Aalto University
School of Science and Technology
Faculty of Information and Natural Sciences
Degree Programme of Bioinformation Technology

Anttu Kurttio

Development of an Adaptive Algorithm for Online Artefact Rejection in Electroencephalographic Recordings

Master's Thesis
Espoo, January 11, 2010

Supervisor: Professor Kimmo Kaski
Instructor: Antti Paukkunen, M.Sc.

Aalto-yliopisto Teknillinen korkeakoulu Informaatio- ja luonnontieteiden tiedekunta Bioinformaatioteknologian tutkinto-ohjelma		DIPLOMITYÖN TIIVISTELMÄ
Tekijä: Anttu Kurttio		
Työn nimi: Mukautuvan häiriönpoistoalgoritmin kehitys reaaliaikaisia aivosähkökäyrämittauksia varten		
Sivumäärä: 36	Päiväys: 11.1. 2010	Julkaisukieli: Englanti
Professuuri: Laskennallinen tiede		Professuurikoodi: S-114
Työn valvoja: Professori Kimmo Kaski		
Työn ohjaaja: DI Antti Paukkunen		

Tiivistelmä:

Työn tavoitteena oli kehittää algoritmi aivosähkökäyrän häiriöiden reaaliaikaiseen poistamiseen. Työ oli osa uuden laitteen kehitysprojektia, jossa pyritään vähentämään tietyntyyppisiin aivosähkökäyrämittauksiin kuluvaa aikaa ja helpottamaan mittausten suorittamista. Mittaukset tehtiin laitteen kahdeksankanavaisella prototyypillä.

Artefaktojen ominaispiirteet määritettiin kokeellisesti. Tärkeimmiksi häiriölähteiksi todettiin silmien räpäytykset, silmien liikkeet, pään liikuttaminen sekä purenta. Ensisijaisesti häiriöiden tunnistamisessa käytettiin laskennallisesti kevyitä virtuaalikanavamenetelmiä, jotka hyödynsivät havaittuja piirteitä. Menetelmiä kehitettiin edelleen useiden koemittausten avulla. Myöhemmissä versioissa algoritmi saatiin mukautumaan erilaisiin mittaustilanteisiin ja muutoksiin mittauksen kuluessa.

Lopullinen algoritmi on huomattavasti tehokkaampi ja luotettavampi kuin aiemmin käytetyt reaaliaikaiset menetelmät. Aiemmat menetelmät ovat perustuneet yksittäiseen raja-arvoon ja niiden hylkäysprosentit ovat korkeintaan 80% käytettäessä samoja kriteereitä kuin tässä työssä. Viimeisimmissä suorituskykykokeissa algoritmi tunnisti ja hylkäsi noin 99% artefaktoista ja hylkäyksistä yli 98% oli oikeaan osuneita. Kokeessa käytettiin useita koehenkilöitä ja mittaustilanne oli mahdollisimman tarkasti laitteen todellista käyttötilannetta jäljittelevä. Tämä osoittaa, että algoritmi on erittäin tehokas ja pystyy mukautumaan sopivaksi kullekin koehenkilölle normaaleissa mittaustilanteissa.

Lopullisessa muodossaan kahdeksankanavainen algoritmi soveltuu mainiosti projektissa kehitettävän laitteen häiriönpoistoalgoritmiksi. Se on tehokas, luotettava ja laskennallisesti verraten kevyt. Mikäli laitteesta kehitetään jatkossa versio, jossa häiriönpoisto tapahtuu sulautetulla prosessorilla, on kehitetty algoritmi varteenotettava ehdokas toteutukseksi. Myös muunlaiset aivosähkökäyrälaitteet ovat potentiaalisia sovelluskohteita algoritmille, sillä häiriönpoisto on eräs niiden yleisimmistä heikkouksista.

Asiasanat: aivosähkökäyrä, artefakta, EEG, ERP, häiriönpoisto, reaaliaikainen,

Aalto University
School of Science and Technology

Aalto University School of Science and Technology Faculty of Information and Natural Sciences Degree Programme of Bioinformation Technology	ABSTRACT OF THE MASTER'S THESIS	
Author: Anttu Kurttio		
Title: Development of an Adaptive Algorithm for Online Artefact Rejection in Electroencephalographic Recordings		
Number of pages: 36	Date: January 11, 2010	Language: English
Professorship: Computational Science	Code: S-114	
Supervisor: Professor Kimmo Kaski		
Instructor: Antti Paukkunen, M.Sc.		

Abstract:

The purpose of the work was to develop an online algorithm for electroencephalograph (EEG) artefact removal. The work was part of a project developing a novel device for easier and faster recording of event related potentials (ERPs). A prototype of the device was used in the recordings involved in the development of the algorithm.

The properties of the artefacts were studied experimentally. Most important artefact sources turned out to be blinks, eye movements, head movements, and jaw muscle activations. The primary methods used in artefact detection were several virtual channel methods that are computationally light and take advantage of the experimentally determined properties. Several developments were made to the methods with the aid of further experimental data. In later versions adaptive features were introduced to the algorithm, allowing it to adjust to changes in measurement conditions without outside interruption.

The final version of the algorithm is more powerful and robust than other online solutions. Earlier solutions have relied on a single potential threshold and have reached only 80% accuracy at best when assessed using the same criteria as the algorithm presented here. In the latest performance tests the algorithm detected and rejected approximately 99% of all artefacts, with over 98% of the rejections being correct. Several test subjects were used in the tests and the recording set-up closely mimicked the set-up where such a device would be used in reality. The tests prove that the algorithm is very powerful and can adapt to different subjects under ordinary but not necessarily identical conditions.

In the final version presented in this work the eight channel algorithm is well suited to remove the artefacts present in the data measured by the device. It is powerful, reliable, and efficient compared to the alternatives. If the device is developed to include an embedded processor for artefact rejection the algorithm is a good candidate for implementation. The algorithm could also be of use in other EEG applications after some minor modifications, because artefact detection is one of the most common weaknesses of the devices.

Keywords: artefact, artefact removal, electroencephalogram, EEG, ERP, online, real-time

Preface

Condensed on these pages is the work of six months. Six months is not that much in the scope of life, so before I get carried away I would like to thank my parents who have supported me for over fifty times that long. Thank you for letting me grow up asking questions and finding answers. You have influenced my thinking more than I can ever know and I am grateful for that. I believe you have done an admirable job equipping me for an enjoyable life. I would also like to thank my little brother who has grown to be a fierce adversary, but only in the most positive sense of the expression. Name a game and I have most likely played it against you and enjoyed the challenge. What better way is there to free the mind of the troubles of work, especially this thesis work?

There have been several difficult moments and also many moments of eureka when everything just seems to fall neatly into place. When I first noticed the open thesis work position I immediately thought that it was a great coincidence. I needed a summer job where I could write my thesis and they needed someone to work on EEG artefacts. I had just held a presentation on the subject during a seminar course, which obviously helped quite a bit during the job interview. I would like to thank Professor Raimo Sepponen for opening the coffers and hiring me to do the work that I now write of. I would also like to thank Antti Paukkunen for immediately believing that I was the right person for the job at hand. I hope and trust you are not disappointed. Consequently, I'd like to apologise to Professor Kimmo Kaski. Sorry that you had delay the start of your summer holiday past 4 pm when I rushed in with the application for the thesis subject at the last minute. And thank you for agreeing to be my supervisor.

The work progressed smoothly, mostly thanks to Antti, who also acted as my instructor and knew well what he wanted me to do. We had a working algorithm in two weeks and by the end of summer it had gone through several developments, most of them surprisingly effective in hindsight. I started writing the thesis in the early September and it proved as challenging as could be expected. During this time I made some more small adjustments to the algorithms, but most of the coding and experimenting was already done when I started to write things down. I have not been able to track how many rewrites, revisions, and smaller fixes the thesis has gone through, but the numbers are not insignificant. This process has made the content and presentation much better than what I could have produced on my own. Thank you, Antti for your most valuable advice.

My girlfriend Hanna is to blame for some of the many sentences ending abruptly, some of which may have survived to the final version. Thank you for pulling me off the computer and asking me for a walk all those times I had been staring idly at the screen for ten minutes. Also, thank you for your understanding when I have not been able to come with you because of one deadline or another. With your help I have been able to live and work 600km apart and still get both done – living a real life and sometimes writing almost around the clock. Hanna, without you I would have probably been a wreck before this work had been finished. Thank you for your unyielding support.

Table of contents

Preface....4
Table of contents....5
List of abbreviations and acronyms....7
Glossary and disambiguation....8
1. Introduction....9
  1.1. Background....9
  1.2. Ongoing research project....10
  1.3. Research objectives....10
2. Preparation of the data and post-processing....12
  2.1. Baseline offset....12
  2.2. Baseline correction methods....12
  2.3. Rapid baseline shifts....13
  2.4. Noise removal....13
  2.5. The Daubechies 4 wavelet....13
  2.6. Multilevel wavelet decomposition and reconstruction....14
  2.7. Adaptive hard thresholding....14
3. Artefact characteristics....15
  3.1. Experiment to determine artefact properties....15
  3.2. Ocular artefacts....15
    3.2.1. Blinking....15
    3.2.2. Horizontal saccades....15
    3.2.3. Vertical saccades....15
  3.3. Myographic artefacts....16
    3.3.1. Biting....16
    3.3.2. Other head and neck muscle artefacts....16
    3.3.3. Heartbeat and other muscle-based activity in other parts of the body....16
  3.4. Equipment-related artefacts....18
4. Artefact Detection....19
  4.1. Eye blinks....19
    4.1.1. Simple virtual channel threshold....19
    4.1.2. "Difference of long and short averages" – method....19
  4.2. Horizontal saccades....20
    4.2.1. "Squared difference and threshold" – method....20
    4.2.2. "Smoothed absolute value of difference and threshold" – method....20
  4.3. Bites....20
    4.3.1. Simple mastoid variance threshold....21
    4.3.2. Adaptive threshold method....21
  4.4. Other muscle activity....21
    4.4.1. Simple dipole magnitude threshold....21
    4.4.2. "Smoothed virtual channel absolute value and threshold" – method....21
    4.4.3. Adaptive muscle artefact detection....22
  4.5. Rapid baseline shift detection....22
  4.6. Performance testing....23
  4.7. Developmental aspects....25
5. Further developments....27
  5.1. Improvement of reliability....27
    5.1.1. Improved blink detection method....27
    5.1.2. Improved saccade detection method....27

5.1.3. Improved bite detection method....27
5.1.4. Other improvements....28
5.2. Improvement of efficiency....28
4.3. Performance testing....30
5.4. On optimization....32
6. Summary and conclusions....33
7. References....35

List of abbreviations and acronyms

CT	Computer tomography
EEG	Electroencephalogram
ERP	Event related potential
fMRI	Functional magnetic resonance imaging
ICA	Independent component analysis
MEG	Magnetoencephalogram
MMN	Mismatch negativity
PCA	Principal component analysis

Glossary and disambiguation

Algorithm	A combination of methods and other procedures that is designed to clean EEG recordings of artefacts and noise.
Artefact	Anything besides brain activity that influences the electric potentials recorded by the device.
Artefactual	A sample that is recorded whenever an artefact is present is artefactual.
Benchmark	Accuracy, precision, recall, sensitivity or specificity. Different measures of how well a method or an algorithm has removed the artefacts.
Bite	Jaw muscle activity. Does not necessarily cause any movement, but if it can be registered on the electrodes it is a bite.
Detector	A synonym for a method.
Method	A detector for a single artefact type.
Noise	Background brain activity that influences the electrode potentials. Can also contain some electronic noise due to the electrode connections. Both sources have wide enough frequency bands and reasonably uniform power spectra, so noise can be considered approximately white.
Recording	An EEG recording session or the resulting data set.
Saccade	A rapid eye movement. Unless otherwise specified a horizontal saccade. One of the artefact types.
Sample	Potentials recorded during a single timestep. A sample contains a single potential value from each channel.
Thresholding	Repeatedly comparing a value that changes during a recording to a threshold.
Timestep	Time between recorded samples. The inverse of the sampling frequency. The device used has a timestep of 5ms.
Virtual channel	A re-referenced and mathematically modified EEG channel.

