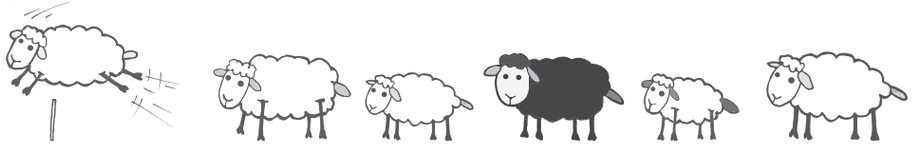
Home Sleep Monitoring



Pedro Fonseca

the **TU/e** sleep series

HOME SLEEP MONITORING

PEDRO FONSECA

Home Sleep Monitoring

PROEFSCHRIFT

ter verkrijging van de graad van doctor aan de Technische Universiteit Eindhoven, op gezag van de rector magnificus prof.dr.ir. F.P.T. Baaijens, voor een commissie aangewezen door het College voor Promoties, in het openbaar te verdedigen op woensdag 18 April 2018 om 11:00 uur

door

Pedro Miguel Ferreira dos Santos da Fonseca

geboren te Lissabon, Portugal

Dit proefschrift is goedgekeurd door de promotoren en de samenstelling van de promotiecommissie is als volgt:

voorzitter:	prof.dr.ir. A.B. Smolders
1e promotor:	prof.dr. R.M. Aarts
co-promotor:	prof.dr. S. Overeem
leden:	prof.dr. R. van Ee (RUN)
	prof.dr. D.A.A. Pevernagie (Universiteit Gent)
	prof.dr.ir. L.M.G. Feijs
reserve:	prof.dr.ir. J.W.M. Bergmans

Het onderzoek dat in dit proefschrift wordt beschreven is uitgevoerd in overeenstemming met de TU/e Gedragscode Wetenschapsbeoefening.

Home Sleep Monitoring

Pedro Fonseca

Home Sleep Monitoring

A catalogue record is available from the Eindhoven University of Technology Library
ISBN: 978-90-386-4460-8

Cover design by Pedro Fonseca

© Copyright 2018, Pedro Fonseca

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, without the prior written permission from the copyright owner.

One day, some years ago, I arrived home and told you I wanted to do this. You asked me, "are you sure?" but your eyes added, "after what I've been through to get mine?"

And yet, it were the examples of your dedication, your perseverance, and your endless pursuit for improvement that guided me to this point.

So this one, dearest Ellen, is for you.

Summary

Sleeping habits have changed throughout history following human progress and technological development. While we slept an average of nine hours per day in 1900, today we hardly get six and a half. To make matters worse, epidemic stress and obesity are contributing to the quick rise of sleep disorders such as insomnia and sleep apnea, with an important percentage of the population either suffering from a clinically relevant disorder, or nonetheless reporting dissatisfaction with sleep.

Luckily, perception and recognition of the importance of sleep is changing. Many of the symptoms of sleep deprivation and other sleep disorders are now recognized as actual sleep-related issues, leading to an earlier diagnosis and timely treatment. The assessment of sleep disorders is traditionally performed in dedicated sleep laboratories, where trained technicians apply a plethora of sensors to monitor overnight physiological processes of the subjects under investigation. However, it has some limitations, such as lack of applicability for extended periods of time, or difficulty to use in a home environment. Progress in sensor technology over the last decades has partially alleviated this problem, with the development of ambulatory monitors for some sleep-related parameters. The simplest and arguably most popular of such ambulatory monitors, actigraphy, consists of the measurement of gross body movements by means of a limb-worn accelerometer. Given its relative comfort and ease of use, actigraphy has met relative success in the clinical assessment of sleep-wake disturbances. Interestingly, it was also the first technique to make its way to non-clinical, consumer-grade sleep tracking (CST) devices.

The recent and growing interest in sleep by the general public has led to an increasing demand for sleep tracking devices which can be used without assistance or intervention of a physician. Although the attempt to provide users with information about a period mostly characterized by unconsciousness can seem commendable, there is light and shadow when it comes to the widespread use of these devices. They have the potential to play an invaluable role in educating consumers on the importance of having enough sleep time, and on how to introduce and maintain healthy sleep routines. But the fact remains that they are based on a surrogate measure of sleep, and furthermore that most of their marketing claims lack proper validation or are outright implausible. This aspect in particular may have seriously negative consequences for individuals with sleep-related concerns - but yet undiagnosed disorders - who will seek an explanation with these devices before consulting with medical professionals, or worse, for asymptomatic individuals with a hidden primary sleep

disorder.

Despite the low level of evidence for CST devices, the physiological basis behind modern sensors which have the ability to measure cardiorespiratory activity during the night is, in theory, solid. Research has shown a strong link between sleep, sleep disorders, and autonomic nervous system (ANS) activity, which is, in turn, correlated with measures such as heart rate variability (HRV) and respiratory rate variability (RRV) that could be derived from such sensors. In particular, the topic of automatic detection of different sleep stages, usually divided in Wake, REM sleep, 'light sleep' (N1 and N2) and 'deep sleep' (N3) has received increasing attention over the last decade, but performance remains modest. Accordingly, the first objective of this thesis is to develop methods to increase the performance of sleep stage classification and bring it closer, for healthy subjects, to the gold standard polysomnography (PSG). This work is separated in three parts.

The first part deals with the analysis of physiological signals that allow features discriminative of sleep stages to be extracted. Chapter 2 describes a method to precisely localize QRS complexes in electrocardiography (ECG) signals, essential for the accurate measurement of HRV. The proposed algorithm achieves a very low localization error, and is robust to severe signal degradation, common in ambulatory acquisitions of ECG. Chapter 3 describes a method which allows body movements to be estimated from artifacts in ECG and respiratory inductance plethysmography (RIP) signals. It alleviates the need to wear an additional sensor in case cardiac and/or respiratory measures are performed. Equally important, it allows retrospective data sets for which actigraphy was not registered to be used in the development and validation of sleep staging algorithms which make use of body movement information.

The second part focuses on methods for automatic sleep stage classification. Chapter 4 describes an algorithm which uses a large set of HRV and RRV features extracted from ECG and RIP, achieving a moderate agreement for four-class sleep stage classification (Wake/REM/N1 and N2/N3) of healthy subjects.

The third part explores the use of advanced probabilistic methods to further improve classification performance. Chapter 6 explores the use of Markov chains which exploit the structured nature of sleep and can use features discriminative not only of different sleep stages, but also of stage transitions and sleep stage continuity. This chapter shows that discriminative models like conditional random fields (CRF) are better suited for this task when compared to other popular probabilistic classifiers such as the simple epoch-based linear discriminants used in Chapter 4, and Hidden Markov Models (HMM) often described in literature. CRF classifiers are further extended in Chapter 7 to the problem of sleep stage classification, achieving a performance improvement over the classifier described in Chapter 4. In addition, this chapter shows how these classifiers are better suited for the more complex task of five-class classification (Wake/N1/N2/N3/REM), the standard for sleep scoring in traditional PSG.

The second objective of the thesis is to prove that these methods can work with sensors that are

less obtrusive. Although ECG and RIP provide the means to measure ANS characteristics during sleep, they are not exactly compatible with long-term, unassisted monitoring. Chapter 5 validates a sleep stage classifier which uses HRV features derived from a photoplethysmography sensor, achieving a reasonable performance on four-class classification of healthy middle aged adults and better performance than actigraphy for Sleep/Wake classification.

Finally, and given the limited amount of evidence regarding the use of this technology in clinically relevant populations, the third and final objective of the thesis is to evaluate whether these methods can be applied in subjects suffering from a sleep disorder. This will provide first evidence of whether the technology has the potential to ever be used in clinical practice. Chapter 7 explores the performance of CRF using HRV and RRV features on patients diagnosed with obstructive sleep apnea, achieving a reasonable performance for four-class classification.

With a veritable explosion in the offer and adoption of wearable monitors, many of which emphasize sleep and sleep quality as one of the key health factors, together with an increasing awareness that sleep is a fundamental element in balanced living, it is inevitable that home sleep monitoring technology will reach widespread use in the next few years, perhaps before it is even adopted and endorsed by (medical) sleep associations. The research agenda in this area should focus on validating and further improving it on a wider patient population and also on understanding its fundamental limitations. Furthermore, it should be redirected from a simple monitoring and reporting function to a tool that can give users in general, and disordered patients in particular, actionable suggestions, providing the means to help the millions of people suffering from sleeping disorders or complaints.

Nederlandse Samenvatting

Slaapgewoontes zijn door de geschiedenis heen veranderd onder invloed van menselijke vooruitgang en technologische ontwikkelingen. Terwijl we in 1900 gemiddeld negen uur per nacht sliepen, slapen we vandaag nog amper zes en half uur. Daarbij komt nog dat verzwarende omstandigheden zoals epidemische stress en obesitas extra bijdragen tot de snelle toename van slaapstoornissen zoals slapeloosheid en slaapapneu en dat daarnaast een belangrijk percentage van de populatie ofwel aan een klinisch relevante stoornis lijdt ofwel rapporteert dat men ontevreden is met zijn of haar slaapkwaliteit.

Gelukkig is de perceptie en de erkenning van het belang van slaap aan het veranderen. Vele van de symptomen van slaap tekort en andere slaapstoornissen worden nu erkend als slaapgerelateerde problemen waardoor men tot een vroegere diagnose en meer tijdige behandeling komt. De beoordeling van slaapstoornissen wordt traditioneel uitgevoerd in gespecialiseerde slaapcentra waar getraind personeel een verscheidenheid aan sensoren gebruikt om nachtelijke fysiologische processen te monitoren. Hieraan zijn echter beperkingen verbonden zoals de onmogelijkheid om deze sensoren toe te passen over langere tijdsperioden of de inherente uitdagingen om ze te gebruiken in een thuissituatie. Vooruitgang in sensor technologie in de laatste tientallen jaren heeft dit probleem gedeeltelijk opgelost in de vorm van de ontwikkeling van ambulante monitors voor slaapgerelateerde parameters. De eenvoudigste en, bediscussieerbaar, de meest populaire van deze ambulante monitors, actigrafie, bestaat uit het registreren van de bewegingen van de ledematen met behulp van een accelerometer. Door het relatieve comfort en gebruiksgemak is actigrafie uitgegroeid tot een veel gebruikte methode voor het klinisch beoordelen van slaap-waak stoornissen. Interessant is dat het ook de eerste methode is die gebruikt wordt in niet-klinische consumer-grade sleep tracking (CST) apparaten.

De recente en groeiende interesse van het grote publiek in slaap heeft geleid tot de toenemende vraag naar slaap monitoring apparaten die kunnen gebruikt worden zonder de hulp of interventie van een arts. Hoewel de inspanning om gebruikers te informeren over een dagdeel dat gekarakteriseerd wordt door laag bewustzijn lovenswaardig is, zijn er voor- en nadelen als het aankomt op het wijdverspreide gebruik van deze apparaten. Ze hebben het potentieel om een rol van onschatbare waarde te spelen bij de educatie van consumenten over het belang van voldoende slaap en daarnaast hoe de consument gezonde slaaproutines introduceert en in stand houdt.

Het blijft echter een feit dat deze observaties gebaseerd zijn op een surrogaat meting van slaap; en bovendien ontbreken de meeste marketing claims aan voldoende validatie of zijn ze ronduit onaannemelijk. Vooral dit aspect kan zware negatieve consequenties hebben voor individuen met slaapgerelateerde zorgen -zonder gediagnosticeerde stoornissen- die een verklaring zoeken in deze apparaten zonder een arts raad te plegen, of erger, voor asymptomatische individuen met een nog niet vastgestelde primaire slaapstoornis.

Ondanks het weinige bewijs voor het nut van CST apparaten is de fysiologische basis achter moderne sensoren die cardio-respiratoire activiteit kunnen meten 's nachts, in theorie, solide. Onderzoek heeft een sterkte link aangetoond tussen slaap, slaapstoornissen en de activiteit van het autonome zenuwsysteem (AZS) die op zijn beurt gecorreleerd is met factoren zoals heart rate variability (HRV) en respiratory rate variability (RRV), beide observaties die afgeleid kunnen worden van zulke sensoren. Het onderwerp van automatische detectie van de verschillende slaapstadia, normaal onderverdeeld in waak, REM slaap, 'lichte' slaap (N1 en N2) en 'diepe' slaap (N3), heeft toenemende interesse gewekt in de laatste tien jaar, maar de prestaties blijven beperkt. Overeenkomstig hiermee is het eerste doel van dit proefschrift om een methode te ontwikkelen die de prestatie van slaapstadium classificatie verbetert en deze dichter brengt, althans voor gezonde proefpersonen, bij de gouden standaard polisomnografie (PSG). Dit werk werd verdeeld in drie delen.

