Preterm infants may spend months in the neonatal intensive care unit. During this time, their brain wires itself for the rest of their lives. Any adverse events related to neurological illness or challenges in treatment may disturb this process.

This Thesis develops methodology for bedside monitoring of brain function in preterm infants by electroencephalography (EEG). The main feature of preterm EEG is intermittently occurring spontaneous activity transients (SAT), which drive the development of neuronal connections. The work in this Thesis developed and optimized an algorithm that automatically detects these events from EEG. Further, it is shown that measures based on SAT detection have clinical correlates. The methodology may contribute to better brain care of preterm infants by providing real time information about brain health to treating clinicians.
Event detection in preterm electroencephalography

Kirsi Palmu

Doctoral dissertation for the degree of Doctor of Science in Technology to be presented with due permission of the School of Science for public examination and debate in Auditorium F239a at the Aalto University School of Science (Espoo, Finland) on the 30th of January 2015 at 12 noon.

Aalto University
School of Science
Dept. of Biomedical Engineering and Computational Science
Supervising professor
Academy Prof. Risto Ilmoniemi

Thesis advisor
Docent Sampsa Vanhatalo

Preliminary examiners
Prof. Sabine Van Huffel, KU Leuven, Belgium
Docent Petro Julkunen, Kuopio University Hospital, Finland

Opponent
Docent Ari Pääkkönen, Kuopio University Hospital, Finland

Aalto University publication series
DOCTORAL DISSERTATIONS 215/2014

© Kirsi Palmu

ISSN-L 1799-4934
ISSN 1799-4934 (printed)
ISSN 1799-4942 (pdf)

Unigrafia Oy
Helsinki 2014

Finland
Preterm infants may spend months in neonatal intensive care units (NICU). Progress in neurological care of these infants depends on the ability to adequately monitor brain activity during NICU treatment. Brain monitoring is most commonly performed using electroencephalography (EEG). The preterm EEG signals are qualitatively different from EEG signals of older individuals, their distinguishing characteristics are the intermittently occurring spontaneous activity transients (SAT), which are believed to be crucial to early brain development. Automated detection of SATs might offer new tools for a neuroscientifically reasoned monitoring of infant brain in the NICU.

In this Thesis, a commercially available algorithm was tested for its applicability in detecting SATs. Because the algorithm was found to be suboptimal, an improved algorithm was developed and its parameters were optimized. Optimization and validation were done systematically, using a gold standard composed of unanimous detections by three human raters. The optimized algorithm was then used to calculate event-based measures in two clinical studies, one studying SAT occurrence in sleep stages, and the other comparing brain activity to structural brain growth.

In leave-one-out crossvalidation, the optimized algorithm showed excellent performance (sensitivity 96.6±2.8 %, specificity 95.1±5.6 %). In the clinical studies conducted, the proportion of EEG covered by SATs (SAT%) was shown to differ between sleep states, providing a possibility for developing an EEG-based measure of brain activity cycling in preterm infants. Finally, brain activity indices derived from EEG recordings shortly after birth were shown to correlate with subsequent structural growth of the brain during preterm life.

The findings together show that an SAT event detector can be constructed for the brain monitoring in NICU, and that indices based on event detection may offer important insight to brain function in the clinical research.
Tiivistelmä


Tässä väitöskirjatyössä testattiin erään kaupallisesti saatavilla olevan algoritmin soveltuvuutta aktiviteettipurskeiden tunnistukseen. Testeissä havaittiin, että algoritmi ei ollut tarkoitukseen optimaalinen, ja tästä syystä tunnistukseen kehitettiin uusi algoritmi, jonka parametrit optimoitiin. Optimointi ja validointi tehtiin järjestyksellisesti, hyödyntäen vertailuaineistona vain sellaisia EEG-jaksoja, jotka kolme asiantuntijaa olivat luokitelleet samalla tavoin. Optimoituna algoritmeja käytettiin sitten tapahtumapohjaisten muuttujien laskemiseen kahdessa kliinisessä tutkimuksessa, joista toisessa tutkittiin aktiviteettipurskeiden esiintymistä eri univaiheissa ja toisessa vertailtiin aktiviteettipurskeiden esittävyyttä aivojen aktiivisuudella ja aivojen rakenteelliseen kasvuun.

Optimoidun algoritmin suorituskyky todettiin ristiinvalidoinnissa erinomaiseksi (sensiitiivisyys 96,6±2,8%, spesifisyys 95,1±5,6%). Sillä tehdyissä kliinisissä tutkimuksissa sijoitettiin, että aktiviteettipurskeiden osuus EEG:stä (SAT%) on erilainen eri univaiheissa. Tulos tarjoaa mahdollisuuden kehittää EEG:hen hohjautuvaa muuttuja keskkosten aivojen aktiivisuuden jakottaisen vaihtelun tutkimiseen. Lopuksi sijoitettiin, että pian syntyvän jälkeen tehdyistä EEG-mittauksista laskutetut aivojen aktivisuuutta kuvaavat muuttujat korreloivat aivojen rakenteellisen kasvun kanssa keskoksaina.

Tulokset osoittavat, että aktiviteettipurskeiden automaattinen tunnistus on mahdollista ja sitä voidaan käyttää aivojen toiminnan monitorointiin vastasyntyneiden teho-osastolla. Tunnistukseen hohjautuvat muuttujat voivat  mahdollisesti antaa tärkeää uutta ymmärrystä aivojen toiminnasta.

Avainsanat Keskonen, vastasyntyneet, keskosuus, aivosähkökäyrä, EEG, spontaani aktiviteettipurske, purske, automaattinen tunnistus, algoritmi, optimointi, validointi

ISSN-L 1799-4934 ISSN (painettu) 1799-4934 ISSN (pdf) 1799-4942
Julkaisupaikka Helsinki Painopaikka Helsinki Vuosi 2014
Work on this Thesis first began in August 2007 when I approached Dr. Sampsa Vanhatalo to discuss a master thesis project. The analysis of pre-term EEG turned out to be so invigorating that I chose to continue working on the subject in my doctoral dissertation. Sampsa’s outstanding guidance made my decision easy. Renowned for his superb grasp of the field and his innovative ideas for its improvement, he is genuinely interested in the thoughts and results of the members of his research group, and always makes time to discuss them. Sampsa, thank you! I could not have asked for a better advisor.

I was very fortunate to have Prof. Risto Ilmoniemi as my supervising professor. I am very grateful for all the support and encouragement through the years.

I am greatly indebted to Dr. Matias Palva and Dr. Nathan Stevenson and their tireless advice on planning details of detection and optimization. Their role was absolutely crucial to the success of my work.

This Thesis is the result of intense international co-operation and afforded me the opportunity to work with leading researchers and analyze exceptional datasets. This has been a real privilege! I want to thank Prof. Lena Hellström-Westas and Dr. Sverre Wikström who collected the dataset in Publications I and II and, as experts, conducted the manual marking of spontaneous activity transients in the same publications. I am indebted to Prof. Geraldine Boylan for her valuable comments on our first publication.

My gratitude is owed to Dr. Turkka Kirjavainen for the dataset and co-authorship in Publication III. Sleep is a fascinating world of its own and I hope to learn more about it! I am grateful to Susanna Stjerna not only for the joint effort in statistical analysis in Publication III, but especially for her friendship and all the discussions between heaven and earth.

Publication IV resulted from a co-operation with Dr. Manon Benders and Prof. Petra Hüppi and their colleagues. It is very rare that both EEG and MRI measurements are conducted on preterm babies. Thank you for the opportunity to study this exceptional dataset!

I am grateful to Dr. Harri Valpola for his support in the early phases of this work. Of my physicist colleagues I must single out and thank Eero Hippeläinen and Eero Ahtola for discussions and sharing their Matlab expertise with me. In Publication I we benefited greatly from Lars Johan Ahnlide’s support and access to the definition of the detection algorithm as implemented in NicOne.
The work on this Thesis overlapped with my medical physics residency in Helsinki and Uusimaa Hospital District. My journey through the different specialties was guided by chief physicist, Prof. Sauli Savolainen, whom I want to thank for keeping the challenging physicist resident puzzle together! During my residency I had the opportunity to work with a great number of devoted professionals of different occupations. Thank you for all the openness and kindness with which you welcomed me, often a novice in your specialty, to your departments! Very special thanks are due chief doctor, Prof. Juhani Partanen and chief physicist, Dr. Mika Kortesniemi. Your unbounded enthusiasm for your field and research made working in your teams a real pleasure.

My residency was made special by our group of “young” physicists. This bunch has been of great help and joy on many occasions, be it everyday work challenges, studying sessions for our final medical physicist examination, or our annual Christmas parties. Thanks colleagues!

I’m grateful for research funding by the Juselius Foundation, the Pediatric Research Foundation and the Emil Aaltonen Foundation as well as the hospital district, which made my work financially possible.

I greatly appreciate the preliminary examiners Prof. Sabine Van Huffel and Dr. Petro Julkunen for their expertise and suggestions, which helped me to further improve this manuscript.

I want to thank my friends for being there. A special thank goes to Dr. Lara Day, a saving angel both when a proof reader is needed with a tight schedule, or when the kids’ room is longing for a painted tree on its wall.

Finally, I am enormously grateful to my family. Dear parents, brothers, Heikki and all the other relatives: thank you for all the years, all the support, all the good moments with you! Heikki and mother, thank you for being with Olli this year when I was writing the summary. Without you, this Thesis would not exist.

Helsinki, 16th December 2014,

Kirsi Palmu
### Contents

Acknowledgements ................................................................................... i  
List of Publications ................................................................................ vii  
Author's Contribution ............................................................................. ix  
List of Abbreviations and Symbols ......................................................... xi  
1. Introduction .......................................................................................... 1  
2. Physiological Background ................................................................... 4  
   2.1 Preterm EEG .................................................................................. 4  
   2.2 Measuring occurrence and properties of SATs and inter-SATs .......... 7  
   2.3 Dynamics of SATs and inter-SATs .................................................. 8  
      2.3.1 Maturation .............................................................................. 9  
      2.3.2 Sleep state ............................................................................ 9  
      2.3.3 Illness and medication ........................................................... 10  
3. Automated methods for analysis of preterm EEG ............................. 11  
   3.1 Amplitude integrated EEG ............................................................ 11  
   3.2 Pattern recognition in preterm EEG ............................................ 12  
      3.2.1 Fixed or variable segment length ......................................... 18  
      3.2.2 Classifier design ................................................................. 20  
      3.2.3 Fixed and adaptive thresholds ............................................. 20  
      3.2.4 Feature selection ................................................................. 20  
      3.2.5 Performance evaluation ...................................................... 21  
      3.2.6 Optimization ....................................................................... 22  
      3.2.7 Validation ............................................................................ 22  
   3.3 Further use of detection results .................................................... 23  
      3.3.1 Classification of continuous and discontinuous segments .... 23  
      3.3.2 Measuring continuity and maturation .................................. 24  
      3.3.3 Sleep state classification ...................................................... 24  
   3.4 Handling of artifacts .................................................................... 24
6.4 Use of detection results in the analysis of vigilance stages ..... 46
6.4.1 Features for SWC analysis .................................................. 47
6.4.2 Analysis of brain activity cycling ....................................... 48
6.5 Correlations between activity and structural growth ............ 49
6.6 Future perspectives ............................................................. 49
7. Conclusion ........................................................................ 50
References ............................................................................ 53

Appendix: Publications
List of Publications

This Thesis consists of an overview and of the following Publications which are referred to in the text by their Roman numerals.


Author’s Contribution

**Publication I:** Detection of 'EEG bursts' in the early preterm EEG: visual vs. automated detection

The author had the primary responsibility for both the study’s design and its practical implementation. She re-implemented a commercial detection method and analysed both the manual and automated detections. She had the main responsibility of manuscript writing.

**Publication II:** Optimization of an NLEO-based algorithm for automated detection of spontaneous activity transients in early preterm EEG

The author designed the study. She developed the improved detection algorithm and carried out the optimization of its parameters. She had the main responsibility of manuscript writing.

**Publication III:** Sleep wake cycling in early preterm infants: Comparison of polysomnographic recordings with a novel EEG-based index

The author developed all technical components of this study: she designed the data preprocessing and automated artifact detection, and calculated the indices based on the automated detection. She also participated in conception of the statistical analysis. She had the main responsibility of manuscript writing.

**Publication IV:** Early brain activity relates to subsequent brain growth in premature infants

The author participated in conception of the study protocol. She chose and calculated several indices based on the automated event detection. She participated in manuscript writing.
# List of Abbreviations and Symbols

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>μV</td>
<td>Microvolt, $10^{-6}$V</td>
</tr>
<tr>
<td>ADR</td>
<td>Average detection rate</td>
</tr>
<tr>
<td>aEEG</td>
<td>Amplitude integrated EEG</td>
</tr>
<tr>
<td>ANOVA</td>
<td>Analysis of variance</td>
</tr>
<tr>
<td>AR</td>
<td>Auto-regressive</td>
</tr>
<tr>
<td>AS</td>
<td>Active sleep</td>
</tr>
<tr>
<td>BAC</td>
<td>Brain activity cycling</td>
</tr>
<tr>
<td>BBI</td>
<td>Burst-to-burst interval</td>
</tr>
<tr>
<td>BGTh</td>
<td>Basal ganglia / thalami volume</td>
</tr>
<tr>
<td>BNO</td>
<td>Bursts of nested (high-frequency) oscillations within large slow-wave depolarisations</td>
</tr>
<tr>
<td>CA</td>
<td>Conceptional age</td>
</tr>
<tr>
<td>CFM</td>
<td>Cerebral function monitoring</td>
</tr>
<tr>
<td>CPAP</td>
<td>Continuous positive airway pressure</td>
</tr>
<tr>
<td>ECG</td>
<td>Electrocardiography, electrocardiogram</td>
</tr>
<tr>
<td>EEG</td>
<td>Electroencephalography, electroencephalogram</td>
</tr>
<tr>
<td>EOG</td>
<td>Electro-oculography, electro-oculogram</td>
</tr>
<tr>
<td>EMG</td>
<td>Electromyography, electromyogram</td>
</tr>
<tr>
<td>FLD</td>
<td>Fisher’s linear discriminant</td>
</tr>
<tr>
<td>FWE</td>
<td>Frequency weighted energy</td>
</tr>
<tr>
<td>GA</td>
<td>Gestational age</td>
</tr>
<tr>
<td>GLM</td>
<td>General linear model</td>
</tr>
<tr>
<td>h</td>
<td>Hour</td>
</tr>
<tr>
<td>Hz</td>
<td>Hertz, 1/s</td>
</tr>
</tbody>
</table>
HRV  Heart rate variability
IBI  Interburst interval, time period between two bursts
IBImax  Maximal IBI in analysed (150 min) epoch
IBR  Interburst–burst ratio
IEI  Inter-event-intervals
inter-SAT  Time period between two SATs, also called IBI
IS  Indeterminate sleep
LOO  Leave-one-out
LRTC  Long-range temporal correlation
mm  Millimeter, $10^{-3}$ m
MRI  Magnetic resonance imaging
NICU  Neonatal intensive care unite
NLEO  Non-linear energy operator
NN  Neural network
NREM  Non-REM
PCA  Principal component analysis
ppA  Peak-to-peak amplitude
RDS  Respiratory distress syndrome
rEEG  Range-EEG
REM  Rapid eye movement
RMS  Root mean square
ROC  Receiver operating characteristic
s  Second
SAT  Spontaneous activity transient, also called burst
SAT%  SAT-percentage, proportion of time covered by SATs
SAT%avg  Average SAT% in analysed (150 min) epoch
SAT%min  SAT% in a 5 min epoch with lowest SAT%
SAT#  Number of SAT events per minute
SAT#min  SAT# in a 5 min epoch with lowest SAT#
SD  Standard deviation
SVM  Support vector machine
<table>
<thead>
<tr>
<th>Abb</th>
<th>Term</th>
</tr>
</thead>
<tbody>
<tr>
<td>SWC</td>
<td>Sleep wake cycle, sleep wake cycling</td>
</tr>
<tr>
<td>T</td>
<td>Tesla</td>
</tr>
<tr>
<td>TBV</td>
<td>Total cerebral brain volume</td>
</tr>
<tr>
<td>V</td>
<td>Volt</td>
</tr>
<tr>
<td>QS</td>
<td>Quiet sleep</td>
</tr>
</tbody>
</table>
1. Introduction

Clinical motivation

Human infants are usually born after 37–41 weeks of gestation (Blencowe et al. 2012). At this time, many developmental processes have reached a state which makes living and breathing in the outside world possible. Preterm infants, that is infants born before 37 completed gestational weeks, face the challenges of extra-uterine life in a less mature condition. Many of them need special care provided by neonatal intensive care units (NICU). Thanks to continuous improvement of medical care, ever smaller and younger infants survive, but their neurodevelopment is often affected by their early birth. Every third child born prematurely suffers from neurocognitive problems (Mwaniki et al. 2012). For children born extremely preterm (before the 26th week of gestation), the situation is even worse: up to 80% of these children were found to be at least mildly disabled at 6 years of age (Marlow et al. 2005).

