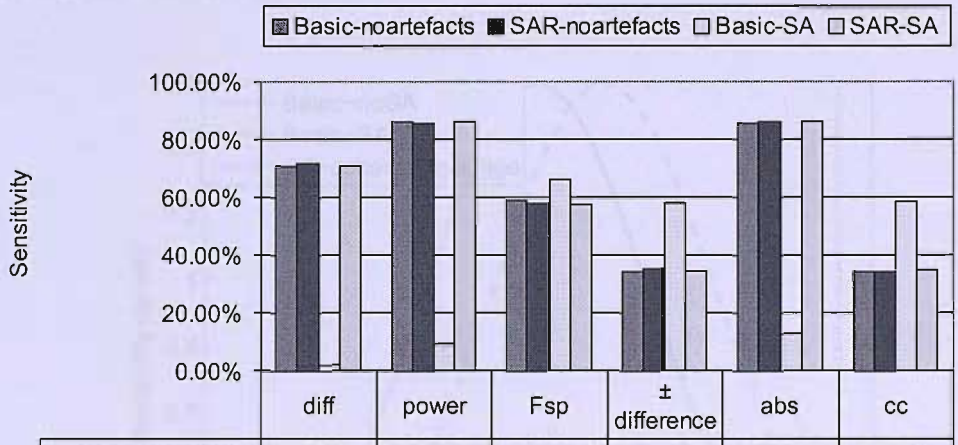
	diff	power	Fsp	± difference	abs	cc
Basic-noartefacts	4.40%	3.20%	3.80%	5.20%	3.20%	5.20%
SAR-noartefacts	3.80%	4.60%	4.00%	5.20%	4.60%	5.20%
Basic-SA	0.00%	0.00%	6.00%	14.40%	0.00%	15.60%
SAR-SA	4.40%	4.60%	3.20%	5.20%	4.20%	5.20%

Figure 6.10: False positive rates for simulations with and without SA achieved by Basic and SAR algorithms, respectively.

6.7.2 Sensitivity for SAR in simulations

The estimation of sensitivity with the SAR bootstrap was performed on two types of simulations including ABR: one with contamination by SA and another without. The sensitivities obtained by both Basic and SAR are shown in Figure 6.11. The results are similar to those for the MAR bootstrap: there were no significant difference

between Basic and SAR when SA was absent. But great differences between these two methods existed when adding SA to the simulations. Basic bootstrap based on parameter *diff*, *power*, and *abs* consistently gave extremely low sensitivities and that based on the other three parameters provided high sensitivities, similar to those from simulations without SA. The reason for these will now be investigated further.



	diff	power	Fsp	± difference	abs	cc
Basic-noartefacts	71.00%	86.20%	58.80%	34.40%	85.60%	34.40%
SAR-noartefacts	71.80%	85.40%	57.80%	35.40%	86.20%	34.60%
Basic-SA	1.80%	9.00%	66.00%	58.00%	12.80%	58.40%
SAR-SA	71.00%	86.00%	57.40%	34.20%	86.40%	34.80%

Figure 6.11: Sensitivity for simulations with and without SA achieved by Basic and SAR algorithms, respectively.

Considering the opposite directions in the tendency of sensitivities for different parameters for the Basic bootstrap method when the SA is present (see the results of the third line in table 6.11), two groups are identified: group 1 (very low sensitivities on parameter *diff*, *power*, and *abs*) and group 2 (high sensitivities on remaining three parameters). First group 1 will be considered, based on the example of *diff*.

The sensitivity is determined by the p-values which are influenced by the value of parameter calculated from the coherent average and the probability density of the parameter from the incoherent averages in bootstrap resampling. The very low sensitivity from the Basic bootstrap with the parameters in Group 1 can be explained by

the Figure 6.12. The parameter estimate (*diff*, vertical line) remains the same with and without SA. However, the distribution from the Basic bootstrap on the signals with SA will shift to the right of the distribution on the signals without SA, because the SA makes an increase of $diff^*$ which results in the shift of the distribution. Therefore, using the Basic bootstrap, the p-value with SA is larger than that without SA, and the sensitivity decreases.

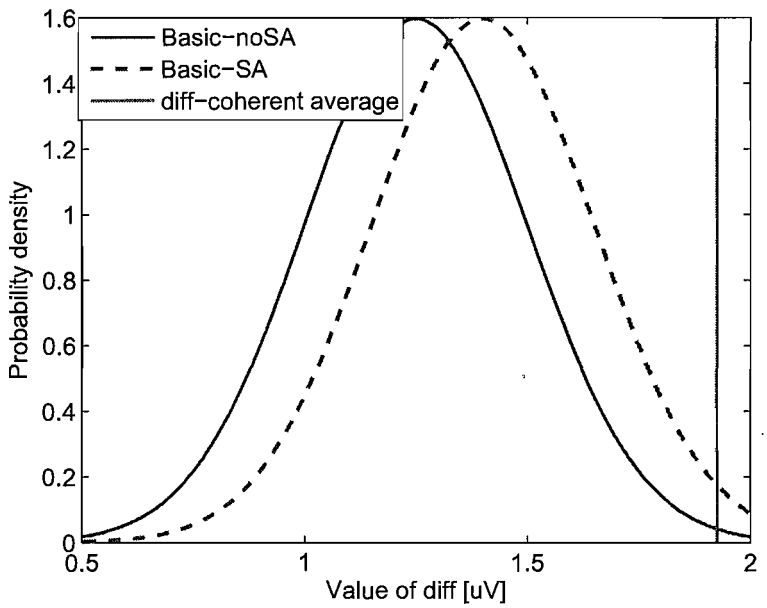


Figure 6.12: The estimation of the p-values with and without SA for the parameter *diff*. The vertical line is *diff* calculated from the coherent average, and the curve is the probability density of the 499 $diff^*$ from incoherent averages of the 'bootstrap' signals. The p-value is given by the area in the tail of curve. This will change when the parameter estimate (vertical line) and the density function are shifted relative to each other. This example shows the parameter estimate does not change, but the density function (dotted line) shifts to the right when the SA is present. This results in an increase of the p-value.

The parameter $\pm$*difference* in group 2 could be taken as an example: here the value of the parameter from coherent averaging again remained the same because SA was out of the time-window for analysis and was not taken into account for parameter calculation, but the probability density function tended to shift to the left, i.e., the

bootstrapped $\pm$*difference* from incoherent averaging decreased. This occurs because the influence of SA on the numerator and denominator are different, increasing the numerator slightly and the denominator more strongly. This leads to the value of $\pm$*difference* decreasing and makes the probability density function of $\pm$*difference* shift to the left. Thus p-values become smaller and sensitivity increases.

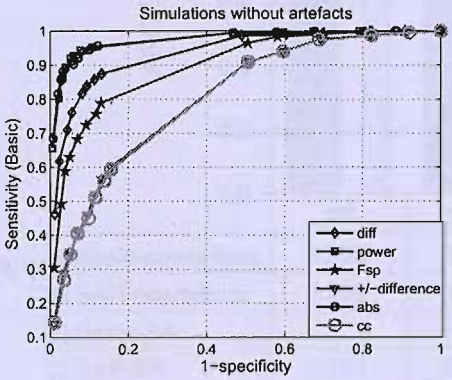
The results from the Basic bootstrap on signals with SA were not acceptable and in cases with SA, only SAR should be considered. The sensitivities with SAR for all six parameters on signals with SA (identified as 'SAR-SA' in Figure 6.11) were similar to those from simulations without SA (identified as 'SAR-noartefact'), and the results for the Basic bootstrap where then also similar. The results indicate that SAR bootstrap worked well in the presence of SA and gave similar sensitivities to the Basic bootstrap in the absence of SA.

The investigation of sensitivity indicated that in the presence of SA, it is necessary to apply SAR bootstrap rather than Basic bootstrap.

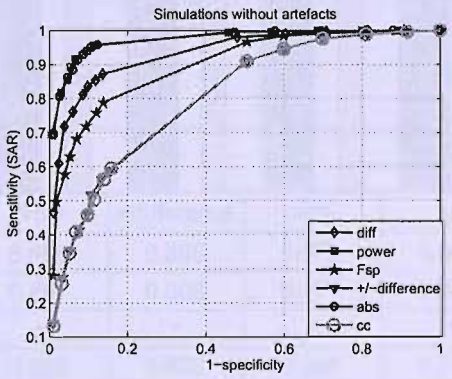
6.7.3 ROC analysis for SAR

The following ROC analysis (ROC curve and its area) gives further evidence of the advantage of SAR bootstrap. Figure 6.13 shows ROC curves for Basic and SAR on simulations with and without SA (Basic-noartefact, SAR-noartefact, Basic-SA, and SAR-SA). Here the movement of ROC-curve away from the top-left corner (see Figure 6.13(c)) for the Basic bootstrap in the presence of SA, is evident. This was consistent with the previous results, indicating the Basic bootstrap failed when SA was present. Thus the analysis of AROC for the Basic bootstrap was no longer useful. Therefore AROC analysis will perform for other three cases.

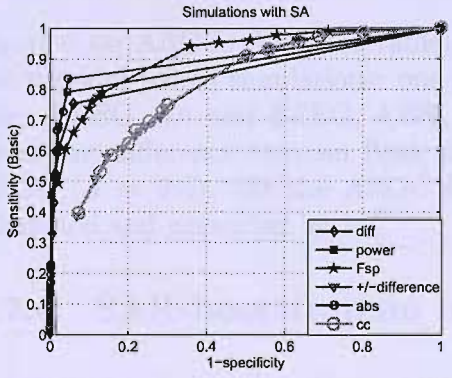
The AROCs in Figure 6.14 provide two main results. Firstly, they show that the performances of Basic and SAR in the absence of SA were similar; this was confirmed by the bootstrap significance test on the AROC ($p > 0.05$). Secondly, it shows that in the presence of SA, SAR gives similar results to the SAR and the Basic bootstrap in the absence of SA.



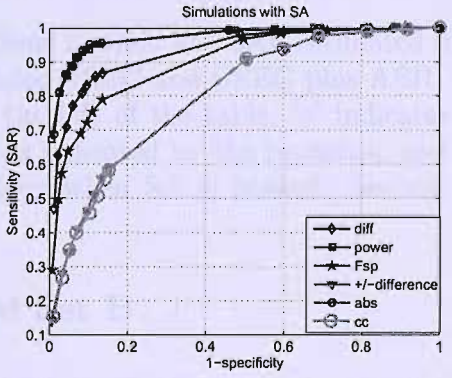
(a) ROC curves for Basic bootstrap method on 500 simulations without artefacts.



(b) ROC curves for SAR bootstrap method on 500 simulations without artefacts.

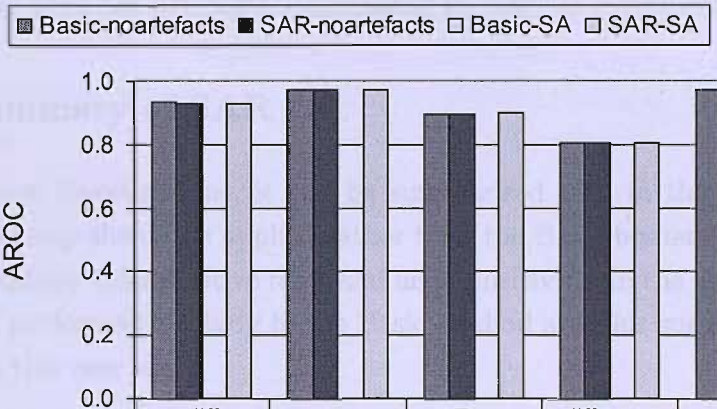


(c) ROC curves for Basic bootstrap method on 500 simulations with stimulus artefacts.



(d) ROC curves for SAR bootstrap method on 500 simulations with stimulus artefacts.

Figure 6.13: ROC curves estimated by Basic and SAR bootstrap methods on simulations with and without stimulus artefacts (SA) for the six parameters introduced in this study.



o		diff	power	Fsp	±difference	abs	cc
o	Basic-noartefacts	0.932	0.972	0.896	0.805	0.970	0.805
o	SAR-noartefacts	0.929	0.970	0.897	0.805	0.971	0.805
	Basic-SA	-	-	-	-	-	-
	SAR-SA	0.929	0.970	0.899	0.802	0.966	0.799

Figure 6.14: AROCs (bars and table) for the Basic method and SAR estimated from two types of paired simulations: one pair includes BEEG and BEEG plus ABR, the other BEEG, SA and BEEG, ABR, SA. On the left of the table, '*o*' indicates no significant difference between Basic and SAR as indicated by the bootstrap test on AROC ($\alpha = 5\%$). As the AROC for the Basic when SA is present, can not be calculated and presented by '−'.

6.7.4 SAR-bootstrap on recorded Set B

The SAR bootstrap was then applied on data Set B. The false positive rate of the five parameters are shown in Figure 6.8 and lie within the range of 1.56% - 8.59% (except $\pm$*difference* with the SAR). The value of 9.38% for $\pm$*difference* exceeded the range by 0.79%, which corresponds to 1 case out of 128.

The sensitivity is shown in Figure 6.9. The results for SAR bootstrap were similar as those obtained by the Basic bootstrap. This might be for either of two reasons: one there was no SA present and the other that even if the SA was present, SAR bootstrap could remove them efficiently and keep the sensitivity at a high level. In the current case, by visual inspection, we found that the quality of the signals was

good and they were not notably contaminated by SA.

6.7.5 Summary of SAR

From the above investigations, it can be summarized that in the presence of SA, the SAR bootstrap should be applied rather than the Basic bootstrap, because SAR provided acceptable false positive rates and high sensitivity. In the absence of SA, the SAR method performed similarly to the Basic method and this suggests that SAR is acceptable in this case also.

While the SAR method was designed to remove the stimulus artefact, it could also be useful in improving sensitivity in detecting some of the smaller waves in auditory evoked potentials. Similar to the motivation for applying SAR, at the latency of smaller waves there will be no contribution from the larger waves (such as wave V in the ABR) in the coherent average, but this wave would contribute to the incoherent average, potentially degrading the sensitivity in a similar manner (but to a lesser extent), than the SA. The SAR scheme would remove this contribution, and thus increase sensitivity.

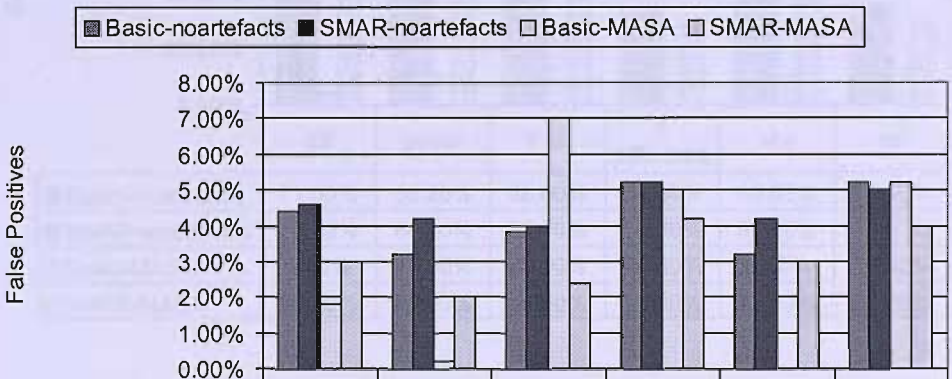
6.8 Application of stimulus and movement artefact rejection scheme

Following the encouraging results with MAR and SAR, these methods were combined. The SMAR was designed to deal with the presence of either or both stimulus and movement artefacts. Similar analysis to the presented above will be carried out, that takes into account possible interactions between the schemes and the effects of the artefacts.

6.8.1 False positives for SMAR in simulations

False positive rates were tested in two simulations (500 signals for each). One only included BEEG and the other BEEG, movement artefact (MA) and stimulus artefact (SA). As expected, false positive rates ('Basic-noartefact' and 'SMAR-noartefact'

in Figure 6.15) were similar for Basic and SMAR with all six parameters, since no artefacts were involved. All results are within the expected range of 3.2% - 6.8%. But when MA and SA were added, rates decreased beyond the range (especially for *diff*, *power* and *abs*), and SMAR clearly provides an improvement. False positive rates higher than the expected $\alpha = 5\%$, as occurred with F_{sp} using the Basic bootstrap method, were however of greater concern than low rates. The latter may even be considered desirable. On the other hand, they may also be associated with low sensitivity in detecting responses when present, and this is investigated in the next set of simulations.