1. Introduction

1.1. Background

Measuring ‘brainwaves’ is a feat that people believe is easy for any modern day practitioner of medicine. In its current state, however, electroencephalogram (EEG) is being slowly replaced by the magnetoencephalogram (MEG), functional magnetic resonance imaging (fMRI), and computer tomography (CT) [17, 18]. These techniques provide high resolution images of the brain instead of the handful of one-dimensional signals that EEG records. There are downsides too – the devices are expensive, they take a lot of space, special expertise is needed to operate them, and in the case of CT the subject also receives a large x-ray dose [19]. Even if costs are ignored the dimensions of the data are not the only reason why the popularity of EEG is diminishing – if EEG could pinpoint the location of a tumour [20] or reliably and quickly tell why a patient is unconscious it would be used a lot more than today. The main problem with EEG is the unreliability of the recording [1-3].

There are two main shortcomings in EEG. First, the signal source is impossible to locate with certainty. Second, to make things worse even that signal is entirely drowned in noise. Because EEG is measured on the scalp and the brain is electrically active in all parts it is impossible to reliably pinpoint the sources of the measured potentials. More importantly, it is impossible to find an explicit solution to a problem where there can be several active sources inside a volume with mostly uniform conductivity (The inverse source problem [8, 9]). Based on these problems, it would be easy to think that EEG is not worth it these days when more accurate tests are available, but there are ways to solve these problems and when used correctly EEG can still be invaluable because it is a cheap and easy way to peer inside the working brain.

Event related potential (ERP) is a characteristic EEG waveform that is the result of the brain processing a given stimulus [4]. They are useful because they provide a way to reduce the background noise that is a big problem in all EEG use. Measuring the response of one group of neurons to a stimulus is like trying to ask a question from someone standing on the other side of a football field – when the field is full of yelling and screaming people all trying to get their own messages across to each other. Fortunately it is known that, in a sense, the brain can be quite simple at times and always answers almost identically to a simple question. So to continue with the analogy, it is possible to ask the question a hundred times and record the response plus some random noise each time. Then all the recordings are summed and the answer is heard clearly enough when it is amplified hundred-fold and the other voices are not.

ERPs have been used and studied extensively and one successful example is Mismatch negativity (MMN). MMN is elicited when the brain detects a change in the incoming stimulus [7, 12]. When the brain expects a standard stimulus and receives a deviant instead the response is very different even if the difference between the two is barely recognizable. Most research into MMN phenomena has used auditory stimuli. This is understandable, because it is easy to alter an auditory stimulus in many subtle ways (duration, pitch, volume, interval etc.), the test subject can quite reliably either focus on or leave unattended auditory stimuli (with the help of visual stimuli to capture attention), and the stimulus rate can be kept high enough to keep the test situation short. MMN is also used in language studies, where the stimuli are words or syllables [5]. The clinical uses of MMN include detecting speech perception problems in infants, predicting coma outcome, and assessing the state of schizophrenics [6].

Although the noise can be handled pretty well in ERP studies through averaging, sudden artefacts [21] can ruin the results if they are not removed. Artefacts are hard to avoid and their impact on the average waveform is significant. Averaging can be used to some extent to reduce artefact impact, but eventually the data quality without artefact removal reaches a maximum and that maximum is not very high assuming normal artefact density. This is why artefact removal in EEG recordings is an area of interest for modern researchers – as it has been for the last fifty years.

1.2. Ongoing research project

This master's thesis was made as a part of a project which aims to improve the clinical applicability of ERP recordings through the development of the instrumentation and means of adaptive signal processing. Currently, a prototype system is has been developed and the signal processing platform is under development.

The prototype system developed in the project uses eight channel electrodes and the sampling frequency of 200Hz, which means that activity up to 100Hz can be accurately measured (Nyquist theorem [11]). The data are provided as a real-time data stream that is transmitted through a wireless connection to the analysis computer. In a typical recording set-up, the signals are recorded from Right mastoid (upper neck, about five cm below inion and to the right), right temporal near location F8, right frontal Fp2, forehead Fz, central Cz, left frontal Fp1, left temporal F7, and left mastoid (see figure 1).

1.3. Research objectives

The purpose of this thesis was to develop an algorithm to reject artefacts in real-time during the recording of EEG. The first research objective was to design a reasonable online algorithm that can process the data while they are collected. Second objective was to achieve adaptivity, making the algorithm respond to changing measurement conditions.

The algorithm cannot be very processor intensive as it has to be feasible online. Most notably any methods incorporating principal component analysis (PCA) or independent component analysis (ICA) are too heavy for the device. Fortunately, most of the simple artefact detection and noise removal methods do not require a lot of processor time. It is therefore possible to simultaneously use many of the simpler methods at each time step even on an embedded processor if that is necessary.

The main constraint of the algorithms presented in this work is to process the raw data faster than the samples are acquired. This means that the algorithms must be able to process over 200 samples per second on average. If the development of such algorithms was successful they could be of use in other similar applications as well.

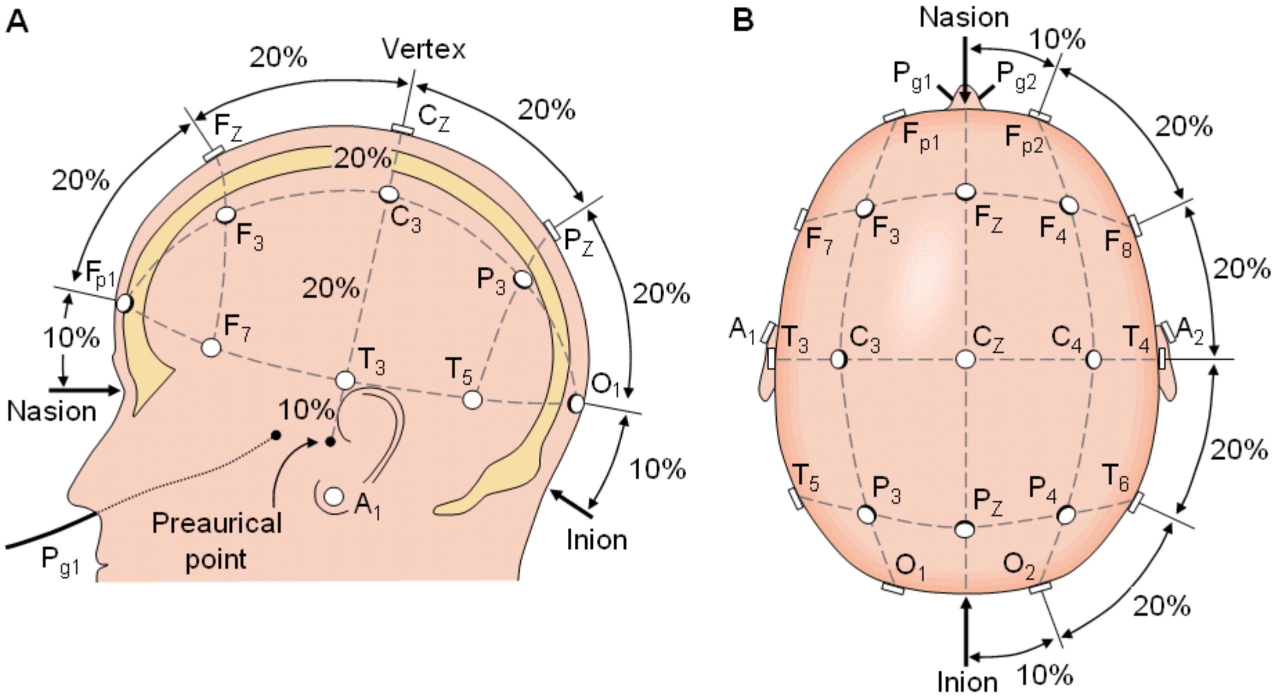

Figure 1: The international 10-20 electrode placement system

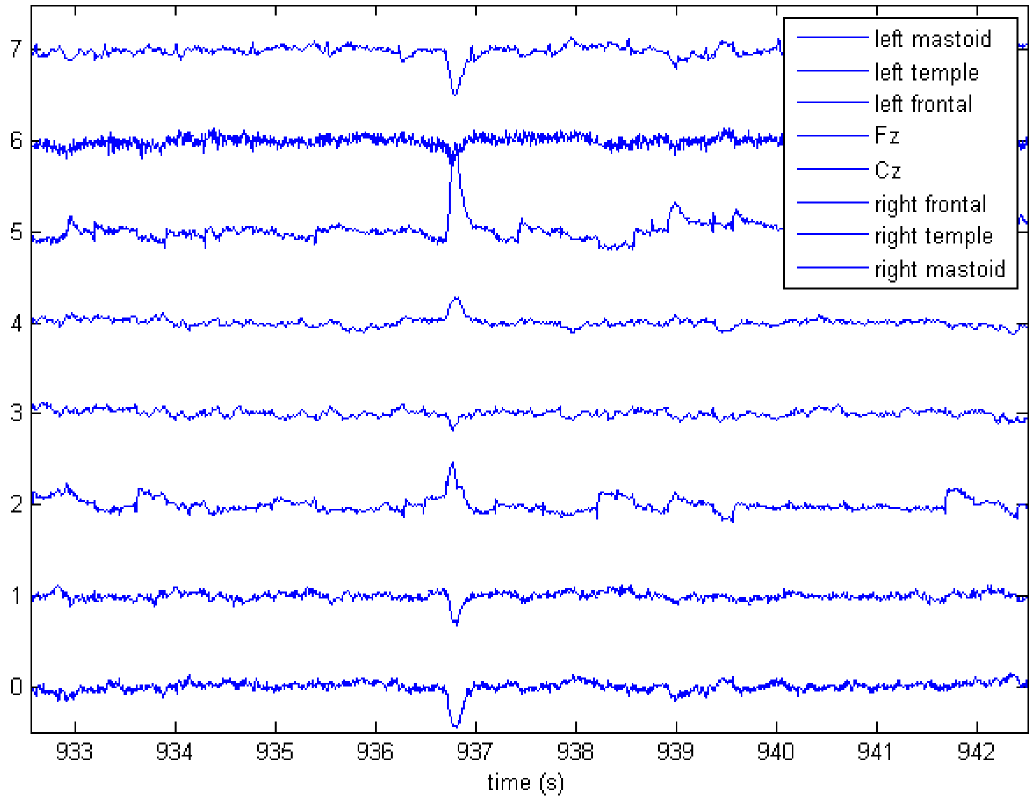

Figure 2: A characteristic blink waveform seen on the eight electrodes. The frontal channels also show some vertical saccades mainly in the 938-939s interval.

2. Preparation of the data and post-processing

Before the real task of artefact detection there are some preliminary tasks to be taken care of. Most importantly the meandering baseline must be corrected. After the data have been recorded the noise must be removed. There are several methods to remove the residual noise left after the averaging. In this case, wavelet filtering is used.