Het eerste deel behandelt de analyse van fysiologische signalen die toelaten om de verschillende kenmerken die slaapstadia bepalen te onderscheiden. Hoofdstuk 2 beschrijft een methode die in staat stelt om nauwkeurig QRS complexen in EKG signalen te lokaliseren wat essentieel is voor de correcte bepaling van HRV. Het vooropgestelde algoritme behaalt een zeer lage foutenmarge in het lokaliseren van QRS complexen en is robuust bestand tegen ernstige signaal degradatie, iets wat vaak voorkomt bij ambulant opgenomen EKGs. Hoofdstuk 3 beschrijft een methode die toelaat om bewegingen in te schatten op basis van artefacten in EKGs en respiratory inductance plethysmography (RIP) signalen. Dit verlicht de nood om een additionele sensor te dragen indien hart en/of respiratoire metingen gewenst zijn. Even belangrijk is dat deze sensor toe laat retrospectieve data sets, waarin geen actigrafie geregistreerd werd, te gebruiken in de ontwikkeling en validatie van algoritmes die gebruik maken van informatie uit bewegingen.

Het tweede deel van het proefschrift behandelt methodes voor automatische classificatie van slaapstadia. Hoofdstuk 4 beschrijft een algoritme dat in staat stelt om in gezonde personen de vier klassen van slaapstadia te definiëren (waak/REM/N1 en N2/N3) op basis van een grote set van HRV en RRV features verkregen uit EKG en RIP.

Het derde deel van het proefschrift exploreert het gebruik van geavanceerde probabilistische methodes om de classificatie prestatie verder te verbeteren. Hoofdstuk 6 verkent het gebruik van Markov chains die de gestructureerde aard van slaap uitbuit en kenmerken gebruikt, niet alleen van de verschillende slaapstadia, maar ook van de transities tussen stadia en de continuiteit van de slaapstadia. Dit hoofdstuk toont aan dat discriminatieve modellen zoals conditional random fields

(CRF) beter geschikt zijn voor deze taak in vergelijking met andere, veel gebruikte, probabilistische classifiers zoals de eenvoudige epoch-gebaseerde lineaire discriminanten, gebruikt in Hoofdstuk 4, en Hidden Markov Models (HMM) die vaak beschreven worden in de literatuur. CRF classifiers worden verder uitgediept in Hoofdstuk 7 om het probleem van de bepaling van slaapstadia te behandelen waarbij een verbetering van de prestaties van de classifier beschreven in Hoofdstuk 4 wordt voorgesteld. Daarnaast beschrijft dit hoofdstuk hoe deze classifiers beter geschikt zijn voor de meer complexe taken zoals vijf stadia classificatie (Waak/N1/N2/N3/REM), de standaard voor slaap scoring in de traditionele PSG.

Het tweede doel van dit proefschrift is om aan te tonen dat deze methoden werken met sensoren die meer comfortabel zijn in het dag dagelijks leven. Hoewel EKG en RIP de mogelijkheid bieden om AZS karakteristieken te meten tijdens slaap, zijn ze niet compatibel met lange termijn, ongeassisteerde monitoring van slaap. Hoofdstuk 5 valideert de slaapstadia classifier die HRV features gebruikt, afgeleid van een photoplethysmographie sensor die redelijke prestaties aflevert wat betreft vier stadia classificatie van gezonde volwassenen van middelbare leeftijd en die een betere prestatie oplevert dan actigrafie voor Slaap/Waak classificatie.

Tenslotte, gegeven de beperkte effectiviteit betreffende het gebruik van deze technologie in klinisch relevante populaties, is het derde en laatste doel van dit proefschrift om te evalueren of deze methodes kunnen toegepast worden op proefpersonen die lijden aan een slaapstoornis. Dit zou een eerste bewijs leveren of de technologie het potentieel heeft om ooit toegepast te worden in de kliniek. Hoofdstuk 7 exploreert de prestatie van CRF gebaseerd op HRV en RRV features op patiënten gediagnosticeerd met obstructief slaapapneu en bereikt daarbij een redelijke prestatie voor vier stadia classificatie.

Met een enorme toename in het aanbod en de toegankelijkheid van draagbare monitors, waarvan vele benadrukken dat slaap en slaapkwaliteit sleutelfactoren zijn voor gezondheid; en de toegenomen bewustwording dat slaap een fundamenteel element is in een gebalanceerd leven, is het onvermijdelijk dat home sleep monitoring technologie een wijd verspreid gebruik zal bereiken in de komende jaren. Er is zelfs een mogelijkheid voor deze technologie om algemeen aanvaard en ondersteund te worden door (medische) slaaporganisaties. De visie in deze onderzoekstak zou zich moeten richten op de validatie en het aanbrengen van verdere verbeteringen door ze toe te passen op een bredere patiënten populatie en eveneens op het begrijpen van zijn inherente beperkingen. Bovendien zal een heroriëntatie moeten gebeuren van eenvoudige monitoring en rapportering tot een applicatie die gebruikers in het algemeen, en patiënten met slaapstoornissen in het bijzonder, advies kan geven van echte actiepunten waardoor de mogelijkheid geopend wordt om miljoenen mensen die lijden aan slaapstoornissen of klachten te helpen.

Abbreviations

AASM	American Academy of Sleep Medicine
AHI	apnea-hypopnea index
ANS	autonomic nervous system
ASMD	absolute standardized mean distance
AT	QRS detector by Afonso, Tompkins, et al
AUC	area under the curve
BCG	ballistocardiography
BFGS	Broyden-Fletcher-Goldfarb-Shanno
BMA	body movement artifact
CI	confidence interval
CNS	central nervous system
CRF	conditional random fields
CRI	cardiorespiratory interaction
CSA	central sleep apnea
CST	consumer sleep tracker
CWT	continuous wavelet transform
DFA	detrended fluctuation analysis
DFW	dynamic frequency warping
DSM-IV	Diagnostic and Statistical Manual of Mental Disorders, 4th Ed.
DTW	dynamic time warping
DVG	difference visibility graph
ECG	electrocardiography
EEG	electroencephalography
EMG	electromyography
EOG	electrooculography
EU	European Union
FPR	false positive rate
HF	high frequency band
HMM	hidden Markov models

HRV	heart rate variability
HSM	home sleep monitoring
HST	home sleep test
IBI	interbeat interval
ICC	intraclass correlation
ICD-10	International Statistical Classification of Diseases and Related Health Problems, 10th Ed.
ICSD-2	International Classification of Sleep Disorders, 2nd Ed.
ICSD-3	International Classification of Sleep Disorders, 3rd Ed.
LD	Bayesian linear discriminant
LF	low frequency band
LoA	limits of agreement
MIT-BIH	MIT-BIH Normal Sinus Rhythm Database
MSE	multi-scale sample entropy
NMI	normalized mutual information
NREM	non-REM
NS	QRS detector by Nygårds and Sörnmo
OLS	ordinary least squares
OSA	obstructive sleep apnea
PAT	peripheral arterial tone
PLMD	periodic limb movement disorder
PNS	parasympathetic nervous system
PPG	photoplethysmography
PPV	positive predictive value
PR	precision-recall
PSD	power spectral density
PSG	polysomnography
PSQI	Pittsburgh Sleep Quality Index
PTBD	PTB Diagnostic ECG Database
QRS	ECG complex defined by the sequence of Q-, R- and S-waves
REM	rapid eye movement
RIIV	respiratory-induced intensity variations
RIP	respiratory inductance plethysmography
RMSE	root mean square error
ROC	receiver operating characteristic
RRV	respiration rate variability
SD	standard deviation
SDB	sleep disordered breathing

SE	sleep efficiency
SFS	sequential forward search
SNR	signal-to-noise ratio
SNS	sympathetic nervous system
SOL	sleep onset latency
STFT	discrete short-time Fourier transform
SVM	support vector machines
SWS	slow wave sleep
TIB	time in bed
TPR	true positive rate
TST	total sleep time
TWT	total wake time
USA	United States of America
VG	visibility graph
VLF	very low frequency band
W/N1-2/N3/R	Wake/N1 and N2/N3/REM
W/N1/N2/N3/R	Wake/N1/N2/N3/REM
W/NR/R	Wake/non-REM/REM
WASO	wake after sleep onset

Contents

Summary vii

Nederlandse Samenvatting xi

List of Figures xxiii

List of Tables xxvii

Introduction 1

1 Sleep assessment 3
- 1.1 Sleep . . . 3
- 1.2 Standards of clinical sleep assessment . . . 5
- 1.3 Home sleep monitoring . . . 8
  - 1.3.1 Home sleep tests . . . 8
  - 1.3.2 Actigraphy . . . 8
  - 1.3.3 Consumer sleep tracking . . . 9
- 1.4 Cardiorespiratory activity during healthy sleep . . . 12
- 1.5 Objective and outline . . . 15

Signal analysis for sleep stage classification 19

2 A low-complexity post-processing algorithm for precise QRS localization 21
- 2.1 Introduction . . . 22
- 2.2 Materials and methods . . . 24
  - 2.2.1 QRS detection . . . 24
  - 2.2.2 Filtering . . . 24
  - 2.2.3 Localization . . . 24
  - 2.2.4 Template matching . . . 27
  - 2.2.5 Evaluation . . . 28

2.3 Results . . . 31
2.3.1 Trigger jitter . . . 31
2.3.2 Localization . . . 32
2.3.3 Template matching . . . 33
2.4 Discussion and conclusions . . . 34
2.5 Appendix A – Theoretical localization error . . . 36
2.6 Appendix B – Computational complexity . . . 38

3 Estimating actigraphy from ECG and respiratory effort 41
3.1 Introduction . . . 42
3.1.1 Body movement artifacts on ECG signals . . . 42
3.1.2 Body movement artifacts on RIP signals . . . 43
3.1.3 Objective . . . 43
3.2 Materials and methods . . . 44
3.2.1 Data sets . . . 44
3.2.2 Continuous Wavelet Transform . . . 45
3.2.3 Feature selection and regression . . . 48
3.2.4 Sleep/Wake classification . . . 49
3.3 Results . . . 51
3.4 Discussion . . . 54
3.5 Conclusions . . . 56

Sleep stage classification based on unobtrusive modalities 59

4 Sleep stage classification with ECG and respiratory effort 61
4.1 Introduction . . . 62
4.2 Materials and methods . . . 63
4.2.1 Data sets . . . 63
4.2.2 Feature extraction . . . 63
4.2.3 Classifier . . . 66
4.2.4 Feature selection . . . 66
4.2.5 Validation and evaluation . . . 68
4.3 Results and discussion . . . 68
4.3.1 Feature selection . . . 68
4.3.2 Cross-validation . . . 71
4.3.3 Comparison with state-of-the-art . . . 72
4.4 Conclusions . . . 73

5 Validation of PPG-based sleep staging in healthy middle aged adults 75
5.1 Introduction . . . 76
5.2 Materials and methods . . . 78
5.2.1 Training data set . . . 78
5.2.2 Validation data sets . . . 79
5.2.3 PPG-based sleep stage classification . . . 81
5.2.4 Comparison with PSG . . . 81
5.2.5 Comparison with actigraphy . . . 82
5.3 Results . . . 83
5.3.1 PPG-based Sleep/Wake statistics . . . 83
5.3.2 PPG-based sleep staging performance . . . 85
5.3.3 Actigraphy-based Sleep/Wake classification performance . . . 86
5.4 Discussion . . . 87

Probabilistic methods for sleep stage classification 91

6 Cardiorespiratory sleep stage detection using conditional random fields 93
6.1 Introduction . . . 93
6.2 Materials and methods . . . 94
6.2.1 Data sets . . . 94
6.2.2 Feature extraction . . . 95
6.2.3 Bayesian linear discriminant . . . 98
6.2.4 Probabilistic graphical models . . . 98
6.2.5 Training and evaluation . . . 106
6.3 Results and discussion . . . 108
6.4 Conclusions . . . 113