In NICU, vital signs (heart and respiratory rate, blood pressure, blood oxygenation) in preterm infants are monitored constantly in order to enable an immediate reaction to any physiological problems. Although the ultimate goal of the monitoring is to protect the infant’s brain direct monitoring of the brain’s wellbeing is not yet a part of standard care procedure. This is unfortunate, particularly because the time preterm infants spend in NICUs closely overlaps with the period in which the main neural connections and sensory organization of the brain are formed (Vanhatalo, Kaila 2010). Adverse events in this period may cause irreversible deficits in brain development. Importantly, some of these events may be caused by the care itself, such as unnecessarily high medication, which prevents normal brain activity.

The preterm brain differs anatomically and functionally from brains of all other age groups including term infants. The most salient feature in preterm electroencephalography (EEG) is the spontaneous activity transient (SAT, Vanhatalo et al. 2005), which is considered crucial to the development of correct nervous connections. Our far-reaching aim is that real-time monitoring of SAT events contributes to a better understanding of the preterm brain and consequently to improved care of vulnerable preterm infants.
Objectives

In this Thesis, I explore the automated detection of SATs. The specific aims of Publications I–IV were:

I To study the agreement of human raters on SAT patterns in preterm EEG. To construct a gold standard dataset consisting of unanimous detections by three human raters. To test a commercial algorithm based on a mathematical feature, non-linear energy operator (NLEO) for its usability in separating SAT events from background EEG.

II To develop an improved version of the commercial algorithm to be used in SAT detection and to optimize its parameters.

III To study the relationship between sleep stages and occurrence of SATs as detected by the optimized algorithm.

IV To study the relationship between brain growth and brain activity as measured by the occurrence of SATs.

Outline

In Chapter 2, I give a brief description of the origin and appearance of SATs. I also summarize present knowledge about measurable properties of SAT events and the time periods between them, inter-SATs.

Chapter 3 is dedicated to the technical possibilities and challenges inherent in constructing an automated SAT detection algorithm. Detection of SATs is considered a pattern recognition task and emphasis is given to methodological issues such as optimization and validation of the algorithm.

Chapter 4 presents data and methods used in the publications of this Thesis. Publications I and II form the core of the Thesis, because they include construction of a gold standard dataset, evaluation of a commercially available algorithm for SAT detection, as well as development and optimization of an improved algorithm, with subsequent evaluation of its generalization performance. Publications I and II use a shared dataset. In Publications III and IV, separate datasets including measurements by other modalities are used to study clinical correlations of indices derived from SAT detection with sleep stages and structural brain growth, respectively.

Chapter 5 presents results. I show that SATs are recognized by human raters with a relatively high inter-rater agreement. The commercial algorithm is found to be suboptimal for detection of SATs in preterm EEG. However, the improved and optimized algorithm presented in Publication II exhibits very high sensitivity and specificity in detecting SATs and inter-SATs, as approximated by leave-one-out cross validation. In Publication III, SAT% is shown to be smaller during deep NREM sleep than during REM sleep. In Publication IV, finally, growth of brain structures is shown to be positively correlated with measures of brain activity as derived from a continuous SAT detection.
In Chapter 6, the results are discussed. Special attention is given to the technical choices made in the development of the algorithm and their relationship to methods used in other automated detection systems. Use of SAT detection in measuring brain activity cycling in a recent paper of our group is described. Finally, some future perspectives are discussed.

Chapter 7 presents the conclusions.
2. Physiological Background

2.1 Preterm EEG

EEG is the most commonly used method for functional measurements of the brain. It measures potential differences caused by synchronous neuronal activation in the brain by electrodes attached to the scalp. Preterm EEG is characterised by large bursts of activity, dominated by a low-frequency wave with superposed higher-frequency oscillations (Fig. 2.1). These distinct EEG events have been called many names, including delta waves or delta brushes referring to a specific frequency range of 0.5–4 Hz named delta in the traditional EEG literature (for an overview on terminology, see Vanhatalo, Kaila 2010). Our group introduced the name spontaneous activity transients (SAT: Vanhatalo et al. 2005) to emphasise the endogenous nature of the transients. An additional reason for the introduction of a new term is that these bursts contain activity in a wide frequency range not necessarily dominated by delta (Vanhatalo, Kaila 2010). Hartley et al. (2012) reason in a similar manner, calling the middle part of the same events “bursts of nested (high-frequency) oscillations within large slow-wave depolarisations” (BNO).

In the remaining chapters, we use the terms “burst” and “SAT” interchangeably, as we do with “inter burst interval” (IBI) and “inter-SAT period”, denoting the time between these events. The terms burst and IBI are used purposefully in the literature review, as they are the most commonly used descriptions in scholarship to date, whereas SAT and inter-SAT are used in reference to our own studies.

SATs are EEG events that only exist for a certain period of development. In the immature brain, they are the main means of communication between brain areas and they are believed to be crucial for the development of correct nervous connections in the brain. Gradually, as the structure of the cortex approaches a more mature state, a qualitatively different, continuous oscillatory activity appears in the EEG of the infants. In the smallest preterm infants, the EEG is dominated by large SATs with nearly flat inter-SAT periods. In infants approaching term age, SATs have diminished in amplitude whereas their structure has become more complex and the inter-SAT periods show oscillations with ever higher amplitudes. Connections between brain areas enable higher synchronisation of the brain activity (Vanhatalo, Kaila 2006).
A schematic presentation of the development of SAT and inter-SAT periods is given in Fig. 2.2. The dramatic effect of brain development on EEG spectra, especially in the lowest frequencies, is shown in Fig. 2.3.

**Figure 2.1.** Preterm EEG with two SATs. EEG shown both without filtering (bottom) and with some conventional filter settings. Conventional high pass filtering heavily distorts the appearance of the SATs. (Palmu 2008)

**Figure 2.2.** Changes in EEG during prematurity. Both the SAT events and the ongoing oscillatory activity during the inter-SAT periods change fundamentally in the period from early preterm to fullterm. Adapted from Vanhatalo, Kaila (2006). The schematic picture does not show the variation in occurrence of SATs between sleep stages: SATs are more common during active sleep.
**Physiological Background**

*Figure 2.3.* Changes in the lowest frequencies of EEG due to maturation. Note logarithmic frequency scale. EEG spectra are dominated by frequencies below 1 Hz especially in the preterm infants (conceptional age 32–36 weeks). Adapted from Vanhatalo et al. (2005).

<table>
<thead>
<tr>
<th>EEG tracing / pattern</th>
<th>André et al. 2010</th>
<th>Haykawa et al. 2001</th>
</tr>
</thead>
<tbody>
<tr>
<td>Continuous</td>
<td>&quot;Physiological for gestational age features, with minimal amplitude of 25 μV, lasting at least 1 min.&quot;</td>
<td>&quot;EEG activity mainly consisting of delta waves &gt; 100 μV that were continuously recognised for more than 20 seconds.&quot;</td>
</tr>
<tr>
<td>Discontinuous</td>
<td>&quot;Bursts of physiological activity according to age, separated by interburst intervals (IBI) of amplitude &lt;25 μV lasting more than 3 s.&quot; For &quot;discontinuous&quot; tracing, the IBIs should cover at least 50% of a 1 min analysis period.</td>
<td>&quot;Bursts of EEG activity separated by low voltage activity &lt; 30 μV for more than five seconds. Bursts were defined as EEG activity with amplitudes more than 100 μV lasting for 2–20 seconds in any of the channels.&quot;</td>
</tr>
</tbody>
</table>

In clinical (visual) interpretation of preterm EEG, continuity is taken as the primary measure. In general, continuity implies EEG activity above certain amplitude, whereas discontinuity is defined as alteration of high amplitude bursts with prominent low-voltage activity or IBIs. The exact definitions of continuous vs. discontinuous pattern vary (for examples see Table 2.1). Often a middle class is also defined for periods that do not fit into either of the main classes. This class might be called, e.g., “semi-discontinuous tracing” (Andre et al. 2010) or “undifferentiated pattern” (Hayakawa et al. 2001).

In preterm infants, the amount of continuity is also associated with vigilance stages. Traditionally, preterm sleep has been divided into periods of “active sleep” (AS) with more continuous tracings, and “quiet sleep” (QS) with more discontinuous tracings (Vecchierini, Andre & d’Allest 2007). Some researchers believe AS and QS are immature forms of the later recognizable sleep differentiation into rapid eye movement (REM) and non-REM sleep but this is still being debated (Grigg-Damberger et al. 2007).

Considering the physiological background of SATs as described above, the distinction between continuous and discontinuous activity seems somewhat arbitrary. In both continuous and discontinuous periods, SATs and inter-SAT
periods follow each other – just their proportions are different: SATs appear with higher frequency during active sleep. Therefore, we believe that numerical measures are needed to describe the quantity of SATs as well as the quality of both SATs and inter-SAT periods.

The appearance of preterm EEG with SATs and inter-SATs has a certain resemblance to burst suppression, a pathological EEG pattern seen in term infants after asphyxia (a period of deficient oxygen supply) and even in adults in some conditions. Both preterm EEG and burst suppression EEG are characterized by an alternation of high-amplitude EEG with rather low activity EEG with abrupt changes between these two states. This similarity is interesting as it might allow for methodological transfer from automated detection algorithms for burst suppression to automated detection of SATs.

A clinically important abnormal event in preterm EEG is seizure. It was estimated that at least 5% of very preterm infants suffer from seizures (Rennie, Boylan 2007). Development of seizure detection algorithms is an active field in EEG research (for a review, see Boylan, Stevenson & Vanhatalo 2013) but is not in the scope of this Thesis. It should be noted, however, that a system for continuous monitoring of EEG in preterm infants should include descriptors for both normal brain activity as well as for potential pathological events such as seizures.

2.2 Measuring occurrence and properties of SATs and inter-SATs

Occurrence of SATs and inter-SATs can be quantified in many ways. The basis of the quantification is always the same: a visual or automated segmentation of the EEG into SAT events and inter-SAT periods. If artefacts are not considered, normal preterm EEG is mostly considered to include only these two patterns. Therefore, there must be (almost) same number of SAT and inter-SAT epochs in each EEG.

Duration characteristics are the most widely used quantification methods of inter-SATs (IBI). Different measures are used: minimum, mean or maximum duration as well as percentage levels (e.g., 10th, 50th, 90th percentile: Victor et al. 2005b). Some researchers calculated burst-to-burst intervals (BBI, Pfurtscheller et al. 2008). Bursts are not quantified as often as IBIs, but the same measures could be used for them.

Also the number of SATs or inter-SATs could be calculated. A third correlated measure is the proportion of time occupied by SATs or correspondingly, the proportion of time occupied by inter-SATs, also called interburst–burst ratio (IBR, Niemarkt et al. 2010).

Using the segmented data, the frequency content of both SATs and inter-SATs could be described.

An example of segmentation of preterm EEG is given in Fig. 2.4 alongside some derived measures. Different measures and statistics used in quantitative analysis of preterm EEG are summarized in Table 2.2.
Physiological Background

**Figure 2.4:** Example of segmentation. A: 5 minutes of preterm EEG. B: A possible segmentation. With this segmentation, SAT%=19 and number of SAT events per minute is 4.

**Table 2.2:** Measures and statistics used in quantitative analysis of SAT occurrence.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Measure</th>
<th>Statistic</th>
</tr>
</thead>
<tbody>
<tr>
<td>IBI (inter-SAT)</td>
<td>duration</td>
<td>min</td>
</tr>
<tr>
<td>burst (SAT)</td>
<td>number</td>
<td>mean</td>
</tr>
<tr>
<td>BBI</td>
<td>proportion of signal</td>
<td>max</td>
</tr>
<tr>
<td>continuity</td>
<td>spectrum</td>
<td>percentile (e.g. 10th, 50th, 90th)</td>
</tr>
</tbody>
</table>

### 2.3 Dynamics of SATs and inter-SATs

Preterm EEG is highly dynamic and its appearance is influenced by many temporal processes of different time scales. Degree of maturation, duration of extra-uterine life and current sleep stage all affect the number of SATs and inter-SATs and their properties. Illness and medication also affect the EEG. These analytically confounding factors make it more difficult to develop detection algorithms to suit all conditions, and to establish normal values or compare results between studies.

Below, general dynamics of occurrence of SATs and inter-SATs in preterm EEG are described. In most of the cited articles, segments were defined manually using certain amplitude and duration criteria. The exact definition of the patterns affects the segmentation results both in manual and automated detection of SATs and inter-SATs.

For example, if the definition of IBI requires the amplitudes to be below 15 μV (Hahn, Monyer & Tharp 1989), the resulting IBIs will be much shorter than in studies in which IBIs were defined as epochs below 30 μV (e.g. Selton, Andre & Hascoet 2000, Hayakawa et al. 2001, Victor et al. 2005b). Some studies to IBIs are summarized in Table 3.
### 2.3.1 Maturation

There is a clear pattern visible in changes due to maturation: IBIs get shorter, bursts get longer and the number of both IBIs and bursts in a given time window decrease with age. Logically, the proportion of signal covered by IBIs decreases at the same time. The energy of EEG during bursts, calculated as root-mean-square (RMS) or spectral power, decreases with age, whereas the energy in IBI periods increases.

Postnatal adaptation has also been shown to affect the occurrence and properties of bursts and IBIs. During the first days of extra-uterine life, preterm EEG changes, gradually exhibiting more continuous activity with shorter IBIs.

For a more detailed description of findings in literature, see Table 2.4.

### 2.3.2 Sleep state

Paul et al. (2003) used a relatively elaborate methodology to study differences in EEG measures between sleep states. They found that generally in preterm infants, the number of quasi-stationary segments was higher in active sleep than in quiet sleep. Fitting well with these results, quasi-stationary segments of the lowest voltage class (corresponding to inter-SATs, see also Chapter 2.2.3) had significantly longer duration in quiet sleep than in active sleep.

Hartley et al. (2012) studied the long-range temporal correlations (LRTC) of “inter-event-intervals” (IEI), which effectively mean almost the same periods as IBI or inter-SAT. They found that, even in the youngest preterm infants (gestational age (GA) 23–30 weeks), the IEI showed LRTCs. These temporal fluctuations could be attributable to varying vigilance stages in the study population.

#### Table 2.3. IBI statistics defined in some studies of preterm infants. Adapted from Victor et al. 2005 and Vecchierini et al. 2007. IBI results given as mean (range) except for Biagioni (mean±SD) and Victor (median, 10th–90th percentile). GA: gestational age.