2.1. Baseline offset

Sometimes all channels are not originally normalized to have zero mean, which is not optimal. It is necessary to remove the baseline offset for wavelet filtering purposes as well. Removing the baseline can be as simple as removing the mean of the latest window of samples or using a highpass filter with a passband starting from below 1Hz. Different methods have different drawbacks and strengths and the right way to remove the baseline effectively without influencing the data is not a trivial choice.

2.2. Baseline correction methods

Removing the baseline offset is one of the first things that are done in almost any kind of signal processing. It makes analysing the data using many mathematical tools possible. It also removes one variable from the differences between recordings – an important step because it is necessary to be able to compare test subjects for the test to have any applicability.

One very basic method for baseline removal is to take the average of a moving window. In this case due to the needs of the online artefact rejection the length of the window is limited to one hundred samples. The mean for each channel is calculated from the samples and then subtracted from the 50^{th} sample that is currently being processed. This method very effectively removes the baseline, because if the baseline changes rapidly the new baseline is completely removed after just 100 samples. On the other hand many interesting or important phenomena can be lost as well as the baseline removal removes low frequency events as well. For example if the subject looks to the side and keeps looking there for more than half a second the change will no longer be visible on the baseline-corrected data. This can be either advantageous or not depending on how the baseline corrected data are used.

For continuous EEG it would be reasonable to use a longer average than the one hundred sample segment mean. In online analysis a longer window is not often viable as it slows down the data processing. Because the baseline changes are usually quite slow the current baseline can be approximated rather well using only the samples preceding the sample that is being processed and extrapolating. This way the baseline variance is smaller, but the method reacts more slowly to changes in the mean. The number of samples used with the method was empirically set to ten thousand, which corresponds to fifty seconds, so during a 15 minute recording where there is one rapid baseline shift and several slower changes the baseline correction works well a little over 14 minutes and fails slightly or more for about fifty seconds. This method is used by the algorithms when the recorded data is not divided into epochs.

When the data are divided into epochs that are to be averaged the usual way to handle baseline correction is to take the average of the prestimulus interval – the time before the stimulus is presented to the subject. In this case the prestimulus interval is 42 samples or 210ms long so the average is a slightly less accurate approximation of the true baseline than the one hundred sample average. There should be no relevant activity during the prestimulus interval, however, so the signal consists of baseline and noise only. This can be a very important detail when the signal to noise ratio (SNR) is calculated. Also, if the prestimulus interval consistently contains some waveform the experimental set-up should be revised, possibly moving the stimuli further apart to avoid overlap from the previous evoked potential and the prestimulus interval of the next. This method is obviously unusable if the experiment does not use stimuli or cannot accurately time them. In ERP recordings prestimulus interval average is the most common method of baseline removal. This method is used by the algorithms when the data are divided into epochs.

It is sometimes viable to use a butterworth filter [15] or some other non-ideal highpass filter [16] for baseline correction. The advantages of the method are the simplicity and universally adopted nature of the IIR filtering methods. At its best a butterworth filter will cause minimal passband ripple and remove the baseline equally from all segments where it is applied. It will not affect any high frequency details if properly used, but slow waves can be slightly deformed by the filter.

2.3. Rapid baseline shifts

If an electrode moves a little it might change the conductivity and therefore cause a rapid change in the potential. When this happens it should be detected and the baseline should adjust accordingly. This is one of the artefacts that can cause the most trouble if it is not rejected, because if an epoch where one channel is baseline shifted a lot moves the resulting average by a significant amount.

2.4. Noise removal

Once the data have been cleaned of the most obvious artefacts the background noise still needs to be minimized without influencing the shape of the responses too much. Wavelet filtering was chosen as the method of choice, because simple averaging was not effective enough in preserving the waveforms [14]. The accepted epochs are first averaged, and the resulting waveforms are wavelet filtered to reduce the noise even more. If the filtering would be done for each epoch before averaging it would alter the shape of the response more, because there would be more noise in the filtered sequences and the more noise there is the less reliable the filtered results are.

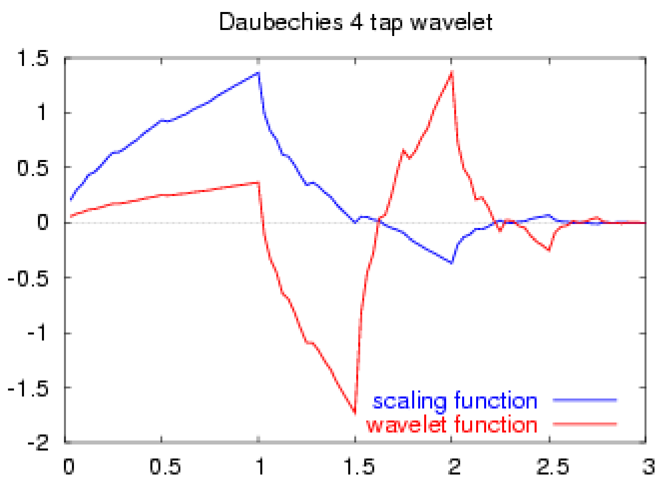

Figure 3: The shape of the wavelet function and scaling function of a Daubechies 4 wavelet.

2.5. The Daubechies 4 wavelet

The wavelet function used in the noise removal method is Daubechies 4, which is a generic wavelet reasonably shaped for ERP filtering. It retains most of the low and medium frequency components of the signal, preserves the peak shape of the waveforms rather well and does not reduce the amplitude. Figure 3 shows the shapes of the wavelet function that covers most of the spectrum and the scaling function used to cover the lowest frequencies.

2.6. Multilevel wavelet decomposition and reconstruction

Multilevel wavelet decomposition divides the data into frequency components. The lowest frequency component is called the approximation while the rest are called detail levels from highest to lowest. In this case when 200Hz data are decomposed using 8-level wavelet decomposition the approximation consists of the 0-2Hz data, the 7th detail level contains the 2-4Hz band, the 6th detail level the 4-8Hz band, and so on until the 1st detail level, which spans from 128Hz to 256Hz, which clearly cannot contain any reliable information because of aliasing effects that make it impossible to accurately detect any activity above 100Hz when using a 200Hz sampling rate.

After discarding some of these components the remaining detail levels and the approximation level are used in the reconstruction, which is simply the inverse process of the decomposition, substituting zero-vectors for the discarded frequency components. This method relies on predetermined thresholds for accepting the coefficients for each detail level. If the thresholds are too low most of the noise is left in the data, but on the other hand if the thresholds are too high most of the information is lost and only the occasional artefactual spike is included in the filtered data – far from the ideal outcome. Of course the highest frequency detail levels can be rejected outright, because even if there were some non-artefactual events at those frequencies they would not be interesting as the MMN phenomena are low to medium frequency events.

2.7. Adaptive hard thresholding

The technique where wavelet coefficients are either accepted as is, or rejected completely is called hard thresholding. In adaptive hard thresholding the threshold is based on some measure of the data or the coefficients and it is periodically recalculated. The advantages of this method are many, as it can, for example, allow all of the approximation and detail levels 7 and 6 to pass, let through the above-median half of each intermediate detail levels coefficients, and reject the rest. This way no matter how noisy the raw data are the filtered data contain about the same amount of each frequency band and different recordings are easier to compare.

Table 1: Overview of the observed artefact characteristics

Artefact type	**Highest amplitude channels**	**Other information**
Blink	Frontal positive, mastoid negative	Fast, uniformly shaped
Horizontal saccade	Temporal, depends on direction	
Bite	Mastoid, temporal	Duration and amplitude vary a lot
Muscle	Depends on the active muscle, most often mastoid, temporal, and frontal	Duration and amplitude vary a lot
Tongue movement		Not noticeable
Opening and closing the mouth		Not noticeable
Turning the head	As a large muscle artefact and several saccades	Eyes seek focal points during the head turn, causing the saccades

3. Artefact characteristics

There are plenty of different artefacts in an EEG recording, even when everything is well prepared. Finding and removing these artefacts is a challenging task that is started by identifying the different types of artefacts.

3.1. Experiment to determine artefact properties

An experiment was conducted in order to have some first-hand information of the artefacts. A test subject repeated a sequence of opening and closing his mouth, biting, blinking, looking from side to side, and moving his tongue up and down. The resulting EEG recording was carefully analysed and each artefact was located if possible. These labelled data were later used as a test data set for the artefact detection methods. Roughly 35% of the samples were artefact-free, which means the artefact density was much higher than in an ordinary recording. In addition to density, also the artefact strength differs from normal. Intentionally caused artefacts are generally stronger than naturally occurring artefacts, because it is difficult to mimic the unconscious small movements that normally contaminate the recordings. These factors make it a little harder to use the experimental data in designing the artefact removal methods, but it is still worth it. Table 1 lists the different artefact types and the information about them gained through the experiment.

3.2. Ocular artefacts

Artefacts related to eye movements can be very large and thereby cause large errors in the averaging process if they are not caught by the filters. The eye movements are also physiologically very fast, although they cannot compete with the electrical artefacts that are understandably even faster.

3.2.1. Blinking

Eye blinks affect the electrode potentials, because the eye is an electric dipole and closing the eyelid forms a conductive pathway from the positive cornea to the forehead [13]. A blink is fast, taking approximately 120-200ms, and there is generally relatively little variability in the artefact waveform even between subjects. A blink shows as a large v-shaped dip on channels below and behind the eye and inversely a positive artefact on the forehead channels. Figure 2 shows a blink surrounded by some non-artefactual activity on all eight channels.

3.2.2. Horizontal saccades

A horizontal saccade (rapid eye movement) causes a symmetric change to the potentials on the left and right hemisphere electrodes. If a subject looks to the left all the left-hand-side electrodes show a rise in potential and all the right-hand-side electrodes show a decline of the same magnitude. The changes are at their largest on the temporal electrodes, because they are closest to the eyes in the direction of movement of the eye dipole. See figure 5 for an example of a horizontal saccade as it is seen on different channels.

3.2.3. Vertical saccades

A vertical saccade is a lot harder to catch than horizontal, because there are no symmetric electrodes above and below the eye to provide an easy way to check if there are inverse changes on different sides. To make things worse a quick glance up can seem almost like a blink on the electrodes, but the duration varies a lot more and the shape of the waveform is far from standardized. Fortunately, vertical saccades are rare and usually very small compared to horizontal – people do not look up or down as often as they glance to the sides. Some vertical saccades can be seen on figure 2.

3.3. Myographic artefacts

Muscle-related artefacts are usually characterized by large amplitude and variance. They do not have as clear waveform shapes or as clearly defined directions as ocular artefacts, but in some cases they can be easier to find and remove because they can be orders of magnitude larger in amplitude than normal brain activity.

3.3.1. Biting

The jaw muscles are some of the strongest muscles in the human body and when they are active it really shows. A moderate bite will increase the activity on all channels tenfold as the muscle-related electric activity drowns the brain activity for a moment. Fortunately this makes large bite-artefacts easy to spot as the variance can be measured and used as an indicator. Unfortunately it is hard to draw the line where small increases in amplitude become artefactual as it is obviously possible to contract the jaw muscles ”just a little” instead of actually biting – for example, if the subject is tense or intensely concentrating on the test he or she can subconsciously bite enough to double the noise level on mastoid channels even if there are no outward signs of tension. This way the muscle activity can be even smaller than the background brain activity and thus not as simple to detect. Figure 4 shows a strong bite that continues to disrupt the potentials for over ten seconds after the subject has seemingly relaxed. It is important to take into account in artefact removal the time it takes for the channel baselines to readjust.