7 A comparison of probabilistic classifiers for sleep stage classification 115
7.1 Introduction . . . 116
7.2 Materials and methods . . . 117
7.2.1 Data sets . . . 117
7.2.2 Feature extraction . . . 118
7.2.3 Multiple-sleep stage classification . . . 119
7.2.4 Evaluation . . . 121
7.3 Results . . . 122
7.4 Discussion . . . 124
7.5 Conclusions . . . 132

Discussion **135**

8 General discussion **137**

8.1 Feature analysis 139

8.2 Subject characteristics and sleep disorders 142

8.3 Sensing modalities 148

8.4 General conclusions and future perspective 149

Bibliography **153**

List of Publications **169**

Acknowledgments **173**

Curriculum Vitae **177**

The TU/e Sleep Series **179**

List of Figures

1.1 Example of a hypnogram of a healthy subject, showing the progression of sleep stages throughout a night . . . 4
1.2 PSG with recording of EEG, EOG, EMG, nasal airflow, respiratory effort and pulse oximetry . . . 6
1.3 HST for diagnosis of sleep disordered breathing . . . 9
1.4 Wrist-worn actigraphy device . . . 10
1.5 Consumer sleep tracking device . . . 11
1.6 Box plot of activity counts and heart rate for different sleep stages of a healthy subject 11
1.7 Autonomic nervous system and some of the cardiovascular and respiratory functions served by the sympathetic and parasympathetic branches . . . 13
1.8 Box plot of HRV frequency characteristics for different sleep stages of a healthy subject 13
1.9 Box plot of respiratory characteristics for different sleep stages of a healthy subject 14
1.10 General framework of sleep stage classification and validation, with indication of the contribution per chapter . . . 16

2.1 Estimating the localization of the R wave peak for a QRS complex with an artifact . 25
2.2 Template matching after QRS detection . . . 28
2.3 Comparison of trigger jitter with and without upsampling . . . 32
2.4 Average localization, trigger jitter and sensitivity before and after localization for combinations of clipping factors and SNR . . . 33
2.5 Trigger jitter versus sensitivity with and without localization and template matching for combinations of clipping factors and SNR . . . 34
2.6 PPV versus sensitivity after template matching using different correlation thresholds 35
2.7 PPV versus false positive fraction after template matching with different thresholds 35
2.8 QRS complex for maximum localization error . . . 39

3.1 Normalized signal spectrum on periods with high and low body movement intensities for ECG and RIP signals . . . 44
3.2 Wavelet functions used for CWT . . . 46

3.3 Example of the effect of a body movement artifact on ECG and RIP signals and corresponding CWT 47
3.4 Bland-Altman plots after log transformation and after restricting the plot to points where both methods agree regarding the presence of body movements 52
3.5 Histogram of actigraphy values for Sleep and Wake classes and agreement on presence of body movements detected with surrogate and reference actigraphy 53
3.6 Confidence intervals for difference in κ obtained with different feature sets with (surrogate) actigraphy using respiratory features 55
3.7 Confidence intervals for difference in κ obtained with different feature sets with (surrogate) actigraphy using respiratory and cardiac features 55

4.1 Relation between ASMD and separation between two normally distributed classes with the same standard deviation 67
4.2 Training performance per iteration and average training performance during feature selection 69
4.3 Examples of reference and predicted hypnograms for the subject with the worst, median and best performance 72

5.1 Bland-Altman plots for SOL and WASO 84
5.2 Bland-Altman plots for TWT and TST 84
5.3 Bland-Altman plots for SE and time in N1-2 85
5.4 Bland-Altman plots for time in N3 and REM 86

6.1 Trellis diagram for process with two classes 99
6.2 Toy examples of features that are only discriminative at the state transitions, and difficult to separate linearly 101
6.3 CRF scores and corresponding classes for toy example of Figures 6.2a and 6.2b . . . 106
6.4 Pooled PR and ROC curves per classifier and classification task, for 'all' recordings . 108
6.5 Mean κ and accuracy, per classifier, per detection task for both data sets 109
6.6 Mean area under the PR and ROC curves, per classifier, per detection task for both data sets 110
6.7 Example reference and predicted hypnograms for subject closest to the average performance 111
6.8 Example reference and predicted hypnogram for subject with the best average performance 112

7.1 Average validation and training κ for a varying number of recordings in the 'regular' and 'OSA' data sets, for the LDtvp, HMM and CRFt five-class classifiers 125

7.2 Average validation and training performance κ for a varying number of recordings in the 'all' data set, for the LDtvp, HMM and CRFt five-class classifiers 125
7.3 Classification example for the subject with the highest κ in the 'regular' data set . . 128
7.4 Classification example for the subject closest to the median κ in the 'regular' data set 129
7.5 Classification example for the subject with the highest κ in the 'OSA' data set 129
7.6 Classification example for the subject closest to the median κ in the 'OSA' data set . 130
7.7 Classification example for the subject with the highest κ in the 'all' data set 130
7.8 Classification example for the subject closest to the median κ in the 'all' data set . . 131

8.1 Absolute Pearson's correlation between each pair of cardiorespiratory features . . . 142
8.2 Hypnogram of the youngest and oldest healthy subjects in our data set 143
8.3 Scatter plot of estimated κ and accuracy based on demographics and sleep statistics 146
8.4 Hypnograms for the subject with the longest SOL and the highest WASO 146
8.5 Hypnograms for the subject with AHI closest to 15, and the subject with the highest AHI 147
8.6 Performance κ versus AHI for overall and dedicated classifier 147
8.7 Classification performance (in κ) for each modality 148

List of Tables

1.1 Major categories of sleep disorders according to the ICSD-3 6

3.1 Feature sets used to train the different classifiers and compare performance with and without actigraphy, and with surrogate actigraphy 51
3.2 Correlation between estimated and reference actigraphy, and number of coefficients used from each signal 52
3.3 Mean and standard deviation of classification performance for healthy subjects, per feature set 54
3.4 Mean and standard deviation of classification performance for all subjects, per feature set 54

4.1 Pooled and 90th percentile ASMD for features selected in all iterations 70
4.2 Cross-validation performance for three and four classes 71
4.3 Confusion matrix after cross-validation 71
4.4 Performance comparison with state-of-the-art 73

5.1 Participant demographics for the data sets used to train and validate the algorithm 78
5.2 PSG sleep statistics, and bias, 95% limits of agreement, and RMSE between PPG and PSG 80
5.3 Epoch-by-epoch agreement between PSG and PPG 86
5.4 Bias, SD and RMSE in Sleep/Wake statistics between actigraphy and PSG, and between PPG and PSG in 53 recordings 87
5.5 Epoch-by-epoch Sleep/Wake classification agreement between actigraphy and PSG and between PPG and PSG in 49 recordings 87

6.1 Demographics and sleep statistics of subjects in the two sets used in the study . . . 96
6.2 Respiratory features used in the study 96
6.3 Cardiac features used in the study 97
6.4 Cardiorespiratory features used in the study 97
6.5 Comparison with classifiers reported in the literature 113

7.1 Demographics and sleep statistics of subjects in the two subsets used in the study . 118
7.2 Classification performance with classifiers trained with three and four classes 123
7.3 Pooled confusion matrix for CRFt and LDtvp classification of 'all' subjects 123
7.4 Pooled confusion matrix for CRFt classification of 'regular' subjects 124
7.5 Pooled confusion matrix for CRFt classification of 'OSA' subjects 124
7.6 Sleep stage classification performance for 'regular' subjects 133
7.7 Sleep stage classification performance for 'OSA' subjects 133
7.8 Sleep stage classification performance for 'all' subjects 134

8.1 Correlation between the most discriminative cardiorespiratory features per class and sleep statistics 141
8.2 Correlation between sleep statistics and five-class staging performance κ and accuracy 144
8.3 Coefficient (standard error) for all variables selected with stepwise regression for κ and accuracy 145

Part 1

Introduction

Dormir! Adormecer! Sossegar! Ser uma consciência abstrata de respirar sossessagadamente, sem mundo, sem astros, sem alma - mar morto de emoção refletindo uma ausência de estrelas.

Sleep! To fall asleep! To have peace! To be an abstract consciousness that's conscious only of breathing peacefully, without a world, without heavens, without a soul – a dead sea of emotion reflecting an absence of stars!

- Bernardo Soares (heteronym), Livro do Desassossego (*transl.* Richard Zenith, The Book of Disquiet)

Chapter 1

Sleep assessment

1.1 Sleep

We spend one third of our lives asleep, attesting to its importance as crucial for physical and mental restoration [1]. Sleep can be defined as a reversible state of perceptual disengagement and decreased responsiveness to environmental stimuli [2]. Although its function remains the subject of research, sleep abnormalities have been linked to cardiovascular disease [3], disturbed metabolism and endocrine function [4], increased risk of diabetes [5], lowered immune response [6], production of hyperalgesic changes and decreased thresholds for acute and chronic pain [7], negative impact on learning capacity and memory formation [8], and decreased neurocognitive performance [9].

Sleep is a highly regulated, heterogeneous process with involvement of extended brain networks. During healthy, normal sleep, which occurs mainly during night-time, brainwave activity exhibits cyclical patterns lasting about 90 minutes, alternating between so-called rapid eye movement (REM) sleep and non-REM (NREM) sleep 4 to 5 times per night (Figure 1.1). The periods of REM become progressively longer and are characterized by encephalographic activity similar to wake, lasting around 20 to 25% of the total sleep time. A REM period typically comprises a tonic state with short periods of intermittent phasic activity associated with rapid eye movements and muscle twitches [2, 10]. NREM sleep is characterized by a progressive decrease in muscle tone and sympathetic activity and an increase in parasympathetic dominance. NREM is usually divided in different stages, and the most recent guidelines by the American Academy of Sleep Medicine (AASM) separate it in N1, N2 and N3 [11]. N1, occurring 2 to 5% of the total sleep time, corresponds to a transitional state from wake to sleep, with electroencephalography (EEG) exhibiting a shift from alpha activity (8-13 Hz) to a mixed frequency with increased theta activity (4-7 Hz), accompanied by a decrease in muscle activity easily observable in electromyography (EMG) [2].

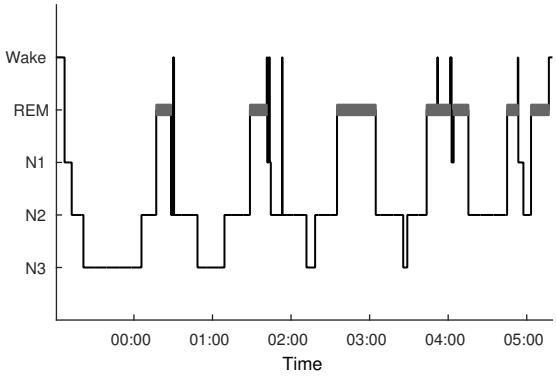


Figure 1.1: Example of a hypnogram of a healthy subject, showing the progression of sleep stages throughout a night

During N2, K-complexes and sleep spindles appear along with the progressive increase of high amplitude, low frequency EEG activity [2, 12]. When N3 appears, the EEG reveals a predominance of synchronized delta waves (0.5-2 Hz), low muscle activity and very limited response to environmental stimuli. N3 intervals decrease in duration as sleep progresses, and account in total for 15 to 20% of the total sleep time, decreasing with increasing age [2]. The length of nocturnal sleep varies highly between individuals and depends on a number of factors such as age, day of the week, and choice and behaviors, with adults averaging between 6 to 8 hours [13].

Unfortunately, and despite increased awareness about its importance, the prevalence of short sleep and sleep disorders continues to increase [14, 15, 16, 17] and this picture of healthy sleep is becoming less common.

The percentage of adults reporting regularly sleeping 6 hours or less averaged nearly 30% in 2004 in the United States of America (USA)[18] and 30.4% in 2012 in the Netherlands [19]. Chronic short sleep (diagnosed as insufficient sleep syndrome in the International Classification of Sleep Disorders, 3rd Ed. (ICSD-3)) has been linked to excessive daytime sleepiness, and also with depressed mood and poor memory or concentration [20].