<table>
<thead>
<tr>
<th>Author</th>
<th>Year</th>
<th>Amplitude threshold (µV)</th>
<th>GA (weeks)</th>
<th>N</th>
<th>Mean IBI (s)</th>
<th>Maximal IBI (s)</th>
<th>IBI / recording time (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Connell</td>
<td>1987</td>
<td>ND</td>
<td>26–27</td>
<td>3</td>
<td>14 (9–17)</td>
<td>60 (35–80)</td>
<td>30 (20–50)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>28–29</td>
<td>4</td>
<td>12 (8–16)</td>
<td>50 (25–70)</td>
<td>20 (10–40)</td>
</tr>
<tr>
<td>Hahn</td>
<td>1989</td>
<td>&lt; 15</td>
<td>26–27</td>
<td>5</td>
<td>5.4 (5–6)</td>
<td>12 (7–19)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>28–29</td>
<td>15</td>
<td>5.6 (5–6)</td>
<td>12.5 (4–31)</td>
<td></td>
</tr>
<tr>
<td>Biagioni</td>
<td>1994</td>
<td>&lt;30</td>
<td>27–28</td>
<td>7</td>
<td></td>
<td>30.7±13.7</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>29–30</td>
<td>10</td>
<td></td>
<td>29.5±20.6</td>
<td></td>
</tr>
<tr>
<td>Selton</td>
<td>2000</td>
<td>&lt;30</td>
<td>26</td>
<td>4</td>
<td></td>
<td>46.4</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>27</td>
<td>9</td>
<td></td>
<td>36</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>28</td>
<td>4</td>
<td></td>
<td>26.6</td>
<td></td>
</tr>
<tr>
<td>Hayakawa</td>
<td>2001</td>
<td>&lt;30</td>
<td>25–26</td>
<td>6</td>
<td>13 (10–16)</td>
<td>44.2 (19–76)</td>
<td>48.4 (19–49)</td>
</tr>
<tr>
<td>Vecchierini</td>
<td>2003</td>
<td>&lt;15</td>
<td>24–26</td>
<td>10</td>
<td></td>
<td>40 (23–59)</td>
<td>45</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>27–28</td>
<td>10</td>
<td>6 (4–8)</td>
<td>14 (7–22)</td>
<td>39 (8–64)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>29–30</td>
<td>3</td>
<td>5 (5–8)</td>
<td>12 (11–23)</td>
<td>19 (18–56)</td>
</tr>
</tbody>
</table>

Physiological Background


### 2.3.3 Illness and medication

In preterm infants with major ultrasound-detected brain lesions, mean and maximum IBI were shown to be longer than in an age-matched control group, whereas mean and minimum burst durations were shorter than in the control group (Conde et al. 2005). Mean IBI was longer in preterm infants with brain injury than in a group without brain injury (Wikström et al. 2008). Medication by morphine prolonged the IBIs (Norman et al. 2013). Medication by phenobarbital, fentanyl and theophylline, all used routinely in NICUs, affected either the length or the time-frequency characteristics of SATs (Malk, Metsäranta & Vanhatalo 2014).

### Table 2.4: Changes in EEG variables due to maturation

#### Maturation

<table>
<thead>
<tr>
<th>Variable</th>
<th>Measure</th>
<th>change</th>
<th>Ref.</th>
</tr>
</thead>
<tbody>
<tr>
<td>burst</td>
<td>min duration</td>
<td>no change</td>
<td>Biagioni et al. 2007</td>
</tr>
<tr>
<td>BBI</td>
<td>duration</td>
<td>↓</td>
<td>Pfurtscheller et al. 2008</td>
</tr>
<tr>
<td>IBI</td>
<td>number</td>
<td>↓</td>
<td>Hahn, Monyer &amp; Tharp 1989</td>
</tr>
<tr>
<td>SAT</td>
<td>number</td>
<td>↓</td>
<td>Vanhatalo et al. 2005</td>
</tr>
<tr>
<td>IBI</td>
<td>proportion of signal</td>
<td>↓</td>
<td>Hahn, Monyer &amp; Tharp 1989, Niemarkt et al. 2010b</td>
</tr>
<tr>
<td>discontinuous activity burst</td>
<td>proportion of signal</td>
<td>↓</td>
<td>Van Sweden et al. 1991</td>
</tr>
<tr>
<td>flat (IBI) spectrum: total power (1.5-25 Hz)</td>
<td>↑</td>
<td>Havlicek, Childiaeva &amp; Chernick 1975</td>
<td></td>
</tr>
<tr>
<td>SAT</td>
<td>RMS during active sleep</td>
<td>↓</td>
<td>Tolonen et al. 2007</td>
</tr>
<tr>
<td>inter-SAT RMS during quiet sleep</td>
<td>↑</td>
<td>Tolonen et al. 2007</td>
<td></td>
</tr>
<tr>
<td>burst</td>
<td>spectrum: total and band powers</td>
<td>↓</td>
<td>Jennekens et al. 2011</td>
</tr>
</tbody>
</table>

#### Extra-uterine life

<table>
<thead>
<tr>
<th>Variable</th>
<th>Measure</th>
<th>change</th>
<th>Ref.</th>
</tr>
</thead>
<tbody>
<tr>
<td>IBI</td>
<td>mean duration</td>
<td>↓</td>
<td>Niemarkt et al. 2010b</td>
</tr>
<tr>
<td>IBI</td>
<td>proportion of signal</td>
<td>↓</td>
<td>Niemarkt et al. 2010b</td>
</tr>
<tr>
<td>discontinuous activity proportion of signal</td>
<td>↓</td>
<td>Van Sweden et al. 1991</td>
<td></td>
</tr>
</tbody>
</table>
3. Automated methods for analysis of preterm EEG

The main part of this chapter deals with automated segmentation and classification of preterm EEG. A currently very popular EEG visualisation method, amplitude-integrated EEG, is introduced briefly as a prelude to the more quantitative methodology. The chapter finishes with a discussion of different attempts to cope with artifacts forming a big challenge for any automated method in the clinic.

3.1 Amplitude integrated EEG

In neonatal intensive care units, information of the brain’s wellbeing is constantly necessary, during all times of day and night. However, staff with expertise in EEG reading is not available all the time. An answer to this challenge is the use of amplitude-integrated EEG (aEEG), first introduced in the sixties by the name cerebral function monitoring (CFM, Maynard, Prior & Scott 1969). At the moment, it is the most commonly used EEG measure in NICUs; aEEG recordings have hence been the basis for many studies on general aspects of preterm EEG.

aEEG is a trend measure which describes the amplitude of EEG oscillations in a condensed form. In contrast to most methods presented in the following sections, aEEG is not based on any segmentation. The EEG is first band-pass-filtered with an asymmetric filter with cut-off frequencies of 2 and 15 Hz. The filtered data undergo a semilogarithmic transformation that emphasises the low-amplitude range. The signal is then rectified and the envelope of this processed signal is plotted with heavy time compression (e.g. 6 cm/h whereas normal EEG is plotted with 3 cm/s).

aEEG is mostly inspected visually, but sometimes quantitative analysis is performed manually. Historically, bursts could be counted as distinctive peaks in the upper margin of aEEG trend. Automated aEEG analysis exists but is not yet widespread (for an example see Bowen, Paradisis & Shah 2010, Niemarkt et al. 2010a).

aEEG’s absolute benefit is that it can be interpreted after relatively short training. However, the method is subjective and sensitive to artefacts, which can lead to false interpretations. The heavy time compression and the low number of electrodes may leave neonatal seizures unnoticed (Rennie et al. 2004).
The filters in aEEG are not well suited to the preterm population: when frequencies below 2 Hz are filtered out, most of the power in preterm EEG is lost. Thus it seems that other methods should be developed instead of aEEG for the special task of preterm EEG analysis (Boylan 2011).

### 3.2 Pattern recognition in preterm EEG

A common component of automated EEG analysis is pattern recognition: we want to recognize certain patterns in order to learn more about underlying data. Central parts of this process are the segmentation of data into epochs that contain only one pattern, and classification of these segments as one of the possible classes (see Fig. 3.2).

Segmentation and classification of preterm EEG can be considered equivalent to the detection of SAT and inter-SAT periods. Some methods state this explicitly, while others simply search for high/low-activity epochs, or pseudo-stationary epochs. Sometimes, more than two EEG classes are defined, such as IBI, burst, and continuous activity (Jennekens et al. 2011).

In the following, some technical aspects of detection algorithms are discussed. An overview of currently available algorithms is given in Tables 3.1A to 3.1C.

**Figure 3.1:** Example of aEEG. In top panel, 4 hours of aEEG trend are shown. In most modern devices it is possible to review raw EEG, too (bottom panel, here about 30 seconds).

**Figure 3.2.** Process of pattern detection. Features that discriminate well between the patterns of interest are either used for both segmentation and classification (A), or the signal is first segmented using some other technique and feature values are calculated for each segment (B), which then are classified.
Table 3.1A: Properties of detection algorithms: Patient group and detector design.

<table>
<thead>
<tr>
<th>Ref.</th>
<th>Patient group</th>
<th>Detecting</th>
<th>Classifier design</th>
</tr>
</thead>
<tbody>
<tr>
<td>Barlow 1985</td>
<td>fullterm</td>
<td>burst and interburst periods in manually selected epochs of trace alternant</td>
<td>adaptive segmentation based on a threshold, clustering of segmentation results by hierarchical clustering</td>
</tr>
<tr>
<td>Krajča et al. 1991</td>
<td>adult</td>
<td>clusters of similar EEG graphoelements</td>
<td>fuzzy c-means clustering of segments found by adaptive segmentation</td>
</tr>
<tr>
<td>Wertheim et al. 1991</td>
<td>preterm</td>
<td>low amplitude intervals (amount of discontinuity)</td>
<td>threshold</td>
</tr>
<tr>
<td>Arnold et al. 1996</td>
<td>fullterm and preterm</td>
<td>burst onsets during manually selected periods of quiet sleep</td>
<td>neural network with adaptive preprocessing unit</td>
</tr>
<tr>
<td>Galicki et al. 1997</td>
<td>neonates</td>
<td>a) burst onset</td>
<td>a) neural network with adaptive preprocessing unit</td>
</tr>
<tr>
<td></td>
<td></td>
<td>b) burst ending</td>
<td>b) threshold</td>
</tr>
<tr>
<td>Sherman et al. 1997</td>
<td>piglet model of hypoxic-ischemic injury in neonates</td>
<td>bursts</td>
<td>threshold</td>
</tr>
<tr>
<td>Atit et al. 1999</td>
<td>piglet model of hypoxic-ischemic injury in neonates</td>
<td>bursts</td>
<td>threshold</td>
</tr>
<tr>
<td>Leistritz et al. 1999</td>
<td>adult burst-suppression</td>
<td>a) suppression</td>
<td>a) neural network</td>
</tr>
<tr>
<td></td>
<td></td>
<td>b) burst, suppression and remaining EEG</td>
<td>b) neural network</td>
</tr>
<tr>
<td>Pan, Ogawa 1999</td>
<td>preterm (CA=30-39w)</td>
<td>a) discontinuity</td>
<td>a) threshold</td>
</tr>
<tr>
<td></td>
<td></td>
<td>b) sleep stage (active sleep, quiet sleep, wakefulness)</td>
<td>b) multivariate discriminant analysis</td>
</tr>
<tr>
<td>Särkelä et al. 2002</td>
<td>adult burst-suppression</td>
<td>burst, suppression and artefacts</td>
<td>sequential thresholds</td>
</tr>
<tr>
<td>Vanhatalo et al. 2005</td>
<td>preterm</td>
<td>multiband activity transients</td>
<td>sequential thresholds</td>
</tr>
<tr>
<td>West et al. 2006</td>
<td>preterm</td>
<td>high amplitude intervals (amount of continuity)</td>
<td>threshold</td>
</tr>
<tr>
<td>Wong, Abdulla 2006</td>
<td>neonate</td>
<td>stationary segments</td>
<td>3 different methods tested. SME and GLR: threshold, NLEO: peak detection.</td>
</tr>
<tr>
<td>Yunhua Wang, Agarwal 2007</td>
<td>neonate</td>
<td>suppression in burst-suppression pattern</td>
<td>sequential thresholds</td>
</tr>
<tr>
<td>Wong, Abdulla 2008</td>
<td>preterm</td>
<td>continuous and discontinuous segments</td>
<td>linear discriminant analysis</td>
</tr>
<tr>
<td>Pfurtscheller et al. 2008</td>
<td>preterm</td>
<td>burst &quot;peaks&quot;</td>
<td>threshold and peak detection</td>
</tr>
<tr>
<td>Krajča et al. 2009</td>
<td>preterm and fullterm</td>
<td>a) clusters of similar EEG graphoelements</td>
<td>a) k-means clustering of segments found by adaptive segmentation</td>
</tr>
<tr>
<td></td>
<td></td>
<td>b) sleep stages (AS vs. QS)</td>
<td>b) threshold</td>
</tr>
<tr>
<td>Vairavan et al. 2009</td>
<td>fetuses and fullterm neonates (MEG)</td>
<td>a) discontinuous MEG patterns</td>
<td>a) sequential thresholds</td>
</tr>
<tr>
<td></td>
<td></td>
<td>b) bursts (and IBI) in segments of discontinuous MEG</td>
<td>b) sequential thresholds</td>
</tr>
</tbody>
</table>
Table 3.1A cont.: Properties of detection algorithms: Patient group and detector design.

<table>
<thead>
<tr>
<th>Ref.</th>
<th>Patient group</th>
<th>Detecting</th>
<th>Classifier design</th>
</tr>
</thead>
<tbody>
<tr>
<td>Niemarkt et al. 2010b</td>
<td>preterm</td>
<td>a) bursts, interburst-intervals</td>
<td>a) As given in NicOne. See Palmu et al 2010a, original algorithm (description in Niemarkt et al. deviates from the technical description by the vendor)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>b) periods of discontinuous activity</td>
<td>b) threshold</td>
</tr>
<tr>
<td>Palmu et al. 2010b</td>
<td>preterm</td>
<td>burst, IBI, (continuous EEG/artifact)</td>
<td>sequential thresholds</td>
</tr>
<tr>
<td>(original algorithm)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Palmu et al. 2010a</td>
<td>preterm</td>
<td>SAT, inter-SAT</td>
<td>sequential thresholds</td>
</tr>
<tr>
<td>Jennekens et al. 2011</td>
<td>preterm</td>
<td>bursts, interburst-intervals and continuous patterns</td>
<td>sequential thresholds</td>
</tr>
<tr>
<td>Bhattacharyya et al. 2011</td>
<td>fullterm</td>
<td>bursts and normal (non-burst) EEG</td>
<td>support vector machine (SVM)</td>
</tr>
<tr>
<td>Hartley et al. 2012</td>
<td>preterm</td>
<td>bursts of nested oscillations (BNO)</td>
<td>threshold</td>
</tr>
<tr>
<td>Matić et al. 2012</td>
<td>&quot;neonate&quot;</td>
<td>IBI's in burst-suppression</td>
<td>adaptive segmentation (Krajča et al), classification of segments based on sequential thresholds</td>
</tr>
<tr>
<td>Mitchell et al. 2013</td>
<td>preterm</td>
<td>delta waves (divided into delta brushes and smooth delta waves and interburst intervals</td>
<td>detection of patterns based on wave morphology utilizing Bayesian probability theory</td>
</tr>
<tr>
<td>Koolen et al. 2014</td>
<td>preterm</td>
<td>bursts (and interbursts)</td>
<td>sequential thresholds</td>
</tr>
</tbody>
</table>

Table 3.1B: Properties of detection algorithms: technical aspects.

<table>
<thead>
<tr>
<th>Ref.</th>
<th>Epoch length</th>
<th>Features</th>
<th>Thresholding</th>
<th>Additional constraint</th>
</tr>
</thead>
<tbody>
<tr>
<td>Barlow 1985</td>
<td>variable</td>
<td>measures of amplitude and frequency as calculated from auto-correlation function</td>
<td>adaptive: segmentation: 120% change in amplitude and frequency measures between test and reference windows</td>
<td></td>
</tr>
<tr>
<td>Krajča et al. 1991</td>
<td>variable</td>
<td>Segmentation: frequency measure FDIF and amplitude measure ADIF. Clustering: average amplitude, variability of the amplitude, maximum positive and negative values, maximum value of the first and second derivative, amplitudes in delta, theta, alpha and beta bands</td>
<td>adaptive (segmentation): local maxima of a difference measure, if it was above an adaptive threshold</td>
<td></td>
</tr>
<tr>
<td>Wertheim et al. 1991</td>
<td>variable</td>
<td>absolute amplitude</td>
<td>fixed: &lt;25 μV</td>
<td>duration &gt; 6s</td>
</tr>
<tr>
<td>Arnold et al. 1996</td>
<td>variable</td>
<td>Broad band and narrow band envelope. Band of narrow band filter is adapted to match the &quot;initial wave&quot; of burst.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Galicki et al. 1997</td>
<td>a) variable</td>
<td>a) mean value of momentary power in 8 subsequents of a 4 s sliding window</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>b) variable</td>
<td>b) &quot;momentary power within a broader frequency band&quot;</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sherman et al. 1997</td>
<td>variable</td>
<td>Teager energy, Kullback-Leibler information criterion, likelihood ratio between five sample points before and after the test point</td>
<td>fixed: 1 for K-L information criterion, 3 for likelihood ratio</td>
<td></td>
</tr>
</tbody>
</table>
### Table 3.1B cont.: Properties of detection algorithms: technical aspects.