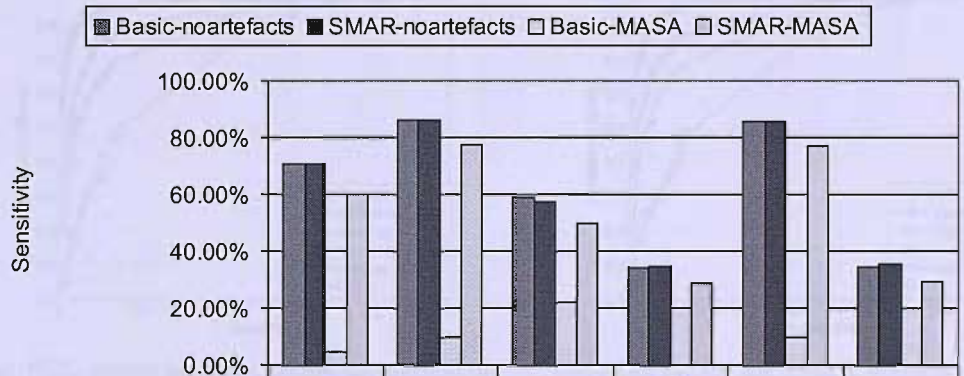
	diff	power	Fsp	± difference	abs	cc
Basic-noartefacts	4.40%	3.20%	3.80%	5.20%	3.20%	5.20%
SMAR-noartefacts	4.60%	4.20%	4.00%	5.20%	4.20%	5.00%
Basic-MASA	1.80%	0.20%	7.00%	5.00%	0.00%	5.20%
SMAR-MASA	3.00%	2.00%	2.40%	4.20%	3.00%	4.00%

Figure 6.15: False positive rates for simulations with and without MA plus SA achieved by Basic and SMAR algorithms, respectively.

6.8.2 Sensitivity for SMAR in simulations

Stimulus responses were added first to the BEEG and then to BEEG, MA and SA. Bootstrap tests were carried out with $\alpha = 5\%$, and results are shown in Figure 6.16. In the absence of artefacts, the Basic and SMAR methods showed very similar

performance, with the highest sensitivity for *power* and the lowest for ± *difference* and *cc*. However, in the presence of artefacts, the Basic method ('Basic-MASA') showed low sensitivity, and SMAR achieved much better results.



	diff	power	Fsp	± difference	abs	cc
Basic-noartefacts	71.00%	86.20%	58.80%	34.40%	85.60%	34.40%
SMAR-noartefacts	70.60%	86.20%	57.40%	34.80%	85.60%	35.20%
Basic-MASA	4.80%	10.00%	22.00%	20.00%	9.80%	19.80%
SMAR-MASA	60.00%	77.60%	49.60%	28.60%	76.80%	29.20%

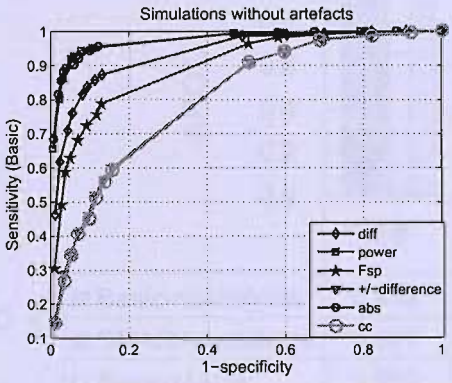
Figure 6.16: Sensitivity for simulations with and without MA plus SA achieved by Basic and SMAR algorithms, respectively.

By comparing the sensitivities of 'SMAR-noartefact' with 'SMAR-MASA', it was found that the six parameters from the latter were always smaller than those of the former. This might be explained as the influence of MAR, which may not have removed all sweeps with MA and also reduced the number of sweeps included in the averaging process (from 2000 to approximately 1800- considering that 10% of sweeps had artefacts added). However, SMAR still clearly provided better results in the presence of MA and SA than the Basic bootstrap method.

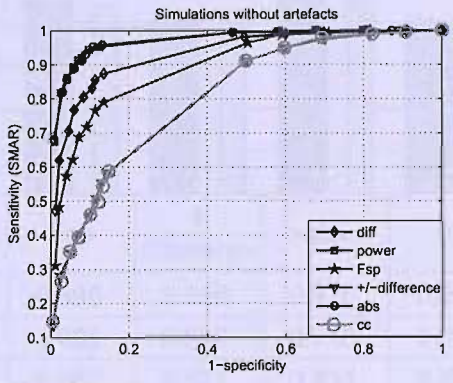
6.8.3 ROC analysis for SMAR

Further investigation of SMAR was carried out by ROC analysis. ROC curves are shown in Figure 6.8.3 and only the curves in Figure 6.17(c) show reduced sensitivity.

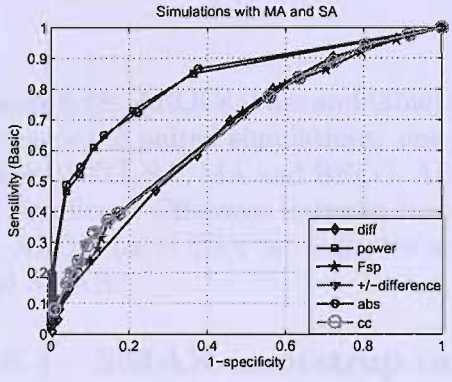
SMAR overcame the limitation of the Basic bootstrap and this was again demonstrated by the higher value of AROC in Figure 6.18.



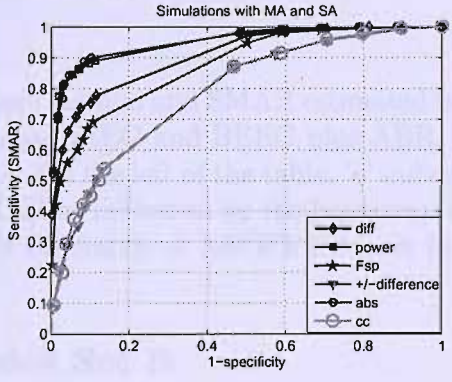
(a) ROC curves for Basic bootstrap method on 500 simulations without artefacts.



(b) ROC curves for SMAR bootstrap method on 500 simulations without artefacts.



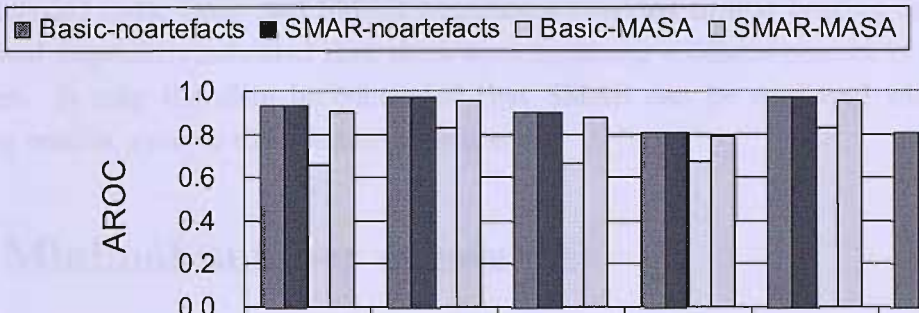
(c) ROC curves for Basic bootstrap method on 500 simulations with stimulus and movement artefacts.



(d) ROC curves for SMAR bootstrap method on 500 simulations with stimulus and movement artefacts.

Figure 6.17: ROC curves estimated by Basic and SMAR bootstrap methods on simulations with and without stimulus and movement artefacts (SA and MA) for the six parameters introduced in this study.

The values of AROC in the absence of MA and SA ('Basic-noartefact' and 'SMAR-noartefact') were similar ($p > 0.05$, bootstrap significance test) and those in the presence of artefacts were significantly different ($p < 0.05$). SMAR did improve the results when artefacts were present.



		diff	power	Fsp	± difference	abs	cc
o	Basic-noartefacts	0.932	0.972	0.896	0.805	0.970	0.805
o	SMAR-noartefacts	0.929	0.967	0.895	0.807	0.968	0.807
*	Basic-MASA	0.660	0.826	0.667	0.671	0.831	0.671
*	SMAR-MASA	0.907	0.950	0.874	0.786	0.950	0.786

Figure 6.18: AROCs (bars and table) for the Basic method and SMAR estimated from two types of paired simulations: one pair includes BEEG and BEEG plus ABR, the other BEEG, SA, MA and BEEG, ABR, SA,MA. On the left of the table, '*o*' indicates no significant difference between Basic and SMAR as indicated by the bootstrap test on AROC ($\alpha = 5\%$); '$*$' indicates a significant difference of AROCs between Basic and SMAR.

6.8.4 SMAR-bootstrap on recorded Set B

The application of SMAR on recorded data Set B indicated similar false positive rates (Figure 6.8) and sensitivity (Figure 6.9) as Basic, SAR or MAR. The reason is probably again that there was no obvious MA and SA present, as evident from visual inspection.

6.8.5 Summary of SMAR

In the presence of artefacts, SMAR greatly increases the sensitivity with all six parameters tested. In the absence of the artefacts, the results from simulation provided similar sensitivity and false positive rates for the Basic bootstrap method and SMAR.

In the recorded data, Basic and SMAR bootstraps provided similar hearing thresholds. Visual inspection indicated that there were no strong artefacts present in these recordings. It may therefore be concluded that SMAR can be employed without degrading results, even in the absence of artefacts.

6.9 Minimal number of sweeps

The bootstrap method applied here is an off-line process and in many conditions an on-line procedure is desirable because at a higher stimulus intensity, e.g., 50 dB SL, the recommended 1000 or 2000 sweeps are more than necessary for detecting the response and therefore stopping the recording as soon as the response is obtained, would save time for recording and further processing.

Previous work discussed in Chapter 5 provided an investigation of the minimal number of sweeps, and this indicated that a relatively smaller number of sweeps are adequate for detection at higher stimulus levels. Here further investigation is carried on in order to identify the required number for different subjects at different stimulus intensities. For each subject at each stimulus intensity, first the bootstrap methods were applied to K=2000 sweeps of this signal, and if $p > 0.05$, the minimal number for this signal was obtained to be 2000, otherwise $p \leq 0.05$. When $p \leq 0.05$, the methods were then applied to 1000 sweeps (half of the previous number of sweeps), and this procedures were continued until $p > 0.05$, the smallest number of sweeps for which $p \leq 0.05$ gave the minimal value required for detection.

The results in Figure 6.19 demonstrate great variability in the minimum number of sweeps between subjects and at the same stimulus intensity. There was no consistent tendency in the minimum number of sweeps required with increasing stimulus intensity. However, it is found that the minimum of the minimal number of the sweeps among the recorded Set B at 30 to 60 dB nHL is 12 sweeps. This means the bootstrap technique still works even with a few sweeps of the recording. Furthermore, the false positive rates are tested on 500 simulated background EEG signals, and each of them only including 12 sweeps. The results are shown in Table 6.2 and most rates are within the range of 3.2% - 6.8%, except those for parameter *power* and *abs* by the Basic and MAR bootstrap methods.

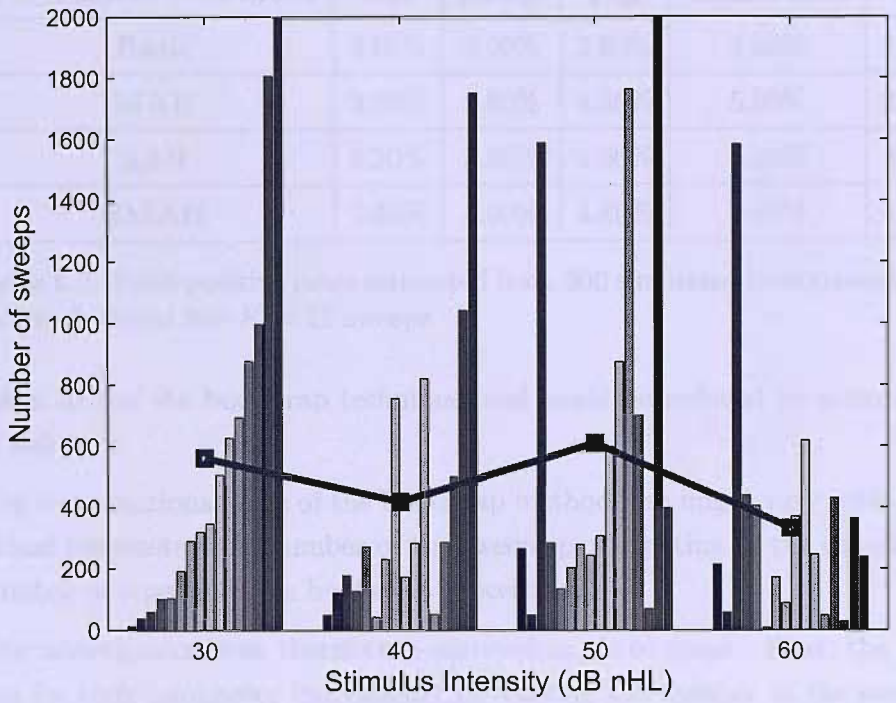


Figure 6.19: Minimal number of sweeps estimated from SMAR based on parameter *diff*. Bars show the minimal number for each subject and the solid line gives the mean value for all the subjects at that stimulus intensity.

The results are encouraging, if K=12 sweeps are applied in clinical settings, a great benefit of the recording and testing time from the bootstrap technique, can be obtained. However, this does not indicate K=12 sweeps are adequate for any individual, according to the variability of the EEG for different individuals.

6.10 Computational time

The computational time required is also of concern, as has been mentioned by other authors (Ozdamar et al., 1990). The program ran on a PC-Pentium (256 MHz) using matlab 7.0 software. The results here only provide a rough indication of the time

Parameters /Methods	diff	power	Fsp	±difference	abs	cc
Basic	3.60%	3.00%	3.80%	4.80%	2.60%	5.00%
MAR	3.80%	2.80%	4.80%	5.00%	2.40%	4.60%
SAR	5.20%	3.60%	4.60%	4.20%	4.00%	4.60%
SMAR	5.40%	4.00%	4.80%	4.40%	4.20%	5.00%

Table 6.2: False positive rates estimated from 500 simulated background EEG signals, and each signal has $K = 12$ sweeps.

taken to run the bootstrap technique, and could be reduced by a more efficient PC or software.

The computational time of the bootstrap method also might vary with different individual parameter, the number of the sweeps participating in the calculation and the number of repeats in the bootstrap process.

The investigation was therefore performed in three steps. First, the program was run for each parameter individually by varying the number of the sweeps (50, 125, 250, 500, 1000, 2000), in order to measure the time taken for different parameters. The results are shown by the bars in Figure 6.20. The paired t-test for any two parameters indicates that there is no significant difference of the computational time between different parameters ($p > 0.05$, t-test). As expected, the computational time increases with the increase of the number of sweeps.

Second, the computational time of the bootstrap methods when simultaneously calculating all six parameters is investigated on the different number of sweeps. This is shown by the dotted line with marks 'o' in Figure 6.20. It is noticed that the computational time for all parameters is slightly longer than that for any individual testing and the average of all individuals (the solid line with mark '□').

Finally, the investigation of the influence of the number of repeats of the bootstrap process on the computational time, is performed using 99, 199, 299, 399, and 499, for the bootstrap methods with six parameters and each signal with 2000 sweeps. The results in Table 6.3(b) demonstrates the testing time increases as expected, with the increase of the number of repeats which is determined by the algorithm of bootstrap technique itself. However, there is no accurate indications in the literature for the

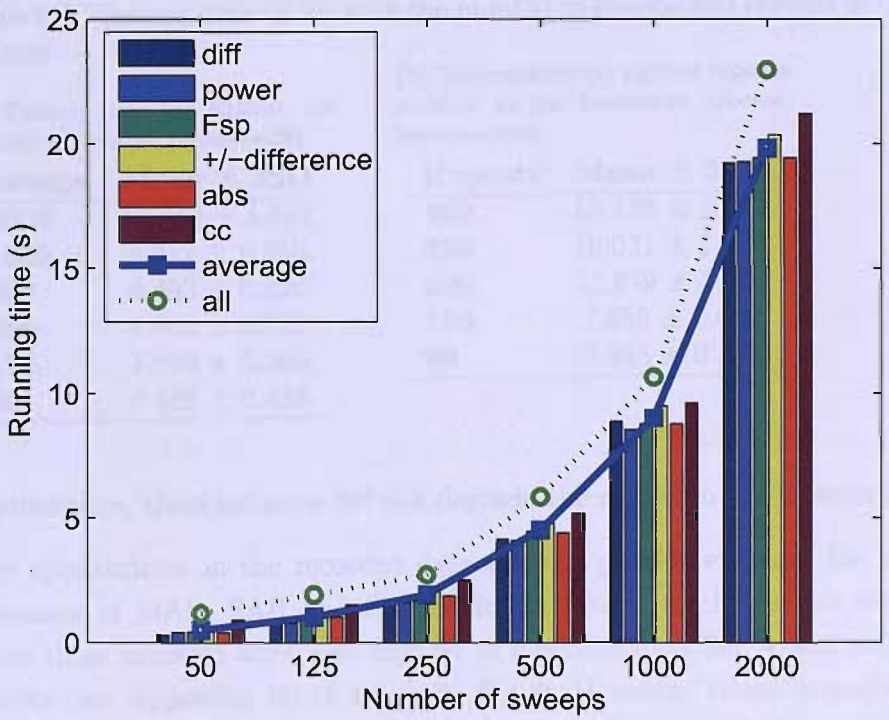


Figure 6.20: Computational time of bootstrap method against the number of sweeps. Bars show the time of bootstrap process based on any individual parameter and the solid line marked by '□' is the average of six individual parameters. Dotted line marked by '*o*' shows the testing time when using six parameters simultaneously in the bootstrap process.

criterion of choosing a suitable number of repeats and more frequently, values between 499 and 999 appear to have been used (Efron and Gong, 1983; Efron, 1993; Zoubir and Boashash, 1998).