Insomnia is considered the most common sleep disorder, although methodological inconsistencies in defining and assessing it make it considerably difficult to estimate its actual prevalence [21]. The symptoms of insomnia according to the Diagnostic and Statistical Manual of Mental Disorders, 4th Ed. (DSM-IV) and the International Classification of Sleep Disorders, 2nd Ed. (ICSD-2) include difficulty in initiating or maintaining sleep, early morning awakenings and non-restorative sleep at least 3 nights per week. An estimated 34.5% of individuals report at least one of these symptoms [22]. Using as criterion symptom presence accompanied by daytime consequences, the percentage drops to 9.8% and using the stricter diagnostic criterion of DSM-IV, the prevalence in several European countries is estimated at 6.6% [22]. A 2008/2009 USA survey showed widely varying estimates in prevalence when using different diagnostic criteria: 22.1% based on the DSM-IV

[23], 3.9% on the International Statistical Classification of Diseases and Related Health Problems, 10th Ed. (ICD-10) and 14.7% on the ICSD-2 [23]. Although there is evidence that a high percentage of people suffering from insomnia symptoms do not seek a physician's help [24, 25], or do not get treatment for insomnia symptoms [25], this disorder has been increasingly linked to serious health consequences, such as risk of accidents, higher likelihood to develop a major depressive disorder [24], and increased risk of cardiovascular disease and metabolic disorders [26].

Another category of highly prevalent sleep disorders is sleep disordered breathing (SDB), which refers to a spectrum of conditions that feature breathing pauses during sleep. The most common SDB is obstructive sleep apnea (OSA) where collapses of the airway cause an obstruction and a consequent reduction (hypopnea) or cessation (apnea) of airflow with continued respiratory effort [27]. This leads to a decrease in blood oxygen saturation, resulting in a cortical arousal and a sudden burst in sympathetic activity with an accompanied increase in heart rate and blood pressure [10]. Repetitive apnea and hypopnea events interrupt sleep continuity and reduce sleep time, which partly explains one of its defining symptoms: excessive daytime sleepiness [28]. Sleep apnea has been associated with an increased risk of cardiac and cerebrovascular diseases such as hypertension, heart failure, arrhythmias, myocardial ischemia and infarction, pulmonary arterial hypertension and renal disease [29], metabolic dysregulation, insulin resistance and lipid disorders [29], stroke, dementia and cognitive impairment in the elderly, and changes in cerebral blood flow and cerebral autoregulation [30]. Sleep apnea patients with daytime sleepiness were also found to be more prone to motor- and work-accidents [31] and are less productive at work [32]. Studies estimated the prevalence of SDB to be 2% for women, and 4% for men [33], but more recent reviews of epidemiological studies in the USA have shown an increase in prevalence to 10% and 17% among 30–49-year-old and 50–70-year-old men respectively, and 3% and 9% among 30-49 year old and 50-70 year old women respectively [34]. In the Netherlands, prevalence is estimated at 8.8% for men, and 5.6% for women [19]. In the context of epidemics of obesity, the dimension of the problem is likely to continue increasing. Despite its high prevalence, more than 85% of USA patients with clinically significant SDB remains undiagnosed [35, 36].

These are just two of the most prevalent in nearly 90 different sleep disorders identified to date [37]. Table 1.1 lists the major categories of sleep disorders according to the ICSD-3. Despite the continuous development of new and more advanced diagnosis techniques, it is also increasingly clear that many patients remain undiagnosed, and therefore untreated, with dire long-term consequences for their health.

1.2 Standards of clinical sleep assessment

The gold standard for the assessment of sleep and diagnosis of many sleep disorders is overnight polysomnography (PSG). This procedure is typically performed in a dedicated sleep laboratory equipped with a bed, and includes the recording of brain wave activity by means of EEG, eye movements by means of electrooculography (EOG) and submental muscle tone by means of EMG.

Table 1.1: Major categories of sleep disorders according to the ICSD-3

Major categories
Insomnia
Sleep-related breathing disorders
Central disorders of hypersomnolence
Circadian rhythm sleep-wake disorders
Parasomnias
Sleep-related movement disorders

Figure 1.2: PSG with recording of EEG, EOG, EMG, nasal airflow, respiratory effort and pulse oximetry [38]

In a clinical context, PSG commonly includes additional sensors to measure breathing, such as oronasal airflow, surrogate measures of respiratory effort with respiratory inductance plethysmography (RIP) or piezoelectric belts around the thorax and abdomen and peripheral pulse oximetry (Figure 1.2). Heart activity is often also recorded by means of electrocardiography (ECG), and limb movement by means of EMG placed on the arms and legs. Video is often registered, and the session is usually attended by trained personnel.

After the recording has been performed, the signals from the different sensors are manually analyzed by trained somnologists, and a report is created describing respiratory, cardiac, movement and other events which occurred during the session. In addition, sleep stages are visually annotated in 30 second epochs based on EEG, EOG and EMG, following well-established standards, such as the scoring guidelines from the AASM [11].

Due to the uniformization of scoring rules and the use of standard registration equipment, allied with the fact that it takes place in a controlled environment assisted by highly qualified personnel which allows the simultaneous registration of a large number of signals, PSG remains invaluable in the clinical evaluation of sleep, in particular for complex or mixed pathologies which would otherwise be difficult to diagnose.

Although PSG has had and will continue having a place in the diagnostics of sleep disorders, it is not without limitations. Even with intensive training, the validity and repeatability of visual scoring

is limited. In fact, inter- and intra-scorer reliability is relatively low for specific sleep stages and events such as cortical arousals and even the agreement in the scoring of some breathing events is, in some cases, limited [39, 40, 41, 42]. A study comparing the annotations of three scorers showed an inter- and intra-scorer positive percent agreement for stage 1 between 19.8% and 28.1%, and between 16.7% and 42.2% respectively, and an intraclass correlation (ICC) in the scoring of the arousal index of 0.54 [39]. Another study compared the annotations of four scorers and showed an inter-scorer ICC between 0.28 and 0.52 for stage 1, between 0.56 and 0.86 for hypopnea index, and between 0.09 and 0.80 for arousal index [42]. Recent advances in the processing of PSG and the analysis of sleep enabled the creation of tools that facilitate this process, by automating to a large extent the sleep stage scoring process and the detection of relevant events [42]. A study showed that an automatic scoring system achieved an ICC against four human annotators comparable to the ICC between the same human scorers for multiple metrics such as apnea-hypopnea index (AHI) (automatic/human: $0.84-0.92$ versus human/human: $0.84-0.94$), hypopnea index ($0.57-0.94$ versus $0.56-0.86$), obstructive apnea index ($0.79-0.88$ versus $0.76-0.95$), central apnea index ($0.80-0.92$ versus $0.78-0.95$), and arousal index ($0.17-0.82$ versus $0.09-0.80$) [42]. Despite advances in automatic scoring systems, manual scoring is still considered a criterion standard for the evaluation of sleep [42].

Another important limitation of PSG, besides that it is a labor-intensive and rather expensive procedure, is related to the comfort of the patient. A well-known phenomenon called first night effect results from the individual's maladaptation to a new sleep environment and to the laboratory equipment and has been shown to impact sleep architecture during the first recording night [43]. The first night effect can even last up to four nights [44] usually translating to a lower sleep efficiency, increased wake time, longer sleep onset and REM onset latency, and decreased overall REM sleep during the monitoring period [43], potentially giving a distorted view of a person's habitual sleep.

Additionally, lab PSG studies have, in some countries, excessively long waiting times, which in the case of sleep apnea, can even exceed one year [45]. Although it seems difficult to pinpoint universal reasons for this, it is suggested that recent years have seen an increase in demand for diagnostic studies, while available sleep laboratory bed capacity remains limited [45].

Finally, and perhaps most importantly, being in a sleep lab rather than in the own bedroom changes habitual pre- and bed-time routines which may be influencing sleep. Furthermore, the complexity and costs involved in this procedure limit its applicability to one or two consecutive nights, hiding night-to-night variability which may be critical in the correct assessment of the severity of sleep disorders such as sleep apnea [46, 47], sleep movement disorders [47, 48], and insomnia [49].

1.3 Home sleep monitoring

Despite these drawbacks PSG remains the most reliable way to assess sleep and its place in sleep diagnostics is unlikely to change in the foreseeable future. However, some of its limitations, such as its lack of applicability for extended periods of time, or its use in a home environment, are inherent to the procedure itself, and cannot not be easily overcome. In an effort to alleviate this, technological advances in physiological sensing have enabled ambulatory monitoring of different sleep parameters, replacing part of the role sleep labs traditionally had in the assessment and diagnosis process. Although in some cases, such as for ambulatory PSG, the recording procedure is still complex and must be set up by specialized technicians, ambulatory or home sleep monitoring (HSM) is, in principle, ecologically valid, and has been shown to alleviate the first night effect, in particular when monitoring insomniacs [50, 51].

1.3.1 Home sleep tests

Home sleep test (HST) are portable devices designed to enable the diagnosis of sleep breathing disorders with unattended recordings, at home. They are designed to be simpler than PSG and comprise a smaller set of sensors to record at least three different physiological signals: peripheral pulse oximetry, respiratory airflow, and respiratory effort (Figure 1.3). Despite their simplicity and increased comfort when compared with PSG, these devices have proven to be a reliable alternative for patients with a high likelihood of suffering from moderate or severe obstructive sleep apnea, showing good diagnostic performance compared to PSG [52]. In addition to the increased simplicity in their use, HST have a much lower cost than PSG. Even when followed-up by PSG in case the HST diagnosis is inconclusive, the costs of HST can be nearly half than what PSG would cost if applied in the first place [53].

The increasing level of evidence for the high diagnostic performance of HST have led AASM to recommend them for patients with a high pretest probability of moderate to severe sleep apnea, without suspicion of significant comorbidities, followed by a PSG in case of a negative result [54].

1.3.2 Actigraphy

Actigraphy consists of the measurement of gross body movements by means of a wrist- (or limb-) worn accelerometer (Figure 1.4). A measure of so-called 'activity counts' quantifying the amount of body movements occurring during an epoch (typically 30 seconds long) is recorded and used to determine whether the subject was awake, resting or asleep during that period [55]. Based on this simple measuring principle, this technique is available for decades and has met relative success in the assessment of sleep-wake disturbances [56]. More recently, it has been indicated by the AASM as a suitable method to assist in the evaluation of patients with circadian disorders and sleep-wake disturbances, and also to assess response to therapy of circadian disorders and insomnia [57].

Figure 1.3: HST for diagnosis of sleep disordered breathing [38]

However simple, actigraphy suffers from an important drawback: whenever subjects are lying quietly awake, without moving, actigraphy will consider that they are asleep, and as a result, underestimate wake time and overestimate total sleep time and sleep efficiency [58, 59]. Additionally, and because it relies exclusively on the measurement of body movements, it simply lacks the ability to describe other aspects of sleep such as sleep architecture or breathing disorder events. Despite these drawbacks, its low cost, simplicity and relative comfort address one of the major limitations of PSG: actigraphy allows long-term monitoring of the active-rest cycles of individuals, potentially highlighting issues which would be hidden with a one- or two-night PSG snapshot of their sleep [60]. For the diagnosis and follow-up of circadian rhythm disorders, for instance, it is important to demonstrate consistency in the pathology for a period of time (7 consecutive days to establish a diagnosis according to the ICSD-2). While this would be prohibitively expensive with PSG, actigraphy is rather well-suited for the task [61]. Regarding insomnia, and despite its known underestimation of sleep onset latency in patients suffering from the condition [62, 63], strong night-to-night variability in sleep parameters often seen with these patients [64] cannot be easily measured with PSG. Furthermore, actigraphy allows a more objective assessment of sleep time than self-report by insomniacs, who might have the propensity to underestimate it [61, 65]. As a last example, actigraphy is also a useful tool to document a patient's sleep routines prior to a visit to a sleep lab, aiding in the distinction between daytime sleepiness due to insufficient night-time sleep as opposed to another sleep disorder [61].

1.3.3 Consumer sleep tracking

Recent years have seen a rapid growth in the market of consumer sleep tracker (CST) devices, more and more attractive to a general population increasingly interested in sleep due to their relative convenience of use, and increased affordability. The most popular CST products can be roughly grouped in three categories:

Figure 1.4: Wrist-worn actigraphy device [38]

- Apps such as Sleep Cycle [66], SleepBot [67] or Sleep As Android [68] - relying exclusively on the sensors available in the smartphone or tablet device where the app is installed and as such, with limited physiological sensing capabilities, usually restricted to the measurement of body movements or sounds, detected when the smartphone or tablet is placed on the bed or night table, close to the subject's body.