<table>
<thead>
<tr>
<th>Ref.</th>
<th>Epoch length</th>
<th>Features</th>
<th>Thresholding</th>
<th>Additional constraint</th>
</tr>
</thead>
<tbody>
<tr>
<td>Atit et al. 1999</td>
<td>variable</td>
<td>as initial features: either Teager energy or squared amplitude. Thresholding based on Kulback-Leibler information criterion and likelihood ratio between five sample points before and after the test point</td>
<td>fixed: 1 for K-L information criterion, 3 for likelihood ratio</td>
<td></td>
</tr>
<tr>
<td>Leistritz et al. 1999</td>
<td>a) variable</td>
<td>a) median amplitude and standard deviation in a sliding window</td>
<td>a) duration &gt;1s</td>
<td></td>
</tr>
<tr>
<td></td>
<td>b) variable</td>
<td>b) spectral edge frequency (SEF95), momentary power within a narrow frequency band, a function describing the distance from last suppression period (bursts could only occur after suppression)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pan, Ogawa 1999</td>
<td>a) variable</td>
<td>a) absolute amplitude</td>
<td>a) fixed: &lt; 25 μV</td>
<td>a) duration &gt;4s</td>
</tr>
<tr>
<td></td>
<td>b) fixed (30s)</td>
<td>b) Minimum Akaike Information criterion, total power (AR-spectrum), delta power (0-3.5 Hz), discontinuity</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Särkelä et al. 2002</td>
<td>variable</td>
<td>smoothed NLEO output in baseline corrected artefact band, difference of smoothed NLEO outputs between baseline corrected EEG and artefact bands, mean amplitude</td>
<td>fixed but not specified</td>
<td>duration (burst &gt; 1 s, suppression &gt; 0.5 s)</td>
</tr>
<tr>
<td>Vanhatalo et al. 2005</td>
<td>variable</td>
<td>normalized amplitude envelope in 10 frequency bands, number of frequency bands simultaneously above the amplitude threshold</td>
<td>adaptive: normalized amplitude envelope &gt; 1.5 SD fixed: number of frequency bands:≥4.</td>
<td></td>
</tr>
<tr>
<td>West et al. 2006</td>
<td>fixed (2s)</td>
<td>amplitude range (rEEG) in 2s segments</td>
<td>fixed: &gt; 50 μV</td>
<td></td>
</tr>
<tr>
<td>Wong, Abdulla 2006</td>
<td>variable</td>
<td>spectral error measurement (SEM), generalized likelihood ratio (GLR), Nonlinear energy operator (NLEO)</td>
<td>For SEM and GLR: fixed but not stated.</td>
<td></td>
</tr>
<tr>
<td>Yunhua Wang, Agarwal 2007</td>
<td>variable</td>
<td>integrated instantaneous amplitude across channels, smoothed by calculating moving average in 0.5s window</td>
<td>fixed (but adjustable): e.g. &lt; 9 μV</td>
<td>duration: suppression 1-60s, burst &gt;0.5s</td>
</tr>
<tr>
<td>Wong, Abdulla 2008</td>
<td>fixed: 10 min epochs with 90 % overlap</td>
<td>Distribution of mean amplitude values in pseudo-stationary segments in 10 min analysis epoch, found by using generalized likelihood ratio. The distribution is modelled by mean and standard deviation of log-transformed mean amplitude values.</td>
<td>fixed: 30th percentile</td>
<td>maxima of SV curves in periods above threshold were detected</td>
</tr>
<tr>
<td>Pfurtscheller et al. 2008</td>
<td>variable</td>
<td>spontaneous variance (SV) in sliding 1 s window</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Krajča et al. 2009</td>
<td>a) variable</td>
<td>a) Segmentation: frequency measure FDIF and amplitude measure ADIF. Clustering: average amplitude, variability of the amplitude, maximum positive and negative values, maximum value of the first and second derivative, amplitudes in delta, theta, alpha and beta bands</td>
<td>a) adaptive: Segmentation: local maxima of a difference measure, if it was above an adaptive threshold</td>
<td></td>
</tr>
<tr>
<td></td>
<td>b) variable</td>
<td>b) smoothed variance of cluster membership in 8 channels</td>
<td>b) fixed but not stated</td>
<td></td>
</tr>
</tbody>
</table>
Table 3.1B cont.: Properties of detection algorithms: technical aspects.

<table>
<thead>
<tr>
<th>Ref.</th>
<th>Epoch length</th>
<th>Features</th>
<th>Thresholding</th>
<th>Additional constraint</th>
</tr>
</thead>
</table>
| Vairavan et al. 2009      | a) fixed: 30s (fetuses), 15s (neonates) 
 b) variable (but performed separately in each 15/30s window) | a) standard deviation of time periods between hilbert phase slips in a 4s window (2s overlap). Signal is full-wave rectified prior to analysis. 
 b) smoothed (0.5s average) full-wave rectified signal. | a) fixed but not specified 
 b) adaptive: Median of the feature is calculated channel wise from time windows corresponding to local minima of the standard deviation in the fixed length segment (15 or 30s). Mean across channels is the threshold. | a) number of channels: =>2 
 b) duration: burst: >1s |
| Niemarkt et al. 2010b     | a) as in NicOne, see Palmu et al. 2010b 
 b) variable | b) interburst-burst-ratio (IBR) 
 b) adaptive: > 20th percentile | 
 | Palmu et al. 2010b (original algorithm) | fixed (1s) | smoothed NLEO output in two different frequency bands | fixed: > 300 μV² (burst), < 40 μV² (IBI), both in EEG band. > 10000 μV² (artifact) | duration (burst > 1-2 s, IBI > 2s) |
| Palmu et al. 2010a | variable | baseline corrected average of NLEO output in a sliding window | fixed: > 1.5 μV² (SAT) | duration (SAT > 1s) |
| Jennekens et al. 2011     | variable | envelope | fixed: 30 μV (below: IBI, above: burst/continuous activity) | duration (burst>20s, continuous activity>20s, IBI>1s), number of channels with high-/low-amplitude activity (burst/continuous activity: 4, IBI: 18). |
| Bhattacharyya et al. 2011 | variable (1 s sliding window with 90% overlap) | ratio of mean nonlinear energy within test window and background reference, ratio of mean absolute voltage within test window and background reference | fixed: >0.8 | duration > 4/22 s, consecutive events within 0.5 s were counted as one |
| Hartley et al. 2012       | variable | product of cube roots of confidence values in three frequency bands (0.5-2 Hz, 8-22 Hz, 2-70 Hz), confidence calculated utilizing amplitude envelopes of the filtered signals, final feature obtained after temporal smoothing | 
 | Matić et al. 2012         | variable | two nonlinear transformations of EEG amplitude | Adaptive (segmentation): maxima of difference signal. Fixed: EEG amplitude transformation in segment < 50. | number of channels (>50%), duration (>3s) |
| Mitchell et al. 2013      | not stated | posterior probability of each pattern based on template matching | 
| Koolen et al. 2014        | fixed (1s, overlap 0.12s) | median of normalized line lengths in different channels | adaptive: > 0.85*mean of line length feature in 150s epoch | amplitude change between successive non-detected and detected point > 0.4 * SD of median line lengths in 150s segment |
### Table 3.1 C: Properties of detection algorithms: optimization and validation

<table>
<thead>
<tr>
<th>Ref.</th>
<th>Optimization</th>
<th>Validation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Barlow 1985</td>
<td>parameters were set empirically based on systematic testing and visual analysis of results</td>
<td>none</td>
</tr>
<tr>
<td>Krajča et al. 1991</td>
<td>thresholds based on visual analysis in simulated data with known segment boundaries</td>
<td>visual analysis of segmentation results in some real EEG</td>
</tr>
<tr>
<td>Wertheim et al. 1991</td>
<td>empirical (mimicing visual analysis)</td>
<td>proportion of visually identified (mixed) discontinuous epochs was linearily correlated with mean proportion of automatically detected low amplitude intervals</td>
</tr>
<tr>
<td>Arnold et al. 1996</td>
<td>system adapts automatically in an empirically set region of parameter values</td>
<td>visual</td>
</tr>
<tr>
<td>Galicki et al. 1997</td>
<td>Training set with 100 bursts and 100 interbursts from 4 neonates. Simultaneous training of adaptive preprocessing unit and neural network.</td>
<td>Test set with 33 bursts and 34 interbursts from one neonate. Confusion matrix of classification results was given.</td>
</tr>
<tr>
<td>Sherman et al. 1997</td>
<td>none</td>
<td>none</td>
</tr>
<tr>
<td>Atit et al. 1999</td>
<td>none</td>
<td>Test set with manual markings</td>
</tr>
<tr>
<td>Leistritz et al. 1999</td>
<td>a) Training set with manual markings</td>
<td>a) Test set with manual markings</td>
</tr>
<tr>
<td></td>
<td>b) Training set with manual markings</td>
<td>b) Test set with manual markings</td>
</tr>
<tr>
<td>Pan, Ogawa 1999</td>
<td>empirical (mimicing visual analysis)</td>
<td>none</td>
</tr>
<tr>
<td>Särkelä et al. 2002</td>
<td>Tuning of thresholds until they are visually optimal in a training set (2x100s)</td>
<td>all data including training set was used for validation. Detection results were compared with manual markings by one expert, results given in confusion matrices</td>
</tr>
<tr>
<td>Vanhatalo et al. 2005</td>
<td>empirical (thresholds between 1 and 2 were tested)</td>
<td>Rate of occurrence of MBATs in real data was compared with detection of MBATs in 200 realizations of randomly shifted surrogate data. The rate in real data exceeded mean +2SD of rates for surrogate data.</td>
</tr>
<tr>
<td>West et al. 2006</td>
<td>thresholds 10, 25, 50 and 100 μV were tried out. 50 μV gave the best spread in dataset.</td>
<td>none</td>
</tr>
<tr>
<td>Wong, Abdulla 2006</td>
<td>none</td>
<td>visual comparison of segment boundaries with discontinuities occurring in time-frequency presentation of the original EEG</td>
</tr>
<tr>
<td>Yunhua Wang, Agarwal 2007</td>
<td>Event wise comparison of detection results with manually marked data from 4 neonates, using different thresholds. 9 μV gave the best combination of sensitivity and specificity.</td>
<td>none</td>
</tr>
<tr>
<td>Pang et al. 2008</td>
<td>A training set is used to define the centers of the data in the 2-dimensional parameter space.</td>
<td>A separate test set exists. However, only a visual example of the detection results is given.</td>
</tr>
<tr>
<td>Pfurtscheller et al. 2008</td>
<td>none</td>
<td>none</td>
</tr>
<tr>
<td>Krajča et al. 2009</td>
<td>b) Parameters were set experimentally.</td>
<td>b) Agreement between detection and visual evaluation were analysed per recording. Definition of &quot;agreement&quot; is not given.</td>
</tr>
<tr>
<td>Vairavan et al. 2009</td>
<td>a) Area under curve (AUC) of ROC curves is maximized using data from 14 neonatal EEGs with manual scoring by two reviewers. Segments were considered discontinuous if either of the reviewers scored it as discontinuous.</td>
<td>a) none</td>
</tr>
<tr>
<td></td>
<td>b) none</td>
<td>b) none</td>
</tr>
</tbody>
</table>
Table 3.1 C cont.: Properties of detection algorithms: optimization and validation

<table>
<thead>
<tr>
<th>Ref.</th>
<th>Optimization</th>
<th>Validation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Niemarkt et al. 2010b</td>
<td>b) none</td>
<td>b) none</td>
</tr>
<tr>
<td>Palmu et al. 2010b</td>
<td>none</td>
<td>performance measured against a gold standard of unanimous markings by 3 clinicians</td>
</tr>
<tr>
<td>(original algorithm)</td>
<td></td>
<td>leave-one-out cross validation, performance measured against a gold standard of unanimous markings by 3 clinicians</td>
</tr>
<tr>
<td>Palmu et al. 2010a</td>
<td>maximization of average detection rate</td>
<td>leave-one-out cross validation, performance measured against a gold standard of unanimous markings by 3 clinicians</td>
</tr>
<tr>
<td>Jennekens et al. 2011</td>
<td>Thresholds based partly on literature, but partly on comparison of sensitivity values against manual notations by two experts.</td>
<td>Validation done on different data than optimization.</td>
</tr>
<tr>
<td>Bhattacharyya et al. 2011</td>
<td>A variety of different features is studied and their performance in discriminating bursts and interbursts is analysed both separately, together, and as different subsets. Best feature set is proposed to be used in the algorithm.</td>
<td>Performance of the proposed algorithm is validated on a separate test set based on manual markings by three experts</td>
</tr>
<tr>
<td>Hartley et al. 2012</td>
<td>threshold based partly on experiments, partly on intended confidence level</td>
<td>detection results with proposed parameters were validated visually by four different researchers</td>
</tr>
<tr>
<td>Matić et al. 2012</td>
<td>Tuning of thresholds until there is a good visual agreement between the detection results and manual markings in a separate training set</td>
<td>Comparison of detected IBIs and their lengths with the manually detected IBIs, separated to different duration classes.</td>
</tr>
<tr>
<td>Mitchell et al. 2013</td>
<td>2 of 14 recordings were used for pattern optimization</td>
<td>12 of 14 recordings (=separate test set) was used to evaluate the performance. Validation method: confusion matrices with algorithm and one of the readers compared to the other reader.</td>
</tr>
<tr>
<td>Koolen et al. 2014</td>
<td>Linelength threshold: Evaluation of ROC curves generated with manual markings by two experts as gold standard. Additional constraint was defined heuristically.</td>
<td>Separate test set. Detection results were compared with manual markings by two experts and reported as confusion matrices.</td>
</tr>
</tbody>
</table>

3.2.1 Fixed or variable segment length

One of the first choices in developing an EEG classification algorithm is to define whether the signal is divided into short, constant-length segments prior to analysis, or whether the segment boundaries may reside anywhere in time, leading to variable segment lengths. In both cases, it is assumed that each segment is pseudo-stationary (also called quasi-stationary), and contains primarily one EEG class.

In order for quasi-stationarity to be sufficiently fulfilled for constant-length segments, the segments must be rather short. Sometimes, a sliding window of constant length is used, giving a higher time resolution than non-overlapping segmentation of the EEG in fixed-length epochs.

Allowing variable segment lengths, the simplest way to segment the data is to apply a threshold to a signal feature such as absolute amplitude. In this approach, segmentation often goes hand in hand with classification: segments with values above threshold are classified into one class, and vice versa (see Fig. 3.3 B).
Figure 3.3: Segmentation strategies. A: Different definitions of reference and test windows in adaptive segmentation. Reproduced based on (Wong 2008). B: Schematic presentation of segmentation based on thresholding. Gray line gives the absolute amplitude and black line the envelope of the same EEG signal as in A. Vertical lines mark the segment boundaries which occur when the feature used in segmentation (here: envelope) crosses the threshold.

A more sophisticated method for variable-length segmentation is adaptive segmentation, which uses local characteristics of the signal to find the pseudo-stationary segments. Adaptive segmentation is based on comparison of two relatively short time windows, a test window and a reference window to which the test window is compared in some mathematical sense. If the dissimilarity between the two windows grows too large, a segment boundary is initiated, and the process is started from the beginning. Especially the definition of the reference window and how it changes as the pseudo-stationary epoch grows differ from case to case (see Fig. 3.3 A).
### 3.2.2 Classifier design

Each segment found in the segmentation step should be classified into one of the possible classes. There exists a multitude of different classifier designs to choose from.

If a proper training set with manually classified segments exists, supervised learning can be used to train a complex classifier such as neural network (NN), support vector machine (SVM) or Fisher’s linear discriminant (FLD). These classifier designs use a set of several features as input. A decision plane in the feature space which best separates the classes is determined based on some measure of distance.