6.11 Discussion

The proposed three artefact rejection schemes for the bootstrap method, i.e., MAR, SAR and SMAR consistently demonstrated their benefit in the presence of the corresponding artefacts (MA for MAR, SA for SAR, and MA plus SA for SMAR).

Table 6.3: Testing time varies with the number of sweeps and repeats in the bootstrap process.

(a) Testing time (s) against the number of sweeps, repeats=499.

Sweeps	Mean ± 2SD
2000	19.810 ± 1.582
1000	8.977 ± 0.910
500	4.493 ± 0.820
250	1.893 ± 0.738
125	1.022 ± 0.366
50	0.488 ± 0.418

(b) Testing time (s) against repeats number in the bootstrap process, sweeps=2000.

Repeats	Mean ± 2SD
499	19.180 ± 1.582
399	16.021 ± 1.284
299	11.979 ± 1.003
199	7.950 ± 0.699
99	3.948 ± 0.338

Furthermore, these schemes did not degrade the results in the absence of artefacts.

The applications on the recorded data did not provide evidence for improved performance of MAR, SAR and SMAR. In additional to the results reported above, these three schemes were also applied to recorded data Set A and provided similar results (see Appendix D) to those on Set B. However, visual inspection suggested that the recorded data was of high quality, and did not include artefacts. In future research, more recordings containing stimulus and/or movement artefacts should be investigated.

The comparison of the three artefact rejection schemes and the Basic bootstrap was of concern in the above discussion. Now the comparison of the six parameters will also be considered. From the above results, the parameter *cc* showed poor performance in detecting ABRs in the recorded signals (Set B) and in simulations. The reason for this is probably that it only takes the shape of the two sub-averages (from the even and the odd-numbered sweeps) into account. As the size of the responses is disregarded, randomly occurring small, but consistent sub-averages in the bootstrap resampling, can lead to relatively high *cc* values, which will increase the p-values for this parameter.

The parameter *cc* quantifies the similarity of the two sub-averages, which would seem an appropriate criterion for the detection of a response, especially when the exact shape of the response is unknown, as it may be in some applications of auditory evoked responses. However, to be more effective, it should be modified to take the

amplitude of the sub-averages also into account, which could be achieved by using the covariance between the two sub-averages, which corresponds to the numerator of *cc*, i.e. without normalizing by the variances. As an alternative to the current approach, the correlation coefficient has been used in comparing coherent averages against a template (Ozdamar et al., 1990). The bootstrap method could readily be adapted to carry out the significance test, but again we would suggest that normalizing by the amplitude of the signals could degrade results.

In addition to the parameter *cc*, a related parameter (cc') was also tested on simulations and recorded data in set A. cc' is the maximum of the cross-correlation function between two replicate estimates of the ABRs (Figure 6.21). The two replicates correspond to the coherent average of the first and second half of the stimulation period (i.e. 1000 stimuli each). The cross-correlation function corresponds to the correlation coefficient between these replicates, with varying time-lag between the signals, and is again calculated over the analysis window of 5-15 ms.

For recording set A, the hearing thresholds from cc' were higher than those from other parameters (e.g. *power*, F_{sp}). The parameter cc' again measures the consistency between ABR replicates, and follows the approach taken in the visual analysis of ABRs, where two replicates are often compared in order to ascertain the presence of a response. For cc' to provide high values, well defined features in the signal are required, However, these are not always evident, as the two examples in Figure 6.22 (stimulation at 50 and 30 dB SL, respectively) illustrate, and in which no significant response was detected. While some consistency in the replicate responses is evident, the slow underlying trend and the evident random variation (noise) lead to peak $cc' \cong 0.7$ and p=0.078 and p=0.076, respectively. Such values of cc' would be highly significant ($p < 0.001$) if the signal analysed were Gaussian white noise (no serial correlation between samples), but with the band-limited signals considered here the degrees of freedom are reduced, and hence increase the p-values. Conventional statistical analysis of this, especially since we are considering the maximum of cross-correlation over a range of lags, would be highly complex, and the bootstrap method provided a robust and simple alternative.

Furthermore, the results for parameter $\pm$*difference* and *cc* were always very similar. For the recorded data, they estimated high hearing thresholds and for simulations, gave low sensitivity. This motivates the investigation of the relationship between these

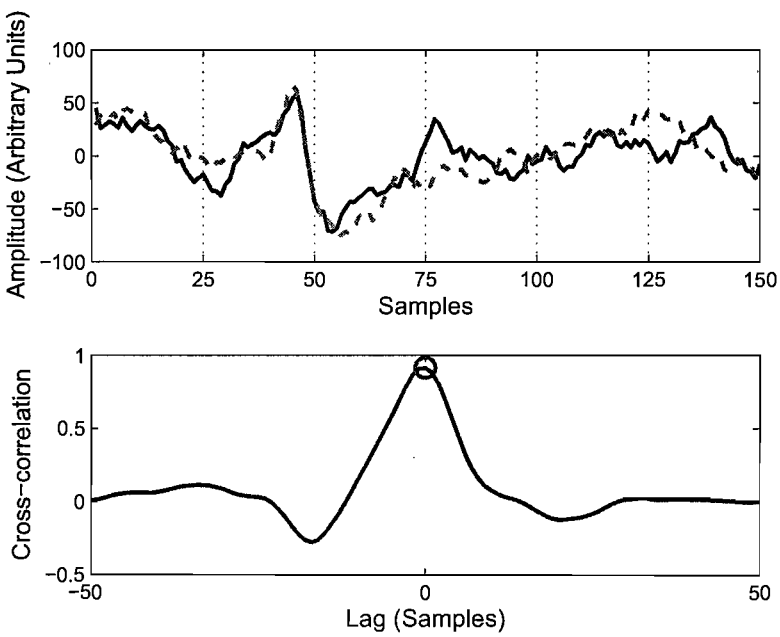


Figure 6.21: An example for calculating the parameter cc'. The top plot shows two replicates of the ABR from the same recording (each obtained by coherently averaging 1000 stimulus responses). The bottom plot gives the cross-correlation function. 'o' indicates the maximum cross-correlation-cc'.

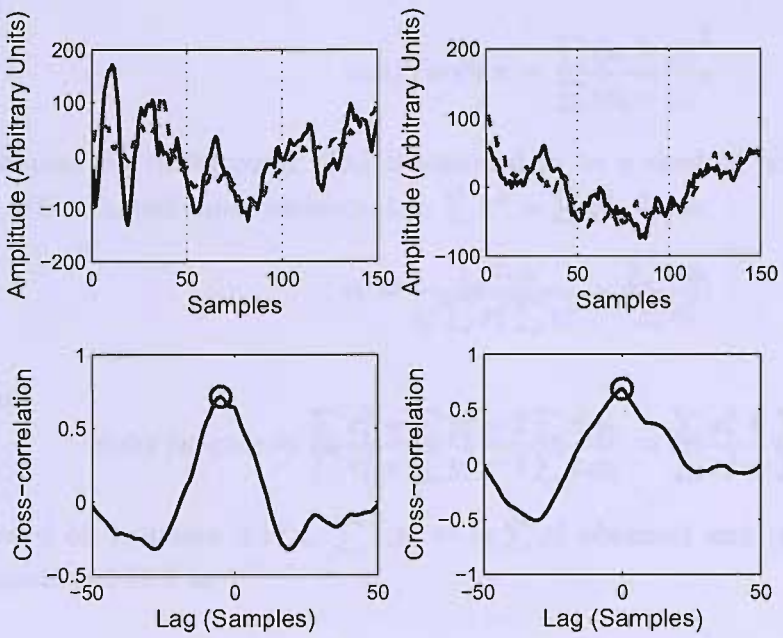


Figure 6.22: The two examples where non-significant cc' was obtained. The top plots show the two replicates, and the bottom plots the corresponding cross-correlation function. 'o' marks the maximum cc'.

two parameters, since both were extracted from even and odd-numbered sweeps. The parameter cc is the correlation coefficient of the two coherent averages, with one average (denoted by x_i) from the even-numbered sweeps, and the other (y_i) from odd-numbered sweeps, where i is the number of the samples in the coherent average. Then

$$cc = \frac{\sum (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{(\sum (x_i - \overline{x})^2)(\sum (y_i - \overline{y})^2)}}$$

where $\overline{x}$ and $\overline{y}$ are average of those coherent averages. The parameter $\pm$*difference* is the ratio of the variance of the coherent average of overall sweeps and the variance of the $\pm$average of the even-numbered and odd-numbered sweeps. This ratio could also be expressed by:

$$\pm difference = \frac{\sum (x_i + y_i)^2}{\sum (x_i - y_i)^2}$$

because the background EEG is assumed to be a random process with zero mean, $\overline{x} \approx \overline{y} \approx 0$, and equal variance, i.e., $\sum x^2 \approx \sum y^2$. Then,

$$cc \approx \frac{\sum x_i y_i}{\sqrt{\sum x_i^2 \sum y_i^2}} = \frac{\sum x_i y_i}{\sum x_i^2} \tag{6.11.1}$$

and

$$\pm difference \approx \frac{\sum x_i^2 + \sum y_i^2 + 2\sum x_i y_i}{\sum x_i^2 + \sum y_i^2 - 2\sum x_i y_i} \approx \frac{\sum x_i^2 + \sum x_i y_i}{\sum x_i^2 - \sum x_i y_i} \tag{6.11.2}$$

based on equation 6.11.1, $\sum x_i y_i \approx cc \sum x^2$ obtained and replacing the $\sum x_i y_i$ in equation 6.11.2 as:

$$\pm difference \approx \frac{\sum x_i^2 + cc \sum x_i^2}{\sum x_i^2 - cc \sum x_i^2} \approx \frac{1 + cc}{1 - cc} \tag{6.11.3}$$

This relationship is illustrated in Figure 6.23: $\pm$*difference* monotonically increases with the increase of *cc*. When the bootstrap method is applied to these two parameters, this relationship is present both for the parameter values from coherent average and the incoherent averages, and thus the p-values are similar. This further leads to similar false positive rates and sensitivities.

In order to compare the performances of the six parameters, ROC analysis was employed. The ROC curve shows the compromise between sensitivity and specificity in detecting the ABRs. The ideal is of course unity values for both sensitivity and specificity. However, usually, when increasing sensitivity (increasing false positive rates in the bootstrap method), some specificity is lost. The optimal operating point depends on the application: in some tasks high sensitivity is required (e.g. in detecting a mid-latency auditory response in depth of anaesthesia monitoring), for others high specificity is essential (e.g. in hearing screening, where false ABR detection may mean that a hearing impairment is missed). The overall performance of the detector was then assessed by the area under the ROC curve (AROC). The bootstrap technique was used here to test the difference between AROCs and indicated parameter *power*

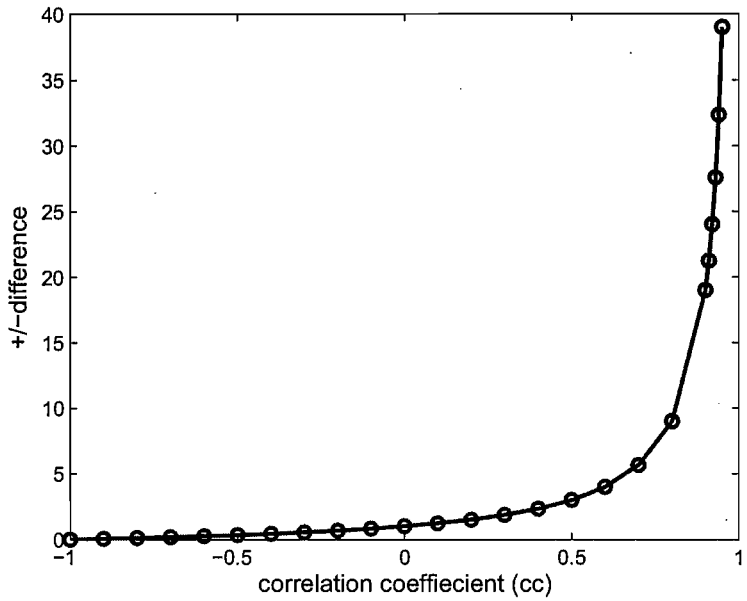


Figure 6.23: The relationship between parameter ±difference and the correlation coefficient (cc).

and *abs* were always very similar in all simulations and in the recordings, and they were the most sensitive parameters compared to the other four. ROC analysis is very simple and useful means to make comparisons between different methods.

Chapter 7

Comparison of Bootstrap with other methods

7.1 Introduction

In the previous two Chapters, the bootstrap method was applied to both simulations and the recorded data and comparative work was carried out on a range of parameters. Among the bootstrap parameters, the widely used methods of Elberling and Don's F_{sp} (Elberling and Don, 1984) and Wong and Bickford's $\pm$ difference (Wong and Bickford, 1980), were investigated, and following the bootstrap resampling process, the estimated distribution of the parameters under the null hypothesis of 'no response present' was obtained. Then the statistical significance of these parameters was compared against this distribution. The detailed procedures on how to use them as parameters within the bootstrap technique was described in section 5.3.2. The F_{sp} and $\pm$ difference methods were initially proposed (Elberling and Don, 1984; Wong and Bickford, 1980) for analysis of evoked potentials and threshold (critical value) for determining whether a response was present or absent was given based on the statistical analysis of the ratio of the variance of the signal and 'noise'. In this chapter, the focus will be placed on the comparison between the bootstrap-F_{sp} (the bootstrap method with the parameter F_{sp}) and F_{sp} as originally proposed (Elberling and Don, 1984) (section 7.2), and the bootstrap-$\pm$difference and the original $\pm$ difference (Wong and Bickford, 1980)(section 7.3).

The bootstrap method proposed and applied here uses resampling of the raw data (signal). Nocera and Ferlazzo (2000) proposed an alternative bootstrap method to assess the within-subject reliability of event-related potentials (ERP). This was achieved by resampling the ensemble instead of the whole signal. This bootstrap was also tested in current work on ABR data and a comparison of the results from the two bootstrap approaches is then made (section 7.4). In order to distinguish between these approaches, our bootstrap method is called 'signal bootstrap' and the other is called 'ensemble bootstrap'.

7.2 Bootstrap method and conventional F_{sp}

The bootstrap-F_{sp} method was described in the Chapter 5 and 6, and the conventional F_{sp} was introduced in section 2.2.5. However, the derivation of the formula of the conventional F_{sp} was not mentioned before. Considering some issues related to the derivation of the formula, an introduction to the theoretical background of the F_{sp} will be given first. Then the comparison of critical values, and detection rates is made.

7.2.1 Theoretical background

Consider the EEG at time t following each stimulus (e.g. click or tone-burst), denominated EEG(t), as consisting of the evoked potential [1], EP (t), and the background noise, usually spontaneous/background EEG, BEEG (t). For each sweep,

$$EEG(t) = EP(t) + BEEG(t) \tag{7.2.1}$$

After N sweeps, the averaged response (coherent average) is

$$\overline{EEG(t)} = EP(t) + \overline{BEEG(t)} \tag{7.2.2}$$

The assumptions for the two components are (Elberling and Don, 1984):

[1] It could represent any evoked potentials, e.g. auditory evoked potential, visual evoked potential, etc.

- EP(t) - is a deterministic signal and has constant latency, amplitude and phase with respect to the stimulus (Wong and Bickford, 1980). Therefore in the normal coherent averaging process, EP(t) remains unchanged.
- BEEG(t) - is a stationary, ergodic random process. These statistical properties were explained in Chapter 4.