- Wearable devices, such as the different models of FitBit [69], JawBone [70], Misfit [71], the now discontinued Basis Peak, or the Philips Health Watch [72] - using sensors integrated in a wrist-band, such as accelerometers to measure gross body movements (FitBit Ultra [73], JawBone Up [74], Misfit Shine [75]), or photoplethysmography (PPG) to measure cardiac activity (Basis Peak, Philips Health Watch).

- Off-body devices, such as the Resmed S+ [76], Withings Aura Sleep Sensor [77], Beddit Sleep Tracker [78], or Sense with Sleep Pill [79] - typically including an array of sensors which, besides measuring environmental characteristics such as light, noise, and ambient temperature, also measure - remotely - physiological characteristics of subjects lying in their radius of action; these range from the measurement of movements of a body in the vicinity of the sensor (Sense with Sleep Pill), to Doppler Radar which measures respiratory movements of subjects at which the sensor is directed (Resmed S+), or ballistocardiography (BCG), which measures, besides the movements of the thorax and abdomen during breathing, also the acceleration of the subject's body associated with the recoil caused by the blood movement in and out of the aorta (Withings Aura or Beddit).

Obviously the performance of the CST devices depends not only on the accuracy of the algorithms used to interpret the recorded data, but first and foremost on the nature of the physiological data measured. Devices which can only measure gross body movements are, in principle, only able to roughly segment periods of sleep and wakefulness. On the other hand, devices which have access to cardiac and/or respiratory activity could provide additional insights into the architecture of sleep, for example by automatically classifying sleep stages into different classes, typically REM, combined N1 and N2, and N3 sleep [80].

Figure 1.5: Consumer sleep tracking device [72]

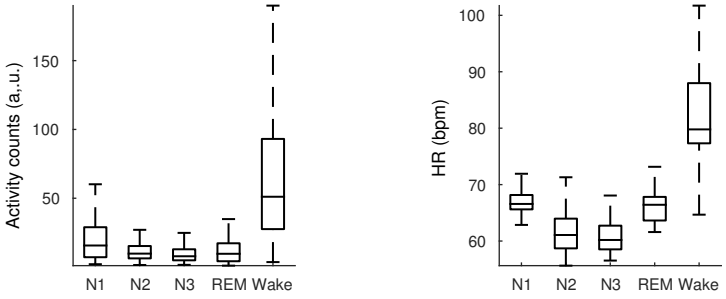


Figure 1.6: Box plot of (a) activity counts and (b) average heart rate measured for different sleep stages of a healthy subject from our data set

Despite the increasing popularity of CST devices, these have been met with skepticism by clinicians and sleep researchers [81, 82]. Although most devices boast a 'this is not a medical device' disclaimer, the marketing claims by these products are often broad and hyperbolic and have the potential to be perceived by consumers as scientifically valid. With the notable exceptions of the JawBone Up [83] and the FitBit Ultra [84, 85], most of the claims have not been properly validated on a healthy population for which these devices are intended, let alone on clinical populations. Some of the claims can be easily recognized by experts as implausible, such as the ability to distinguish sleep stages based on body movements alone: as the example in Figure 1.6a illustrates, there is indeed a substantial difference between the activity counts during Wake and any sleep stage. However, there are no visible differences in body movements between different sleep stages and using this modality alone to automatically identify them does not seem plausible. This mismatch between lack of validation and marketing claims may have particularly negative consequences for individuals with sleep-related concerns - but yet undiagnosed disorders - who will seek an explanation with these devices before consulting with medical professionals, or worse, for asymptomatic individuals with a hidden primary sleep disorder. In the event that the devices do not recognize a medical issue, they may give false reassurance, effectively preventing a timely diagnosis of the person's condition.

1.4 Cardiorespiratory activity during healthy sleep

Although there is a clear motivation and need for HSM devices, current techniques and devices do not suffice for the overall purpose of long-term, unassisted, comfortable sleep monitoring. Actigraphy only gives insights into active/rest patterns, lacking the physiological modalities to assess other aspects of sleep; HSTs are appropriate for the diagnosis of SDB, but are not comfortable for prolonged, unobtrusive sleep monitoring; and current CST devices simply lack the level of evidence needed to be accepted and adopted in clinical practice.

Despite the low level of evidence for CST devices, the physiological basis behind modern devices which, by means of PPG, BCG or Doppler radar, have the ability to monitor cardiorespiratory activity during the night is, in theory, solid. Research has shown a strong link between sleep, sleep disorders, and autonomic nervous system (ANS) activity. The ANS is a branch of the peripheral nervous system which is concerned with the regulation of smooth muscle, cardiac muscle and glands, and consists of two interacting systems: a sympathetic division, or sympathetic nervous system (SNS), and a parasympathetic division, or parasympathetic nervous system (PNS). The SNS controls functions that accompany arousal and expenditure of energy stored in the body, while the PNS controls functions related to activities that support increase or maintenance of the body's energy supply [86]. The cardiovascular autonomic nervous system in particular (Figure 1.7), is responsible for ensuring an adequate cardiac output to the vital organs through continuous adjustments of heart rate and arterial blood pressure and is coupled with the circadian rhythm, the sleep-wake cycle, and ultradian processes such as NREM and REM sleep [87].

In healthy subjects, autonomic cardiovascular regulation varies considerably per sleep stage. As NREM sleep progresses from N1 to N3, there is an accompanying increase in cardiovagal drive and PNS activity and a reduction in cardiac and SNS activity, translating to a decrease in heart rate and increase in the respiratory mediation of heart rate variability, visible in the high frequency band (HF) (0.15 to 0.4 Hz) of heart rate variability (HRV), compared to wakefulness [87]. In contrast, REM is a state of autonomic instability where PNS and SNS activity fluctuates producing abrupt changes in heart rate. The average heart rate and the power in the low frequency band (LF) (0.04 to 0.15 Hz) of HRV is higher during REM than during NREM sleep, and there is a shift of the LF/HF ratio towards sympathetic dominance [87]. These differences are clearly visible in the examples of Figures 1.6 and 1.8 which illustrate some of these characteristics for different sleep stages of a healthy subject.

Regarding respiration, and in the absence of respiratory disorders, breathing is more regular during NREM sleep, whereas REM sleep is characterized by irregular breathing both in amplitude and frequency, higher breathing frequency and substantial variations in breath-to-breath tidal volume [88, 89]. The difference between these characteristics for different sleep stages of a healthy subject are illustrated in Figure 1.9.

The reason why ANS characteristics are attractive for the characterization of sleep is because

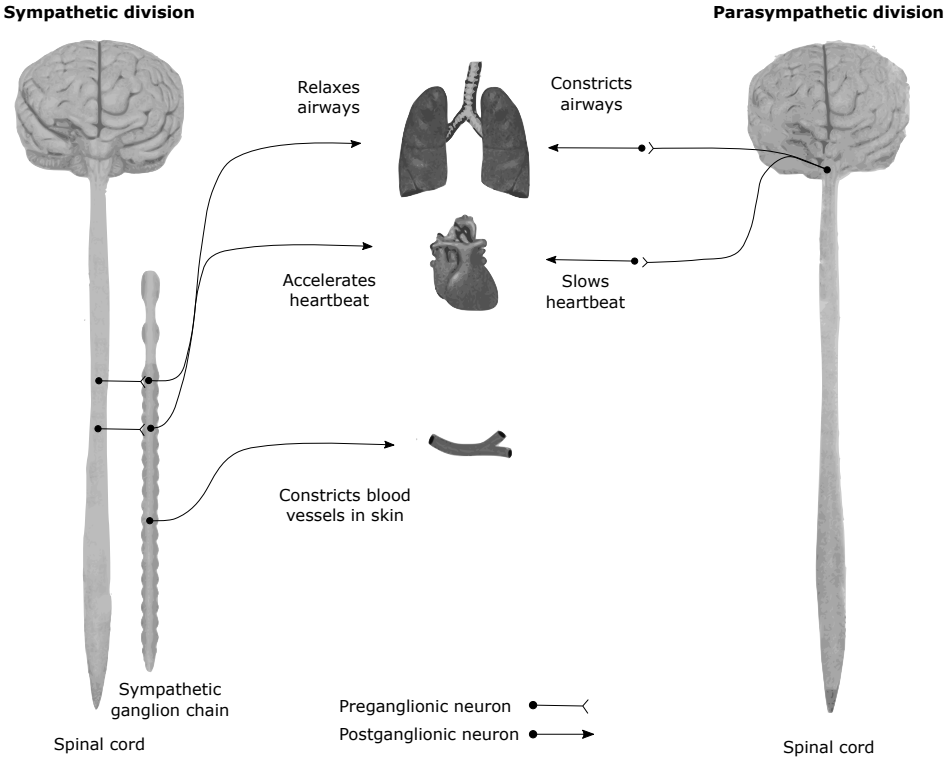


Figure 1.7: Autonomic nervous system and some of the cardiovascular and respiratory functions served by the sympathetic and parasympathetic branches. Adapted from [86]

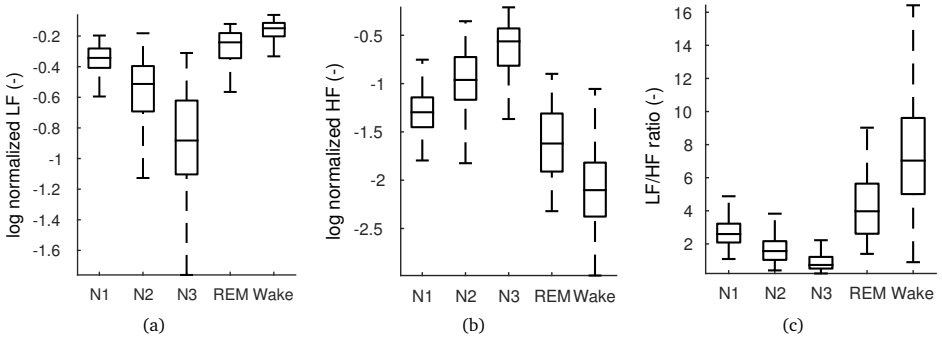


Figure 1.8: Box plot of HRV frequency characteristics for different sleep stages of a healthy subject from our data set: (a) log normalized power in the LF, (b) log normalized power in the HF, (c) ratio between the power in the LF and the HF

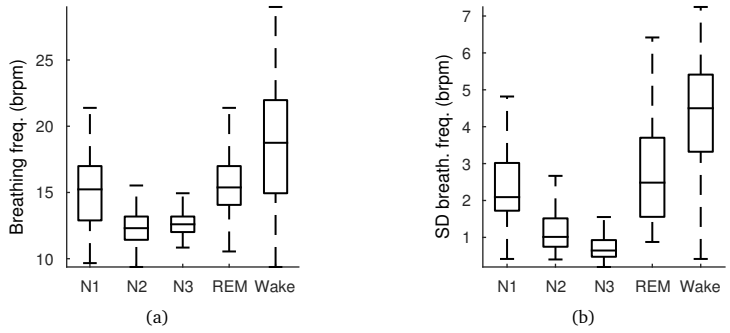


Figure 1.9: Box plot of respiratory characteristics for different sleep stages of a healthy subject from our data set: (a) average breathing frequency, (b) standard deviation of breathing frequency

they are substantially easier to acquire than some of the traditional signals used to score sleep such as EEG, and EOG, which require contact with the scalp and the face. Some of the sensors used in CST devices mentioned earlier can exploit changes in ANS during sleep. PPG, for instance, measures changes in blood volume associated with individual heart beats [90]. By measuring the time interval between consecutive heart beats, HRV characteristics can, in principle, be accurately measured. Modern PPG devices can even be mounted on a wrist-band, making this technology convenient and comfortable to wear during the night [91]. Based on another sensing principle, BCG measures the recoil caused by the ejection of blood to the vessels with each heart beat. Again, by measuring the time interval between successive peaks in acceleration, HRV characteristics can be measured [92]. In addition, when placed under the area of the thorax, BCG can also measure chest movements associated with respiration [92]. BCG devices can be placed on top or even under the bed mattress, and are therefore adequate for night-time sleep monitoring [93]. Doppler radar sensors use the Doppler effect to measure changes in velocity of the subject's body associated with chest movements due to respiratory activity. Doppler radar-based devices can be placed next to the bed, for example on the night-table, pointed at the subject under monitoring and are therefore adequate for the monitoring of sleep at home [94].