Unsupervised learning such as clustering divides the data in several clusters without prior knowledge of art or even number of the classes present. The value of clustering results depends crucially on the relevance of chosen features. Unsupervised learning is not yet often used on preterm EEG and it is therefore not discussed further in this Thesis.

Most often EEG analysis is done with less complex classifiers which are based e.g. on static or adaptive thresholds applied to either a single or a few features. The structure of the classifier might mimic a human interpreter using some predefined, visually available criteria for detection of each class. This often leads to a sequential design where periods of EEG above (below) threshold are classified into a certain class if they fulfil some additional constraints, such as a minimum duration. In multichannel recordings, similar constraints might apply to the number of channels.

### 3.2.3 Fixed and adaptive thresholds

In most detection algorithms developed for preterm EEG, a threshold is used either to detect the segment boundaries or to classify the predefined epochs into one of the classes of interest. Thresholds can be either fixed or adapted.

Fixed thresholds are given as feature values, such as 35 μV. They remain constant from patient to patient. Adaptive thresholds change from patient to patient or even dynamically as a function of one of the features. The use of normalized values as features, with a fixed threshold such as 1.5 standard deviations (SD) is equivalent to the use of adaptive thresholds.

In adaptive segmentation, segment boundaries can be initiated by thresholding the difference function, in which case the decision about fixed or adaptive thresholds also applies to this approach. Sometimes, however, local maxima of difference function are used to determine the segment boundaries, making definition of thresholds unnecessary.

All thresholding needs prior definition of the exact threshold. Thresholds can be defined either by prior knowledge, some trial and error or a more systematic optimization procedure (see also Chapter 3.2.6).

### 3.2.4 Feature selection

Feature selection maybe the most important step in algorithm development. A large number of features could be explored to compare their discriminative
Automated methods for analysis of preterm EEG

power (for an example of this approach in field of neonatal seizure detection, see Greene et al. 2008). An important aspect is the determination the number of features to be used. Systematic approaches exist for the search of best subsets of features given a classifier design (Bhattacharyya et al. 2011).

In most cases of preterm EEG analysis, however, the number and type of features are defined heuristically based on experience and intended classifier structure.

SATs are high-amplitude low-frequency oscillations with embedded high frequency oscillations. The epochs between SATs show much smaller amplitudes. In EEG epochs considered “continuous”, SATs appear more frequently. Therefore, different measures of signal energy are common choices as features in algorithms intended for segmenting and classifying SATs (bursts) and inter-SATs (IBIs) or continuous and discontinuous EEG. Often, some filtering is applied prior to feature calculation in order to extract the information of a certain frequency band. In general, the motivation behind extracting different features from the original signal is the enhancement of differences between the classes.

3.2.5 Performance evaluation

Performance of an algorithm can be evaluated in several ways. A prerequisite for proper performance evaluation is the existence of a dataset with correctly labelled samples or segments of EEG, also called a “gold standard” for the classification.

Traditional variables used in performance evaluation are sensitivity, specificity and accuracy of an algorithm, which are calculated based on the values of a confusion matrix as described in Fig. 3.4. Average detection rate (ADR) is a summary measure defined as the average of sensitivity and specificity.

All these variables are based on comparison of detector output with the gold standard. In case of preterm EEG, gold standards consist of manually marked epochs of SATs and inter-SATs by one or more human markers. The comparison is not straightforward and can be conducted in several ways.

One decision is whether to compare the detector output to a gold standard sample by sample (or segment by segment) or event-wise. In event-wise comparison small differences in beginning and end of an event in human and automated detection are not taken into account. In a sample-by-sample approach (equal to the segment-by-segment approach in fixed segment algorithms), differences in event durations affect the performance measures. The latter approach is more unambiguous because the definition of correct or missed detection requires no additional explanation.

Another decision is how to utilize several experts’ manual markings, if available. One possibility is to calculate performance measures separately between the algorithm and each of the markings. Another possibility, used in this Thesis, is to consider only samples with unanimous markings in the gold standard dataset. In this way, however, some of the data is lost. A third possibility would be to use major voting where data are labelled according to the majority of the votes.
3.2.6 Optimization

Optimization of an algorithm means maximization of its performance. Often this is accomplished by supervised learning: parameters of the algorithm are tuned using a training set with correctly labelled instances of data until the algorithm performance is optimal in some sense.

When constructing complex classification algorithms such as neural networks, training the classifier is an innate part of the process. In threshold-based classifier designs, however, systematic optimization is rare. Often threshold-based classifiers are constructed to mimic a human observer. If, for example, human detections are conducted using an amplitude threshold, the same value can be used in automated detection. Another common method in threshold definition is trying different values and choosing the best based for example on visual comparison of manual and automated detection, suitable distribution of output measures or correlation of some summary measure between manual and automated detection.

In Publication II of this Thesis, we have systematically optimized a SAT detection algorithm. Several papers published after ours also use a more systematic approach in their algorithm optimization, such as maximizing the sensitivity class-wise (Jennekens et al. 2011) or finding the point with best combined sensitivity and specificity on a receiver operating characteristic (ROC) curve (Koolen et al. 2014).

3.2.7 Validation

Validation of an analysis method can be understood in several ways. It might mean testing the usability of the method in real life such as its reliability and value in clinical praxis. Here, ‘validation’ is used to describe the process of testing the algorithm on unseen data (data not used in the optimization), that is: on a separate test set. Using the same data for optimization of the algorithm and its validation gives an overly optimistic perspective on the generalization capability of the algorithm.
Manual marking of EEG is time consuming and the original datasets are often small in several aspects: the number of recordings is small giving a poor sample of EEG variability between individuals, the duration of manually marked signal is short giving a small number of examples per individual, and additionally presenting only part of the intraindividual variability, and, finally, the number of experts performing the marking is small, leading to biased gold standard datasets. Based on these constraints, it is tempting to use all available data for optimization, and again for validation.

An opportunity to circumvent the problem is to use all data for the optimization but to estimate the generalization performance of the algorithm separately. In Publication II, we used this approach and estimated the performance of the algorithm by leave-one-out cross validation, testing the algorithm trained on all but one recordings on the left-out recording.

### 3.3 Further use of detection results

Segmenting and classifying preterm EEG into SAT and inter-SAT is just the beginning of the analysis procedure. Often detection results are summarized in some way for each recording. Examples of such summary measures were given in Chapter 2.2. Another option would be to analyse the detections further and to derive some more elaborate indices. The situation can be compared to electrocardiography (ECG) and the detection of QRS complexes. Those detections can be simply used to describe the number of QRS complexes in a time window (pulse), or they can be used to calculate heart-rate-variability (HRV), which contains more information of the patient than the pulse alone.

#### 3.3.1 Classification of continuous and discontinuous segments

Niemarkt et al. (2010b) used interburst–burst ratio (IBR) as calculated by a commercial algorithm to segment the EEG in periods of continuous and discontinuous activity. The twentieth percentile of IBR values was used as an (arbitrary) threshold: periods above the threshold were considered discontinuous and vice versa. The authors also claim that the periodic variation of IBR values can be attributed to rudimentary sleep-wake cycles. As summary measures for each 4 h recording, mean IBR, total length of discontinuous and continuous activity as well as mean interburst interval during discontinuous activity were calculated.

Wong and Abdulla (2008) segmented the EEG in pseudo-stationary epochs. In a 10-minute sliding window, mean absolute amplitudes of the pseudo-stationary epochs were calculated. The distribution of the amplitude values was highly skewed and was modelled using a log-normal distribution. Two parameters, the estimated mean and standard deviation of the amplitude values in the log space, were then used as features for classification of the 10-minute segments into either continuous or discontinuous pattern. The classifier was designed as a linear discriminant and different training and test sets were used for training and evaluation.
3.3.2 Measuring continuity and maturation

In her doctoral dissertation (Wong 2008), Wong developed the above-mentioned method even further. Most interesting is her idea to use principal component analysis (PCA) to the mean and standard deviation values. Utilizing the age of the infants and visually defined continuity states (continuous, discontinuous and burst suppression) in each epoch of the training dataset, Wong showed that the principal component of the transformation correlated with the amount of discontinuity in the EEG epoch, whereas the values along the minor component of the transformation correlated very strongly with the age of the infant. Hence, from each segment, two different indices could be obtained: one describing the amount of continuity, and the other describing the maturation.

Wong also presented a way to visualize the results by plotting the mean ± standard deviation of the log-transformed distribution over time. The resulting curve exhibited distinct similarity to the dynamics of aEEG, but it was much smoother and less prone to short artifacts.

Wong’s results are promising and hopefully the ideas are developed further and implemented in commercial devices. Considering the details, it may even be unnecessary to segment the data prior to calculation of the amplitude values. In Publication III we have calculated RMS values in short, fixed-length segments of preterm EEG and observed a very similar distribution of the RMS values as described by Wong in relation to the mean amplitudes.

3.3.3 Sleep state classification

Pan and Ogawa (1999) tested varying measures for sleep-stage differentiation in preterm infants. The amount of discontinuity, defined as the proportion of segments with amplitude below 25 μV for at least four seconds at a time, showed significant differences between sleep stages. It was combined with three other measures showing best discriminative power to a multivariate discriminant function which yielded significant differences between sleep stages in all age groups studied (conceptional age (CA) was 30–39 weeks).

3.4 Handling of artifacts

Artifacts are present in all real-life EEG recordings, no matter how carefully conducted. Several technical artifacts such as the line noise (50 Hz) are common in NICU, where the infants are surrounded by a multitude of technical devices. Physiological artifacts arise from muscle or cardiac activity, respiration, hiccups, sucking, movements of the eyes or movements of the infant. Handling of the infant also causes artifacts. (Walls-Esquivel et al. 2007)

Artifacts are one of the main challenges for the wider use of automated analysis methods. Many if not most studies characterizing EEG quantitatively are based on manually selected, artifact-free epochs. In such a setting, automated methods are mere tools assisting the doctor’s offline analysis. For monitoring use, however, the method must be able to cope with epochs containing
artifacts. Some authors claim that their methods are complex enough to work with contaminated data (Flisberg et al. 2011) but most attempts are methods automatically detecting and either rejecting or correcting artifactual epochs.

Artifact detection is a pattern-detection task similar to the detection of SATs. However, because there are multiple reasons for artifacts, their appearance shows much greater variation than the appearance of SATs. Artifact detection could thus be constructed separately for each type of artifact or, more generally, for all subtypes.

SATs are high-amplitude transient events containing both slow- and high-frequency oscillations. This makes preterm EEG very different from more mature EEG, which consists mostly of ongoing activity with rather small amplitudes. The properties of preterm EEG make the use of simple amplitude-based artifact detection methods developed for other age groups questionable. Often in a NICU, only a few channels are recorded, and therefore methods based on component analysis or correlations with reference channels such as ocular channel are also problematic.

Systematic studies on artifact detection necessitate a large dataset with manually marked artifacts of different types. Because the definition of an artifact might not be self-evident especially in the milder cases, manual marking should be done by several markers to obtain a reliable data set for optimization and validation of the methods. However, artifacts are generally not the doctors’ real point of interest. Therefore, systematic studies on artifacts are rare (however, see Bhattacharyya et al. 2013), and even when artifact detection is applied in preterm EEG analysis, it is based mostly on rather straightforward rejection methods. Certain signal features used in artifact rejection in studies of automated detection algorithms are summarized in Table 3.2.
### Table 3.2. Methods used in rejection of EEG epochs contaminated by artifacts.

<table>
<thead>
<tr>
<th>feature</th>
<th>threshold</th>
<th>Ref.</th>
</tr>
</thead>
<tbody>
<tr>
<td>absolute amplitude</td>
<td>fixed (300 μV)</td>
<td>Agarwal et al. 1998, Agarwal, Gotman 2001 (as one step of the artefact detection algorithm)</td>
</tr>
<tr>
<td>absolute amplitude</td>
<td>adaptive (four times the average)</td>
<td>Aarabi, Grebe &amp; Wallois 2007 (as one step of the artefact detection algorithm)</td>
</tr>
<tr>
<td>absolute amplitude</td>
<td>fixed (500 μV)</td>
<td>Yunhua Wang, Agarwal 2007 (as one step of artefact detection)</td>
</tr>
<tr>
<td>standard deviation of absolute amplitudes in a sliding window of 10s</td>
<td>based on probability distribution of feature values in each recording</td>
<td>Schetinin, Jakaite &amp; Schult 2011</td>
</tr>
<tr>
<td>standard deviation of absolute amplitudes in frequency range 30-50 Hz (EMG) in a 5s window</td>
<td>fixed (5)</td>
<td>Yunhua Wang, Agarwal 2007 (as one step of artefact detection)</td>
</tr>
<tr>
<td>difference between amplitude value and mean amplitude in two channels containing the same electrode</td>
<td>fixed</td>
<td>Yunhua Wang, Agarwal 2007 (as one step of artefact detection, aim to detect loose electrodes)</td>
</tr>
<tr>
<td>total power in 1 min epochs</td>
<td>based on distribution of feature values in each recording</td>
<td>Myers et al. 1997</td>
</tr>
<tr>
<td>total power in four frequency bands summed over 8 recording channels in 1s segments</td>
<td>5% of epochs with highest feature values were rejected</td>
<td>Schumacher et al. 2011</td>
</tr>
<tr>
<td>relative power in frequency band above 40Hz in 1s segments</td>
<td>adaptive (four times the average?)</td>
<td>Aarabi, Grebe &amp; Wallois 2007 (as one step of the artefact detection algorithm)</td>
</tr>
<tr>
<td>frequency weighted energy (average of NLEO output in a segment)</td>
<td>dynamic, not specified further</td>
<td>Agarwal et al. 1998 (as one step of the artefact detection algorithm)</td>
</tr>
<tr>
<td>NLEO output in frequency band &gt; 47Hz</td>
<td>fixed (not reported)</td>
<td>Särkelä et al. 2002</td>
</tr>
<tr>
<td>NLEO output in frequency band 47–49Hz</td>
<td>fixed (10000)</td>
<td>Palmu et al. 2010b</td>
</tr>
<tr>
<td>wavelet coefficients, spectral edge frequency, fractal dimension and mean absolute voltage</td>
<td>decision surface in feature space is found by SVM</td>
<td>Bhattacharyya et al. 2013</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>feature</th>
<th>discrimination method</th>
<th>Ref.</th>
</tr>
</thead>
<tbody>
<tr>
<td>absolute amplitude</td>
<td>segments with constant amplitude were considered as saturated signal</td>
<td>Aarabi, Grebe &amp; Wallois 2007 (as one step of the artefact detection algorithm)</td>
</tr>
<tr>
<td>eye blink template</td>
<td>segments with high correlation (&gt;0.8) with template were considered eye blinks</td>
<td>Aarabi, Grebe &amp; Wallois 2007 (as one step of the artefact detection algorithm)</td>
</tr>
<tr>
<td>QRS template</td>
<td>segments with high correlation (&gt;0.8) with template were considered as ECG artefacts</td>
<td>Aarabi, Grebe &amp; Wallois 2007 (as one step of the artefact detection algorithm)</td>
</tr>
</tbody>
</table>
4. Patients and methods

Numerical analysis of the data was performed using Matlab (Version 7.6.0, MathWorks, Natick, MA, USA) and Excel (2003, Part of Microsoft Office).

4.1 Publications I and II

4.1.1 Patients and recordings

The dataset used in both Publications I and II consists of EEG recordings from 18 preterm infants (GA=23–30 weeks, mean 26 weeks). All EEG recordings were conducted at Lund University Hospital by the team of Dr. Lena Hellsström-Westas during the first three days of life with the P3–P4 derivation (midline reference) using a Nervus/NicOne 3.3 EEG system with U16 amplifier (Cardinal Healthcare, Nicolet Biomedical, Madison, WI). The passband of the amplifier started at 0.16 Hz, and our recordings used the sampling rate of 256 Hz. An 11-minute long epoch of good-quality EEG from each infant was selected for further analysis.

In Publication I, this dataset was divided into two groups with 12 extremely preterm infants (GA 23–27 weeks) and 6 very preterm infants (GA 28–30 weeks).