3.3.2. Other head and neck muscle artefacts

The tongue is a large muscle relatively close to the recording and reference electrodes. It would be easy to assume that speaking or otherwise moving the tongue would cause large artefacts. This, however, is not the case. Moving the tongue while other muscles are relaxed does not show markedly on the electrodes. Neck muscles, on the other hand, are directly underneath the mastoid electrodes so moving the head in almost any way always increases mastoid variance and can cause other artefacts as well. Figure 6 shows a small muscle artefact that also shows up on the Fz electrode.

3.3.3. Heartbeat and other muscle-based activity in other parts of the body

Heart is a constantly active strong muscle that causes a strong enough electric potential on the skin to be measured – the EKG is an old and well established clinical tool. It is therefore reasonable to assume that the heart causes changes in scalp potential as well, which it does. The choice of the reference electrode is the key here. If the reference electrode was placed on the wrist, ankle or chest the distance and direction to heart would be quite different than the distance and direction from the recording electrodes to heart. This would cause the heartbeat to show on the channels, similar to an EKG waveform. When the reference electrode is close to the recording electrodes the heart artefact is minimized. The same logic applies to all other distant muscle activity. They do cause potential changes, but the changes to the reference electrode are similar and thus the overall effect is minimal. If, at some point, the device is developed further and the noise levels can be reduced significantly, or additional processor power becomes available so the blind source separation techniques can be implemented these artefacts can become visible and thereby removable.

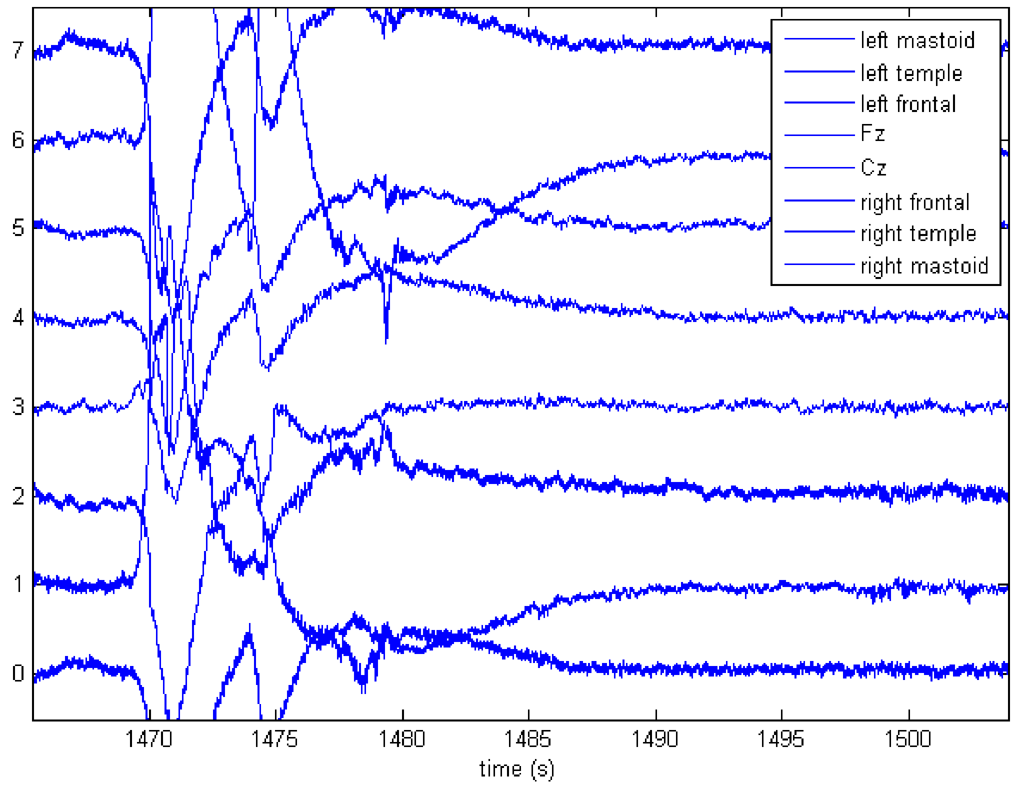

Figure 4: A strong bite. All channels are heavily affected.

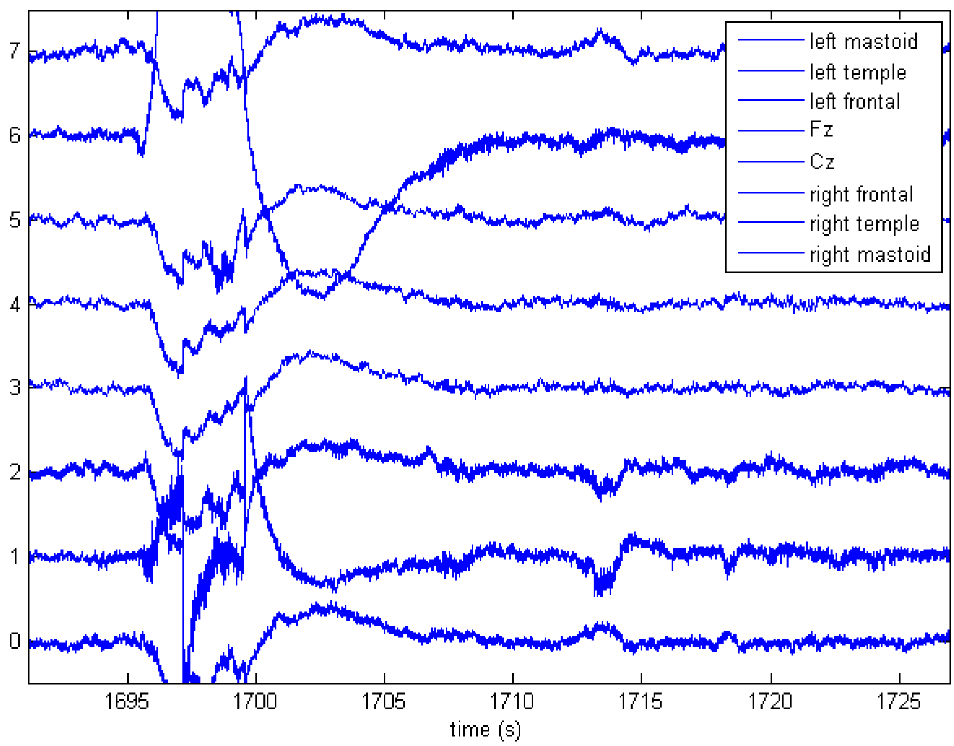

Figure 5: Two horizontal saccades in opposite directions during a bite. The temporal channels show sharp changes where the other channels are not as heavily affected by the horizontal movement.

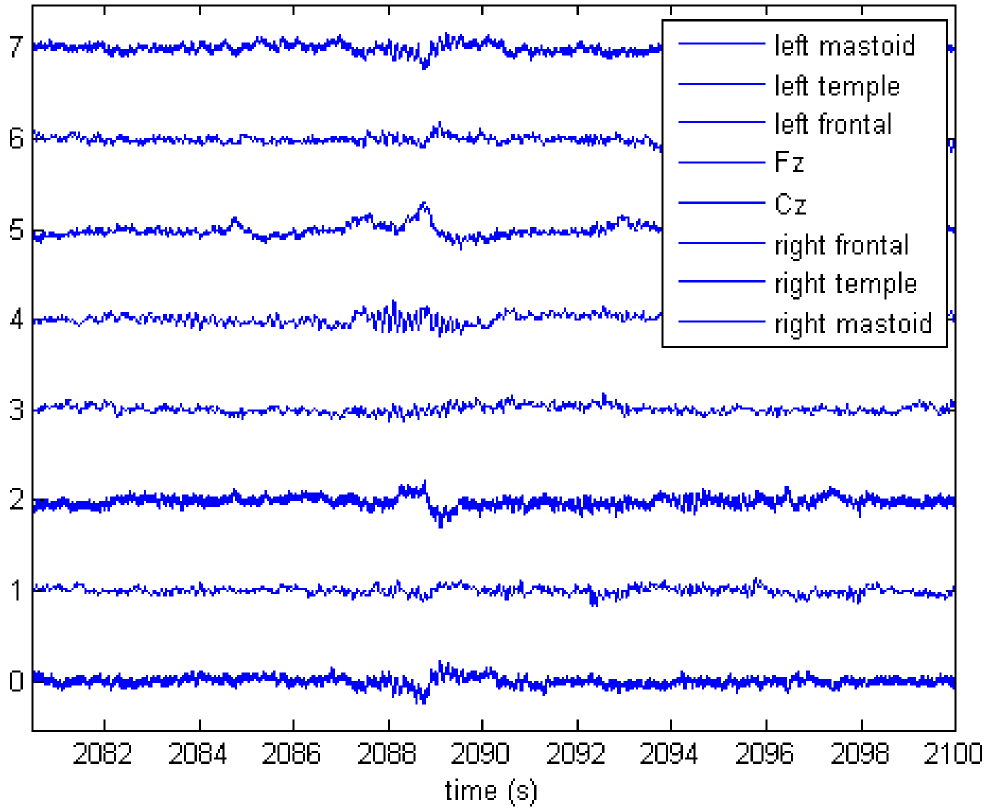

Figure 6: Some small scalp muscle activity during an otherwise clean epoch. It is evident that most of the activity is seen on Fz and at the mastoids. This might indicate a frown or raised eyebrows.

3.4. Equipment-related artefacts

There are many things that can go wrong in an EEG recording. Applying the electrodes can be difficult and they will invariably have different impedances even though the differences are kept at a minimum. To improve the connections varying amounts of a conductive electrode paste is used. The paste is an excellent conductor when it is moist, but it is impossible to keep it from drying, at least to some degree. This can go so far as to detach an electrode even though it has not moved at all. It is of course possible that electrodes move a little when the subject moves and even a small movement can affect the electrode connection greatly. This is one of the reasons patients are asked to remain as still as possible during an EEG experiment. The only thing that can be done if an electrode is completely detached is to reattach the electrode and restart the experiment. If the electrode paste dries significantly the amount of noise on a channel will increase, which must be taken into account if the experiment takes time.

A spike artefact is another type of artefact that can be caused by electrode movement, but also by faulty components or other electronics-related reasons. A spike is a very short, very high amplitude disturbance. They are often easy to remove using a simple amplitude threshold to reject samples where the amplitude goes over the threshold.

4. Artefact Detection

This chapter presents the methods used in the early versions of the algorithm to detect and remove artefacts from EEG data in real-time. The starting point was a technique called amplitude threshold, which means rejecting the samples where the absolute value of the potential for any channel is above a predetermined threshold – for example 75µV. This will reject the most severe artefacts, but lets through a lot of undesired data, because the threshold cannot be set very low in order to avoid rejecting all data. The methods described in this chapter can be used to improve the detection and rejection of specific artefact types. When they are combined they should improve the overall artefact rejection percentages.

Most of the methods use virtual channels to improve the sensitivity of the detectors and to reduce the number of false positives. Virtual channel refers to a re-referenced and mathematically modified EEG channel. Instead of being directly recorded, it is computed from recorded data using simple operations like averaging or subtracting. Virtual channels give the ability to reinforce waveforms that are the hallmarks of specific artefact types and therefore make it easier to detect those artefacts. They also help in noise reduction, because it is possible to average two electrodes that are placed symmetrically or close to each other. This way the influence of background activity on the virtual channel is less than what it would be on a single directly recorded channel.