In order to derive the statistical distribution of F_{sp}, first the variance of both sides of equation 7.2.2 is calculated. The EP and $BEEG$ are regarded as being uncorrelated, which means when a stimulus is presented, an evoked potential does not change the statistical properties of the background EEG. The EP and $\overline{BEEG}$ may however (by chance) show a non-zero estimate for the correlation coefficient, $R(EP, \overline{BEEG})$ or a covariance, $COV(EP, \overline{BEEG})$, different from zero.

$$VAR(\overline{EEG}) = VAR(EP) + VAR(\overline{BEEG}) + 2 \bullet COV(EP, \overline{BEEG}) \qquad (7.2.3)$$

which can be rewritten by inserting the signal-to-noise ratio, SNR [2]:

$$VAR(\overline{EEG}) = \left[SNR^2 + 1 + 2 \bullet R(EP, \overline{BEEG}) \bullet SNR\right] \bullet VAR(\overline{BEEG}) \qquad (7.2.4)$$

The background EEG distribution can be approximated by collecting the values in one single point of each individual sweep and its variance is then calculated from those values. With an increase in the number of the sweeps, the distribution of the sample of single point values converges to the statistical distribution (probability distribution) of the background EEG. The variance, $VAR(SP)$, of the single point sample will be a measure of the variance of the background EEG, $VAR(BEEG)$. The estimated variance, $VAR(\overline{SP})$, of the averaged background EEG can be represented as:

[2]SNR is defined as the RMS-value (root mean square) of the EP divided by the RMS-value of the averaged background EEG, $\overline{BEEG}$:

$$SNR = \frac{RMS(EP)}{RMS(\overline{BEEG})}$$

$$VAR(\overline{SP}) = \frac{VAR(SP)}{K} \tag{7.2.5}$$

where K represents the number of sweeps in the signal (see Chapter 5).

Then a variance ratio related to the F-distribution is defined as:

$$\begin{aligned} F_{sp} &= \frac{VAR(\overline{EEG})}{VAR(\overline{SP})} \\ &= [SNR^2 + 1 + 2R(EP, \overline{BEEG})SNR]\frac{VAR(\overline{BEEG})}{VAR(\overline{SP})} \end{aligned} \tag{7.2.6}$$

The variance ratio, $VAR(\overline{BEEG})/VAR(\overline{SP})$, under certain conditions (i.e., Gaussian/Normal distribution of the signal) follows the F distribution, $F_{(\nu_1,\nu_2)}$, with ν_1 and ν_2 being the degrees of freedom for the numerator and denominator, respectively. For the F_{sp} method, a critical value was then determined (Elberling and Don, 1984) based on worst-case assumptions regarding the correlation between samples in the signal. This assumption affects the number of degrees of freedom for the numerator which is less than the number of samples. The value given for 250 sweeps was $F_{sp} - crit = 2.25$ for $\alpha = 5\%$ and $F_{sp} - crit = 3.09$ for $\alpha = 1\%$.

7.2.2 Critical values

In addition to providing the p-values in testing for a response in each recording, the bootstrap method can also readily provide the critical values for each parameter for the rejection of the null-hypothesis. This can be obtained from the bootstrap sampling distribution by finding the values of the parameter at $\alpha = 5\%$(or any other desired significance level). Figure 7.1 showed the critical values from the bootstrap-F_{sp} at $\alpha = 5\%$ (upper plot) and $\alpha = 1\%$ (bottom plot), respectively. At each stimulus intensity, the boxplot was obtained from the 16 subjects in data Set B. The mean value for all the critical values was 1.46 for $\alpha = 5\%$ and 2.08 for $\alpha = 1\%$ when averaging 250 sweeps. With an increase in the number of sweeps to 2000, these decrease to 1.43 and 1.97, respectively. The results also show a considerable variation of critical values between individuals.

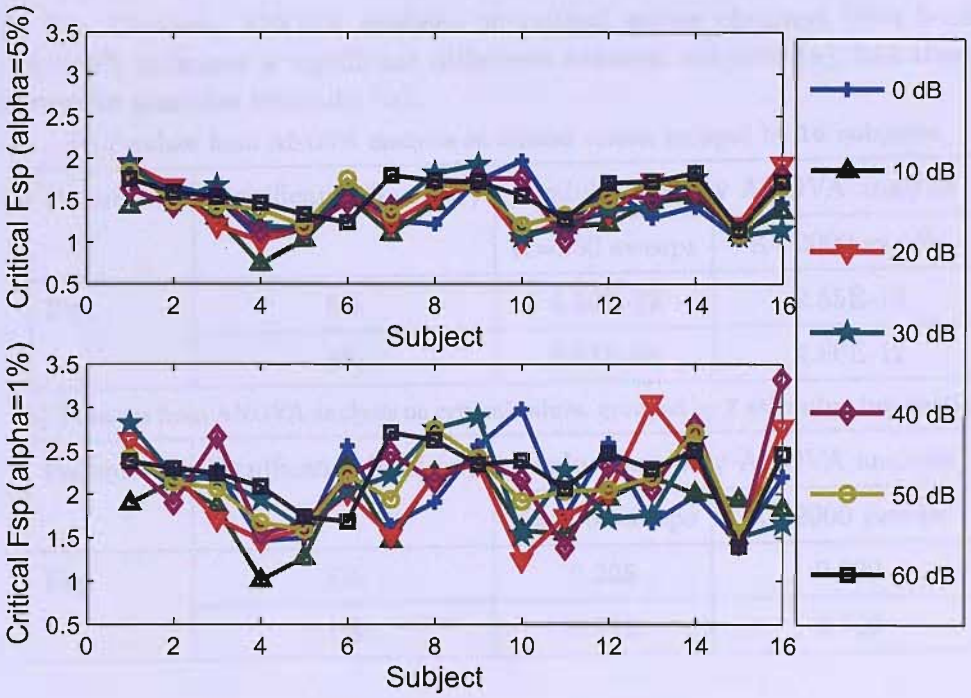


Figure 7.1: Critical values of F_{sp} from the bootstrap cumulative probability distribution with $\alpha = 5\%$ (upper) and $\alpha = 1\%$ (lower). These were obtained with 250 sweeps. For different stimulus intensities, the critical values did not vary greatly. For different subjects, these varied greatly.

This figure lead us to investigate further the reason for the variation among critical values. One-way ANOVA analysis provided a statistical tool to test the two possible influencing factors of subject and stimulus intensity. ANOVA analysis was performed on critical values grouped by 16 subjects and by 7 stimulus intensities, respectively. The results from ANOVA analysis for both 250 sweeps and 2000 sweeps consistently indicated that the critical values varied between subjects ($p < 10^{-8}$ in Table 7.1(a)) and did not vary with stimulus intensity ($p > 0.05$, Table 7.1(b)).

From the bootstrap method (with SMAR), the critical values were found to be generally lower than those given previously by Elberling and Don (1984) (see Figure 7.1), and furthermore, they differed considerably between individuals, ranging from 0.76

Table 7.1: One-way ANOVA analysis on critical values obtained from bootstrap-F_{sp}. $p < 5\%$ indicates a significant difference between subjects (a), but there is no difference by stimulus intensity (b).

(a) P-values from ANOVA analysis on critical values, grouped by **16 subjects**.

Parameter	Significance level (α)	p value (one-way ANOVA analysis)	
		K=250 sweeps	K=2000 sweeps
Fsp	5%	4.20E-12	2.55E-15
	1%	6.83E-08	4.86E-12

(b) P-values from ANOVA analysis on critical values, grouped by **7 stimulus intensities**.

Parameter	Significance level (α)	p value (one-way ANOVA analysis)	
		K=250 sweeps	K=2000 sweeps
Fsp	5%	0.225	0.229
	1%	0.312	0.110

to 2.07 for $\alpha = 5\%$ and from 1.16 to 3.57 for $\alpha = 1\%$. The critical values given by Elberling and Don (1984) are similar to the maximum critical values determined here by the bootstrap method for the sample of subjects investigated, and thus appear to represent a valid 'worst case' value. It is clear that there is no single $F_{sp} - crit$ is valid for all recordings. From theory (Elberling and Don, 1984) it is clear that this critical value depends on the signal characteristics of each recording (in particular the spectrum of the signal), and this differs between subjects. It is therefore also not surprising that the critical value was approximately constant in the repeated recordings from the same subject (see ANOVA analysis), since different stimulus intensities would not greatly change the spectrum of the recorded signals. Similar to the results presented here for SMAR, we also found that the critical values obtained with the Basic bootstrap technique were generally lower than those given previously by Elberling and Don (1984), and also differed considerably between individuals. A critical value of $F_{sp} - crit = 2.0$ was suggested by Lutman and Sheppard (1990) based on click-evoked otoacoustic emission. Thus while previous methods based their estimates of critical values either on worst-case assumptions for the signals (Elberling and Don, 1984) or a large sample of subjects (Lutman and Sheppard, 1990), the bootstrap method

allows the critical value to be determined subject by subject, with and without the implementation of artefact rejection. Another implication of the current work is that a fixed critical value will lead to differing false-positive rates and sensitivities for different subjects. Thus while the alternative methods of Elberling and Don (1984) and Lutman and Sheppard (1990) may provide the expected average false positive rates over a group of subjects, they cannot ensure these in each individual.

7.2.3 Pass and Refer rate

In this section, 'pass' and 'refer' rates are investigated. If parameter (F_{sp}) exceeds the selected threshold (criterion) value, a response is considered to have been detected from the recording and this is denoted as 'pass'. Otherwise the clinician would 'refer' the patient for further investigation. For the bootstrap-F_{sp} method, the statistical significance p-value provides the criterion, and the null-hypothesis is rejected at $p <= 0.05$, and a significant response and 'pass' is recorded. According to Elberling and Don (1984) the criteria recommended are those shown in Table 7.2.

	Pass		**Refer**	
	K=250 sweeps	K=2000 sweeps	K=250 sweeps	K=2000 sweeps
Fsp	Fsp $\geq$ 2.25	Fsp $\geq$ 3.09	Fsp $<$ 2.25	Fsp $<$ 3.09
Bootstrap - Fsp	p $\leq$ 0.05		p $>$ 0.05	

Table 7.2: Criteria for conventional F_{sp} and bootstrap-F_{sp} under different conditions, for K=250 sweeps and K=2000 sweeps.

Table 7.3 shows that for K=250 sweeps, there is 75% agreement (the pass rate where both methods agree, plus the refer rate for both) between the F_{sp} and Bootstrap-F_{sp}. Disagreements are, as expected, all in the direction of over-referral by F_{sp}. In 28 cases, there were significant responses detected by the Bootstrap-F_{sp} and not detected via the conventional F_{sp}. When K increased from 250 to 2000 sweeps (Table 7.4), agreement between F_{sp} and Bootstrap-F_{sp} increased to 90.18%. Even though the F_{sp} threshold increased, i.e., the criterion for detection became tighter as expected, the 'pass' rate dramatically improved mainly because of the larger number of sweeps

which make the stimulus response clearer by reducing the effect of the background EEG.

K=250	**Fsp**	
Bootstrap_Fsp	Pass	Refer
Pass	**23** (20.54%)	**28** (25%)
Refer	**0** (0%)	**61** (54.46%)

Table 7.3: The performance of F_{sp} and Bootstrap-F_{sp} ('pass' and 'refer' rates) estimated from the first 250 sweeps of the recordings.

7.2.4 Discussion

In selecting the critical value for the F_{sp} based on theory, the problem lies in identifying the degrees of the freedom ν_1 (Elberling and Don, 1984). If the background EEG were truly white noise, the degrees of freedom, ν_1 would approximate the number of samples within the analysis window. However, the narrower the band of dominant frequencies, the lower the degrees of freedom, ν_1. Elberling and Don (1984) evaluated ν_1 experimentally in the no-stimulus condition. ν_2 was determined by the number of sweeps (e.g., 250 in the original work), considering each sweep being independent.

In order to investigate the effect of ν_1, it was increased from the suggested value of 5 to the number of samples in the analysis window (5-15 ms), 51 in our study. For each value of ν_1, a threshold exceeding 95% ($\alpha = 5\%$) of the samples from an F distribution, was determined. Figure 7.2 shows the degree of freedom of the numerator, ν_1 against

K=2000	F_{sp}	
Bootstrap-F_{sp}	Pass	Refer
Pass	**81** (72.32%)	**11** (9.82%)
Refer	**0** (0%)	**20** (17.86%)

Table 7.4: The performance of F_{sp} and Bootstrap-F_{sp} ('pass' and 'refer' rates) estimated from 2000 sweeps of the recordings.

the threshold values at 95%. With the increase of ν_1, the values decrease. The 'true' degree of freedom for the numerator cannot readily be obtained. Therefore, the bootstrap method again shows its benefit for estimating the critical value.

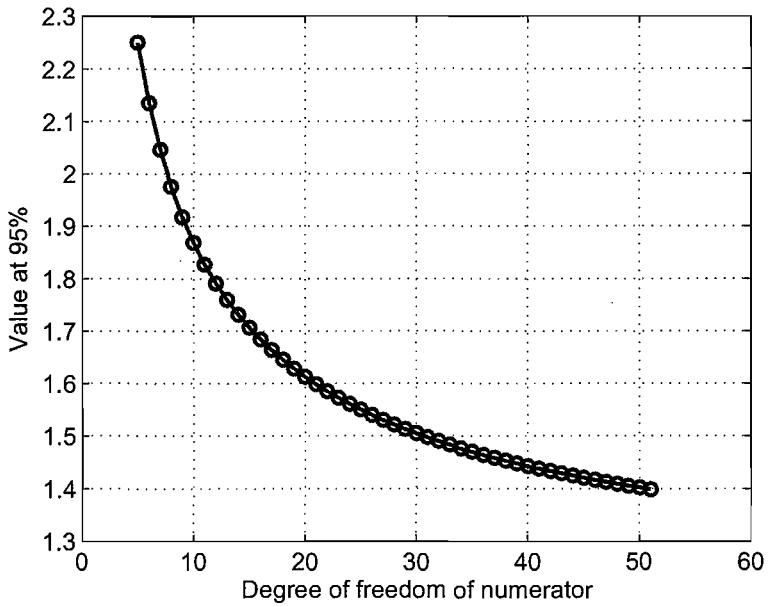


Figure 7.2: Degrees of freedom in the numerator against the values that should exceed 95% of the samples from an F distribution with ν_1 degrees of freedom in the numerator and ν_2 degrees of freedom in the denominator. Here in order to demonstrate the influence of the degree of freedom in the numerator on that value, ν_1 varied from 5 (assumed worst case) to 51 (number of samples in the analysis window) and ν_2 remained as 250 (number of sweeps).

7.3 Bootstrap method and conventional ±difference

The ± difference method is an alternative popular method (Wong and Bickford, 1980) in the detection of the ABR. The critical values of the bootstrap-±difference will again be compared with those suggested in the literature (Wong and Bickford, 1980). The latter were identified based on the assessment of the ratio of the variance of the signal to the variance of the noise. In addition, the pass and refer rates will be compared.

7.3.1 Theoretical background

Similarly to the F_{sp}, the theory of $\pm$*difference* method will first be considered. The description of the signal and the assumptions are the same as those described for the conventional F_{sp}. $\overline{EEG'(t)}$ is defined as:

$$\overline{EEG'(t)} = \frac{1}{K}[\sum_{i} \{EEG_i(t) \bullet (-1)^i\}] \tag{7.3.1}$$

i=1 to K, K even

An alternative variance ratio is then formed and denoted as $\pm$ *difference*:

$$\pm difference = \frac{VAR(\overline{EEG})}{VAR(\overline{EEG'})} \tag{7.3.2}$$

Wong and Bickford (1980) suggested that for human data, the empirical value of $\pm$*difference* > 2, based on the experimental results, might be used as a criterion in the detection of the ABR.

7.3.2 Critical values

Following the same methods as for the bootstrap-F_{sp}, the critical value of bootstrap-$\pm$difference was obtained in each recording from the bootstrap sampling distribution, by finding the values of the $\pm$*difference* at $\alpha = 5\%$ (or other desired significance level, e.g., $\alpha = 1\%$). Figure 7.3 shows the critical values from the bootstrap-$\pm$difference at $\alpha = 5\%$ and $\alpha = 1\%$, respectively, based on 250 sweeps. The critical values were obtained from 16 subjects (x-axis) and each marked line corresponded to one stimulus intensity. One-way ANOVA was again performed on critical values grouped by 16 subjects and by 7 stimulus intensities, respectively. The results from ANOVA for both 250 sweeps and 2000 sweeps consistently indicated that the critical values varied between subjects (Table 7.5(a)) and did not vary by stimulus intensity (Table 7.5(b)).