However attractive, these modalities present a substantial challenge when compared with traditional, visual sleep scoring of PSG signals: it is substantially more difficult for a human to associate patterns in ANS activity to sleep stages. Individual HRV and respiratory characteristics provide less separation between sleep stages than EEG, EOG or EMG patterns, and they are often accompanied by large within- and between-subject variability. Visual scoring of HRV and respiratory features would require the simultaneous visualization of dozens of feature traces, and the rules to do so would be impractically complex. For this reason, and given the increased robustness and availability of these sensors, fully automatic cardiorespiratory-based sleep staging has been gaining increased attention in the last decade. Starting from Sleep/Wake classifiers [95, 96], pro-

gressing with more detailed analysis of sleep architecture, differentiating Wake, REM, and NREM [80, 93, 97, 98, 99], between Wake, combined N1 and N2, N3 and REM [100, 101, 102] and even between Wake, N1, N2, N3 and REM [103][104][105]. Despite the increasing level of evidence for the feasibility of cardiorespiratory sleep staging, most of these studies still use signals measured from ECG and thoracic respiratory effort and only a few are actually based on sensors which can be plausibly used for long-term home monitoring, such as BCG [93] and Doppler radar [94]. Although in theory the same information about HRV and respiratory activity can be derived from the other sensors mentioned above, the level of evidence for CST devices, in particular with existing products, remains scarce.

Another challenge in the use of ANS characteristics to characterize sleep is related to the possible presence of certain sleep disorders. SDB, for instance, is defined precisely by the occurrence of respiratory events which are known to disrupt sleep. A measure of breathing regularity would be obviously impaired by the occurrence of OSA events [27], while phenomena such as Cheyne-Stokes breathing and often associated central sleep apnea (CSA) events [106] would impact any respiratory frequency or amplitude measure during such a period. Furthermore, arousals which usually follow these events are associated with bursts of sympathetic activity which elicit sudden increases in heart rate and in LF power, and decreases in HF power [27, 107], otherwise uncharacteristic of some sleep stages, such as N3, during which SDB events can occur. Insomniacs, on the other hand, often show increased SNS activity with increased heart rate and elevated power in the LF band, and blunted HF response [108]. The presence of these conditions may impair the discriminative power of ANS characteristics, especially if classification algorithms are modeled exclusively with healthy subjects. In contrast with the relatively large body of literature on healthy subjects, studies about cardiorespiratory-based sleep staging in sleep disordered patients is limited to a few articles with relatively poor performance [98, 102]. If this technology is to ever be adopted in clinical practice this limitation must be resolved.

1.5 Objective and outline

The topic of cardiorespiratory sleep staging has received increasing attention over the last decade, but performance remains modest. The first objective of this thesis is to develop methods which increase the performance of sleep staging and bring it closer, for healthy subjects, to the gold standard PSG. In addition, most of the studies on cardiorespiratory sleep staging use ECG and respiratory activity measured with belts worn around the thorax. Although these sensors provide accurate means to measure ANS characteristics during sleep, they are not exactly compatible with long-term, unassisted monitoring at home. The second objective of this thesis is to prove that these methods can work with a PPG sensor which is more adequate for this type of monitoring. Finally, and given the limited amount of evidence regarding the use of this technology in sleep disordered patients, the third objective is to evaluate the extent to which these methods can be applied in subjects suffering from such a disorder. This will help establish whether the technology has the

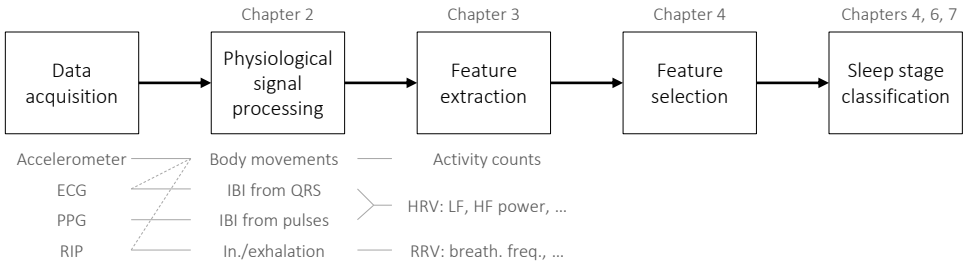


Figure 1.10: General framework of sleep stage classification and validation, with indication of the contribution per chapter, and sensors, physiological characteristics and examples of extracted features described throughout the thesis

potential to ever be used in clinical practice, and complement PSG in the assessment, diagnosis and treatment follow-up of sleep disorders.

The work described in this thesis follows the general machine learning framework illustrated in Figure 1.10. The first two blocks, data acquisition and in particular, physiological signal processing, are sensor specific and cannot, generally speaking, be trivially used for different cardiac or respiratory sensors. However, once certain physiological characteristics, e.g. interbeat intervals (IBIs), are computed, and provided that the level of accuracy is comparable across different sensors, the remaining methodology is sensor-agnostic. Respiration rate variability (RRV) features, for example, only need information about inhalation and exhalation intervals. Comparably, HRV features are fully based on the analysis of IBI information, regardless of the sensor used to compute them. This is an advantage, as it allows not only the adoption of a general methodology for different sensors, but perhaps more importantly, it allows data sets collected with different sensors to be interchangeably used for training and validation. Following this framework, the thesis comprises three parts and is organized as follows.

The first part describes methods for signal analysis and extraction of features which can be used for sleep staging. **Chapter 2** describes a technique developed for the accurate localization of QRS complexes from ECG. Estimation of interbeat intervals is of importance to properly characterize HRV during sleep, and insufficient accuracy can hamper the extraction of differentiating features. **Chapter 3** describes a method to estimate body movements from non-actigraphy signals, such as ECG and RIP. Despite that body movements are valuable for sleep staging - in particular to distinguish between Sleep and Wake - they are usually not part of standard PSG lab set-ups, and therefore not commonly available in retrospective databases with sleep data. Although this technique is not actually used for PPG-based sleep staging, because most PPG devices usually also comprise an accelerometer, it does allow existing data sets to model and train classifiers for that purpose.

The second part describes methods for unobtrusive sleep stage classification using cardiores-

piratory features. In **Chapter 4** we present a machine learning method which uses a large set of cardiac and respiratory features derived from ECG and RIP signals to perform sleep staging. In **Chapter 5** we evaluate the feasibility of cardiac sleep staging based on a truly unobtrusive sensor, PPG, and establish whether the features and machine learning methods implemented in previous chapters are truly sensor agnostic.

The third part is devoted to the exploration of more advanced probabilistic methods for sleep stage classification. Exploiting the structured nature of sleep, **Chapter 6** evaluates whether a conditional random fields (CRF) classifier, a type of discriminative Markov model, is better suited to the task of sleep stage detection when compared with a single-epoch based, commonly used classifier such as Bayesian linear discriminant (LD) classifier, or a generative Markov model such as hidden Markov models (HMM) classifier. In **Chapter 7** this work is extended to multiple-class classification, and further evaluated in the automatic classification of Wake, N1, N2, N3, and REM sleep on healthy subjects and on patients with obstructive sleep apnea.

Finally, **Chapter 8** discusses the work presented in this thesis and evaluates the research objectives. It also discusses the limitations of the current work and highlights areas where cardiorespiratory sleep staging should be improved and further validated before it can be effectively deployed in clinical practice for home sleep monitoring.

Part 2

Signal analysis for sleep stage classification

Se eu um dia pudesse adquirir um rasgo tão grande de expressão, que concentrasse toda a arte em mim, escreveria uma apoteose do sono. Não sei de prazer maior, em toda a minha vida, que poder dormir. O apagamento integral da vida e da alma, o afastamento completo de tudo quanto é seres e gente, a noite sem memória nem ilusão, o não ter passado nem futuro.

Were I ever granted a flash of expressive power so great that it concentrated all art in me, I would write a eulogy to sleep. I know no greater pleasure in life than that of being able to sleep. The total snuffing out of life and the soul, the complete banishment of all beings and people, the night without memory or illusion, the absence of past and future.

- Bernardo Soares (heteronym), Livro do Desassossego
(*transl.* Richard Zenith, The Book of Disquiet)

Chapter 2

A low-complexity post-processing algorithm for precise QRS localization

Adapted from P. Fonseca, R. M. Aarts, J. Foussier, and X. Long, A novel low-complexity post-processing algorithm for precise QRS localization, *Springerplus*, vol. 3, no. 1, p. 376, 2014.

Abstract: Precise localization of QRS complexes is an essential step in the analysis of small transient changes in instant heart rate and before signal averaging in QRS morphological analysis. Most localization algorithms reported in literature are either not robust to artifacts, depend on the sampling rate of the ECG recordings or are too computationally expensive for real-time applications, especially in low-power embedded devices. This chapter proposes a localization algorithm based on the intersection of tangents fitted to the slopes of R waves detected by any QRS detector. Despite having a lower complexity, this algorithm achieves comparable trigger jitter to more complex localization methods without requiring the data to first be upsampled. It also achieves high localization precision regardless of which QRS detector is used as input. It is robust to clipping artifacts and to noise, achieving an average localization error below 2 ms and a trigger jitter below 1 ms on recordings where no additional artifacts were added, and below 8 ms for recordings where the signal was severely degraded. Finally, it increases the accuracy of template-based false positive rejection, allowing nearly all mock false positives added to a set of QRS detections to be removed at the cost of a very small decrease in sensitivity. The localization algorithm proposed is particularly well-suited for implementation in embedded, low-power devices for real-time applications.

2.1 Introduction

Automatic QRS detection is a basic and widely used technique in the analysis of electrocardiography (ECG) recordings and a significant amount of research effort has gone into developing automatic algorithms with high sensitivity and positive predictive value (PPV) [109]. However, despite being able to correctly detect the presence of QRS complexes, most algorithms fail to provide their precise location. In fact, standards for evaluating the performance of QRS detectors are often not strict enough. For example the ANSI/AAMI EC57 norm [110] recommends a margin of 150 ms when evaluating QRS detections. In a recording where the subject has an average heart rate of 60 beats per minute (bpm), this margin corresponds to 15% of the average length of a beat. This criterion is clearly meant as a guideline for evaluating the detection performance. Although it may suffice in applications where the correct detection of QRS complexes is more important than their localization, there are many areas in which the detection of small transient changes in heart rate is crucial.

For example, in the field of sleep research, Catcheside et al. [111] reported how the heart rate increases by approximately 14% within two seconds of the onset of arousal-inducing tones during sleep. This was further confirmed by Bangash et al. [112] who observed an increase of 10% (during rapid eye movement (REM) sleep) and 15% (during non-REM (NREM)) three seconds after an arousal-inducing tone, followed by a decrease of 15% (REM) and 25% (NREM) five seconds after the tone onset.

Another example comes from the area of QRS morphology analysis in high-resolution ECG. If the localization of QRS complexes before signal averaging[1] is not precise enough, the resulting signal will suffer from a low-pass filtering effect which may hide relevant high-frequency low-level potentials [113, 114]. Because these potentials can help, for instance, in the analysis of ventricular late potentials in patients recovering from myocardial infarction, guidelines have been proposed describing the desired precision of QRS localization in terms of trigger jitter (standard deviation of the localization error), which should be below 1 ms [113].

A third area in which the precise localization of QRS complexes is of paramount importance is in heart rate variability (HRV) analysis. HRV analysis is a widely used technique to assess different aspects of the autonomic nervous system (ANS) and has been shown to be clinically relevant for many different applications, such as a predictor of cardiovascular disease and mortality [115], to assess possible ANS dysfunctions in patients with chronic obstructive pulmonary disease [116], and for screening of sleep disorders and evaluation of sleep quality [117]. HRV time series are usually computed from the intervals between consecutive QRS complexes and it is recommended that the ECG is sampled at high enough sampling rates (250-500 Hz or higher) to guarantee a small trigger jitter in the localization of the R wave fiducial point, needed for the accurate estimation of HRV parameters [118].

[1] Signal averaging is a technique used in QRS morphological analysis whereby a number of QRS complexes in an ECG recording are first aligned in the time domain and then averaged to improve the signal-to-noise ratio of the resulting signal.