All artefact detection methods require thresholds to which the electrode potentials or virtual channel values are compared. The thresholds in this chapter were selected based on the experiment described in section 3.1. Once reasonable thresholds had been established the methods were tested on the test data set to find out how well they performed. Adaptive bite and muscle artefact thresholds are discussed in sections 4.3.2 and 4.4.3.

4.1. Eye blinks

The blink detection methods use a virtual vertical dipole channel, which is calculated by taking the average of the mastoid channels and subtracting the frontal channel average. The dipole channel was chosen because it takes advantage of both the averaging that removes some noise from the channel pairs and the fact that blinks show up as negative changes on mastoid channels and positive changes on frontal channels. The overall blink effect is larger on the virtual channel than on either the mastoid average or the frontal average alone.

4.1.1. Simple virtual channel threshold

The simplest way to detect artefacts is to set a predetermined threshold and reject each sample where the potential or the absolute value of the potential crosses the threshold. In the case of blinks this is a viable method, but it provides no way to differentiate between blinks and other types of artefacts. It also detects only the very peaks of blink artefacts, even though that is often enough when whole epochs are rejected if any artefacts are present. It is still possible for the onset of the artefact to be accepted if it occurs at the very end of an epoch, or similarly for the diminishing late part of the blink to cause trouble in the beginning of an epoch.

4.1.2. “Difference of long and short averages” – method

An improved method for blink detection is to measure a slightly longer segment of virtual channel data – one hundred samples in this case – and then take a long and short average around the data point that is to be classified. An average of the one hundred samples is a smooth curve that follows the long duration effects of the channel, but generally stays closer to zero than the raw data. The short average is calculated for 30 samples around the data point to be classified, which corresponds to 150ms. This means that most of the noise is removed, but the average still reacts quickly to most changes, including blinks.

A blink is clearly visible on the difference of the long and short averages. Before the blink starts there is a small dip below zero, then a rapid large rise, and finally an equally rapid decline back below zero once the blink is over. This method is used by setting a suitable threshold that mostly just the blinks cross and rejecting the samples during which this happens. All slower artefacts cause the shorter average to follow the longer average closely and do not cause the threshold to be crossed. The method is not foolproof – sometimes a bite or muscle activity causes so fast changes in the vertical dipole that the threshold is crossed and those samples get falsely rejected as blinks. This is largely inconsequential when only rejecting the artefactual samples counts, but when striving to classify the artefacts as well as possible it should be avoided. Finally, the need to correctly classify blinks will be extremely important if the algorithm is revised to include online blink compensation. For that to happen, the compensation must become viable without specialized EOG electrodes or prohibitive computing power requirements.

4.2. Horizontal saccades

Horizontal saccades are easy to detect, because they affect the difference between symmetrically located electrodes. The effect is at its largest on the temporal electrodes, so a virtual channel that is formed by subtracting the left temporal channel from the right is excellent for detecting saccades.

4.2.1. “Squared difference and threshold” – method

Simplest saccade detection method is to just take the square or the absolute value of the virtual channel and then reject samples where this value crosses a predetermined threshold. The method removes the ability to differentiate between saccades to the left and to the right, but that is rarely relevant. The correct choice between taking the second power or not depends largely on whether the other methods use the magnitude by which a classification is made as a tool for classifying artefacts. If the saccade detection method is compared to other methods that do not use the second power then using the squared value can cause events to be falsely labelled as saccades even when they strongly exhibit the properties of other artefact types. For example an event where the muscle threshold is exceeded tenfold can cause the saccade threshold to be exceeded fivefold. Such an event can be classified as a saccade if the saccade channel is squared, as a muscle artefact if both detectors use absolute values, or as both if such comparisons are not used.

4.2.2. “Smoothed absolute value of difference and threshold” – method

A slight improvement on the method described above, the point of this method is to take an average of the absolute values of one hundred virtual channel values around the point of interest and then classify the data point as artefactual if the average crosses a predetermined threshold. This removes most of the spikiness of the previous method that can easily classify five consecutive samples as ‘rejected–clean–rejected–clean–rejected’, which does not usually represent the actual physiological facts. This method can also classify as artefactual the moment when the eyes are looking straight ahead when the subject swiftly shifts their gaze from one side to the other.

4.3. Bites

Removing bite artefacts is more difficult than removing ocular artefacts, because the artefact lacks a clear waveform or a simple amplitude threshold that would always indicate its presence. It is also one of the strongest artefact types so it is necessary to have a best possible method for removing epochs contaminated by biting from the data, even if the method borders on overly sensitive.

The jaw muscle activity detection methods presented here rely on variance estimates. Variance can be calculated for any signal, but not for single values. This means using a moving window of some sort if data from only one channel are to be included in the calculation. The requirement is not really a problem as most of the other methods also require a moving window for calculating averages.

4.3.1. Simple mastoid variance threshold

This method uses the one hundred latest samples from mastoid channels and calculates the variance for both of them when categorizing the 50th latest sample. The average variance of the channels is compared against a threshold and if it is exceeded the sample is rejected. A hundred samples are definitely enough to get a reliable estimate for the variance, but it makes it impossible to accurately estimate the onset and end of the artefact. The larger problem is the large variability in the bite amplitudes. The variance threshold should be set as low as possible, but if it is set too low some recordings could be entirely rejected because the connections are not good enough. The better the electrode connections and the more relaxed the subject the lower the background noise levels are and the lower the threshold can be without causing trouble. Unfortunately in an online recording the background noise level cannot be predetermined and the threshold must be left quite high – just in case everything does not go optimally.

4.3.2. Adaptive threshold method

A natural expansion of the simple threshold method is to adaptively move the threshold if needed. This method is otherwise identical to the simple mastoid variance technique, but the threshold is moved up or down if the average variance of the accepted samples becomes too high or too low compared to it. This way the same method can be used for all recordings without any adjustments and all bite classifications will be comparable. It is noteworthy that this increases the number of false positives slightly, but the few additional detected bites make it worth it. The method can be adjusted by setting the minimum and maximum levels of the threshold compared to the average variance. If the levels are set higher the number of rejections decreases, which might decrease the number of false positives, but could also cause some real jaw muscle activity to go undetected.

4.4. Other muscle activity

The virtual dipole channel presented earlier as a method of blink detection is used in muscle activity detection as well. If the subject turns his or her head or there is other activity in the head or neck muscles the dipole channel amplitude increases.

4.4.1. Simple dipole magnitude threshold

Once again the simple yet surprisingly effective method is to monitor the absolute value of the dipole and compare it to a predetermined threshold. The problem with the approach is that it is prone to alternate between short rejected and accepted segments. It also cannot properly differentiate between blinks, vertical saccades, and muscle artefacts. It is in any case better to monitor the absolute value of the virtual channel than the absolute values of each of the separate channels, because the lower noise level causes less false positives.

4.4.2. “Smoothed virtual channel absolute value and threshold” – method

As the name would suggest, this improvement on the previous method consists of taking the absolute value of the virtual channel values and averaging over a time window – in this case the usual 100 samples. One of the advantages of this approach is that the blinks are no longer often misclassified as muscle activity as well, because they are so short. Also, the beginning and the end of each artefact are usually included in the rejection, which improves the quality of the accepted epochs. One remaining problem with the method is that there are countless different muscle artefact types that all produce different sizes and shapes of artefacts. The predetermined threshold faces the same kind of problems as in the case of bite detection methods – the changes between subjects and changing conditions during recording cannot be taken into account.

4.4.3. Adaptive muscle artefact detection

Adaptivity is added to the above method similarly to the adaptivity for the variance threshold used in bite classification. In this case the threshold is adjusted if the average vertical dipole magnitude for accepted samples changes too much, but otherwise the principles described earlier apply.

4.5. Rapid baseline shift detection

Sometimes the baseline on one or all of the channels changes rapidly. This can be caused by anything that causes spike type artefacts or electrode connection problems. The shift is a strong artefact so it is troublesome if the detectors do not catch it, but mostly the difference between the old and the new baseline is so large that at least the bite and muscle artefact thresholds are exceeded. If the shift happens on the Cz or Fz channels it can go undetected a lot easier. This is why a simple baseline shift detection method was incorporated in the system. It calculates a mean of the one hundred sample window and of a two hundred sample window including the latest hundred. If the difference between the two is very large samples are rejected until the difference returns to normal.

4.6. Performance testing

This section covers the performance of both the methods described in this chapter (summarized in table 2) and the algorithms formed by combining them (summarized in table 3). Baseline removal methods and rapid baseline shift detection are not compared, because there were no rapid baseline shifts in the test data and the baseline wander was minimal. Different noise removal methods are not compared, because noise is removed after artefacts have been rejected and it would thereby not affect the percentages at all.

Table 2: Comparing the performance of artefact removal methods. Accuracy is the fraction of correctly classified samples out of all samples. Precision is the fraction of correctly classified artefactual samples. Recall is the fraction of true artefacts out of rejected samples.

Blink removal method	**Accuracy**	**Precision**	**Recall**	**Notes**
Simple virtual channel thresholding	**89,8%**	47,6%	19,4%	2 / 77 blinks were missed entirely
Difference of long and short averages	89,6%	**68,4%**	**23,6%**	2 / 77 blinks were missed entirely
Saccade removal method	**Accuracy**	**Precision**	**Recall**	**Notes**
Squared difference and threshold	68,7%	60,4%	50,7%	many false positives
Smoothed absolute value of difference and threshold	**77,2%**	**88,3%**	**59,6%**	no saccades missed
Bite and muscle removal methods	**Accuracy**	**Precision**	**Recall**	**Notes**
Simple mastoid variance and dipole magnitude thresholds	69,3%	76,7%	**59,8%**	
Simple mastoid variance threshold and smoothed absolute value of dipole magnitude threshold	**69,5%**	78,4%	59,7%	
Adaptive methods	68,6%	**79,6%**	58,6%	

Table 3: Combined results of three progressively more advanced algorithms. The percentage of samples in the test data set that were not rejected is listed in the second column. Precision is the fraction of correctly classified samples out of the samples classified as artefactual. Accuracy is the fraction of correctly classified samples, both artefactual and clean. Sensitivity is the rejected fraction of artefactual samples and specificity is the fraction of clean samples that was not rejected.

Methods	**Accepted**	**Precision**	**Accuracy**	**Sensitivity**
Algorithm 1: All use simple thresholds, bite from average mastoid variance, others use virtual channels	33.0%	70.2%	70.4%	83.0%
Algorithm 2: Smoothed virtual channels, bite still simple threshold	33.3%	74.8%	76.5%	**88.1%**
Algorithm 3: Adaptive bite and muscle thresholds added to algorithm 2	33.5%	**75.1%**	**76.7%**	**88.1%**

The blink results show that even though simple virtual channel peak detection is effective in finding most of the blinks it also gives a lot of false positives. It can also be sensitive to electrode connections and the small personal variations in blinks between subjects unless it is recalibrated for each subject. In this case the threshold was set after the recording and an attempt was made to make sure it was low enough to detect most blinks and not low enough to reject an inordinate amount of samples. Of course, this would not be feasible in a real online recording device, but for purposes of comparison it was a necessity. The difference of long and short averages method has similar accuracy and recall, but one key difference is the duration of a single rejection, which is usually double or triple the number of samples compared to the simple threshold for any given blink. This makes the length of the rejections match the duration of the true blinks much better. There are still too many false positives, but as they are all longer the actual number of false rejections is less than half the number of falsely rejected events in simple thresholding.