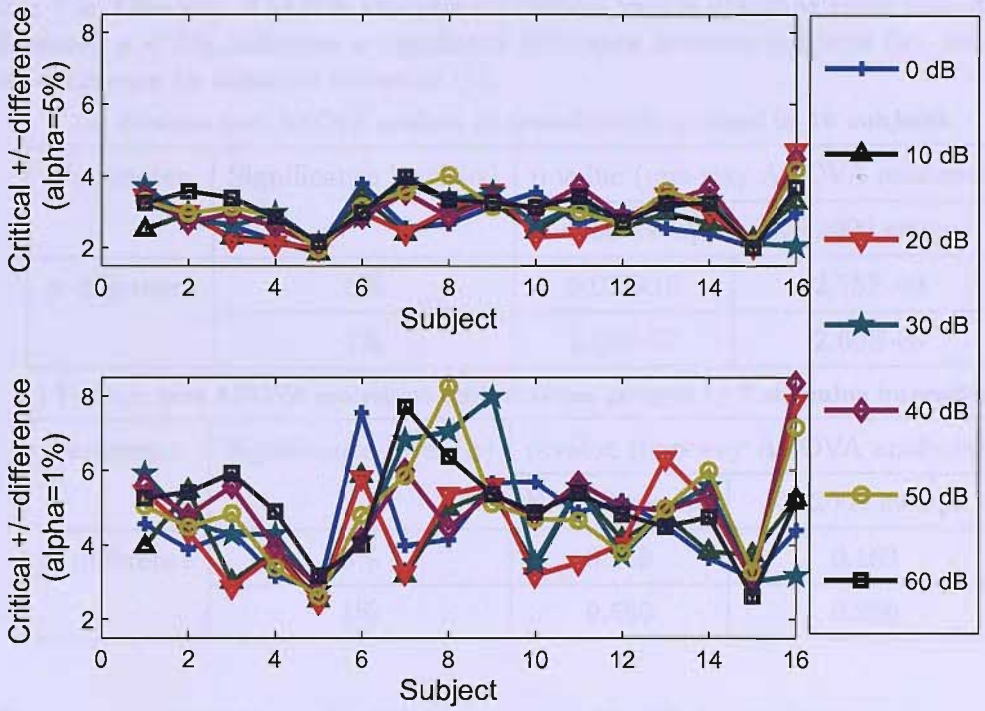


Figure 7.3: Critical values of ± *difference* from the bootstrap cumulative probability distribution with $\alpha = 5\%$ (upper) and $\alpha = 1\%$ (lower). These were obtained with 250 sweeps.

Therefore the critical value of ± difference method empirically determined as 2 by experts (Wong and Bickford, 1980), might not be suitable in all cases. Because ANOVA analysis (see Table 7.1(a)) indicated a significant difference between individuals, and also from theory, it is clear that this critical value depends on the signal characteristics (in particular the spectrum of the signal) and this differs between individuals (subjects).

7.3.3 Pass and Refer rate

The criteria of the bootstrap-± difference and ± *difference* are shown in Figure 7.6. Tables 7.7 and 7.8 show that agreement between the bootstrap-±difference and ±

Table 7.5: One-way ANOVA analysis on critical values obtained from bootstrap-± difference. $p < 5\%$ indicates a significant difference between subjects (a), but there is no difference by stimulus intensity (b).

(a) P-values from ANOVA analysis on critical values, grouped by **16 subjects**.

Parameter	Significance level (α)	p value (one-way ANOVA analysis)	
		K=250 sweeps	K=2000 sweeps
± difference	5%	5.07E-10	2.15E-09
	1%	1.78E-07	2.60E-08

(b) P-values from ANOVA analysis on critical values, grouped by **7 stimulus intensities**.

Parameter	Significance level (α)	p value (one-way ANOVA analysis)	
		K=250 sweeps	K=2000 sweeps
± difference	5%	0.158	0.163
	1%	0.590	0.256

difference was 86.61% for K=250 sweeps and 90.18% for K=2000 sweeps, respectively. The improvement in 'pass' rates with both methods is very evident. This again demonstrates that the increase in the number of sweeps greatly increased the probability of detecting the response.

Table 7.6: Criteria for ±different and bootstrap-±difference methods under different conditions, K=250 sweeps and K=2000 sweeps.

	Pass		**Refer**	
	K=250	K=2000	K=250	K=2000
±difference	±difference > 2		±difference ≤ 2	
Bootstrap - ±difference	p ≤ 0.05		p > 0.05	

K=250	**±difference**	
Bootstrap-±difference	Pass	Refer
Pass	**22** (19.64%)	**0** (0%)
Refer	**15** (13.39%)	**75** (66.96%)

Table 7.7: Pass and refer rate for K=250 sweeps.

K=2000	**±difference**	
Bootstrap-±difference	Pass	Refer
Pass	**72** (64.29%)	**0** (0%)
Refer	**11** (9.82%)	**29** (25.89%)

Table 7.8: Pass and refer rate for K=2000 sweeps.

7.3.4 Discussion

The results from the above showed that when K=250 sweeps, there were 15 cases referred by the bootstrap-±difference and passed by ±difference and 11 cases when K=2000 sweeps. This means that ± difference method is more sensitive than bootstrap-± difference. This is not surprising, given that the threshold is lower. However, the determination of the criterion for the ± difference might be not reliable. The means for estimating the degree of the freedom of $VAR(\overline{BEEG})$, ν_1 can also be used in the statistical evaluation of the ±difference, which was defined as a ratio of $VAR(\overline{BEEG})$ to $VAR(\overline{BEEG'})$. From the expressions of the conventional F_{sp} and ±difference, it is clear that the difference between them is the denominator of the expressions, as the numerators are the same. Therefore for the ±difference, the degree of the numerator under the worst case assumptions suggested in Elberling and Don (1984) could be $\nu_1 = 5$, and that of the denominator, an alternative estimates of the background EEG, $VAR(\overline{BEEG'})$ should have the same degrees of the freedom as the normal averaged background EEG, $\overline{BEEG}$. As mentioned before, the ratio, $VAR(\overline{BEEG})/VAR(\overline{BEEG'})$ follows the F-distribution. Similarly to estimating critical values for the F_{sp}, that for ±difference can be estimated from the F distribution with $\nu_1 = \nu_2 = 5$ (the smallest degrees of freedom for ν_1). The

corresponding upper 95% percentile is $F_{(5,5)} = 5.05$. Elberling and Don (1984) discussed that in estimation of the response quality, this means that compared with the single point method, the use of the ±difference would increase the uncertainty approximately by a rate of 2 (5.05/2.25) (Elberling and Don, 1984). From the above analysis, the critical value of 5.05 might be more reliable for the ±difference.

The comparison of the experimental results between K=250 sweeps and K=2000 sweeps show that the pass rates increase greatly. These were as expected, since larger number of sweeps will help to recover the evoked potentials and make them easier to detect.

7.4 Signal bootstrap and ensemble bootstrap

In the bootstrap method proposed and applied in previous chapters, the raw data (signal) is randomly resampled with replacement such that the 'sweeps' are no longer synchronized with the stimuli.

Nocera and Ferlazzo (2000) proposed an alternative bootstrap process to assess the within-subject reliability of experimental modulation effects on event-related potentials (ERPs). Basically, the bootstrap method was then used to resample the sweeps of the ensemble, and thus each resample was synchronized with the stimuli, unlike the bootstrap method proposed in this thesis. The latter will now be referred to as the 'signal bootstrap', and the former as 'ensemble bootstrap'.

Considering the principles of the ensemble bootstrap method, it may provide an alternative in detecting the ABR. An introduction of the initial ensemble bootstrap will be given. Then two ensemble bootstrap methods will be introduced for detecting the ABR. The false positive rate will be tested and problems with the ensemble bootstrap will then be discussed. Finally, a discussion based on the problems and a comparison of the ensemble bootstrap and the signal bootstrap will be provided.

7.4.1 Introduction to the ensemble bootstrap

The ensemble bootstrap was originally proposed to compare the difference of event-related potentials (ERPs) in response to 'old' and 'new' words (old/new effect) (Nocera and Ferlazzo, 2000). In memory tasks, ERPs have been reported to differ in terms of whether the stimuli evoking them had been previously presented or not (Rugg, 1995). The old/new effect refers to a larger positive peak at about 400 ms after stimulation shown by the ERPs to old items relative to the ERPs to new items. The mean value of ERPs is different after the presentation of the words. In order to test whether the difference in the ERPs is significant or not in each subject, the ensemble bootstrap method was applied.

In accordance with the null-hypothesis, it was supposed there was no difference between the responses to the new and old words. In the ensemble bootstrap method first the responses to the new and old words are pooled together, and from this pool two groups with the same number of stimuli are randomly selected, and coherently averaged. The difference in mean amplitude of these averages provides an estimate of this difference under the null-hypothesis of 'no difference present'. This process was repeated 1000 times, and the distribution of this difference was then estimated. By comparing this with the difference obtained between 'old ' and 'new' words, the significance value (p-value) could be found in accordance with the usual bootstrap procedure. That will indicate the acceptance or rejection of the null-hypothesis at the pre-defined significance level (e.g. $\alpha = 5\%$).

7.4.2 Ensemble bootstrap for ABR detection

In order to apply the ensemble bootstrap method for detecting the ABR, two points should be considered: the null-hypothesis and the parameter. As for the signal bootstrap method, the null-hypothesis is 'no response present' (H_0). The parameter under this null-hypothesis can be estimated from the no-stimulus signal or the stimulus signal. For the former, the basic principle of the ensemble bootstrap is to compare the ensemble following stimuli and one acquired without stimulation, thus two signals are actually needed for each subject. For the alternative approach, the signal without stimuli is no longer required, and the parameter under the null-hypothesis should be

estimated from the signal acquired during stimulation.

As already discussed, when presenting acoustic stimulation (clicks or tone-burst) to the normal-hearing ear, there should be a dominant peak (wave V) at the latency of 5-10 ms after the onset of the stimulus in the ABR recording. A parameter is therefore required to represent this feature, and considering that *power* was found to be the most powerful one in previous chapters, here it is again used to test the ensemble bootstrap method. When the parameter under the null-hypothesis (denoted by θ_0) is estimated from the no-stimulus signal, the method is called 'two-ensemble bootstrap' and the alternative 'one-ensemble bootstrap'.

The procedures for the two-ensemble bootstrap are: firstly, θ_0 is calculated from the coherent average of the no-stimulus recording under the null-hypothesis. Then the signal obtained during stimulation is coherently averaged to calculate the observed value of θ. Following the ensemble bootstrap method, a 'new' ensemble is then generated by randomly resampling the sweeps of the observed signal with the same number as in the original signal, and is averaged to obtain an estimated parameter (θ^*). This bootstrap process is repeated 499 times (the same as in the signal bootstrap method). Finally a distribution of θ^* is achieved. By calculating the percentage of θ_0 greater than θ^*, the significance value (p-value) is found. Again, choosing a significance level (α), if $p \leq \alpha$, a significant response is considered to be present.

The two-ensemble bootstrap method was tested on 500 simulated signals (as in Chapter 6) without stimulation and $\alpha = 5\%$, in order to evaluate the false positive rate. This rate was 22.2% and outside the range given by the binomial distribution with 500 trials and a success rate of 5%. The reason leading to this was an inherent 'error' resulted from the ensemble bootstrap resampling process with replacement, as can be explained from theory.

A no-stimulus or stimulus signal can be written in a matrix as described in equation 2.2.1 and $x_{k,m}$ represents one sample value. $k = 1, 2, ..., K$, where K refers to the number of the sweeps, and $m = 1, 2, ..., M$, where M is the number of the sample in each sweep. The mean power of the ensemble (P_x) can be calculated as:

$$P_x = \frac{1}{N} \sum_{M} \sum_{K} x_{k,m}^2 \tag{7.4.1}$$

where $N = K \times M$. The mean power of the coherently averaged signal ($P_{\bar{x}}$) can be estimated as:

$$\begin{aligned} P_{\bar{x}} &= \frac{1}{M}\sum_{M}\bar{x}_m^2 \\ &= \frac{1}{M}\sum_{M}(\frac{1}{K}\sum_{K}x_{k,m})^2 \\ &= \frac{1}{M}\frac{1}{K^2}\sum_{M}(\sum_{K}x_{k,m})^2 \end{aligned} \tag{7.4.2}$$

Then expand the components in the parentheses, separately. The square of the sum includes the sum of the square and the product of two sweeps.

$$\begin{aligned} (\sum_{k=1}^{K}x_{k,m})^2 &= (x_{1,m}+x_{2,m}+x_{3,m}+...+x_{K,m})(x_{1,m}+x_{2,m}+x_{3,m}+...+x_{K,m}) \\ &= x_{1,m}^2+x_{2,m}^2+x_{3,m}^2+...+x_{K,m}^2 \\ &\quad +x_{1,m}x_{2,m}+x_{1,m}x_{3,m}+...+x_{1,m}x_{K,m}+...+x_{K-1,m}x_{K,m} \end{aligned} \tag{7.4.3}$$

Since each sweep in the ensemble is approximately uncorrelated, the products of any two sweeps ($x_{1,m}x_{2,m}$, $x_{1,m}x_{3,m}$, ...) in equation 7.4.3 are zero. When calculating the value of power from the no-stimulus signal ($P_{\bar{x}0}$ under the null-hypothesis of 'no response present'), and the power ($P_{\bar{x}}$) from the coherently averaged signal of the stimulus signal, are zero. However, the cross-terms for the estimated power from the resamples will be a positive value, since a sweep is possibly selected more than once, and the part of $x_{1,m}x_{2,m}$ becomes $x_{1,m}^2$. Therefore, the distribution of the estimated power will be consistently greater than the expected values as shown in Figure 7.4. For the values of $P_{\bar{x}0}$ (dotted vertical line in Figure 7.4) and $P_{\bar{x}}$ (solid vertical line), the shift of the estimated distribution to the right of the expected (realistic) distribution leads to a decrease of the p-value. Decreased p-value will increase the false positive rate. However, in detecting a response (when present), that will also increase the sensitivity (detection rate). For the parameter *power*, the ensemble bootstrap thus provides a 'biased' result, because the parameter involves second order terms that

use the correlation between samples. Resampling with replacement in this ensemble bootstrap form thus has an inherent problem.

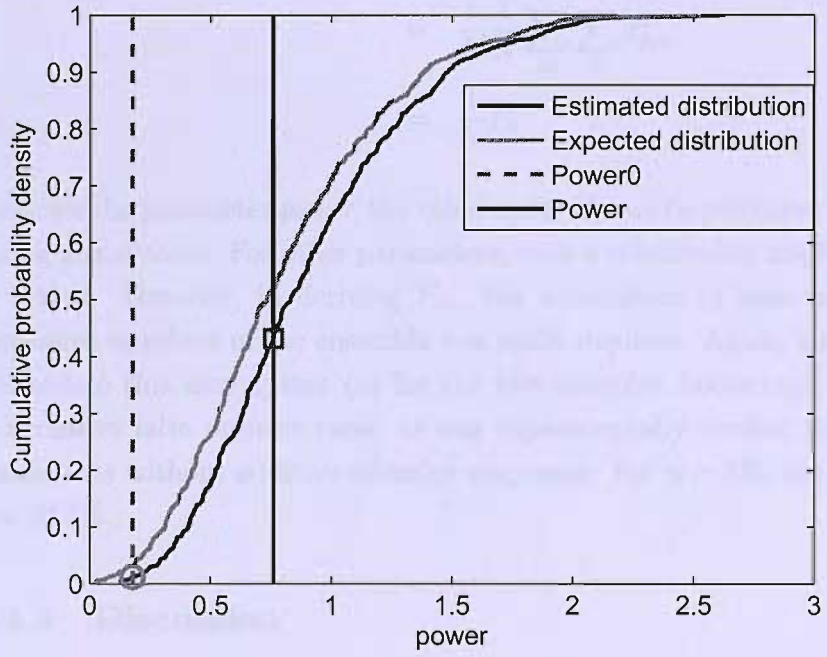


Figure 7.4: An illustration of the cumulative distribution of the power as estimated from the ensemble bootstrap method and as expected from theory (usually not available) under the null-hypothesis. The estimated distribution is shifted to the right of the expected one because of the influence of the resampling process.

The one-ensemble bootstrap method was then considered. Under the null-hypothesis (no stimulus response), the power can be estimated from P_x based on the following derivation. Again the product of the samples from two different sweeps is zero (expected value), as successive sweeps are considered uncorrelated as in equation 7.4.2 and 7.4.3.

$$(\sum_{k=1}^{K} x_{k,m})^2 = \sum_{K} x_{k,m}^2 \tag{7.4.4}$$

$$\begin{aligned} P_{\bar{x}0} &= \frac{1}{M}\frac{1}{K^2}\sum_M\sum_K x_{k,m}^2 \\ &= \frac{1}{K}\frac{1}{N}\sum_M\sum_K x_{k,m}^2 \\ &= \frac{1}{K}P_x \end{aligned} \tag{7.4.5}$$

Thus, for the parameter *power*, the value under H_0 can be predicted from the recording during stimulation. For other parameters, such a relationship might be more difficult to derive. However, in deriving $P_{\bar{x}0}$, the assumption of zero correlation between successive members of the ensemble was again required. Again, ensemble resampling will isolate this assumption (as for the two-ensemble bootstrap). This again leads to excessive false positive rates, as was experimentally verified in 500 Monte Carlo simulations without additive stimulus responses. For $\alpha = 5\%$, the false positive rate was 20.4%.