Several localization methods have been proposed to address the issue of precise QRS localization. However, most either depend on the sampling rate of the original ECG (such as cross-correlation- and normalized integrals-based methods [114]), are too sensitive to noise or to signal clipping (such as threshold based methods [114]) or are computationally complex and not adequate to real-time applications (such as interpolation and curve fitting [119]). Other techniques such as vectorcardiographic loop alignment, depend on the recording of multiple ECG leads [120]. The increase in the computational power of modern computers allows techniques such as cross-correlation localization to overcome limitations in regard to the sampling rate by simply up-sampling the ECG signal. However, this step renders these algorithms unsuitable for low-power, embedded processors for wearable or even implantable devices. Furthermore, template-based localization algorithms such as cross-correlation are not adequate for real-time processing because they require a template to first be built from QRS complexes before localization can finally take place.

The objective of this chapter is to propose and evaluate a post-processing method for precise localization of QRS complexes in single-lead ECG recordings. The proposed method, henceforth referred to as the *slope* algorithm, should address the following requirements:

- **Complexity** - the algorithm should have low complexity, $O(n)$, to enable real-time processing with modest hardware requirements and integration in low-power embedded devices.

- **Trigger jitter** - the algorithm should yield at least the same trigger jitter as the more computationally complex cross-correlation-based method in high resolution ECG, and better trigger jitter at lower sampling frequencies, to enable a high localization precision in ECG recorded with lower sampling frequencies.

- **Localization error** - the algorithm should have low average localization error, below 10 ms (1% of the average length of a beat for a recording where the average heart rate is 60 bpm), to enable applications (such as sleep arousal detection) which depend on the detection of small transient changes in interbeat interval length.

- **Robustness** - the algorithm should be robust to the presence of noise and movement artifacts, to enable its application in recordings performed in uncontrolled (non laboratory) conditions; in addition, the algorithm should handle signal clipping, a common occurrence in recordings during sleep [98].

- **Agnostic** - the algorithm should be agnostic to the QRS detection algorithm used, and perform equally well regardless of the method.

Additionally, it will be shown how template matching can be used to reduce the number of false positive detections and that the localization step performed by this post-processing algorithm is essential for that purpose. Template matching can be used, for instance, to exclude false positive

detections which lead to erroneously short estimates of interbeat intervals, or to exclude beats with aberrant QRS morphology before averaged complexes are analyzed [120].

2.2 Materials and methods

One of the main problems of most QRS detectors is that in the presence of noise, movement artifacts or signal clipping, they do not provide precise estimates of the location of the peaks. Consequently, a post-processing localization step is usually performed to obtain a more precise location of detected complexes [109].

2.2.1 QRS detection

The localization algorithm described in this chapter can be used after any of the vast number of QRS detectors reported in literature [109, 121, 122]. For the purpose of evaluating its performance, three popular detectors described in literature were used: the Hamilton-Tompkins (HT) detector [123, 124], an envelope-based detector by Nygårds and Sörnmo (NS) [125] and a filter bank-based detector by Afonso, Tompkins et al. (AT) [126].

2.2.2 Filtering

Before the signal is processed, baseline wander is first removed with a linear phase high-pass filter using a Kaiser window of 1.016 sec, with a cut-off frequency of 0.8 Hz and a side-lobe attenuation of 30 dB [127]. The coefficients of the impulse response were determined by computer-aided filter design with the software Matlab R2012b (The MathWorks Inc., Massachusetts).

2.2.3 Localization

The level of noise, the amount of artifacts and the possible presence of signal clipping, common in recordings during sleep, all have a significant impact in the localization performance of most QRS detectors. Consider the example of a QRS complex with a motion artifact which distorts the R peak, as illustrated in Figure 2.1a (highlighted with a circle). Although most QRS detectors will correctly identify the presence of this complex, they will usually yield an incorrect location, usually the local maximum. This common type of problems is first addressed by observing that the Q-to-R and R-to-S amplitudes of a QRS complex are normally much larger than the amplitude of measurement noise. They are usually also larger than the amplitude of artifacts which do not distort the shape of the complex beyond the point that the R wave is no longer distinguishable [122]. Note that this might not hold in the presence of intense body movements. However, it is arguable whether the QRS detectors used before post-processing would be able to detect the presence of such peaks anyway.

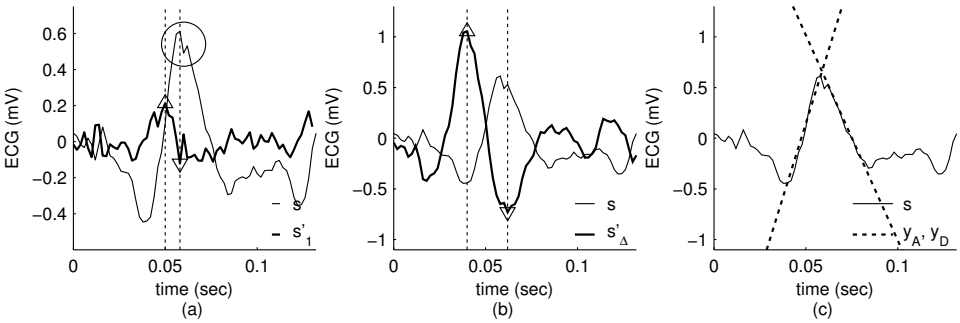


Figure 2.1: Estimating the location of the R wave peak for a QRS complex with an artifact: (a) signal s with a motion artifact (indicated with a circle) and differentiated signal with a step size of 1 sample, s'_1 (in bold), with maximum and minimum derivative values indicated by up- and down-pointing triangles, respectively; (b) differentiated signal with a step size Δ of 10 samples, s'_Δ (in bold); (c) tangents to the ascending and descending slopes of the QRS complex, y_A and y_D, with the peak localized by their intersection.

Because the R waves are very steep, signal differentiation usually gives the ascending and descending slopes as local maxima and minima, respectively. This simple approach has been widely used [109], and remains the basis of some of the most popular algorithms to date [124, 128]. However, it suffers from some drawbacks, notably in the presence of artifacts which cause sudden, drastic changes in the signal amplitude. In such cases it often occurs that the local maxima of the differentiated signal for these artifacts is larger than the maxima which correspond to the actual slopes of the R wave (Figure 2.1a). This problem can be solved simply by observing that QRS complexes have a predictable length: in the absence of a medical (cardiac) condition, they last in average about 0.08 seconds [129]. Instead of differentiating the ECG signal on a sample-to-sample basis, a larger step can be used to more accurately calculate the slope of the tangent to the signal (Figure 2.1b). Local maxima and minima will indicate the beginning of the slopes which are simultaneously steep and have at least a given duration. The intersection of the tangents to these slopes yields the location of the R wave peak (Figure 2.1c). The *slope* algorithm is formally described as follows:

- Differentiate the filtered ECG signal with a step size Δ to obtain s'

$$s'[n] = \frac{s[n+\Delta] - s[n]}{\Delta}. \tag{2.1}$$

- For each base peak location p_i estimated by a QRS detector:
  - Find the location of the local maximum $\hat{p}_i$ in a window centered around p_i,

$$\hat{p} = \arg\max_{n \in P_i} s[n] \tag{2.2}$$

with

$$P_i = \left[p_i - \frac{W-1}{2}, p_i + \frac{W-1}{2} \right], \tag{2.3}$$

where W corresponds to the expected length (in samples) of the QRS complex (0.08 sec).

– Find the location of the local maximum and the local minimum of s' in a window respectively before and after $\hat{p}_i$,

$$M_i = \underset{n \in [\hat{p}_i - W, \hat{p}_i]}{\arg\max}\ s'[n] \tag{2.4}$$

$$m_i = \underset{n \in [\hat{p}_i, \hat{p}_i + W]}{\arg\min}\ s'[n]. \tag{2.5}$$

– Compute the slope and the y-intercept of the tangent to the ascending slope, y_A

$$y_A[n] = a_A \cdot n + b_D \tag{2.6}$$

with

$$a_A = s'[M_i] \tag{2.7}$$

$$b_A = s[M_i] - a_A \cdot M_i \tag{2.8}$$

and to the descending slope, y_D

$$y_D[n] = a_D \cdot n + b_D \tag{2.9}$$

with

$$a_D = s'[m_i] \tag{2.10}$$

$$b_D = s[m_i] - a_D \cdot m_i. \tag{2.11}$$

– Compute the location of the intersection n_I between the two tangents,

$$n_I = \frac{b_D - b_A}{a_A - a_D} \tag{2.12}$$

$$y_I = a_A \cdot n_I + b_A. \tag{2.13}$$

Note that the tangent intersection n_I does not necessarily correspond to the exact (integer) location of a sample in the discretized ECG. This means that the location of the peak can be computed with sub-sample precision and has as an important consequence, as will be shown, that the localization error is to a certain extent independent of the sampling rate of the ECG signal.

The differentiation step Δ is an important parameter in the algorithm. As explained in Appendix A, this factor should be as large as possible to minimize the localization error. However, it should not be larger than the length of the slope. In order to increase the robustness of the algorithm to clipping artifacts which cause the slopes to be shortened not only in amplitude but also in duration, a step size corresponding to approximately half of the expected length of each slope was chosen. Considering that each QRS complex lasts about 80 ms, and that each slope lasts around 40 ms, a step size of 20 ms was chosen and experimentally found to be adequate. Note that when analyzing unclipped signals this parameter can be increased in order to further reduce the trigger jitter, as explained in Appendix A; lowering the parameter increases the trigger jitter in the presence of noise. Figure 2.1a illustrates (in bold) the signal obtained after single-sample differentiation. The artifact at the center of the R wave introduces additional local minima which do not correspond to the descending slope of the QRS complex. Figure 2.1b illustrates the signal obtained after differentiating the ECG with a Δ factor of 10 samples (20 ms at a sampling frequency of 500 Hz). The maximum and minimum values now correspond to the beginning of the steepest parts of the slopes. After identifying the parameters of the tangent to the ascending and descending slopes, their intersection yields the location of the R wave peak (Figure 2.1c).

As explained in Appendix B, this algorithm has a complexity of $O(n)$, lower than that of cross-correlation-based methods.

2.2.4 Template matching

Template matching can be used to reduce the number of false positive detections in a recording. The template should be chosen such that it represents the morphology of the QRS complexes in an ECG recording and can be obtained by averaging the signal around the location of each peak in a window with a length equal to the median duration of the beats in each recording.

After the template is estimated, the correlation between each detected complex and the template can be computed. Under the assumption of monomorphic QRS complexes, false positive locations should have a lower correlation value (Figure 2.2a) than undistorted complexes. Complexes slightly distorted by noise should also yield a lower correlation value, but as long as they retain some of their characteristics the correlation should still be higher than of false positives (Figure 2.2b). As an example of the impact of body movement artifacts, Figure 2.2c illustrates the correlation obtained with complexes detected and localized in an ECG recorded simultaneously with actigraphy (Actiwatch Spectrum, Philips Electronics) during a full night. As it can be easily seen, the correlation is inversely proportional to the amplitude of the actigraphy peaks.

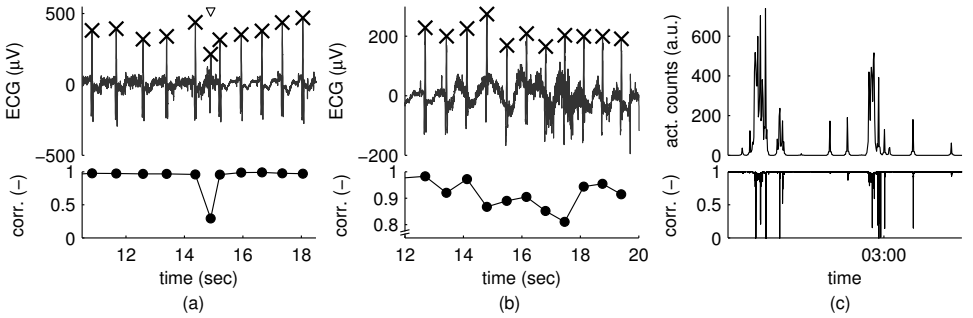


Figure 2.2: Template matching after QRS detection (with the Hamilton-Tompkins detector): (a) in the presence of a false positive (indicated with a downward arrow); (b) in the presence of complexes heavily distorted by noise; (c) compared with actigraphy for a full night recording.