Saccade detection methods have a clear winner – the smoothed rejection method is superior on all the scales. This is partly due to the wider rejections, which means fewer artefactual samples in the start and end of a saccade slip through. The most important reason why smoothing improves the method is because all the tiny differences between temporal channels are not saccades. When an average over one hundred samples is taken, these minor differences in background activity do not cause rejections but the true saccades and some other sizeable events do. The difference between the two methods is best seen in figures 7 and 8. Shown on the figures is a short sequence from the test data set, containing two saccades and a bite. The most striking difference between the two is the number of rejections and the shape of each rejection curve. Most of the rejections by the simple thresholding method are short spikes, many of which are false positives, while the smoothed method only fails in that it mislabels the bite as a saccade.

Bite and muscle detection methods have been grouped together because it would be inconvenient trying to separate the two types of muscle artefacts when the indicators used in them partially overlap. Every time mastoid variance increases when the subject bites the vertical dipole values also fluctuate more, which causes the muscle artefact detectors to fire more often as well. The simple threshold methods used together provide a reasonable result in this case. There are some problems with spikiness, false positives and false negatives, but generally the results are far from terrible. The smooth version of muscle artefact detection slightly reduces false positives and improves all three benchmarks a little. When adaptivity is added to the methods the changes seem minor in light of these values, but the main advantage is that the manual process of searching for the correct threshold becomes unnecessary. In this recording where all thresholds are set manually close to optimal the advantage of an automatically adjusting threshold all but disappears. However, the results show that the adaptation is efficient and finds a good threshold value rather quickly.

In addition to the method specific results some overall results are of interest. Table 3 lists the methods used in three different versions of the artefact removal algorithm and the results from using the test data set as a performance measure. A sample is considered correctly labelled if it is clean and it is not rejected by any detector or if it is artefactual and rejected by at least one detector. This way even if the correct detector fails to notice the artefact but some other detector rejects the sample by accident the end result is counted as a correct classification.

The results clearly indicate that with better methods the combined performance is also better. There is a marked difference between algorithms 1 and 2. All the benchmarks indicate that algorithm 2 performs better than its predecessor, which proves that averaging the virtual channels over time is an improvement that reduces the number of artefacts in the accepted data. As the only difference between algorithms 2 and 3 is the way the threshold values for bites and muscle artefacts are set it is natural that the results are very similar. The only real difference between the algorithms is the amount of work involved. Algorithm 2 has two threshold values that need to be carefully balanced between too few rejections and too many false positives, while algorithm 3 does the balancing automatically during the recording.

The methods do not provide exactly identical fractions of rejected samples. This makes comparing them to each other difficult, as the test data set only has a certain number of samples labelled to contain each type of artefacts. These results reflect a situation where each algorithm accepts between 33-34% of the samples in the test data set. Narrowing down the overall ratio makes comparing the overall results more reliable. It does not necessarily mean that the individual methods are rejecting similar shares of samples. For instance, if one algorithm had a blink detector that rejects half the samples the other methods can only cause the rejection of up to 17% of all samples. Another algorithm could reject only 10% of the samples as blinks and have 57% left to share between the other methods. Obviously the example is exaggerated, but it should be kept in mind that the overall acceptance ratio does not tell the whole truth. Table 2 shows evidence of the balancing of the methods in this study. Special care was taken to ensure that any improvements to the results did not come at the cost of reducing the performance of methods that were not being assessed. As an example, algorithms 2 and 3 use the same methods for blink and saccade detection, so the thresholds for those detectors must be equal for the bite and muscle artefact detection results to be comparable.

4.7. Developmental aspects

It is evident that more advanced methods provide better overall results in all artefact categories. Even the most basic methods often detect an artefact at its peak amplitude or during the highest variance window, but the more sophisticated methods can often correctly classify the artefact from the beginning to the end. This is relevant, because the performance of the methods is estimated by comparing the classifications they provide to manual sample-by-sample classifications. The manual classifications are far from perfect, but they give an adequate meter to estimate the performance of the methods.

Of course, there are some shortcomings in all of the methods. When the methods are used together there are many types of artefact waveforms that are classified simultaneously into several different artefact categories. This shows that the method specificity is not very good. It would be ideal if each artefact waveform would always be classified into the correct category. With these methods most strong artefact waveforms trigger all the detectors, raising questions about the reliability of the methods. More reliable classifications could allow the implementation of artefact compensation for some artefact types. The problem with the compensation methods is that they often require heavy computations and in the worst case they could remove brain activity-based waveforms as well. If successfully implemented, however, such compensation would shorten the recordings as a higher percentage of the data could be used.

In their current state, none of the methods use the Cz channel for artefact detection. This leads to the question if the channel is actually needed, and if it is not, how much time can be saved if it is removed. In reality removing one channel is not a very large improvement, but if several channels could be removed the savings in preparation time could become remarkable. The losses in artefact detection reliability would lengthen the recording, but the whole test might still be over sooner.

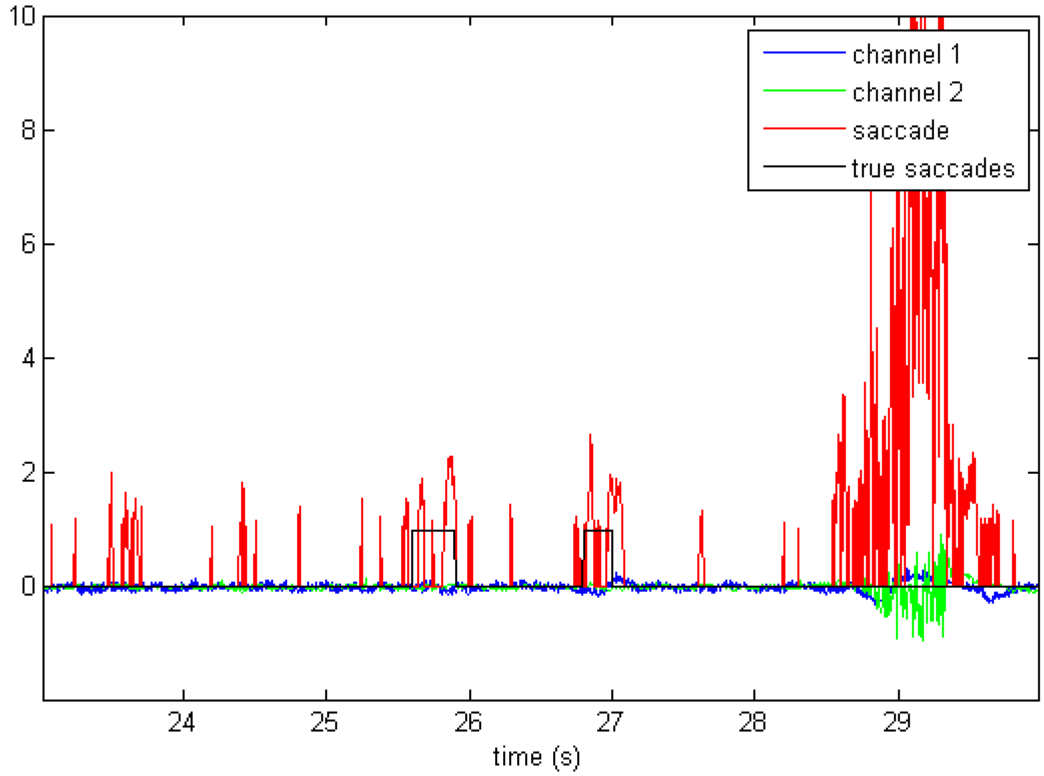

Figure 7: Absolute value of temporal difference and threshold - method. X-axis is scaled so that 1 is the threshold for rejection.

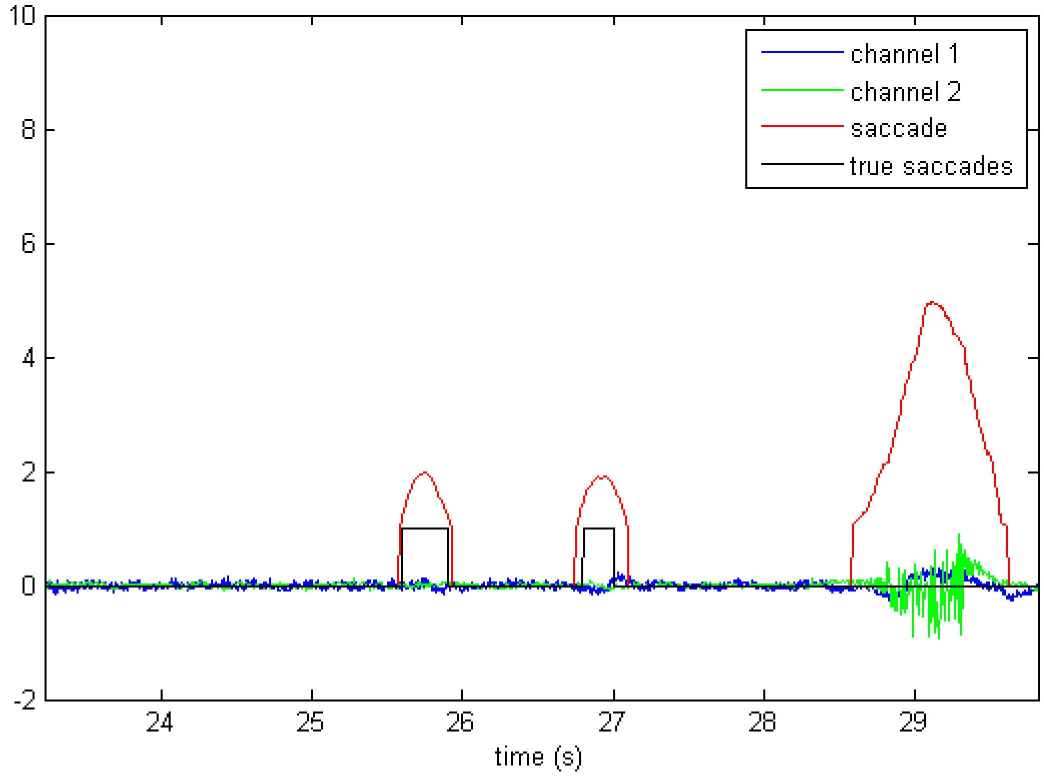

Figure 8: Smoothed absolute value of temporal difference and threshold - method. The X-axis scaling is the same as above.

5. Further developments

This chapter builds on the results of the methods described in the previous chapter. It focuses on two possible ways to improve the overall efficiency of the device. Section 5.1 focuses on developments to improve the reliability of the methods. Section 5.2 discusses the possible ways to reduce the preparation time before each recording. The better the artefact rejection algorithm the less time the recording takes. And the less time the recording takes the more time is spent on preparations compared to actually recording. This makes both approaches valid and necessary.

5.1. Improvement of reliability

In this section the best methods presented in the previous chapter are refined in an attempt to improve the reliability of the rejections. More reliable rejections allow the thresholds to be set lower to detect even smaller artefacts without fear of too many false positives. This improves the quality of the data and shortens the recording as less data are needed to make the results reliable.