7.4.3 Discussion

The two ensemble bootstrap methods demonstrated an inherent problem deriving from the resampling process with replacement, for the parameter *power*. That parameter was selected, because it quantifies the main feature of the signal and is easily evaluated.

However, there are two significant advantages of the ensemble bootstrap. First, the testing time was fairly short: $2s$ for each signal (2000 sweeps), whereas that for the signal bootstrap was about $19s$ on the same PC. There is a big difference between these, because the ensemble resampling is done sweep by sweep, but signal resampling is carried out on the signal. In clinical applications, the reduction of the testing time is clearly desirable.

Another advantage of the ensemble bootstrap is related to the artefact rejection scheme. For the ensemble bootstrap method, stimulus artefact rejection does not need to be considered because the stimulus artefact is never involved in the calculation of the parameter when setting an analysis window to exclude the stimulus

artefact, say 5-15 ms. The stimulus artefact always appears during the first 2 ms after the onset of the stimuli. The movement artefact rejection scheme (MAR) is also much easier to apply in the ensemble bootstrap than in the signal bootstrap. In the signal bootstrap, at the stage of estimating the observed parameter, MAR is applied to the raw signal, and sweeps exceeding the rejection threshold are removed from the signal. In the bootstrap process, this procedure has to be applied again to each resampled signal. Thus prolonging the testing time.

In assessing a method from the detection task, the false positive rate is a core factor. Thus, even though the benefits from the ensemble bootstrap method are evident, the ensemble bootstrap method can not replace the signal bootstrap method for detecting the ABR, when the parameter *power* is employed. For other parameters (e.g. first order such as difference), the ensemble bootstrap may work, though mathematical analysis would be more complex.

Chapter 8

Conclusions and Future Work

8.1 Conclusions

Bootstrap methods have become a very widely used tool for statistical analysis. They have also been extensively exploited in many areas of signal processing, as introduced in Chapter 2. To the best of our knowledge, this is the first proposed use of the bootstrap method for detecting evoked potentials.

Conventionally, the statistical analysis of a signal parameter is performed on repeated signals to establish the distribution of the parameter, or based on some assumptions about the signal (i.e., Gaussian distributed) or the parameter. However, in this application, it is not feasible to repeat the experiment many times, because the recording time would be too long for the patients to tolerate. Moreover, intra-subject variability is present and thus makes the estimates more unreliable. The bootstrap method allows the statistical significance of arbitrary signal parameters to be assessed and thus provides a very powerful tool for the future development of evoked-response analysis, including the selection of new and optimized parameters for response detection. It allows responses to be detected at a user-defined false-positive rate, for an arbitrary number of stimuli, and takes the statistical characteristics of each individual recorded signal into account. In the current work, the results on six parameters were presented, but the bootstrap method could be applied to other parameters.

Bootstrap methods facilitate the analysis of data subject-by-subject. Thus, rather than performing comparison between groups of subjects, it is often possible to perform

statistical tests on each case individually. This avoids the requirement for the often questionable assumption such as that signals recorded from different subjects have similar statistical characteristics. In the investigation of the critical values (discussed in Chapter 7), the variations between subjects are present. Therefore independent analysis for individual case is a great benefit of the bootstrap method.

In Chapter 5, the Basic bootstrap was tested on simulations and recorded data Set A. The results for 500 simulated background EEG signals provided acceptable false positive rates (within the range of 3.2% - 6.8%), and those for 500 simulated signals with a 'stimulus response' gave a sensitivity function for a range of SNR levels. The results on the recorded Set A indicated that the Basic bootstrap method provided slightly lower hearing thresholds than those from audiologists using traditional visual inspection.

In this work, the bootstrap method was tested on the ABRs. It also can be applied in other modalities (e.g. visual, somatosensory, and event-related). Bootstrap methods have been used previously in finding confidence limits for the SNR and inter-ocular amplitude ratio in visual evoked potentials (Fortune et al., 2004), and various parameters in somatosensory evoked potentials (Adams and Kunz, 1996), as well as in assessing ROC curves (Valdes et al., 1997) for steady-state auditory evoked potentials. However, it does not appear to have been used previously for detecting the presence of an evoked response. In the current work we are not proposing that the bootstrap method should replace established statistical criteria for detecting responses (Mauricio et al., 2001; Simpson et al., 2000). However, the bootstrap method can be applied in testing the significance of parameters that are not readily analysed by conventional statistical approaches, such as the F_{sp} or $\pm difference$.

In order to overcome the limitations of the Basic bootstrap method, three artefact rejection schemes MAR, SAR and SMAR were proposed in Chapter 6. The results on the simulated signals showed that MAR, SAR and SMAR can eliminate the influence of the corresponding artefacts and provide higher sensitivity and acceptable false positive rates. There was no significant improvement when employing MAR, SAR and SMAR on the recorded data Set A and B, but neither did these techniques degrade results, compared to the Basic bootstrap method.

The Basic bootstrap method provided similar false positive rates and sensitivity for

the recorded data Set B, compared to MAR, SAR and SMAR, since visual inspection indicated that no obvious artefacts were present. However, when artefacts were present (e.g. in simulated data), the Basic bootstrap method performed poorly. In particular, greatly reduced or unreasonably increased sensitivity was observed (Figure 6.4, 6.11 and 6.16). The false-positive rate (in the absence of a stimulus response) was too low or too high compared to the nominal 5%. Too high false positive rate is of some concern, as in hearing screening more than the expected 5% of subjects with impairment may be missed. This might harm the subject, because he/she may not receive suitable treatment as quickly as possible. Therefore the conclusion can be drawn that the modified bootstrap method (SMAR) should replace the Basic method.

Of course the benefits of the bootstrap method come at a cost. These approaches usually cost a few seconds, and this might prevent them from some real-time applications. However, compared to the time of data acquisition or pre-processing, this may be negligible and the computational cost is no longer a good reason to discard bootstrap methods. More efficient implementation of the algorithm could greatly reduce computational time. The other limitation associated with the random feature of the bootstrap process, is that results are not precise: repeat runs will provide slightly different results. But this variation can be reduced by using a large number of resamples. This variation can be considered in context: when the experiment and data acquisition were repeated, even conventional methods would possibly produce different results to some extent.

In this study, the detected fraction of the simulated signals with different SNR (Figure 5.6 in Chapter 5), and ROC analysis demonstrated the good performance of bootstrap technique for detecting the response. But bootstrap estimates, like all statistical methods, have inherent errors (Efron, 1993) shown in Figure 8.1. These errors come from two distinct sources: sampling variability, due to the fact that only a sample of size n rather than the entire population, is available, and bootstrap resampling variability, due to the fact that only B bootstrap samples rather than an infinite number, are taken. In the current work, 2000 sweeps contributed to the calculation of the coherent average from which the parameters were extracted. Thus the sample size of 2000 is fairly big to approximate the population and greatly avoid the sampling variability. And 499 repeats used here to estimate the distribution of the parameter might not lead to a great resampling variability.

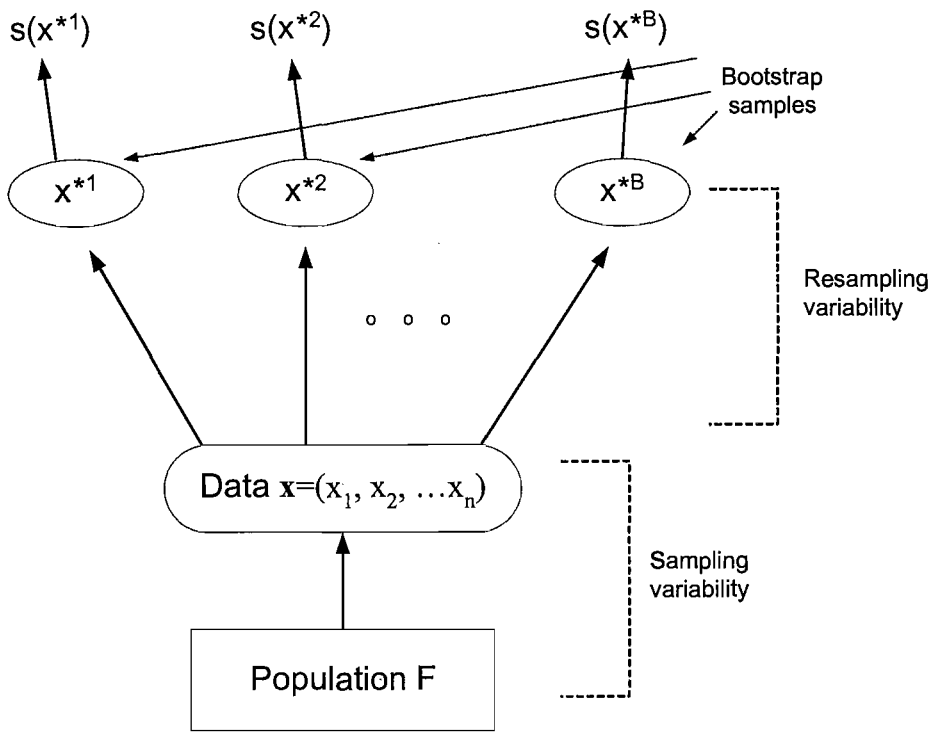


Figure 8.1: The schematic shows the sampling and resampling variability. The sample variability is due to the finite sample size n which can not represent the entire population. The resampling variability results from the finite number of bootstrap resamples which can not ideally demonstrate the sample data x.

Independence between observational units is often an important assumption in data analysis and is usually present in bootstrap-based inferences. Lack of independence can reduce the accuracy of interference: see Hampel et al. (1986) for discussion of this. In this work, the parameter $\pm$*difference* and *cc* always worked poorly, compared to other four parameters. A contributing factor could be the dependence of the signals. For example, when calculating the correlation coefficient of the two signals, they are assumed to be independent. When the bootstrap resampling process performs, one sample may be selected many times for different sweeps, and this sample can be possible in any position of a signal. Thus the 'incoherent' average of these sweeps will not be independent between samples, and correlation coefficient of these two 'incoherent' averages will be biased.

8.2 Future work

Issues arising from the current work deserve further study. In this section, firstly, two possible alternative implementations of bootstrap methods will be introduced, in order to improve the efficiency in computation time. Secondly, a more specific protocol for determining the hearing threshold will be described. Thirdly, bootstrap methods could be applied to additional parameters, which can be extracted from both time or Fourier (frequency) presentations of the signal. Fourthly, the further investigation of bootstrap methods on other signals will be mentioned. Finally, some considerations on bootstrap methods applied in clinical applications will be given.

8.2.1 Efficiency in computation time

In the work described here, analysis was carried out off-line. On a PC-Pentium based computer, it took approximately 19s to analyse a recording containing 2000 sweeps, when programmed in Matlab®. A more efficient implementation would considerably reduce this time. Using an extended form of this approach, on-line methods for could also be developed, in which the p-value could be continuously updated, and the recording stopped, once a threshold p-value was reached. Such a procedure could considerably shorten the time required for hearing tests, and follows the suggestions of Don and Elberling (1996).

Alternatively, a critical value of the interested parameter can be estimated from bootstrap resamples (based on one recording) under the null-hypothesis of 'no response present'. Then this parameter of any recording, calculated from the coherent averaged signal, is compared with the critical value. If this is greater than the critical value, a response is considered to be present. In addition, two points should be considered. As known that characteristics of recordings from different subjects vary and affect the critical value, and thus different critical values for different subjects should be considered. In addition, the critical value is influenced by the number of sweeps. Therefore, a suggestion is that for each subject, the critical value is determined with a known (or fixed) number of sweeps and 'test' recordings should have the same number of sweeps.

8.2.2 Protocol of hearing threshold

In the current work, the hearing threshold was determined by finding the minimal stimulus intensity at which $p < 0.05$ (with $p < 0.05$ for all higher stimulus intensities). There are some cases where the responses in the recordings from stimulations at 10 dB, 30 dB or higher stimulus intensities can be detected, however, no response is found in the 20 dB-recording, with p-value slightly greater than 0.05 (threshold p-value). Following the current protocol, the hearing threshold would then be determined as 30 dB. It is possible that the bootstrap method on the replicate 20 dB-recording provides $p < 0.05$, and a response might have been detected. Then 10 dB will be the hearing threshold. According to this, a more specific protocol can be proposed. For example, for the same case, bootstrap methods on more recordings at 20 dB stimulus intensity are performed, and then it is determined whether the 20 dB-recording has a significant response, by statistical analysis. This further test may give a relatively lower hearing threshold than the current approach.

8.2.3 Other parameters

Bootstrap methods are tested mainly on six parameters extracted from the representation of the signals in the time domain. In the future, additional parameters can be built up in the time or frequency domain.

It was already discussed in Chapter 6 that the parameter *cc* always performs poorly because only the shape of the two sub-averages is taken into account. Therefore, a modified parameter *cc* can be made by taking the amplitude information into account by using un-normalized coefficient, i.e., the covariance between the sub-averages. Alternatively, a modification of cc' can performed as follows: firstly, a 'template' (typical) signal is obtained by averaging recordings at higher stimulus intensity, e.g. 60 dB from many subjects. Then cc' (as discussed in Chapter 6) would be calculated between the 'template' and the coherent average from a new recording. Following that, *ccs* are computed between the 'template' and 'incoherent' averages obtained from bootstrap resamples. The significance p-value is again used to determine whether a response is present or absent.

Furthermore, the parameters can be extracted from the Fourier representations of

the signal, using the amplitude and/or phase information at one or more frequencies. The selection of parameters can take MSC, PC (or Rayleigh test), circular T^2 etc., as references.

8.2.4 Bootstrap methods on other signals

The proposed bootstrap methods can be readily applied on other signals, such as middle latency response (MLR), steady-state evoked potentials, visual evoked potentials and so on. The issue is to select suitable parameters which can appropriately represent the signal. For example, if the bootstrap method is used for detecting 40 Hz steady-state evoked potentials, the parameter may come from the Fourier representation at a frequency of 40 Hz, where the amplitude should be dominant and the phase is clustered.

Moreover, when developing new applications, it is strongly recommended that the techniques are first tested in Monte Carlo simulations to generate data that is as realistic as possible but has known and well-controlled characteristics. The poor performance on the simulations can reveal shortcomings in the implementation or the underlying principle, e.g. the definition of the null-hypothesis.

8.2.5 Considerations of clinical application

A long-term aim based on the current work is to apply bootstrap methods in clinical application, e.g. hearing screening, surgical monitoring. Although the evaluation of the method has been made on both Monte Carlo simulations and recordings from normal-hearing subjects, it is also desirable to be further evaluated on recordings collected in hospitals, from different subjects (adults, children), with and without hearing loss and audiological pathologies, and possibly with artefacts. With a complete evaluation including sensitivity and false positive rate and an indication of the good performance, bootstrap methods may be used in commercial products which will be applied in clinical settings.

Appendix A

Screening Questionnaire

Do you think that your hearing is normal?

Have you ever had any persistent problems with your ears or hearing, for example discharging ears or earache?

Do you suffer from troublesome tinnitus?

Have you been exposed to loud noises, for example at work, gunfire or explosives?

Are you suffering from or recently had a cold?

Have you ever had attacks of dizziness or loss of balance related to vestibular (balance) disorder (if known)?

Have you ever suffered from high or low blood pressure?

Have you ever had an epileptic attack with convulsions or loss of consciousness?

Have you ever suffered with heart trouble?

Are you receiving any medical treatment or medication that may affect your hearing?

Appendix B

Consent Form

Consent form to be completed by adult subjects taking part in an experiment (Adults are 18 years of age or older)

Exposure Number: .....

University of Southampton

Institute of Sound and Vibration Research

Before completing this form, please read the objectives of the experiment which has been provided by the experimenter on the next page of this form.

This consent form applies to a subject volunteering to undergo an experiment for research purposes. The form is to be completed before the experiment commences.

I, ...
of .....
(address or department) consent to take part in '**Recordings of auditory brainstem response (ABR) with different stimulus levels and without stimulation**' to be conducted by Miss Jing Lv in ....., November, 2004.