4.1.2 Markings and gold standard

Three medical doctors, all experienced EEG readers, marked all SATs in the dataset on a visual basis. The marking did not involve any prior epoch definition: the markings could start and stop anywhere in time. Inter-rater agreement of the doctors was assessed sample by sample using confusion matrices. Additionally, the proportion of overall agreement (Fleiss 1971) was evaluated; this measure is useful as a single measure of agreement between multiple markings.

Epochs where all three markers agreed on the definition of the EEG as either SAT or inter-SAT were used as the gold standard in the validation and further development of the automated SAT detection algorithm in both Publications I and II. All comparisons between gold standard and algorithm were made sample-by-sample. Therefore, the apparently shorter duration of individual SAT and inter-SAT events in the gold standard did not have any effect on optimization or validation.
4.1.3 Validation of a commercially available method

In NicOne devices, a method originally developed for the detection of adult burst suppression during anaesthesia (see Särkelä et al. 2002) is implemented. This method has also been used for the assessment of SATs and inter-SATs in preterm infants (Niemarkt et al. 2010b, Wikström et al. 2008).

The method uses non-linear transformation of the signal amplitude in two separate frequency ranges as features. Särkelä et al. (2002) found that this non-linear transformation, called non-linear energy operator (NLEO, Plotkin, Swamy 1992), improved the detector performance in comparison to pure amplitude criteria.

The origin of NLEO was a paper published in 1990 in which Kaiser proposed a simple algorithm to calculate the ‘energy’ $E$ of the signal, and called it Teager’s algorithm. The energy is defined as

$$ E(n) = x(n)^2 - x(n + 1)x(n - 1) $$

where $x(n)$ is the $n$th sample of the signal being analysed. Kaiser showed that in case of a single-component signal sampled with frequency of at least 8 times the frequency of the oscillation, $E$ approximates the square of the product of signal amplitude $A$ and frequency $\Omega$,

$$ E \approx A^2 \Omega^2 $$

Agarwal et al. (1998) called the output of Teager’s algorithm in single-component signal frequency weighted energy (FWE) to differentiate it from classical definition of energy as mean-square error of the signal.

In case of a multi-component signal, the algorithm output is no longer a direct approximation of signal energy but includes additional, time-varying cross terms generated by the non-linear operation.

Plotkin and Swamy (1992) showed that Teager’s algorithm is a particular case of a more generalized NLEO

$$ \Psi_g (x(n)) = x(n - l)x(n - p) - x(n - q)x(n - s), \quad l + p = q + s, \quad (4.3) $$

where $n, l, p, q$ and $s$ are sample indices and $x(n)$ is the signal of interest.

Agarwal et al. (1998, 1999) were the first to use NLEO on EEG. They studied the effect of different choices of $l, p, q$ and $s$ on the output of NLEO in case of a multi-tone input signal with additive-white Gaussian noise. They showed that for $l+p$ and $q+s$ the expected value of the output of NLEO did not contain a component reflecting the input noise while still maintaining the FWE property. Särkelä et al. (2002) used $l=0, p=3, q=1$ and $s=2$. The effect of changing frequency and amplitude of the signal on output of NLEO with these settings is demonstrated in Fig. 4.1.
Patients and methods

Figure 4.1. Effect of varying amplitudes and frequencies on NLEO output. In the example, a single component signal is used.

The method for the detection of burst suppression by Särkelä et al. (2002) was based on calculating the output of NLEO in a sliding window of 1 s from EEG filtered into two frequency bands. The high-frequency component (>47 Hz) was used for artifact detection, while the difference between lower-frequency component (<8 Hz) and high-frequency component was used to detect bursts and suppression, each of which had its own amplitude and time thresholds. In Särkelä et al. (2002) these thresholds were optimized by using short training data from anesthetized adult patients. The mean value of unfiltered data was used in baseline correction and classification. The amount of suppression was described by, for example, burst-suppression ratio defined as the proportion of suppression time in the analysed epoch.

Based on information obtained from the vendor, the implementation of the algorithm in NicOne devices differs from the algorithm in Särkelä et al. (2002) in several aspects. For example, in the commercial algorithm the mean value of unfiltered data is not utilized. Also, the smoothed NLEO value of EEG band (0.1–8 Hz) is used as a stand-alone feature, and not after the subtraction of the NLEO value in artifact band (>47 Hz) as in Särkelä et al (2002). In Publication I, the exact definition of the algorithm as used in NicOne was published for the first time. The method was implemented in Matlab and validated utilizing the gold standard as described in 4.1.2. Sensitivity, specificity and accuracy of the automated detection, calculated on sample-by-sample basis, were used as performance measures.

4.1.4 Further development of the method

In Publication II, the above-mentioned algorithm was streamlined and some of its issues resolved. The algorithm has several numerical parameters, the choice of which affects the detection results. One of these, the minimum duration of SATs, was set to one second using the knowledge from the manual markings. Four other parameters of the improved algorithm (low and high cut off frequency of the EEG filters, length of a smoothing window and threshold
for SAT detection) were set by systematic probing of different parameter combinations. For each of the recordings, SAT detections were calculated by the algorithm with 3150 different combinations of the parameters (see Table 2 in Publication II). Average detection rate (ADR, average of sensitivity and specificity) was assessed by comparing the detection results to the gold standard. Finally, the ADR results from different recordings were averaged. In this way, the impact of each recording on overall results was the same, irrespective of the proportion of unanimous detections or disregarded, artifactual epochs. The parameter combination maximizing the ADR was considered optimal.

Real performance values of the algorithm when used on unknown data were estimated by using leave-one-out (LOO) cross validation. The optimization was repeated 18 times, each time leaving one of the recordings out and finding the parameter combination that maximized the ADR in remaining recordings. This parameter combination was then used to detect SATs in the left-out recording. Average performance values from LOO were used as an estimate of the performance of the algorithm on unseen data.

In this publication, all 18 EEGs were considered as one group.

4.2 Publication III

4.2.1 Patients and recordings

The dataset of Publication III consists of 18 polysomnographic recordings from 12 preterm infants. The data were first collected by Dr. Kirjavainen and his colleagues at Turku and Helsinki University Hospitals, as part of a study on ventilator strategies; thus all infants suffered from respiratory distress syndrome (RDS) and were either mechanically ventilated or had continuous airway pressure (CPAP) ventilation support. The infants were born with mean GA of 27.2 (range 24.7–30.3) weeks and the recordings were conducted at mean CA of 29.3 (range 25.9–32.7) weeks. The mean duration of the recordings was 6.6 h (range: 3.0–12.4 h).

Three different devices (Amlab, Amlag Technology Pty Ltd., Sydney, Australia; Siesta, Compumedics, Abbotsford, Australia; Embla, ResMed, Australia) were used in the recordings. Several polygraphic and EEG channels were recorded; however, in this study, only two EEG signals (C3–A2 and O2–A1), eye movements (electro-oculogram, EOG) and electromyogram (EMG) were analysed. The sampling rate was 100–200 Hz. All data were converted to European Data Format prior to analysis.

4.2.2 Sleep-stage analysis

A pediatric pulmonologist and sleep specialist scored the vigilance stages in all recordings manually using Somnologica software (Medcare, Reykjavik, Iceland). Scoring was done in 20-s epochs. Each epoch was scored either as wakefulness, REM-sleep, light non-REM (NREM) sleep or deep NREM sleep.
4.2.3 EEG analysis

For EEG analysis, epochs with reasonably good signal quality were selected visually. Three recordings were excluded because of excessive artifacts and in four cases, only one of the EEG channels could be analysed. The remaining dataset consisted of 14 epochs from C3–A2 (mean duration 3.7 h; range 0.9–12.3 h) and 12 epochs from O2–A1 (mean duration 2.5 h; range 0.9–8.4 h).

The recordings were carried out with different devices, sampling frequencies and derivations than the optimization of our SAT detection algorithm. Therefore, preprocessing of the data was deemed necessary. The data were first up-sampled to 256 Hz. Then, the amplitude levels were adapted by equalizing the mean RMS levels of the current datasets with those of the optimization dataset.

After preprocessing, automated SAT detection was run with the optimized algorithm as presented in Publication II. The proportion of time covered by SATs (SAT%) was calculated in the same 20-s epochs as used in sleep scoring.

4.2.4 Artifact detection

In Publication III we present an artifact rejection scheme based on RMS values in 5-s epochs. Trials using datasets from both Publications I and II and from Publication III showed that the distribution of RMS values in 5-s epochs is highly similar between recordings, if the RMS values are normalized by their median value (see also Supplementary Fig.s S2 and S3 in Publication III). In the optimization dataset, which contained only artifact free data, only very few values (3.8%) were higher than five times the median. This value was defined as a threshold and all epochs with normalized RMS values higher than the threshold were discarded from further analysis.

4.2.5 Differences in mean SAT% values between sleep stages

Statistical analysis was conducted to find out whether there are differences in mean SAT% values between sleep stages. All analysis was conducted separately for C3–A2 and O2–A1 derivations.

For each individual recording, analysis of variance (ANOVA) was used. The aim of the analysis was to show whether the mean SAT% values differ between sleep stages on individual level. Such differences would be the prerequisite for the clinical use of SAT% in the analysis of fluctuations of vigilance stages.

For completeness, pooled data from all recordings from each derivation was also analysed. Here, a general linear model (GLM) was used.

The Bonferroni correction for multiple comparisons was used in all pairwise comparisons.
4.3 Publication IV

4.3.1 Patients

The dataset consists of data from 32 preterm infants. The study protocol includes magnetic resonance imaging (MRI) shortly after birth and at term equivalent age as well as EEG recordings from the first 72 h after birth. In some cases only part of the protocol was accomplished. There were 21 infants with two MRI assessments for whom amplitude-based EEG measures could be calculated. Event-based EEG measures were calculated for 20 infants but for 3 of these MRI assessment was unavailable and 2 had only undergone the first MRI. The dataset was collected at the Children's Hospital in Geneva by Dr. Hüppi, Dr. Benders, and their colleagues. The infants’ GA was 29.3 ± 2.5 (range 25.6–35.6) weeks. Six of the 32 infants received morphine as medication and eight had minor, punctate MRI lesions.

4.3.2 EEG recordings

During the first 72 hours of life, EEG was recorded using a BRM3 brain monitor (Natus, USA, former Brainz monitor). The sampling frequency was 256 Hz and the signal was filtered using a combination of first-order high-pass filter at 1 Hz and fourth-order Butterworth low-pass filter at 50 Hz prior to storage. Although three channels were recorded, in this publication we considered only the cross-cerebral derivation P3–P4.

4.3.3 Amplitude-based EEG measures

Representative 1-h epochs between 20 and 24 hours after birth were selected. The range EEG (rEEG) paradigm was used to estimate peak-to-peak amplitudes (ppA) in 2-s intervals. These ppA values were then divided to different amplitude “bands”. Band A (ppA 0–10μV) reflects an almost flat EEG signal, whereas bands D and E (combined together, ppA > 50 μV) capture the bursts in preterm EEG. The proportions of each amplitude band in the 1-h epoch were used as measures in the correlation analysis.

4.3.4 Event-based EEG measures

Representative 2.5-h epochs (mean postnatal age at epoch onset 29 ± 11 h) were selected and imported into Matlab. Automated detection of SAT events was performed using the algorithm optimized in Publication II. Based on the algorithm output (classification of each EEG sample as either SAT or inter-SAT), several measures were calculated. Based on whole epoch, longest inter-SAT (traditionally called IBI) duration (IBImax) and proportion of SAT events (SAT%avg) were determined. Additionally, the number of SAT events per minute (SAT#) and the proportion of SAT events (SAT%) were calculated in successive 5-min epochs, resulting in 30 values for each measure in the 2.5-h recording. Different statistics such as minimum, maximum, standard deviation or median value of SAT# and SAT% measures were used in correlation analy-
sis. However, to reduce complexity, only the most promising indices (minimum of SAT#, SAT#min and minimum of SAT%, SAT%min) were reported in Publication IV.

4.3.5 MRI recording and analysis

Infants underwent two MRI examinations on 1.5-T or 3-T devices: the first one as soon as possible after birth and the second one at term equivalent age. Coronal images with high spatial resolution (0.7 x 0.7 x 1.5 mm$^3$ or 0.8 x 0.8 x 1.2 mm$^3$) were acquired using anatomical sequences. Total cerebral brain volumes (TBV) and basal ganglia/thalami (BGTh) volumes were evaluated using a tissue classification tool based on the k-nearest-neighbours method (Hüppi et al. 1998).

Additionally, the degree of cortical gyrification was evaluated in the first MRI examination. Using an approach based on above-mentioned segmentation (Dubois et al. 2008b, Dubois et al. 2008a), the interface between the developing cortex and white-matter zone (“inner cortical surface”) was reconstructed in three dimensions. Using the estimated surface curvature, sulci were defined as connected components of negative curvature. Finally, the total closed surface was calculated, and the sulcation index (proportion of sulci of total closed surface) was determined.

In addition to MRI-based metrics calculated at distinct timepoints, we calculated the rate of structural brain growth, defined as (Volume at term MRI – Volume at early MRI) / (weeks in between). This measure enabled us to correlate early brain activity with subsequent brain growth.

4.3.6 Statistical analysis

The aim of the study is to relate EEG-based metrics measuring brain activity to MRI-based metrics measuring brain structure. Many of the measures are also correlated with GA and may be affected by medication or lesions. Therefore, partial correlation analysis as implemented in SPSS was used, controlling for GA, medication and lesions. In cases of missing data, pairwise exclusion was used.
5. Summary of the results

5.1 Inter-rater agreement (I)

There was a high agreement on SATs between the expert markers. The proportion of overall agreement was 86% in the group of extremely preterm infants and 81% in the group of very preterm infants. 80% of the extremely preterm data were marked unanimously as either SAT or inter-SAT by all three experts. For the very preterm infants, the unanimous markings covered 71% of the recording time. Epochs of unanimous markings formed the gold standard.

5.2 SAT detection

5.2.1 Commercial algorithm (I)

The accuracy (proportion of correctly classified samples) of the NicOne algorithm was on average 87% and 79% for extremely preterm and very preterm infants, respectively. At 64% and 69%, the sensitivity was low. The specificity of the algorithm varied very greatly between infants and was on average 96% and 88%.

5.2.2 Optimized algorithm (II)

Optimized settings for the algorithm were the following: low cut-off frequency 0.5 Hz, high cut-off frequency 10 Hz, averaging window length 1.5 s and threshold for SAT detection $1.5 \mu V^2$.

In leave-one-out cross-validation, the ADR of the optimized algorithm was estimated as $95.8 \pm 2.3 \%$ (mean ± SD). The sensitivity was estimated as $96.6 \pm 2.8 \%$ and specificity as $95.1 \pm 5.6 \%$. In the optimization, all infants were studied as one group.

5.2.3 Differences in mean SAT% values between sleep stages (III)

Analysed on the individual level, there were significant differences ($p<0.001$) in mean SAT% values between the sleep stages in every EEG of the dataset. In pairwise comparisons, mean SAT% values in deep NREM sleep were significantly smaller than mean SAT% values in REM sleep in all recordings in both derivations C3–A2 and O2–A1 ($p<0.05$). Differences between light NREM sleep and REM sleep were significant ($p<0.05$) in 22 of the 26 analysed EEGs with light NREM sleep showing smaller mean SAT% values than REM sleep.
In 20 of the 26 EEGs, the difference between deep NREM and light NREM sleep was significant too, with smaller mean SAT% values in deep NREM sleep. Mean SAT% values of epochs scored as wakefulness showed no constant relationship to SAT% values in other sleep states.

In pooled data, the differences in estimated mean SAT% values were significant. All pairwise comparisons between sleep stages showed a highly significant ($p<0.001$) difference in estimated mean SAT% values.

### 5.2.4 Correlations between EEG and MRI measures (IV)

Several EEG-based measures showed a significant correlation with structural measures based on MRI. All findings are summarized in Tables 5.1 and 5.2.

More activity, reflected as smaller proportion of band A (ppA 0–10 μV), bigger proportion of band C (ppA 25–50 μV) or D+E (ppA >50μV), shorter IBI-max or bigger values of SAT#min, SAT%min or SAT#avg, was correlated with more developed brain structure, reflected as higher values of cortical sulcation or surface, volumetric measures of total brain or basal ganglia and thalami, or even with more rapid growth of brain volumes.