5.1.1. Improved blink detection method

This blink detection method is similar to the "difference of long and short averages" – method presented in the previous chapter, but it incorporates several extra criteria that must also be met for a sample to be rejected. First, the duration of the middle wave is relevant, because natural blinks are always quite standard in length. The maximum number of blink samples and the maximum zero-to-zero time can be used to limit the length of the artefacts classified as blinks. This way the events that are shorter or longer than a blink can be easily classified into a different artefact category. Such artefacts are usually muscle activity or spike artefacts. This method requires the muscle artefact detection method to be adjusted accordingly, so that if an artefact is not classified as a blink but still crosses the blink threshold it must be rejected as it is obviously not brain activity.

5.1.2. Improved saccade detection method

One problem with the earlier saccade detection methods was that they do not actually detect eye movements but rather eye direction. In a recording situation the subject can be sitting slightly to the side of the screen or not directly facing it. In such case both the head and the eyes are naturally turned slightly towards the screen to minimize the stretching of muscles. The slight asymmetry and tension in the neck is not entirely unimportant, but the eye direction is far more relevant for the saccade rejection methods. This method adds an adaptive mean value for the temporal difference channel. The smoothed temporal difference channel is then compared to the mean instead of always assuming the neutral eye direction is straight ahead (zero difference). When calculating the mean any samples that have been labelled as artefactual are ignored. This way the reliability of the gaze direction is considerably better. Through testing the number of samples included in calculating this eye direction variable was set at two thousand, corresponding to ten seconds in time. This number allows the method to react quickly enough to changes in the subject's posture, but is not prone to excessive fluctuations.

5.1.3. Improved bite detection method

Mastoid variance is not all that changes when a subject is biting. The single sample variance across channels also increases dramatically and the temporal channel variance increases slightly even before the bite has truly started, then peaks in the middle of a bite and eventually quiets down over several seconds once the bite is over. These properties, combined with the knowledge that horizontal saccades cause temporal variance increase but do not affect mastoid variance or sample

variance much, make it easier to distinguish between the different types of artefacts.
The two alternate criteria for sample rejection in this method are: Either the average mastoid variance and the average temporal variance are over their respective thresholds, or the sample variance is above its threshold and the smooth temporal difference channel is below the saccade threshold or relatively less above it than the sample variance is above its own threshold. These criteria combined allow the method to identify the artefact onset and end points much better than the mastoid variance threshold method alone can. The criteria also make separating bite and saccade artefacts possible while improving the reliability of the methods.

5.1.4. Other improvements

After strong artefacts the channel baselines take a while to readjust, as was shown on figure 4, which can cause trouble if samples are accepted too soon after such an artefact. The artefacts also tend to start quickly and quiet down relatively slowly. Both of these factors give a reason to question the fully symmetric artefact detection methods that make no difference between before the artefact and after it. The simplest solution is to extend the rejection period after the last sample that crosses the threshold. This extension should reflect the nature of the artefact – so a blink has a shorter extension than a bite, for instance. Experimenting with the length of the extensions led to the following lengths: blink – 10 samples, bite – 30 samples, saccade – 60 samples, and muscle – none. It could be possible to develop this aspect of the rejection algorithm further by taking into account the strength of the artefact in determining the minimum number of samples rejected after it. For muscle artefacts the extension would be unnecessary, because the virtual channel averaging is sufficiently robust in rejecting the tail ends of muscle artefacts.

The methods are also made to work together by adding interdependencies to the algorithm. Basically this means two things. First, the relative amount by which a threshold is crossed is taken into account when classifying the artefacts. For example if a sample barely triggers the saccade detector, but also crosses the bite threshold tenfold it would only be classified as a bite, because that would be the dominant artefact type. Second, when the first way of reducing overlap fails the situation can be analysed and in some cases the correct rejections can be found. Some artefacts always cause many rejections, some of which are unnecessary. The shapes and sizes of the rejections are similar as long as the artefact waveforms in question are relatively standard. Knowing this, the correct classification can be applied after the artefact detection methods have done their work. This approach depends on well known and defined artefact characteristics. For example, a large blink causes a small muscle rejection when the amplitude is at its highest. With this information it is possible to search the rejections for sequences containing first a blink rejection, then both a high blink rejection and a low muscle rejection, and finally a blink rejection. When such a sequence is found the muscle rejection is removed, which improves muscle artefact detection reliability when there are fewer false positives caused by blinks.

5.2. Improvement of efficiency

Another way to improve the overall efficiency of an EEG recording is to minimize the preparation time. This includes reducing the number of electrodes, simplifying the process of attaching them to the skin and removing any unnecessary hold-ups in the recording procedure. These changes can all influence the methods. The number and locations of the recorded channels influence how the artefacts can be detected, and is therefore central to all methods of artefact removal. All of the methods presented can, in principle, be adjusted to work with only three channels if the channels are correctly selected. This algorithm was tested using the mastoid channels and a forehead channel near Fz.

Blink detection was easily adapted to only one forehead channel. The vertical dipole channel values did not change radically when frontal channel average was replaced with Fz. Noise level increased slightly, but the blink detection method has a built-in noise reduction in the form of averaging. Saccade detection faced larger problems as the temporal channels had to be replaced with the mastoids that are much further away from the eyes. In practice the larger saccades were still strong enough to be reliably detected, but the error rate increased slightly. Bite detection was actually improved by using the frontal channel variance and single sample variance in addition to the average mastoid variance. These three measures were monitored as has been described in section 5.1.3. Finally, muscle artefact detection used the same virtual channel as blink detection. Increased noise influences muscle artefact detection even less than blink detection due to the longer average used in the method.

The tests on the three-channel version were conducted using the same test data set as with all other algorithms. The data were pre-processed by removing five of the eight recorded channels, so the remaining data faithfully simulate a recording with only three electrodes. All of the improvements to the basic methods discussed in section 5.1 that were possible to implement using only three channels were incorporated in the three-channel algorithm. If the improvements were considerable this would allow the three-channel version to outperform the basic eight channel methods, which would warrant further development of the three-channel algorithms.

Attaching the electrodes more quickly would speed up the whole process somewhat, but would cause increased noise in the recordings and an increased chance of electrode detachment. It was not tested at this time, even through simulation, because noise removal or noise level normalization before artefact detection would require new methods outside the scope of this work. Currently the most time consuming preparation necessary is attaching the electrodes and testing the connection impedances. If the number of electrodes was reduced by five the time would be approximately halved, from ten to five minutes. If the number of extra epochs that could be recorded in that time is taken into account it might be acceptable to have a slightly lower artefact removal performance.

4.3. Performance testing

The performance of each new method is summarized on table 4. Both new algorithms achieved success, as can be seen on table 5. Table 6 shows the artefact detection rates of a set of test recordings using algorithm 4.

Table 4: Artefact type specific results of the new algorithms compared to algorithm 3. The highest value for each artefact type and benchmark combination has been emphasized.

Algorithm 3: Best basic methods	**Accuracy**	**Precision**	**Recall**	**Notes**
Blink detection	89,6%	68,4%	23,6%	2 / 77 missed blinks
Saccade detection	77,2%	88,3%	59,6%	
Muscle and bite detection	68,6%	79,6%	58,6%	
Algorithm 4: Improved methods	**Accuracy**	**Precision**	**Recall**	**Notes**
Improved blink detection	**89,9%**	**74,6%**	**25,1%**	One missed blink
Improved saccade detection	**83,0%**	**89,1%**	**67,8%**	
Improved muscle and bite detection	**75,3%**	**80,8%**	66,3%	
Algorithm 5: three-channel versions	**Accuracy**	**Precision**	**Recall**	**Notes**
Improved blink detection	**89,9%**	61,8%	22,8%	11 missed blinks
Improved saccade detection	76,4%	78,1%	59,9%	
Improved muscle and bite detection	74,9%	73,1%	**68,1%**	Few false positives

Table 5: Overall performance of the new algorithms. See table 3 for details of the benchmarks.

Algorithm	**Accepted**	**Precision**	**Accuracy**	**Sensitivity**	**Specificity**
Algorithm 3	33.5%	75.1%	76.7%	88.1%	61.7%
Algorithm 4	33.9%	**79.6%**	**82.2%**	**92.7%**	**68.8%**
Algorithm 5	33.6%	76.0%	77.9%	89.1%	63.2%

Table 6: Results of applying algorithm 4 to three evaluation recordings

Recording	**Correct detections**	**False detections**	**Undetected artefacts**	**Percentage of artefacts detected**	**Percentage of correct detections**
Evaluation 1	1038	45	12	98,9%	95,8%
Evaluation 2	519	2	5	99,0%	99,6%
Evaluation 3	1145	6	3	99,7%	99,5%
Total	2702	53	20	99,3%	98,1%

Comparing the improved methods used in algorithm 4 to the best earlier method of each category (tables 4 and 5), it is evident that there are many improvements. All the benchmarks have improved, which is an excellent result. Part of the success is explained by the better artefact type identification – as there are fewer samples that are mistakenly classified into several categories the accuracy and recall of the individual methods increases. The rest is explained by the methodological improvements. Bite detection has improved dramatically as other measures besides mastoid variance have been included in the process. Likewise, saccade detection is a lot more reliable and produces a lot less false positives when it is based on eye direction change monitoring instead of just the eye direction. Blink detection gains most from removing extra false positives at samples where another stronger artefact is present. There are still many false blink detections, but they are unavoidable as long as the goal of the methods is to catch each blink if at all possible.

The results for the methods using only three recorded channels are also interesting. Blink detection suffers from the loss of frontal channels – which is evident in the precision result. It is much harder to detect a blink when the distance from the eyes to the electrode doubles. The two electrodes in the eight channel version helped reduce noise in blink detection through averaging, which is not possible using just one forehead electrode. The increased noise is seen as many false positives, and to keep the fraction of rejected samples within reason the rejection threshold had to be raised. Blink detection accuracy is still very good, though, which means there are not too many falsely rejected samples compared to the other algorithms. Saccade detection faces similar problems; two dedicated temporal electrodes has become two virtual channels dependant on the difference of the mastoid potentials. Mastoids are a lot further away from the eyes than the temporal electrodes, so the effect of eye movements on them is not as large. This is evident in how much worse all the benchmarks are compared to even algorithm 3. The overall results for the three channel version are positive, however, as supported by the percentages on table 5. The muscle artefact detection method works better than in algorithms 1-3. This can be credited to the performance gains brought by adding sample variance and temporal channel variance monitoring to the method. In this case the temporal variance is replaced with the frontal channel variance, but the overall effect seems to be almost as good as in the eight channel version. The three-channel method is not as accurate or precise, but it does not suffer from as many false rejections as the eight-channel version.

All the algorithms have comparable running times of around two minutes when used on the test data set. That means they perform approximately three times faster than necessary on a slightly dated PC (dual core 2.2GHz processor) running the algorithms in Matlab. The time includes running several non-critical parts such as plotting the results, but it gives a good approximation of how fast the algorithms are. If the computational costs were too high the algorithms could be rewritten in C or some other programming language. This will be a necessary improvement if the algorithms are to be used in an embedded system in the future.

The combined improved methods in algorithm 4 are clearly superior to the old methods in this light as well. All benchmarks have improved approximately five percentage units, which is remarkable given the far from perfect manual classification the algorithms are compared to. The performance of the combined three-channel versions of the improved methods in algorithm 5 is also surprisingly good. The blink and saccade detectors left a lot to be hoped for, but the overall results are consistently better than the relatively good earlier methods. One of the reasons why this is the case is because blinks are very short in duration, so the number of blink contaminated samples in the test data set is but a fraction of the muscle or saccade contaminated samples. This does not explain the difference completely as saccade performance was poor as well. It would seem that the results for algorithm 5 should be interpreted as a success on the part of the new bite detectors.