The purpose and nature of this experiment have been explained to me. I understand that the investigation is to be carried out solely for the purposes of research. I

am willing to act as a volunteer for that purpose on the understanding that I shall be entitled to withdraw this consent at any time, without giving any reasons for withdrawal. My replies to the above questions are correct to the best of my belief, and I understand that they will be treated by the experimenter as confidential. I also agree that the data obtained in this experiment can be used in the research and PhD thesis of Miss Jing Lv.

Date: ..... Signed: .....
(Subject)

I confirm that I have explained to the subject the purpose and nature of the investigation which has been approved by the Human Experimentation Safety and Ethics Committee.

Date: ..... Signed: .....
(Researcher in charge of experiment)

This form must be submitted to the Secretary of the Human Experimentation Safety and Ethics Committee on completion of the experiment.

Appendix C

Formulae Derivation

Correlation coefficient

The different formulae for correlation coefficient (r) are derives as follow:

$$r = \frac{\sum (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{(\sum (x_i - \overline{x})^2)(\sum (y_i - \overline{y})^2)}}$$

In order to explain the procedures of transformation clearly, the nominator is denoted as A, the first sum of the denominator about variable x is denoted as B and the last sum of denominator about variable y is denoted as C. Let us calculate step by step in the order of A, B and C.

Because $\overline{x}$ has the same value of each of the n observations and could be taken as a constant and put outside the sign of sum.

$$\begin{aligned}
A &= \sum (x_i - \overline{x})(y_i - \overline{y}) \\
&= \sum (x_i y_i - x_i \overline{y} - \overline{x} y_i + \overline{xy}) \\
&= \sum x_i y_i - \overline{y} \sum x_i - \overline{x} \sum y_i + \overline{xy} \sum 1 \\
&= \sum x_i y_i - \frac{\sum y_i}{n} \sum x_i - \frac{\sum x_i}{n} \sum y_i - \frac{\sum y_i}{n} \frac{\sum x_i}{n} n \\
&= \sum x_i y_i - \frac{\sum x_i \sum y_i}{n}
\end{aligned}$$

$$\begin{aligned} B &= \sum (x_i - \overline{x})^2 \\ &= \sum (x_i^2 - 2x_i\overline{x} + \overline{x}^2) \\ &= \sum x_i^2 - 2\overline{x} \sum x_i + \overline{x}^2 \sum 1 \\ &= \sum x_i^2 - 2\frac{\sum x_i}{n} \sum x_i + (\frac{\sum x_i}{n})^2 n \\ &= \sum x_i^2 - \frac{(\sum x_i)^2}{n} \end{aligned}$$

Equivalently,

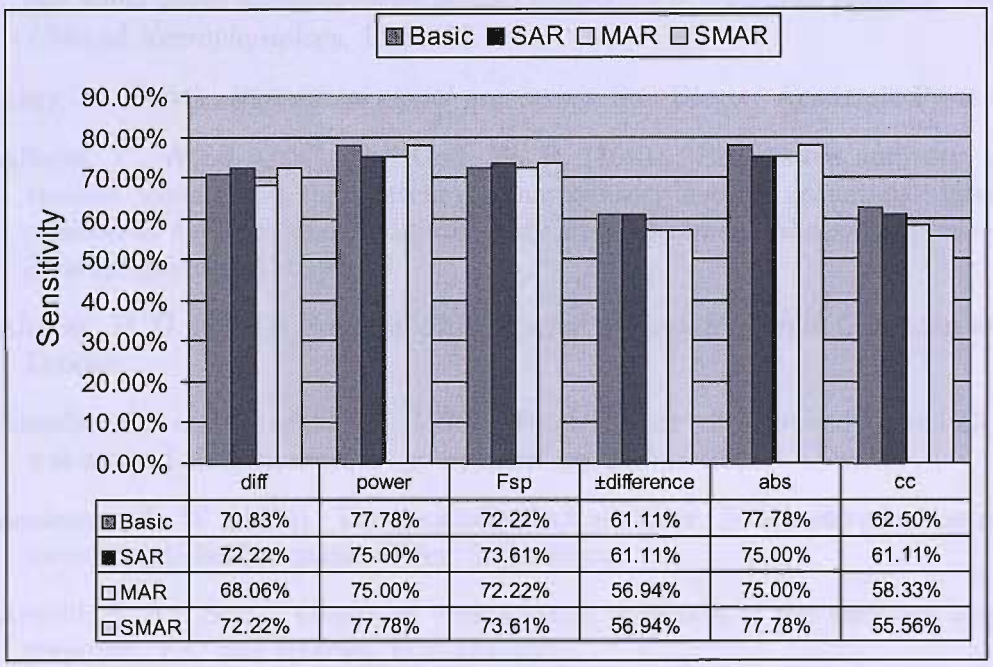
$$\begin{aligned} C &= \sum (y_i - \overline{y})^2 \\ &= \sum y_i^2 - \frac{(\sum y_i)^2}{n} \end{aligned}$$

Therefore, correlation coefficient could be represented in another expression by replacing the original A, B, and C by the new formulae as follows:

$$r = \frac{\sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n}}{\sqrt{(\sum x_i^2 - \frac{(\sum x_i)^2}{n})(\sum y_i^2 - \frac{(\sum y_i)^2}{n})}}$$

Appendix D

Sensitivity for data Set A



	diff	power	Fsp	±difference	abs	cc
Basic	70.83%	77.78%	72.22%	61.11%	77.78%	62.50%
SAR	72.22%	75.00%	73.61%	61.11%	75.00%	61.11%
MAR	68.06%	75.00%	72.22%	56.94%	75.00%	58.33%
SMAR	72.22%	77.78%	73.61%	56.94%	77.78%	55.56%

Figure D.1: Sensitivity estimated from Set A by Basic, MAR, SAR and SMAR-bootstrap methods.

References

Aceto, P., Valente, A., Gorgoglione, M., Adducci, E., and Cosmo, G. D. (2003). Relationship between awareness and middle latency auditory evoked responses during surgical anaesthesia. *British Journal of Anaesthesia*, 90(5):630–635.

Adams, H. P. and Kunz, S. (1996). Inter- and intraindividual variability of posterior tibial nerve somatosensory evoked potentials in comatose patients. *Journal of Clinical Neurophysiology*, 13:84–92.

Akay, M. (1994). *Biomedical signal processing.* San Diego : Academic Press.

Allison, T., Wood, C. C., and Goff, W. R. (1983). Brain stem auditory, pattern-reversal visual, and short-latency somatosensory evoked potentials: latencies in relation to age, sex, and brain and body size. *Electroencephalography and Clinical Neurophysiology*, 55:619–636.

Altman, D. G. (1991). *Practical Statistics for Medical Research.* Chapman and Hall, London.

Amadeo, M. and Shagass, C. (1973). Brief latency click-evoked potentials during waking and sleep in man. *Psychophysiology*, 10:244–250.

Anderson, T. W. (1984). The generalized t2 -statistic. In *An introduction to multivariate statistical analysis.* Wiley, 2nd edition.

Arnold, S. A. (1985). Objective versus visual detection of the auditory brain stem response. *Ear and Hearing*, 6(3):144–150.

Beer, N. A. M., Hooff, J. V. V., Brunia, C. H. M., Cluitmans, P. J. M., Korsten, H. H. M., and Beneken, J. E. W. (1996). Midlatency auditory evoked potentials as indicators of perceptual processing during general anasthesia. *British Journal of Anaesthesia*, 77:617–624.

Bell, S. L. (2003). *Improving Acquisition of Auditory Evoked Potentials for Clinical Diagnosis and Monitoring.* Thesis/dissertation.

Bendat, J. S. and Piersol, A. G. (1986). *Random data. Analysis and measurement procedures.* John Wiley and Sons, Inc., New York, second, revised and expanded edition.

Bland, M. (2000). *An introduction to medical statistics.* Oxford University Press Inc., Oxford and New York, third edition.

Blegvad, B. (1975). Binaural summation of surface recorded electrocochleographic responses in normal hearing subjects. *Scandinavian Audiology*, 4:233–238.

Bobbin, R. P., May, J. G., and Lemoine, R. L. (1979). Effects of pentobarbital and ketamine on brain stem auditory potentials. *Archives of Otolaryngology*, 105:467–470.

Borda, R. P. and Frost, J. D. (1968). Error reduction in small sample averaging through the use of median rather than the mean. *Electroencephalography and Clinical Neurophysiology*, 25:391–392.

Boston, J. R. (1981). Spectra of auditory brainstem responses and spontaneous eeg. *IEEE Transactions on Biomedical Engineering*, 28:334–341.

Cane, D. A. (2002). *A comparison between the click and the chirp in threshold auditory brainstem response measurements.* Thesis/dissertation.

Cebullar, M., Sturzebecher, E., and Wernecke, K. D. (1996). Objective detection of auditory evoked potentials. comparison of several statistical tests in the frequency domain by means of monte carlo simulations. *Scandinavian Audiology*, 25:201–206.

Cebullar, M., Sturzebecher, E., and Wernecke, K. D. (2000). Objective detection of auditory brainstem potentials: comparison of statistical tests in time and frequency domains. *Scandinavian Audiology*, 29(1):44–51.

Cerutti, S., Liberati, D., and Mascellani, P. (1985). Parameter extraction in eeg processing during riskful neurosurgical operation. *Signal Processing*, 9:25–35.

Champlin, C. A. (1992). Method for detecting auditory steady-state potentials recorded from humans. *Hearing Research*, 58:63–69.

Chávez, M., Martinerie, J., and Quyen, M. V. (2003). Statistical assessment of nonlinear causality: application to epileptic eeg signals. *Journal of Neuroscience Methods*, 124(2):113–128.

Cochran, W. G. (1950). The comparison of percentages in matched samples. *Biometrika*, 37:256–266.

Cooper, W. A. J. and Parker, D. J. (1981). Stimulus artefact reduction systems for the tdh-49 headphone in the recording of auditory evoked potentials. *Ear and Hearing*, 2(6):283–293.

Darvas, F., Pantazis, D., Kucukaltun-Yildirim, E., and Leahy, R. M. (2004). Mapping human brain function with meg and eeg: Methods and validation. *NeuroImage*, 23(SUPPL. 1):S289–S299.

Darvas, F., Rautiainen, M., Pantazis, D., Baillet, S., Benali, H., Mosher, J. C., Garnero, L., and Leahy, R. M. (2005). Investigations of dipole localization accuracy in meg using the bootstrap. *NeuroImage*, 25:355–368.

Davison, A. C. and Hinkley, D. V. (1997). *Bootstrap methods and their application.* Cambridge : Cambridge University Press.

DeLong, E. R., DeLong, D. M., and Clarke-Pearson, D. L. (1988). Comparing the area under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, 44:837–845.

Deupree, D. L. and Jewett, D. L. (1988). Far-field potentials due to action potentials traversing curved nerves, and crossing boundaries between cylindrical volumes. *Electroencephalography and Clinical Neurophysiology*, 70:355–362.

Dobie, R. A. and Wilson, M. J. (1989). Analysis of auditory evoked potentials by magnitude-squared coherence. *Ear and Hearing*, 10:2–13.

Dobie, R. A. and Wilson, M. J. (1993). Objective response detection in the frequency domain. *Electroencephalography and Clinical Neurophysiology*, 88:516–524.

Dobie, R. A. and Wilson, M. J. (1994). Objective detection of 40 hz auditory evoked potentials:phase coherence vs. magnitude-squared coherence. *Electroencephalography and Clinical Neurophysiology*, 92:405–413.

Dobie, R. A. and Wilson, M. J. (1996). A comparison of t test, f test, and coherence methods of detecting steady-state auditory-evoked potentials, distortion-product otoacoustic emissions, or other sinusoids. *Journal of the Acoustical Society of America*, 100(4):2236–2246.

Don, M., Allen, A. R., and Starr, A. (1977). Effect of click rate on the latency of auditory brain stem responses in humans. *Annals of Otolaryngology*, 86:186–195.

Don, M. and Elberling, C. (1996). Use of quantitative measures of auditory brainstem response peak amplitude and residual background noise in the decision to stop averaging. *Journal of Acoustic Society American*, 99(1):491–499.

Don, M., Elberling, C., and Waring, M. (1984). Objective detection of averaged auditory brainstem responses. *Scandinavian Audiology*, 13:219–228.

Don, M., Ponton, C. W., Eggermont, J. J., and Masuda, A. (1994). Auditory brainstem response (abr) peak amplitude variability reflects individual differences in cochlear response times. *Journal of the Acoustical Society of America*, 96:3476–3491.

Durka, P. J., Zygierewicz, J., Klekowicz, H., Ginter, J., and Blinowska, K. J. (2004). On the statistical significance of event-related eeg desynchronization and synchronization in the time-frequency plane. *IEEE Transactions on Biomedical Engineering*, 51(7):1167–1175.

Efron, B. (1979a). Bootstrap methods. another look at the jackknife. *The Annals of Statistics*, 7:1–26.

Efron, B. (1979b). Computers and the theory of statistics: Thinking the unthinkable. *SIAM Review*, 4:460–480.

Efron, B. (1981a). Censored data and the bootstrap. *Journal of the American Statistical Association*, 76:312–319.

Efron, B. (1981b). Nonparametric standard errors and confidence intervals (with discussion). *The Canadian Journal of Statistics*, 9:1–26.

Efron, B. (1993). *An introduction to the bootstrap.* Chapman and Hall, London.

Efron, B. and Gong, G. (1983). A leisurely look at bootstrap, the jackknife, and cross-validation. *American Statistician*, 37(1):36–48.

Egan, J. P. (1975). *Signal detection theory and ROC analysis.* ACADEMIC PRESS INC. LTD, London, United Kingdom.

Elberling, C. and Don, M. (1984). Quality estimation of averaged auditory brainstem responses. *Scandinavian Audiology*, 13(3):187–197.

Elberling, C. and Wahlgreen, O. (1985). Estimation of auditory brainstem response, abr, by means of bayesian inference. *Scandal Audiology*, 14:89–96.

Fisher, N. I. and Hall, P. (1989). Bootstrap confidence regions for directional data. *Journal of the American Statistical Association*, 84:996–1002.

Fisher, N. I. and Hall, P. (1990). New statistical methods for directional data - i. bootstrap comparison of mean directions and the fold test in palaeomagnetism. *Geophysical Journal International*, 101:305–313.

Fisher, N. I. and Hall, P. (1991). General statistical test for the effect of folding. *Geophysical Journal International*, 105:419–427.

Fortune, B., Zhang, X., Hood, D. C., Demirel, S., and Johnson, C. A. (2004). Normative ranges and specificity of the multifocal vep. *Documenta Ophthalmologica*, 109:87–100.

Fowler, C. G. and Noffsinger, D. (1983). Effects of stimulus repetition rate and frequency on the auditory brainstem response in normal cochlear-impaired, and viii nerve/brainstem-impaired subjects. *Journal of Speech and Hearing Research*, 26:560–567.

Fria, T. J. and Doyle, W. J. (1984). Maturation of the auditory brain stem response (abr): Additional perspectives. *Ear and Hearing*, 5:361–365.

Friedman, M. (1937). The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association*, 32:675–701.

Gennaro, L. D., Ferrara, M., Ferlazzo, F., and Bertini, M. (2000). Slow eye movements and eeg power spectra during wake-sleep transition. *Clinical Neurophysiology*, 111(12):2107–2115.

Gorga, M. P., Worthington, D. W., Reiland, J. K., Beauchaine, K. A., and Goldgar, D. E. (1985). Some comparisons between auditory brainstem response thresholds, latencies and the pure-tone audiogram. *Ear and Hearing*, 6:105–112.

Graimann, B., Huggins, J. E., Levine, S. P., and Pfurtscheller, G. (2002). Visualization of significant erd/ers patterns in multichannel eeg and ecog data. *Clinical Neurophysiology*, 113(1):43–47.

Gross, J., Timmermann, L., Kujala, J., Salmelin, R., and Schnitzler, A. (2003). Properties of meg tomographic maps obtained with spatial filtering. *NeuroImage*, 19(4):1329–1336.

Hall, J. M. (1992a). Effect of acquisition factors. In *Handbook of Auditory Evoked Responses*, book chapter 5, pages 177–220. Boston: Allyn and Bacon.

Hall, J. W. (1992b). *Handbook of Auditory Evoked Responses.* Boston: Allyn and Bacon, third edition.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., and Stahel, W. A. (1986). *Robust statistics: the approach based on influence function.* Wiley, New York.

Hashimoto, I., Ishiyama, Y., Yoshimoto, T., and Nemoto, S. (1981). Brain-stem auditory-evoked potentials recorded directly from human brain-stem and thalamus. *Brain*, 104(4):841–859.

Hayes, M. (1996). *Statistical Digital Signal Processing and Modeling.* John Wiley and Sons, Inc.