**Figure 5.1:** Mean SAT% values in different sleep stages. Results are presented separately for each montage (C3–A2 and O2–A1) and recording, with x-values corresponding to the CA of the infant at time of the recording. Note that if available, values of different montages are similar. Values of REM sleep are highest and values of deep NREM sleep lowest in each recording.
### Summary of the results

Table 5.1: Correlations between EEG metrics and brain structure.

<table>
<thead>
<tr>
<th>GA</th>
<th>early MRI</th>
<th>term MRI</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>cortical sulcation</td>
<td>cortical surface</td>
</tr>
<tr>
<td>Amplitude based measures</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Band A</td>
<td>ns</td>
<td>ns</td>
</tr>
<tr>
<td>Band C</td>
<td>ns</td>
<td>0.95, p&lt;0.001</td>
</tr>
<tr>
<td>Band D+E</td>
<td>ns</td>
<td>ns</td>
</tr>
</tbody>
</table>

Table 5.2: Correlations between EEG metrics and brain growth rate

<table>
<thead>
<tr>
<th></th>
<th>TBV growth</th>
<th>BGTh growth</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amplitude based measures</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Band A</td>
<td>ns (-0.458, p=0.07)</td>
<td>ns</td>
</tr>
<tr>
<td>Band C</td>
<td>ns</td>
<td>ns</td>
</tr>
<tr>
<td>Band D+E</td>
<td>0.85, p&lt;0.01</td>
<td>0.82, p&lt;0.05</td>
</tr>
</tbody>
</table>

|          |          |          |
| Event based measures |          |          |
| IBlmax    | -0.678, p<0.05 | ns (-0.667, p=0.05) |
| SAT#min   | ns         | 0.927, p<0.001 |
| SAT%min   | ns         | 0.922, p<0.001 |
| SAT%avg   | ns (0.53, p=0.07) | 0.882, p=0.001 |
6. Discussion and future perspectives

6.1 Inter-rater agreement

Inter-rater agreement is a scarcely studied subject in EEG analysis, even though developing automated detection methods for EEG events mostly involves a comparison with human markings. To our knowledge, we were the first group studying inter-rater agreement in visual detection of SATs. Our results show that the agreement is relatively high. However, for around 20 or 30% of time (extremely preterm and very preterm infants, respectively), the three raters disagreed on the definition of EEG as either SAT or inter-SAT. The difference arises mainly from the exact definition of the onset and end of each SAT (see also Fig. 2 in Publication I) and not from discrepancies in definition of certain waveforms as SAT events. The different visual “thresholds” for SATs affect the duration of both the SATs and the inter-SATs. This is notable as much scholarship of preterm EEG is based on manual definition of inter-SATs (IBIs) by a single marker.

Referring to our study (Publication I), Jennekens et al. (2011) compared manual markings of bursts and IBIs by two observers. They report 66±3% and 70±6% agreement for bursts and IBIs, respectively. Exact definition of “agreement” is not given, but the values seem comparable or slightly worse than those obtained in our study. Events were marked with precision of 1 s.

Mitchell et al. (2013) compare the two markings of their study event-wise, using one rater at a time as gold standard. Possible event classes are delta waves and IBIs. The raters agreed on 579 delta waves, whereas the first rater had marked 709 and the second rater 754 delta waves. This means that 65% of all delta waves marked were marked in agreement by both raters. For IBIs, event markings were in agreement in 48 cases whereas the raters had marked 65 and 62 IBIs, meaning that agreement was 61%.

Also Koolen et al. (2014) have two manual markings available, with possible classes of burst and IBI. The paper’s general methodology suggests that agreement is calculated sample-by-sample. The inter-rater agreement is found to be 86.2% which is a very high value and somewhat better than in our data. However, since our comparison included three raters, the total agreement between them can be expected to yield lower values. Pairwise inter-rater agreement in our data was 88.1%, 86.3% and 85.1% for extremely preterm infants (rater 1 vs. rater 2, rater 2 vs. rater 3 and rater 1 vs. rater 3, respectively) and
Discussion and future perspectives

82.8%, 74.0% and 85.2% for very preterm infants (unpublished results). These values are comparable with results of Koolen et al (2014).

6.2 SAT detection algorithm

The main aim of this Thesis was to develop a reliable detection algorithm for SATs in preterm EEG. In this chapter, technical choices met in development process are discussed and compared with the choices made by other groups.

6.2.1 NLEO as feature

NLEO is a non-linear transformation with an output proportional to squared amplitude and squared frequency of the original signal (see Equation 4.2). In Kaiser (1990) and Agarwal et al. (1998), the sampling rate is not mentioned to affect the proportionality, given a signal with frequency content much below the sampling rate. It is, however, obvious that absolute values of NLEO output, based on four adjacent sample values, are highly dependent on sampling rate. For example, if a smooth signal is sampled at twice the original sampling rate, its apparent amplitude (based on amplitude differences between samples) is 50% of the original amplitude, leading to NLEO output of 25% of the values obtained with original sampling rate. The effect of sampling rate on smoothed NLEO values in case of real EEG is shown in Fig. 6.1.

In all our studies, the EEG signal was resampled to 256 Hz as used in the optimization of the algorithm prior to analysis. This is especially important, as our algorithm includes fixed thresholds.

In developing a detection algorithm, the choice of features to be used in segmentation and classification is crucial. In our case, the NLEO as feature was given as it was part of the commercial algorithm validated in Publication I. After Publication I and II of this Thesis, two other, equally simple detection algorithms have been proposed. They use the envelope (Jennekens et al. 2011) and line length (Koolen et al. 2014) as features. In each of the algorithms, the feature values are calculated in a short (1–1.5s) sliding window.

To test the discriminative power of each of these three features in separating SATs and inter-SATs, we calculated their feature values in our optimization dataset using a 1-s sliding window. As preprocessing step, the data were filtered to 0.5–20 Hz. This is a compromise between the different filter settings used in each of the algorithms. Utilizing the gold standard (unanimous markings of SATs and inter-SATs), a SAT/inter-SAT ratio was calculated for each feature. This ratio can be seen as signal-to-noise ratio of the feature in detecting SATs.

Median values (range) of SAT/inter-SAT ratio were 39.1 (16.5–89.6), 7.6 (4.8–13.4) and 5.3 (3.9–7.4) for smoothed NLEO, envelope and line length, respectively. The values of the three methods differed significantly (p<0.001) when tested with the non-parametric Kruskall-Wallis one-way analysis of variance. In pairwise comparisons with the non-parametric Mann-Whitney test, NLEO showed significantly higher values than envelope and line length (in both tests, U=0, n=18, p<0.001). SAT/inter-SAT ratio for envelope was
significantly higher than for line length ($U=58$, $n=18$, $p=0.001$). Values for SAT/inter-SAT ratio for each infant are shown in Fig. 6.2, and examples of feature values in a short epoch of EEG are shown in Fig. 6.3. Based on the results, NLEO clearly outperforms the two other features as the basis for automated SAT detection. It should, however, be emphasized that this statement holds for the feature values prior to any postprocessing such as thresholding. The performance of the detection algorithms utilizing the features depends also on several other decisions made in the algorithm design (see Chapter 6.2.5).

![Figure 6.1: Influence of sampling frequency on NLEO output. A: Two minutes of preterm EEG. B: Smoothed NLEO output calculated from original signal (256 Hz, red) and the same signal downsampled to 0.5*fs (128 Hz, green) and upsampled to 2*fs (512 Hz, blue). C: Same as B but with logarithmic scale. D: Ratio of NLEO values. Blue: 512 Hz vs. 256 Hz. Red: 256 Hz vs. 128 Hz. Doubling the sampling rate from 128 Hz to 256 Hz (red) or from 256 Hz to 512 Hz (blue) leads to NLEO values only ¼ as high as with lower sampling frequency.](image-url)
Discussion and future perspectives

Figure 6.2. SAT/inter-SAT ratio of different features. Ratios of NLEO, envelope and line length feature values during unanimously marked SATs vs. inter-SATs in 18 recordings used in Publications I and II.

Figure 6.3. Example of feature values of three different features in a short EEG epoch. A: Two minutes of preterm EEG (blue) with manual markings by one expert (red line, higher values correspond to SAT epochs). B–D: Smoothed feature values of NLEO (B), envelope (C) and line length (D), normalized by mean feature value in this recording to make traces better comparable, overlaid with the manual marking. All features show similar elevations during SATs. NLEO shows relatively smaller feature values during inter-SATs.
6.2.2 Variable-length segmentation

Preterm EEG is highly nonstationary, containing many quasi-periodically occurring spontaneous transients, the SATs. SATs are rather short, often lasting only 1–3 s (see Publication I), and inter-SATs are usually shorter than 10 s (see Table 2.3). In such data, fixed-length segments seem a rather suboptimal choice. Even with a short segment length of 1 s, segments often contain several classes of EEG leading to feature values reflecting this mixture and making the classification of the segment difficult. In cases of adjacent fixed-length segments, the segment length also defines the time resolution of the algorithm. For example, with 1-s segments all event durations are multiples of 1 s, which is a coarse measure of the underlying phenomena.

Some algorithms use a sliding window with fixed length for feature extraction and classification. The overlap between the windows varies. Given a reasonably short window length, it is the overlap that defines the time resolution of the algorithm. It could be said then that the algorithm does not classify each window but instead finds the transitions between the classes. For example, if the window length is 1 s and overlap is 90%, the time resolution of the algorithm is 0.1 s and the event durations are multiples of 0.1 s.

In our algorithm (see Publication II), we have made classification sample by sample, giving our algorithm a time resolution equal to the sampling frequency of 256 Hz. However, the feature which we used in classification is the smoothed value of NLEO output calculated as average in a 1.5-s window. Hence, our approach could be equally described as a sliding window of 1.5 s with overlap of 255/256.

As Galicki et al. (1997) point out the EEG during SATs is not stationary as such. Therefore, methods based on finding pseudo-stationary epochs of EEG (e.g. Krajča et al. 1991) might lead to division of each SAT into several segments. Galicki et al. (1997) circumvented this problem by searching for typical temporal power profiles. The simple methods proposed in recent years (Publications I and II, Jennekens et al. 2011, Koolen et al. 2014) seem not to have the problem of splitting the SATs into several segments. This might be due to the temporal smoothing used in most methods, or to the binary classification designs which search only for high- and low-energy epochs.

6.2.3 Fixed threshold

Amplitude values in EEG are influenced by many factors including the electrode distance in bipolar recordings and the position of the electrodes relative to different brain regions, as shown for aEEG by Quigg and Leiner (2009). Technically, different amplifiers use quite different initial filters and the low-frequency cut-off of the device in particular have a dramatic effect on SAT amplitudes. Physiologically, the dynamic changes in appearance of SATs and inter-SATs, due to maturation for example, are significant. The distinction between SAT and inter-SAT waveforms becomes gradually more unclear until SATs finally disappear after term age.
All these factors make the use of fixed thresholds in the detection of SATs and inter-SATs a problematic choice. At least limitations should be clear to the users of the algorithm, and it should be used only for infants of appropriate age. In case of differences in montage or inbuilt filters of the recording device, normalization procedure such as used in Publication III prior to analysis may be helpful.

However, adaptive thresholds are also problematic. They are often based on some statistic measure of the feature used in classification, calculated separately in each patient or even dynamically during the recording (see e.g. Koolen et al. 2014). The assumption is that the signal contains both SATs and inter-SATs. In very sick neonates however, the EEG may be quite flat and contain hardly any SATs. Use of adaptive thresholds on such data could lead to classification of noise or artifacts as SATs. Thus, the algorithm would not work in a robust way. Fortunately, a nearly flat EEG should be visually striking enough to lead to further inspection irrespective of values generated by SAT detection or a similar algorithm describing the EEG characteristics.

One aspect affecting the choice of threshold is the contrast between the classes using the chosen feature. In Chapter 6.2.1 we have shown that the output of NLEO shows very high contrast between SAT and inter-SAT periods (median SAT/inter-SAT ratio 39.1). High contrast makes the detection algorithm more robust: one fixed thresholds can work well with different amplitude levels.

Therefore, when using a high contrast feature in classification of SATs in recordings by one device and montage only, the use of a fixed rather than an adaptive threshold is, in my opinion, preferable.

6.2.4 Optimization and validation

Proper optimization and validation of any segmentation or classification algorithm requires a gold standard dataset with known segment labels. In studies preceding our Publication I, such datasets exist. However, none of these known early studies evaluate the accuracy of manual markings. They thus rely on markings by one expert as the ground truth.

In Publication I, we showed that there were significant differences in the perception of SATs and inter-SATs between the three experts involved. Especially the beginnings and ends of SATs were defined variably. In case of several SATs in the sequence, this also affected the number of SATs and inter-SATs detected. We hence constructed our gold-standard dataset based only on samples marked unanimously either as SATs or inter-SATs by all experts. We also performed all performance measurements sample by sample and not event-wise, because it is very difficult to define systematically what the correct detection of an event is.

Gratifyingly, in burst-detection algorithms published after our Publication I, manual markings are mostly done by two (Jennekens et al. 2011, Mitchell et al. 2013, Koolen et al. 2014) or three (Bhattacharyya et al. 2011) experts.

In the optimization of the algorithms, different approaches are followed. In Publication II, we maximized the ADR of the algorithm by comparison to the
Discussion and future perspectives

gold standard. Different combinations of the four adjustable parameters were studied in a systematic way. Jennekens et al (2011), with three possible patterns (IBI, burst and continuous activity), maximized the joint sensitivity of the algorithm in its detecting of each of the classes. Similar to Publication II, detections were compared with agreed annotations of the markers. Three parameters were optimized, each separately. It is not stated which values were used for the other two parameters when optimizing one of them. Also Bhattacharyya et al. (2011) use unanimous burst detections as gold standard. However, the “correct” burst duration is defined as the time between the beginning of the first marking in time and the end of the last marking in time. Different performance metrics are used to find the best possible structure of SVM and the best possible subset of features for classification. Mitchell et al. (2013) modified the patterns used in template matching “to give the best agreement between the algorithm and the electroencephalographers”; no further details are given. To optimize their algorithm Koolen et al. (2014) use ROC curves calculated separately for each of the raters. Only one of the parameters of the algorithm is optimized, other parameters are defined heuristically.

As pointed out in Publication II, an algorithm should be evaluated with data not used in the training of the algorithm. In this way, generalization performance of the algorithm is not exaggerated. In Publication II, we used leave-one-out cross validation to evaluate the generalization performance of the algorithm. Optimization was repeated 18 times, each time leaving out one of the recordings, which then was used as test set for the current, optimal parameter values. In this way, all available data could be utilized for both optimization and performance evaluation.

In most recent publications, evaluation is conducted using a separate test set, which often consists of only a few recordings. Jennekens et al. (2011) use four recordings and let two experts validate the detection results of the algorithm. Accuracy of the onset or offset of an event was not considered. Events are considered correctly classified if both experts agree with the automated classification. Mitchell et al. (2013) report confusion matrices based on the comparison of automated classification with each of the two raters separately as well as on the comparison of automated classification with those events rated in agreement by both raters. Bhattacharyya et al. (2011) report several performance metrics but it is not quite clear how the evaluation is done. It is probable that the detections were compared with gold standard event-wise. Koolen et al. (2014) perform the analysis sample-by-sample, reporting sensitivity, specificity and accuracy both separately and pooled together for training set and validation set. They say that performance evaluation is conducted separately using one set of manual markings at time as gold standard, however, only one value for each metric in each dataset is reported. It can be speculated that results using the two markings were averaged for reporting.

In conclusion, the optimization and validation methodology used in Publication II has had a positive impact on the methodology used in subsequent publications on burst detection. Compared with the recent papers discussed
above, Publication II uses the available data more efficiently and studies the parameters more systematically.

6.2.5 Comparison of detection performance between methods

The performance of the detection algorithm in Publication II was found to be exceptionally good with mean sensitivity of 97% and mean specificity of 95%. Therefore, it would seem to outperform the algorithm by Jennekens et al. (2011), which classified the EEG in three possible classes, IBI, burst and continuous EEG, and obtained for these mean sensitivity of 80%, 90% and 97%, respectively.