Comparing the results of this work to the results in literature would have been impossible without a shared performance measure. To that end, three controlled recordings were conducted. Artefacts in the data were manually located to compare the results of algorithm 4 to other generally tested online algorithms. Table 6 lists the percentages based on artefacts, not single samples. Most of the artefacts were blinks or saccades, with some muscle activity and a few bites in all of the recordings.

The results are encouraging as they show that the detectors work well on unintentionally caused artefacts. They also show that there are a lot less false positives in the real recordings than in the test data set. This phenomenon could be due to the difficulty of manually categorizing blinks that occur during other artefacts, and as the test data set was almost entirely filled with intentional artefacts the only parts where the artefacts were easy to find were the parts where they were intentional. The intentionally introduced artefacts are usually stronger than natural unintentional artefacts, which makes them easier to detect manually. Computational methods also find intentional artefacts very easily, but the length of the artefacts is generally more difficult to assess, which could cause false positives if the results are tallied sample by sample, but would not affect the count of detected artefacts. One final difficulty for the automatic methods comes from the frequency of the artefacts. Ordinarily, a test subject blinks between two and thirty times a minute during an EEG recording, but in the test data set there are segments with more than three blinks per second. In this light it is understandable that an algorithm that missed one blink out of seventy seven difficult ones that were manually labelled would not miss many when they are naturally shaped and separated by artefact-free segments.

5.4. On optimization

The trade-off between finding a higher percentage of the artefacts and having less false positives is a matter of optimization. The drop in data quality caused by one undetected artefact is compared to an additional 3.55 seconds of measurement. 3.55 seconds, because every fifth stimulus on average is a deviant, and there are so many more standard epochs that a few small artefacts will not seriously affect the standard waveform. Thereby, for each undetected artefact there is a 20% probability of it reducing the quality of the deviant waveform average. Similarly, for each false positive there is a 20% probability of the rejection happening during a deviant epoch. When that happens the recording lasts five epochs longer on average as the number of clean epochs needed for averaging is approximately constant. Five epochs equals 3.55 seconds. Generally, it takes several averaged clean epochs to reduce one artefact to the level of background noise. Exactly how many clean epochs depends on many complex aspects (how large is an average undetected artefact of each type, how many epochs are needed for reliable noise removal, how long does the electrode paste last without drying on average, and so on).

On the subject of the current state of the three-channel algorithm. Eleven missed blinks are too many, even considering that the test data set contains a lot of very difficult-to-detect blinks that are close to each other. The saccade detector also requires some further development if the three-channel method is to be used in real recordings. However, these results are a good start, because the methods used were just conversions of methods designed for eight channels. With further development and more accurate assessment of the possible gains through reducing the number of channels new methods could be developed specifically for it. That way it might be possible to improve the algorithm to such a degree that the three-channel device could become a viable competitor to the current set-up. Finding the optimal locations for the three electrodes is also an open question – and one that is more complicated than it would seem at first. There is a fine balance between finding the clinically interesting information and finding the artefacts if the number of electrodes is severely restricted.

6. Summary and conclusions

The purpose of this work was to find a solution to the problem of rejecting EEG artefacts in real time. The key limitation that was constantly kept in mind was that the algorithm had to be light enough to be run by an embedded processor in some future version of the system being built in the main project. The purpose was fulfilled using adaptive methods that are ready to use without preparation or calibration. The final version of the algorithm was developed through step-by-step improvements to simple artefact detection methods. First, the relevant artefact types were identified. Preliminary testing agreed with the literature: muscle artefacts, blinks and eye movements were the most important artefact sources. Second, a way to separate the artefacts from the data was selected. Virtual channel approach was a logical choice. It is light enough and provides good results. After that the improvements followed a process of identifying a problem with the existing method and then attempting to fix it with a simple solution.

The development of the different methods was not always as smooth as it could have been. There were plenty of unnamed methods that were tested and found lacking. Some of the methods ended up working for completely other artefact type than the one they were first tried on. Some methods worked, but not in the way that was expected. There were methods that seemed to work on one data set but not on all of them. Some of these offered insights into what would later be developed into a working method, but others were archived as curiosities at best. Overall there was always a clear sense of progress and development. There are many little details in this work that have not, to my knowledge, been published before. None of the methods can be claimed to be entirely novel, but the final algorithms are effective and could perhaps be of interest to those who work in the field of online EEG artefact rejection.

Compared to the previous work in this field, the performance of the algorithms was surprisingly good. the majority of the EEG signal processing is done offline and lighter methods of this type have not been developed much. Most published results using specialized EOG channels and computationally heavy algorithms achieve near perfect blink compensation and good eye movement compensation. This means that no matter how the subject moves his or her eyes the potentials are unaffected. The problem with such methods is the possibility that they can remove both artefactual waveforms and brain activity of the frontal lobes from the data. The results of such methods are also difficult to compare to artefact rejection methods. There are very few published results of computationally light artefact rejection, none of which use a moving window to reduce noise and improve the percentage of correct detections. It would seem that the reliability of the final algorithm version is closer to the reliability of the heavy algorithms that use specific channels for artefact detection than to the most common amplitude threshold approach.

Regarding the further development of the three-channel version, the future is wide open. Every detail is open for further testing and development starting from the very basics of electrode placement and even the number of electrodes. If a fourth electrode would improve the artefact detection performance a lot then it could be a worthy addition. Finding an optimal placement scheme for three or four electrodes would be a project in itself.

In addition to further methodological improvements that are certainly possible it would be reasonable to attempt to reduce the computational cost of the algorithms. Through simplifications and translation to a more efficient programming language the program could be made much lighter. That would eventually reduce the cost of the device when a smaller embedded processor would suffice. It might also be possible to combine some of the many vector operations into matrix operations to save time. Optimizing the rate at which the adaptive thresholds are checked could bring further savings, as could stretching the intervals between other supporting operations like the eye direction assessment.

The algorithms are working better than expected. For the purposes of the project, they are at least adequate on all accounts. In some cases they surpass the performance of any similar algorithms by a large margin. That makes the work a resounding success.

7. References

[1] **Electroencephalography: Basic Principles, Clinical Applications and Related Fields**. Ernst Niedermeyer, Fernando Lopes da Silva; Lippincott Williams & Wilkins; 2004

[2] **Enhanced detection of artifacts in EEG data using higher-order statistics and independent component analysis;** Arnaud Delorme, Terrence Sejnowski, Scott Makeig; *NeuroImage 34 (2007) 1443–1449;* http://sccn.ucsd.edu/papers/neuroimage07_delorme.pdf

[3] **Detection of the EEG Artifacts by the Means of the (Extended) Kalman Filter** Martina Roháľová, Peter Sykacek, Miloš Koska and Georg Dorffner; *Measurement science review*, volume 1, number 1, 2001; http://www.measurement.sk/Papers2/Rohalova.pdf

[4] **Event-related brain potentials: an introduction** Coles, Michael G.H., Michael D. Rugg; *Electrophysiology of Mind* Oxford Scholarship Online Monographs 1996; pp. 1–27.

[5] **Processing of an inflectional affix by the human brain as revealed by the Mismatch Negativity (MMN)**; Shtyrov, Y., Pulvermüller, F.; *European Journal of Neuroscience,* 15, 1085-1091; 2002

[6] **Mismatch Negativity: Clinical and Other Applications** Näätänen, R.; *Audiology and Neuro-Otology;* Vol. 5, No. 3-4; 2000 http://content.karger.com/ProdukteDB/produkte.asp?Doi=13874

[7] **The mismatch negativity (MMN) in basic research of central auditory processing: A review** Näätänen, R., Paavilainen, P., Rinne, T., & Alho, K.; *Clinical Neurophysiology*, 118, 2544-2590; 2007

[8] **Review of Methods for Solving the EEG Inverse Problem** Roberto Domingo Pascual-Marqui; *International Journal of Bioelectromagnetism;* 1999

[9] **Review on solving the inverse problem in EEG source analysis** Grech, Roberta; Cassar, Tracey; Muscat, Joseph; Camilleri, Kenneth P.; Fabri, Simon G.; Zervakis, Michalis; Xanthopoulos, Petros; Sakkalis, Vangelis; Vanrumste, Bart; *Journal of NeuroEngineering and Rehabilitation;* November 7, 2008

[10] **Independent Component Approach to the Analysis of EEG and MEG Recordings** Ricardo Vigário, Jaakko Särelä, Veikko Jousmäki, Matti Hämäläinen, and Erkki Oja; *IEEE Transactions on biomedical engineering;* vol. 47, no. 5; may 2000

[11] **Certain topics in telegraph transmission theory** H. Nyquist; *Trans. AIEE*, vol. 47, pp. 617-644, Apr. 1928; http://www.loe.ee.upatras.gr/Comes/Notes/Nyquist.pdf

[12] **Auditory Evoked Potentials to Unpredictable Shifts in Pitch** Judith M. Ford, Walton T. Roth, Bert S. Kopell; *Psychophysiology;* Volume 13 Issue 1 , pp. 32 – 39; January 1976

[13] **Clinical Neurophysiology, Volume 2: EEG, Paediatric Neurophysiology, Special Techniques and Applications**; Colin D. Binnie, Raymond Cooper, F. Mauguiere, John W. Osselton, Pamela F. Prior, B. M. Tedman; *Elsevier*; 2003

[14] **Wavelet Filtering before Spike Detection Preserves Waveform Shape and Enhances Single-Unit Discrimination**; Alexander B. Wiltschko; Gregory J. Gage; Joshua D. Berke; *Journal of Neuroscience Methods*; 173(1): pp. 34–40; August 2008; http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2602872/

[15] **Digital Signal Processing** Editors: Lawrence R. Rabiner, Charles M. Rader; IEEE; 1972

[16] **Electronic filter simulation & design** Giovanni Bianchi, Roberto Sorrentino; McGraw-Hill Professional; 2007

[17] **Magnetic resonance imaging of brain in people at high risk of developing schizophrenia** Lawrie S. M.; Whalley H.; Kestelman J. N.; Abukmeil S. S.; Byrne M.; Hodges A.; Rimmington J. E.; Best J. J. K.; Owens D. G. C.; Johnstone E. C.; *Lancet;* vol. 353, n°9146, pp. 30-33; 1999

[18] **Accurate, noninvasive diagnosis of human brain tumors by using proton magnetic resonance spectroscopy**; Mark C. Preul; Zografos Caramanos; D. Louis Collins; Jean-Guy Villemure; Richard Leblanc; Andre Olivier; Ronald Pokrupa; Douglas L. Arnold; *Nature Medicine 2*, pp. 323 – 325; 1996
[19] **Computed Tomography — An Increasing Source of Radiation Exposure**
David J. Brenner; Eric J. Hall; N Engl J Med. Nov 29; 357(22):2277-84; 2007; http://content.nejm.org/cgi/content/full/357/22/2277
[20] **Simultaneous EEG source localisation and artifact rejection during concurrent fMRI by means of spatial filtering**; Matthew J. Brookes; Karen J. Mullinger; Claire M. Stevenson; Peter G. Morris; Richard Bowtell; *NeuroImage;* Volume 40, Issue 3, pp. 1090-1104, 15 April 2008
[21] **Removal of EEG Noise and Artifact Using Blind Source Separation**
Fitzgibbon, S P.; Powers, D M. W.; Pope, K J.; Clark, C R.; *Journal of Clinical Neurophysiology*; Volume 24; Issue 3; pp. 232-243; June 2007