Haykin, S. (1986). *Adaptive filter theory.* Prentice-Hall.

Haynor, D. R. and Woods, S. D. (1989). Resampling estimates of precision in emission tomography. *IEEE Transactions on Medical Imaging*, 8:337–343.

Herrmann, K. R., Thornton, A. R., and Joseph, J. M. (1995). Automated infant hearing screening using the abr: Development and validation. *American Journal of Audiology*, 4(2):6–14.

Hood, L. J. (1998). *Clinical applications of the auditory brainstem responses.* Singular Publishing Group, Inc., Aan Diego London.

Hotelling, H. (1931). The generalization of student's ratio. *The annals of Mathematical Statistics*, 2(3):360–378.

Hyde, M. L. (1985). Instrumentation and signal processing. In Jacobson, J. T., editor, *The Auditory brainstem response*, pages 33–48. Taylor & Francis.

Isaksson, A., Wennberg, A., and Zetterberg, L. H. (1981). Computer analysis of eeg signals with parametric models. *Proceedings of the IEEE*, 69:451–461.

James, C. J. and Lowe, D. (2003). Extracting multisource brain activity from a single electromagnetic channel. *Artificial intelligence in medicine*, 28:89–104.

Jasper, H. H. (1958). The ten-twenty electrode system of the international federation. *Electroencephalography and Clinical Neurophysiology*, 10:371–375.

Jerger, J., Chmiel, R., Frost, J. D., and Coker, N. (1986). Effect of sleep on the auditory steady state evoked potential. *Ear and Hearing*, 7(4):240–245.

Jerger, J. and Hall, J. (1980). Effects of age and sex on auditory brainstem response. *Archives of Otolaryngology*, 106:387–391.

Jervis, B. W., Nichols, M. J., Johnson, T. E., Allen, E., and Hudson, N. R. (1983). A fundamental investigation of the composition of auditory evoked potentials. *IEEE Transactions on Biomedical Engineering*, 30:43–50.

Jewett, D. L., Romano, M. N., and Williston, J. S. (1970). Human auditory evoked potentials:possible brainstem components detected on the scalp. *Science*, 167:1517–1518.

Jewett, D. L. and Williston, J. S. (1971). Auditory evoked far fields averaged from the scalp of humans. *Brain*, 94:681–696.

Jones, T. A., Stockard, J. J., and Weidner, W. J. (1980). The effects of temperature and acute alcohol intoxication on brain stem auditory evoked potentials in the cat. *Electroencephalography and Clinical Neurophysiology*, 49:23–30.

Kannurpatti, S. S. and Biswal, B. B. (2005). Bootstrap resampling method to estimate confidence intervals of activation-induced cbf changes using laser doppler imaging. *Journal of Neuroscience Methods*, 146(1):61–68.

Katz, J. (2001). *Handbook of clinical audiology.* Lippincott Williams and Wilkins, Philadephia, fifth edition.

Kileny, P. R. (1988). New insights on infant abr hearing screening. *Scandinavian Audiology Supplement*, 30:81–88.

Kraus, N., Ozdamar, O., Heydemann, P. T., Stein, L., and Reed, N. L. (1984). Auditory brain-stem responses in hydrocephalic patients. *Electroencephalography and Clinical Neurophysiology*, 59:310–317.

Kukreja, S. L., Galiana, H. L., and Kearney, R. E. (2004). A bootstrap method for structure detection of narmax models. *International Journal of Control*, 77:132–143.

Landis, J. R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33:159–174.

Liberati, D., DiCorrado, S., and Mandelli, S. (1992). Topographic mapping of single sweep evoked potentials in the brain. *IEEE Transactions on Biomedical Engineering*, 39:943–951.

Lutkenhoner, B. (1991). Theoretical considerations on the detection of evoked responses by means of the rayleigh test. *Acta Otolaryngologica (Supplement)*, 491:52–60.

Lutman, M. E. and Sheppard, S. (1990). Quality estimation of click-evoked otoacoustic emissions. *Scandinavian Audiology*, 19:3–7.

Lv, J., Bell, S. L., and Simpson, D. M. (2004a). A statistical test for the detection of auditory evoked potentials. *IPEM meeting: Signal Processing Applications in Clinical Neurophysiology.*

Lv, J., Simpson, D. M., and Bell, S. L. (2004b). Objective tests for the detection of auditory evoked potentials. In L Sweetman, O. J. G. and James, C. J., editors, *Third IEEE EMBSS UK and RI Postgraduate Conference in Biomedical Engineering and Medical Physics*, pages 1–2. IEEE EMB Student Society UKRI.

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

Margolis, D. J., Bilker, W., Boston, R., Localio, R., and Berlin, J. A. (2002). Statistical characteristics of area under the receiver operating characteristic curve for a simple prognostic model using traditional and bootstrapped approaches. *Journal of Clinical Epidemiology*, 55:518–524.

Marple, S. L. (1987). *Digital spectral analysis with applications.* Prentice Hall.

Martin, M. and Moore, E. (1977). Scalp distribution of early (0 to 10 msec) auditory evoked responses. *Archives of Otolaryngology*, 103:326–328.

Mason, J. A. and Herrmann, K. R. (1998). Universal infant hearing screening by automated auditory brainstem response measurement. *Pediatrics*, 101(2):221–228.

Mason, S. M. (1984). On-line computer scoring of the auditory brainstem response for estimation of hearing threshold. *Audiology*, 23:277–296.

Massof, R. W. and Emmel, T. C. (1987). Criterion-free parameter-free distribution-independent index of diagnostic test performance. *Applied Optics*, 26(8):1395–1408.

Mauricio, A., de Sa, F. L. M., Infantosi, A. F., and Simpson, D. M. (2001). A statistical technique for measuring synchronism between cortical regions in the eeg during rhythmic stimulation. *IEEE Transactions on Biomedical Engineering*, 48(10):1211–1215.

Mendelson, T., Salamy, A., Lenoir, M., and McKean, C. (1979). Brain stem evoked potential findings in children with otitis media. *Archives of Otolaryngology*, 105:17–20.

Menon, V., Freeman, W. J., Cutillo, B. A., Desmond, J. E., Ward, M. F., Bressler, S. L., Laxer, K. D., Barbaro, N., and Gevins, A. S. (1996). Spatio-temporal correlations in human gamma band electrocorticograms. *Electroencephalography and Clinical Neurophysiology*, 98(2):89–102.

Michalewski, H. J., Thompson, L. W., Patterson, J. V., Bowman, T. E., and Litzelman, D. (1980). Sex differences in the amplitudes and latencies of the human auditory brain stem potential. *Electroencephalography and Clinical Neurophysiology*, 48:351–356.

Mizrahi, E. M., Maulsby, R. L., and Frost, J. D. (1983). Improved wave v resolution by dual-channel brain stem auditory evoked potential recording. *Electroencephalography and Clinical Neurophysiology*, 48:351–356.

Moller, A. R., Jannetta, P., and Moller, M. B. (1981). Intracranially recorded responses from the human auditory nerve: new insights into the origin of brain stem evoked potentials (bseps). *Electroencephalography and Clinical Neurophysiology*, 52(1):18–27.

Moore, B. R. (1980). A modification of the rayleigh test for vector data. *Biometrika*, 67(1):175–180.

Nagaoka, S. and Amai, O. (1991). Estimation accuracy of close approach probability for establishing a radar separation minimum. *Journal of Navigation*, 44:110–121.

Neelon, M. F., Williams, J., and Garell, P. C. (2006a). The effects of attentional load on auditory erps recorded from human cortex. *Brain Research*, 1118:94–105.

Neelon, M. F., Williams, J., and Garell, P. C. (2006b). The effects of auditory attention measured from human electrocorticograms. *Clinical Neurophysiology*, 117(3):504–521.

Nocera, F. D. and Ferlazzo, F. (2000). Resampling approach to statistical inference: Bootstrapping from event-related potentials data. *Behavior Research Methods, Instruments, and Computers*, 32(1):111–119.

Ozdamar, O., Delgado, R. E., Eilers, R. E., and Urbano, R. C. (1994). Automated electro-physiologic hearing testing using a threshold-seeking algorithm. *Journal of the American Academy of Audiology*, 5:77–88.

Ozdamar, O., Delgado, R. E., Eilers, R. E., and Widen, J. E. (1990). Computer methods for on-line hearing testing with auditory brain stem responses. *Ear and Hearing*, 11(6):417–429.

Ozdamar, O. and Kalayci, T. (1999). Median averaging of auditory brain stem responses. *Ear and Hearing*, 20:253–264.

Ozdamar, O. and Kraus, N. (1983). Auditory brainstem responses in infants recovering from bacterial meningitis: Neurologic assessment. *Archives of Neurology*, 40:499–502.

Ozdamar, O., Kraus, N., and Stein, L. (1983). Auditory brainstem responses in infants recovering from bacterial meningitis: Audiological evaluation. *Archives of Otolaryngology*, 109:13–18.

Pantev, C. and Khvoles, R. (1984). Comparison of the efficiency of various criteria for artifact rejection in the recording of auditory brain-stem responses (abr). *Scandinavian Audiology*, 13:103–108.

Park, S. H., Goo, J. M., and Jo, C. H. (2004). Receiver operating characteristic (roc) curve: practical review for radiologists. *Korean Journal of Radiology*, 5:11–18.

Picton, T. W. and Hillyard, S. A. (1974). Human auditory evoked potentials: Ii. effects of attention. *Electroencephalography and Clinical Neurophysiology*, 36:191–199.

Picton, T. W. and Hillyard, S. A. (1976). Human auditory evoked potentials: Ii effect of attention. *Electroencephalography and Clinical Neurophysiology*, 40:418–426.

Picton, T. W., Vajsar, J., Rodriguez, R., and Kampbell, K. B. (1987). Reliability estimates for steady-state evoked potentials. *Electroencephalography and Clinical Neurophysiology*, 68:119–131.

Politis, D. N. and Romano, J. P. (1994). The stationary bootstrap. *Journal of the American Statistical Association*, 89(428):1303–1313.

Pool, K. D. and Finitzo, T. (1989). Evaluation of a computer-automated program for clinical assessment of the auditory brains tem response. *Ear and Hearing*, 10:304–310.

Ramos, E. G., Simpson, D. M., Panerai, R. B., Nadal, J., Lopes, J. M. A., and Evans, D. H. (2006). Objective selection of signals for assessment of cerebral blood flow autoregulation in neonates. *Physiological Measurement*, 27:35–49.

Robert, F. B. and Carrie, S. (2001). Overview of auditory evoked potentials. In Katz, J., editor, *Handbook of Clinical Audiology*, book chapter 14, pages 233–273. Lippincott Williams and Wilkins, fifth edition.

Rodriguez-Rivera, A., Baryshnikov, B. V., Veen, B. D. V., and Wakai, R. T. (2006). Meg and eeg source localization in beamspace. *IEEE Transactions on Biomedical Engineering*, 53(3):430–441.

Rossi, L., Bianchi, A. M., Merzagora, A., Gaggiani, A., Cerutti, S., and Branchi, F. (2007). Single trial somatosensory evoked potential extraction with

arx filtering for a combined spinal cord intraoperative neuromonitoring technique. *BioMedical Engineering OnLine*, 6(2):http://www.biomedical-engineering-online.com/content/6/1/2.

Rugg, M. D. (1995). *Electrophysiology of mind : event-related brain potentials and cognition.* Oxford : Oxford University Press.

Sanders, R. A., Duncan, P. G., and McCullough, D. W. (1979). Clinical experience with brain stem audiometry performed under general anesthesia. *Journal of Otolaryngology*, 8:31–38.

Schimmel, H., Rapin, I., and Cohen, M. M. (1974). Improving evoked response audiometry with special reference to the use of machine scoring. *Audiology*, 13:33–65.

Simpson, D. M., Panerai, R. B., and Ramos, E. G. (2004). Assessing blood flow control through a bootstrap method. *IEEE Transactions on Biomedical Engineering*, 51(7):1284–1286.

Simpson, D. M., Tierra-Criollo, C. J., Leite, R. T., Zayen, E. J. B., and Infantosi, A. F. C. (2000). Objective response detection in an electroencephalogram during somatosensory stimulation. *Annals of Biomedical Engineering*, 28:691–698.

Sokal, P. R. and Rohlf, F. J. (1995). *Biometry : the principles and practice of statistics in biological research.* W.H. Freeman, New York, 3rd edition.

Starr, A. and Achor, L. J. (1975). Auditory brain stem responses in neurological disease. *Archives of Neurology*, 32:761–768.

Starr, A., Amlie, R. N., Martin, W. H., and Sanders, S. (1977). Development of auditory function in newborn infants revealed by auditory brainstem potentials. *Pediatrics*, 60:831–839.

Starr, A. and Hamilton, A. E. (1976). Correlation between confirmed sites of neurological lesions and abnormalities of far-field auditory brainstem responses. *Electroencephalography and Clinical Neurophysiology*, 41:595–698.

Stegeman, D. F., Oosterom, A. V., and Colon, E. J. (1987). Far-field evoked potential components induced by a propagating generator: computational evidence. *Electroencephalography and Clinical Neurophysiology*, 67:176–187.

Stockard, J. J. and Rossiter, V. S. (1977). Clinical and pathologic correlates of brainstem auditory response abnormalities. *Neurology*, 27:316–325.

Stockard, J. J., Stockard, J. E., and Sharbrough, R. W. (1978). Nonpathologic factors influencing brainstem evoked potentials. *American Journal of EEG Technology*, 18:177–209.

Sturzebecher, E., Cebulla, M., and Wernecke, K. D. (1999). Objective response detection in the frequency domain: comparison of several q-sample tests. *Audiology and Neurotology*, 4:2–11.

Sturzebecher, E. and Cebullar, M. (1997). Objective detection of auditory evoked potentials. comparison of several statistical tests in the frequency domain on the basis of near-threshold abr data. *Scandinavian Audiology*, 26:7–14.

Tauxe, L., Kylstra, N., and Constable, C. (1991). Bootstrap statistics for paleomagnetic data. *Journal of Geophysical Research*, 96:11723–11740.

Tibshirani, R. (1996). A comparison of some error estimates for neural network models. *Neural Computation*, 8:152–163.

Valdes, J. L., Perez-Abalo, M. C., Martin, V., Savio, G., Sierra, C., Rodriguez, E., and Lins, O. (1997). Comparison of statistical indicators for the automatic detection of 80 hz auditory steady state responses. *Ear and Hearing*, 18:420–429.

Valds-Sosa, M. J., Bobes, M. A., Perez-Abalo, M. C., Perera, M., Carballo, J. A., and Valdes-Sosa, P. (1987). Comparison of auditory-evoked potential detection methods using signal detection theory. *Audiology*, 26:166–178.

van Straaten, H. (1999). Automated auditory brainstem response in neonatal hearing screening. *Acta Paediatra Supplement*, 432:76–79.

Victor, J. D. and Mast, J. (1991). A new statistic for steady-state evoked potentials. *Electroencephalography and Clinical Neurophysiology*, 78:378–388.

Vidler, M. and Parker, D. (2004). Auditory brainstem response threshold estimation: subjective threshold estimation by experienced clinicians in a computer simulation of the clinical test. *International Journal of Audiology*, 43:417–429.

Viera, A. J. and Garrett, J. M. (2005). Understanding interobserver agreement: the kappa statistic. *Family Medicine*, 37(5):360–363.

Weber, B. A. and Fletcher, G. L. (1980). A computerized scoring procedure for auditory brainstem response audiometry. *Ear and Hearing*, 1:233–236.

Whitcher, B., Craigmile, P. F., and Brown, P. (2005). Time-varying spectral analysis in neurophysiological time series using hilbert wavelet pairs. *Signal Processing*, 85(11):2065–2081.

Wong, P. K. H. and Bickford, R. G. (1980). Brain stem auditory evoked potentials: the use of noise estimate. *Electroencephalography and Clinical Neurophysiology*, 50:25–34.

Zoubir, A. M. and Boashash, B. (1998). The bootstrap and its application in signal processing. *IEEE Signal Processing Magazine*, 15(1):56–76.

Zoubir, A. M. and Bohme, J. F. (1995). Multiple bootstrap tests: an application to sensor location. *Ieee Transactions on Signal Processing*, 43:1386–1396.

Zurek, P. M. (1992). Detectability of transient and sinusoidal otoacoustic emissions. *Ear and Hearing*, 13:307–310.

Zygierewicz, J., Durka, P. J., Klekowicz, H., Franaszczuk, P. J., and Crone, N. E. (2005). Computationally efficient approaches to calculating significant erd/ers changes in the time-frequency plane. *Journal of Neuroscience Methods*, 145(1-2):267–276.