The best way to compare detection performance between algorithms is, however, to apply them on the same dataset of EEG and gold-standard markings. Two attempts to compare our method with methods published thereafter exist (Bhattacharyya et al. 2011, Koolen et al. 2014).

A general problem arising from this kind of comparison is that when algorithms are re-implemented based on publications important details may be overlooked. A more open publication of the scripts would be of great benefit to the scholarly community.

Bhattacharyya et al. (2011) include an NLEO-based method in their detector comparison quoting both Särkelä et al. (2002) and Publication I of this Thesis. This is understandable, as the commercial algorithm studied in Publication I is said to be an implementation of the algorithm in Särkelä et al. (2002). There are however important differences. The method used in Bhattacharyya et al. (2011) seems to be an implementation of the method in Särkelä et al. (2002). Unfortunately, the improved and optimized method as presented in Publication II of this Thesis is not included in the comparison.

Bhattacharyya et al. (2011) obtain a very low sensitivity of 29% and a high specificity of 92% for the NLEO-based method. The poor performance is assumed to be caused by the static thresholds of the method. The criticism of static thresholds is adequate; however, two technical aspects should be considered. Part of the dataset used in Bhattacharyya et al. (2011) was sampled with 2000 Hz and a part with 150 Hz. As mentioned in 6.2.1, changes in sampling frequency affect the values of NLEO output, resulting in smaller values when higher sampling frequencies are used. Additionally, the feature used in classification in Bhattacharyya et al. (2011) is the sum and not the average of NLEO outputs in a sliding window of constant length. Hence, the sensitivity of the algorithm might be entirely different if all data were resampled to 256 Hz prior to analysis.

Koolen et al. (2014) compare their line-length-based method with the optimized NLEO method in Publication II, as well as with the envelope-based method of Jennekens et al. (2011). Koolen et al. (2014) report a very high sensitivity of 93% and 95% (in training and test sets, respectively), and a rather low specificity of 68% and 74% for the NLEO method. Respective values for the line-length method are 85% and 82% (sensitivity) and 83% and 89% (specificity), and for envelope-based method 81% and 85% (sensitivity) and 78% and 80% (specificity). It should be noted that the training set was used to op-
timize the parameters of the line-length method but that for NLEO and envelope based methods, fixed thresholds as published previously were used.

The performance values in Koolen et al. (2014) indicate that all three studied algorithms can be used to detect SATs with relatively high accuracy. The conclusion of the superiority of the line-length method over NLEO should, however, be questioned. In the comparison, the NLEO-method had superior sensitivity but relatively low specificity. As an example of low specificity, an EEG segment with almost ongoing high-amplitude activity is shown (Fig. 6b of Koolen et al. 2014). The NLEO detected these periods as bursts while the clinicians did not. Here, it seems that low specificity is a result of different definitions of burst and IBI in our group and the group of Koolen rather than a real failure of the algorithm. Another example of varying definitions is that Koolen et al. (2014) only consider IBIs longer than 2s. In our algorithm, no such constraint exists.

Each of the three simple algorithms recently proposed for SAT detection (Publication II, Jennekens et al. 2011, Koolen et al. 2014) includes different preprocessing steps such as filtering and a multitude of additional constraints such as minimum duration. Every detail of the algorithm affects the detection results and it is not easy to identify which part of the algorithm is necessary or superfluous. However, the input feature of the algorithm is definitely an important factor in defining the detector performance. Chapter 6.2.1 shows that NLEO outperforms both envelope (Jennekens et al. 2011) and line length (Koolen et al. 2014) in separating periods of SAT from inter-SAT periods.

Finally, in all studies quoted above, detection performance was defined as the proportion of correctly classified samples or events in a comparison with a manually marked gold standard. A real validation test of the methods will be whether they are clinically usable and useful as well as robust in different situations. None of the proposed methods is widely used as yet and it remains to be seen whether one can gain the critical mass of users necessary for broad clinical testing.

### 6.3 Artifact rejection

For the usability of an automated detection of SATs, a way of handling the artifacts is essential. As summarized in Chapter 3.4, several attempts of artifact rejection have been made; however, most of the approaches are heuristic, and include no optimization.

Many artifacts stand out by high amplitudes and/or quick transients. In Publication III, we developed a scheme for rejecting short periods of EEG based on RMS values. The threshold of the artifact rejection was defined in relation to the median RMS of the studied epoch as a whole. The reason for this choice was the strikingly similar, skewed distribution of normalized RMS values in all studied recordings. The advantage of using the median is that the method adapts to the data and the threshold is not affected by artifacts as long as they cover less than 50% of the data. However, the median can only be calculated
Discussion and future perspectives

afterwards and therefore the method is not directly applicable to real-time monitoring, necessitating further improvement to the method.

Jennekens et al. (2011) implemented a very similar method based on a skewed distribution in total band power values in different frequency bands. They decide to prune the data by rejecting 5% of the 1-s segments with highest band powers. The problem inherent in this approach is that it does not adapt to signal quality. Even if there are no artifacts, 5% of the data are rejected, while, on the other hand, if the recording contains lots of artifacts, only the coarsest are rejected.

Bhattacharyya et al. (2013) have constructed a rare dataset of manually labelled artifacts and hope to find the features best separating neonatal bursts from artifacts when using a SVM for classification. After an elaborate, systematic study of a large number of features, they propose the combined use of 10 features as inputs to an SVM separating bursts and artifacts. The artifacts are thus assumed to occur only in periods previously detected as bursts by a different SVM.

Notably, Bhattacharyya et al. (2013) compare their proposed methodology for combined burst and artifact detection with the method of Särkelä et al. (2002), again quoted together with Palmu et al. (2010b) (see above, 6.2.5). The proposed method is found to be superior in performance. This result might well be true. However, the technical concerns discussed in 6.2.5 continue to apply.

6.4 Use of detection results in the analysis of vigilance stages

In our analysis of long-term EEG recordings by our SAT detection method, we noticed that the SAT% values show remarkable fluctuation with time, sometimes resembling an almost periodic signal. This made postulate that SAT% might reflect the fluctuation of vigilance states, often referred to as sleep wake cycling (SWC). The ability of a preterm infant to maintain SWC is an important prognostic feature. SWC is seen in infants with relatively stable clinical status (Hayes et al. 2007, Olischar et al. 2007, Kidokoro et al. 2010, Weisman et al. 2011). Physiologically, sleep wake cycling is only possible if wide brain networks work together (Villablanca 2004, Karlsson et al. 2005). Appearance of SWC has also a relationship with later neurodevelopmental outcomes (Bowen, Paradisis & Shah 2010, Klebermass et al. 2011, Kidokoro et al. 2012, Wikström et al. 2012, Natalucci et al. 2013). However, visual analysis of SWC based e.g. on aEEG is highly subjective and dependent on the expertise of the staff available.

Automated analysis of SWC could bring additional benefit from the use of long-term EEG monitors in NICU. A prerequisite for SWC analysis is the identification of one or several features with discriminative power between vigilance stages.
6.4.1 Features for SWC analysis

In Publication III, we showed that mean SAT% values calculated in 20-s epochs differed significantly between vigilance stages. Statistical differences could be shown on both individual and group levels.

Pan and Ogawa (1999) completed a similar analysis of several features with the goal to find those with highest discriminative power. In their study, vigilance stages were scored visually based on respiratory signal as well as body movements and rapid eye movements analysed from video recordings. Pan and Ogawa scored 30-s epochs as either quiet sleep (QS), active sleep (AS), indeterminate sleep (IS) or wakefulness (AW). Seven EEG features were calculated, based partly on an auto-regressive (AR) model. Resulting feature values were compared between vigilance stages separately for different age groups.

Pan and Ogawa showed significant differences in Minimum Akaike information criterion, total power and discontinuity between QS and AS as well as between AS and AW in all age groups. Power in the delta band showed significant differences in most age groups.

Discontinuity comes very near to the SAT% as used in Publication III. Discontinuity was shown to be greater in QS than in AS. This supports our result that SAT% is smaller in deep NREM sleep than in REM sleep. Pan and Ogawa also show constantly smaller discontinuity values in AW than in AS. This differs from our results, because in our dataset the SAT% values during wakefulness could be either larger or smaller than during REM sleep. It should be noted, however, that the dataset of Pan and Ogawa includes recordings from infants with CA of 30–39 weeks, while in Publication III the recordings were done at CA 26–32 weeks. Since the EEG of preterm infants changes rapidly during maturation, the results of the studies are not directly comparable.

Turnbull et al. (2001) observed higher power values in the theta (4–8 Hz) frequency band co-occuring with manually marked epochs of trace alternant. The paper does not, however, give any suggestions of how to quantify the fluctuations of power values.

Paul et al. (2003) studied several quantitative features of preterm and fullterm EEG to find which differed between active sleep and quiet sleep. The EEG was first segmented into quasi-stationary segments and then the values of the features, calculated from each segment, were averaged in three voltage classes. Since preterm EEG consists of SAT and inter-SAT periods, the results of the voltage class with smallest voltages can be associated with inter-SATs and the voltage class with highest voltages can be associated with SATs, although this was not the writers’ intention. The results of the middle-voltage class can contain data from both SATs and inter-SATs.

In preterm infants, power in delta, theta and alpha bands was found to be smaller in quiet sleep than in active sleep in all channels in lowest voltage class (inter-SATs). This result corresponds to the schematic presentation in Fig. 2.2, in which the oscillations in inter-SAT periods have much smaller amplitudes during non-REM (quiet) sleep. In the highest voltage class (SATs), the power of several frequency bands was significantly higher in quiet sleep than active
Discussion and future perspectives

This result could be interpreted as showing that SATs have higher amplitudes during quiet sleep, even though they appear more seldom.

The proportion of time covered by the lowest voltage class (inter-SATs) was significantly smaller in AS than in QS, and correspondingly, the proportion of time covered by the highest voltage class was significantly larger in AS than in QS. Paul et al. (2003) also found the number of high-voltage class segments to be larger in AS than in QS.

Myers et al. (1997) calculated high frequency power (12–24 Hz) of preterm EEG and low frequency power (0.03–0.20 Hz) in rectified versions of the same signal. The low frequency band corresponds to cycle length of 5–30 s and hence 2–12 bursts per minute. Values of high- and low-frequency power and their ratios were compared in AS and QS. Myers et al. (1997) found higher high-frequency power values in AS than in QS, while the low-frequency power was found to be higher in QS than in AS. The ratio between low-frequency and high-frequency power was statistically an even more powerful feature.

In conclusion, several features calculated from preterm EEG have shown to vary with sleep stages. One feature studied in several publications is the proportion of time covered by bursts, called SAT\% in Publication III and time percentage of highest voltage class occurrence in Paul et al. (2003). It is shown to be smaller in QS than in AS. Equivalently, discontinuity is shown to be greater in QS than in AS (Pan, Ogawa 1999).

### 6.4.2 Analysis of brain activity cycling

After finding one or more features exhibiting differing values between vigilance stages, a measure reflecting the stability and strength of vigilance-stage fluctuations is necessary. In this context it is important to note that the EEG does not need to be segmented into periods of different vigilance stages. It is not the stages but their fluctuation that is of interest.

In Publication III, we discovered that fluctuation of SAT\% match strikingly with the fluctuation of vigilance stages defined by polysomnographic recording (see Fig. 3 in Publication III). We hypothesized that these fluctuations could be measured utilizing SAT\% curves calculated from adjacent short segments of EEG.

In Stevenson et al. (2014) our group further developed this idea. SAT\% curves were calculated in 4-h epochs of EEG with temporal binning of 5 minutes. Six different frequency-domain representations were calculated from each curve. These were described by three different statistics, giving 18 possible measures of brain activity cycling (BAC). Each of the SAT\% curves was also evaluated visually to define degree of cyclicity on a scale 1–10.

The capacity of each measure to describe BAC was studied by linear and in some cases non-linear correlation analysis with grades of visually assessed cyclicity. There were strong correlations. Three measures showing strongest correlations were band energy estimated from a periodogram, relative band energy calculated from nonstationary frequency marginal (Stevenson et al. 2012) and g-statistic (normalized maximum of frequency domain presentation) calculated from nonstationary frequency marginal. The frequency band
of interest was defined as $f = 1/3600 \text{Hz} - 3/3600 \text{Hz}$, corresponding to cycle length of 20–60 minutes. For a real-life implementation of the method it is of particular importance that measures mentioned above were shown to work in a robust way even in compromised situations with missing data. In long-term recordings in NICUs, missing data due to artifacts, care procedures etc. are almost inevitable components of every recording.

Thus the measures of BAC developed in Stevenson et al. (2014) are highly promising. Further clinical testing is needed to prove their potential in monitoring preterm infants.

### 6.5 Correlations between activity and structural growth

In Publication IV, we correlated measures calculated from EEG during the first postnatal days with measures of brain structure and growth calculated from serial MRI acquisitions. Even though the dataset was relatively small, we found several significant correlations. In general more activity was correlated with better brain growth. Our findings emphasize the importance of monitoring the brain activity in preterm infants in the vulnerable time frame they spend on NICU. EEG metrics such as those proposed in Publication IV might help in assessing the activity and even in optimizing the individual care of the infants including medication in a way that gives them the best possible opportunities for unbiased development.

### 6.6 Future perspectives

The SAT detection method developed in this Thesis has been proven to be useful in the analysis of early brain activity that relates to both functional and structural maturation of the preterm brain, as well as to the sleep states of the preterm brain. There are, however, still several barriers to this detection method’s wide use in clinical trials.

One such necessary step is the validation of derived measures and their correlation with other measures such as GA and CA or later neurocognitive outcome using a larger dataset. Collection of a large dataset of healthy preterm infants is already under way by our collaboration partners (Dr. Katrin Klebermass and colleagues) in Vienna. Repeated long-term EEG recordings are carried out on each infant and neurocognitive development of the infants is followed until they reach two years of age. The dataset may enable us to publish normative values for several SAT derived indices as well as measures of BAC.

Even without normative values, quantitative measures based on SAT detection could be used to monitor changes in the state of an individual infant. For this sake, however, the analysis methods should be implemented in a real-time manner and made accessible in NICUs. As long as the measures are still in the testing phase, a separate computer could be used to analyse the EEG and to visualize the results in a user-friendly manner. In the long term, the analysis should become part of basic recording software of major vendors of EEG devices.
7. Conclusion

Spontaneous activity transients (SATs) are a distinctive feature of preterm EEG. In this Thesis, an algorithm was developed that automatically detects SATs with high accuracy.

Special attention was given to optimization and validation of the algorithm. Several methodological decisions, such as the use of more than one manual marking in construction of gold standard, have already been adopted by other research groups.

Reliable automated detection of SATs is a starting point. In the future the detection of SATs may form a routine part of the monitoring paradigms in NICUs, with SAT% serving both as a trend measure and a feature for the detection of brain activity cycling. Brain monitoring may become as normal as monitoring a heart rate, improving the care of vulnerable preterm infants, and thus their prospects for a life without disability.
References


Palmu, K 2008, Spontaanien aktiviteettipurskeiden automaattinen tunnistaminen keskoslasten aivosähkömittauksesta (Detection of spontaneous activity transients in electroencephalography of preterm infants), Helsinki University.

References


Paul, K, Krajiča, V, Roth, Z, Melichar, J & Petranek, S 2003, ‘Comparison of quantitative EEG characteristics of quiet and active sleep in newborns’, *Sleep Medicine*, vol. 4, no. 6, 543–552.


Wong, L 2008, Quantitative continuity feature for preterm neonatal EEG signal analysis, University of Auckland.


Preterm infants may spend months in the neonatal intensive care unit. During this time, their brain wires itself for the rest of their lives. Any adverse events related to neurological illness or challenges in treatment may disturb this process.

This Thesis develops methodology for bedside monitoring of brain function in preterm infants by electroencephalography (EEG). The main feature of preterm EEG is intermittently occurring spontaneous activity transients (SAT), which drive the development of neuronal connections. The work in this Thesis developed and optimized an algorithm that automatically detects these events from EEG. Further, it is shown that measures based on SAT detection have clinical correlates. The methodology may contribute to better brain care of preterm infants by providing real-time information about brain health to treating clinicians.