2.2.5 Evaluation

Data sets

The *slope* algorithm was tested in two publicly available data sets. The first comprises 18 long-term ECG recordings (lead II) of subjects with no significant arrhythmias from the MIT-BIH Normal Sinus Rhythm Database (MIT-BIH) [130, 131]. Each recording has a sampling rate of 128 Hz and an average length of 24.3 ± 0.83 hour. This data set was chosen because it comprises full day recordings, including periods with intense motion artifacts, naturally occurring in this type of monitoring scenarios.

The second data set comprises 79 high-resolution ECG recordings (lead II) of subjects annotated as healthy controls in the PTB Diagnostic ECG Database (PTBD) [130, 132]. Each recording has a sampling rate of 1000 Hz and an average length of 118.9 ± 3.4 sec.

Cross-correlation localization

As discussed, localization has been traditionally performed for the purpose of signal averaging in the area of QRS morphology analysis where it is crucial to have a low trigger jitter. One of the most successful methods uses cross-correlation between a template (built by averaging QRS complexes detected with a QRS detector) and each complex to improve the estimated location. Because this is one of the most successful localization algorithms reported in literature, especially when used on high-resolution (or upsampled) ECG signals, it will be used to establish the ground-truth locations of QRS complexes in our data set and as reference to evaluate the *slope* algorithm.

An implementation of this method, henceforth referred to as the *xcorr* algorithm, is described as follows:

- Remove baseline wandering using the filtering procedure described in an earlier section.

- Compute a template of the QRS complexes based on the initial locations, using the template algorithm described in a later section.
- For each complex, find the location which maximizes the cross-correlation between the template and the ECG signal around the original location.
- Build a new template based on the adjusted locations.
- For each complex, store as the final location of its R wave peak the location that maximizes the cross-correlation between the new template and the ECG signal around the adjusted location.

Ground-truth

The MIT-BIH includes annotated QRS locations per recording. Unfortunately these only indicate the location of the QRS complex, and not always the precise location of the R wave peak. On the other hand, the PTBD does not include annotations with the location of QRS complexes, so these were first estimated using the HT detector. In order to obtain a ground-truth for both data sets, each recording was first upsampled to 10000 Hz, after which the *xcorr* algorithm was used to obtain precise locations. A similar procedure was previously used by Shaw and Savard [133] to obtain precise ground-truth locations in their data set.

Trigger jitter

To compare the trigger jitter obtained with the *slope* algorithm with the more computationally complex *xcorr* algorithm, both were tested on high-resolution ECG recordings from the PTBD, after downsampling[2] them to different sampling rates (500, 200, 100, and 50 Hz). To illustrate that the localization precision is not only bound by the time resolution of the signal, but also by its bandwidth which is in turn limited by the sampling rate, both algorithms were also tested after upsampling each (downsampled) recording to the original sampling rate of 1000 Hz. Note that this procedure increases the time resolution but does not change the spectral content of the signal. For each algorithm and sampling rate, the pooled trigger jitter σ_e^p was computed as

$$\sigma_e^p = \sqrt{\frac{\sum_{i=1}^{k}(n_i - 1)\sigma_{e_i}^2}{\sum_{i=1}^{k} n_i - 1}} \tag{2.14}$$

where k is the number of recordings in the data set and σ_{e_i} is the standard deviation of the localization error (trigger jitter) for all n_i peaks localized within 40 ms of any ground-truth location[3]

[2] A low-pass filter with a cutoff frequency of $0.8 \cdot (fs/2)$, with fs as the desired frequency, was used before downsampling to prevent aliasing.

[3] This is the same margin as that indicated in the guidelines for analysis of ventricular late potentials using signal-averaging technique [113] and corresponds to half of the average length of QRS complexes [129].

for recording i.

Localization

In order to evaluate the robustness of the *slope* algorithm in the presence of artifacts and noise, the ECG signals in the MIT-BIH were degraded with clipping artifacts and Gaussian noise. Clipping artifacts were introduced to simulate saturation artifacts [134] and digital clipping. The following procedure was used:

- Remove baseline wandering using the filtering procedure described in an earlier section.
- Compute the median, p_m, and the 99.9^{th} percentile p_M of the amplitude of the ECG.
- Determine the clipping threshold *ct* for a given clipping factor *cf*

$$ct = p_m + (p_M - p_m) \cdot cf. \tag{2.15}$$

- Clip the ECG signal,

$$s_c[n] = \begin{cases} ct & \text{if} s[n] \geq ct \\ s[n] & \text{otherwise.} \end{cases} \tag{2.16}$$

In order to simulate a wide range of clipping and saturation artifacts, the following clipping factors were used: 1 (no clipping), 0.6, and 0.3. The last factor is meant as a lower bound for this type of artifacts, as it is unlikely that a segment of a recording where the ECG is distorted to that degree can still allow any useful analysis.

In addition to clipping, the localization precision was tested in the presence of Gaussian noise added to the ECG signal to obtain specific signal-to-noise ratio (SNR): 20 dB, 10 dB, and 5 dB. For simplicity, the noise power in the original signal was considered negligible in comparison with the added noise.

The localization precision of the *slope* post-processing algorithm was evaluated with each of the three QRS detectors listed earlier on the ECG recordings of the MIT-BIH degraded with the conditions described above. For each recording, the mean and standard deviation of the localization error (distance between detected/localized peaks and the ground-truth locations) before and after post-processing were computed.

Template matching

The impact of localization on template matching was evaluated by comparing its effectiveness in rejecting false positive detections before and after the *slope* algorithm was used. Mock 'false positive detections' were added to the list of base QRS detections, after which template matching was

applied. Complexes with a correlation below certain thresholds were rejected, and the resulting PPV and sensitivity were computed. The following procedure was used:

- Use the HT detector to detect the base locations in each ECG recording of the MIT-BIH.
- Add a percentage f_i of randomly located false positives ('mock false positives') to the base list of locations. For example, a 'mock false positive percentage' f_i of 100% means that a number of mock false positives equal to the number of original locations will be added.
- Compute the median length B_j between consecutive localized QRS complexes for each ECG recording j.
- Compute a template as the signal average of all windows of length B_j for each recording j.
- Compute the correlation between the template of each recording and the ECG signal for each location on that recording.
- Determine the threshold T such that the sensitivity obtained after removing locations with a correlation lower than T remains above a specified limit. Compute the corresponding PPV ($\text{post}_{\text{none}}$).
- Use the *slope* algorithm to obtain a list of adjusted locations.
- Compute a new template and the corresponding correlations with the adjusted locations.
- Determine the thresholds T such that the sensitivity obtained after removing locations with a correlation lower than T remains above a specified limit. Compute the corresponding PPV ($\text{post}_{\text{slope}}$).

The choice of using a varying threshold instead of a fixed one serves the purpose of simultaneously evaluating the effect on the sensitivity and PPV of the algorithm. In turn, this allows an assessment of its performance for different scenarios where for example, sensitivity is more important than PPV, or vice-versa.

2.3 Results

2.3.1 Trigger jitter

Figure 2.3 illustrates the pooled trigger jitter obtained with the *slope* and with the *xcorr* algorithms with and without upsampling, after having downsampled the original high-resolution ECG signal to different sampling rates. As it can be observed, the trigger jitter obtained with the *slope* algorithm for (down)sampling frequencies above 50 Hz was always lower than 1 ms, even without upsampling the signal, which means that the algorithm meets the strict requirement of 1 ms specified by the guidelines for QRS localization [113]. In contrast, the trigger jitter obtained with the *xcorr*

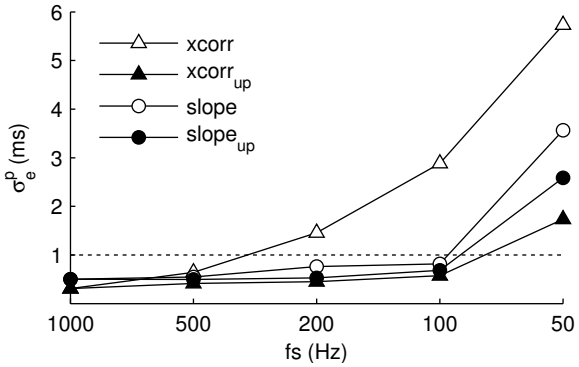


Figure 2.3: Comparison of trigger jitter obtained using the *slope* algorithm with and without upsampling ('slope' and 'slope$_{up}$') and using the *xcorr* algorithm with and without upsampling ('xcorr' and 'xcorr$_{up}$')

algorithm increased beyond 1 ms unless the signal was first upsampled. The localization error obtained with *xcorr* at 1000 Hz is not zero because the ground-truth was obtained after upsampling the original recordings at 10000 Hz.

The reason why the trigger jitter increased when the signal was downsampled to 50 Hz is related to the spectral properties of the QRS complexes. As reported by Thakor et al. [135], QRS complexes have spectral components up to 40 Hz. When the signal is downsampled below 80 Hz, these components are affected, changing the morphology of the signal and causing a decrease in localization precision.

2.3.2 Localization

The localization error and the trigger jitter were computed for all clipping artifact and noise conditions. The results, illustrated in Figure 2.4, show that the average localization error obtained after post-processing is, overall, lower than that obtained with the original locations output by the three QRS detectors, and always below 2 ms. In addition, the trigger jitter is also lower than 8 ms and lower than of the original locations with the exception of the AT detector. However, it should be noted that the sensitivity of this detector in these conditions was very low. After post-processing, sensitivity increased substantially, reflecting the increased precision in the localization of the complex.

To emphasize the improvement over the base HT performance, Figure 2.5 illustrates the trigger jitter obtained with and without *slope* localization after HT detection and after template matching. Varying the clipping factor for a fixed SNR of 20 dB (Figure 2.5, above) shows that the trigger jitter after localization (black-filled markers) is always lower than without localization (white-filled markers), highlighting the robustness of the algorithm to clipping artifacts. This is true for the whole range of sensitivities obtained using different thresholds with template matching.

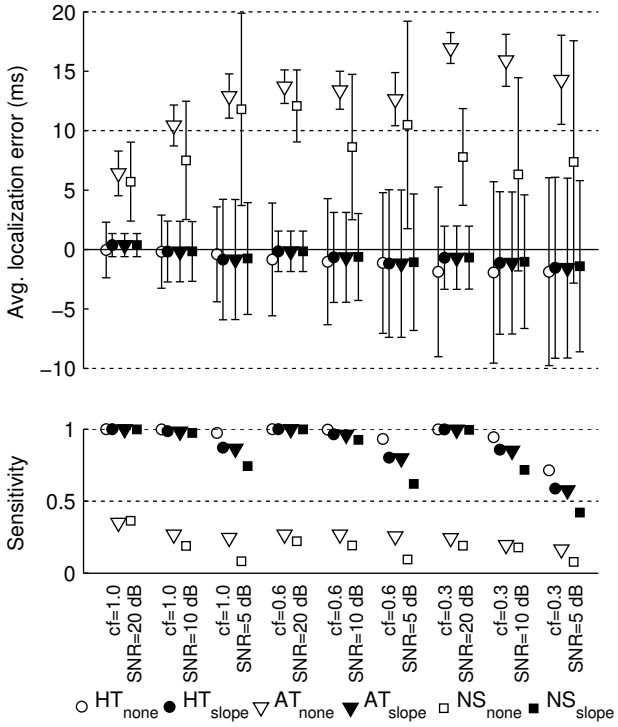


Figure 2.4: Average localization error and trigger jitter for three QRS detection methods, before and after localization (subscripts 'none' and 'slope' respectively) for combinations of clipping factors and SNR (above), and corresponding sensitivity (below)

Varying the SNR without clipping (Figure 2.5, below) shows that with the exception of SNR = 5 dB, the trigger jitter is also always lower with *slope* localization, highlighting the robustness of the algorithm to noise, especially in moderate conditions. Regarding the condition SNR = 5 dB and if the analysis is restricted to complexes that are less distorted by noise (at the cost of a lower sensitivity), it is clear that below a sensitivity of 0.8 the trigger jitter with localization is also lower. Applications with strict requirements in terms of trigger jitter require this template matching step anyway to minimize the localization error. For such applications, the template matching procedure, with a complexity of $O(n)$, does not increase the overall complexity of the *slope* algorithm.

These results are in line with the theoretical localization error derived in Appendix A.

2.3.3 Template matching

Figure 2.6 illustrates the sensitivity and PPV obtained after rejecting locations with a correlation value below a varying threshold for mock false positive percentages, $f_i = 100\%$ and $f_i = 200\%$. In both cases, increasing the threshold leads to the rejection of an increasing number of locations

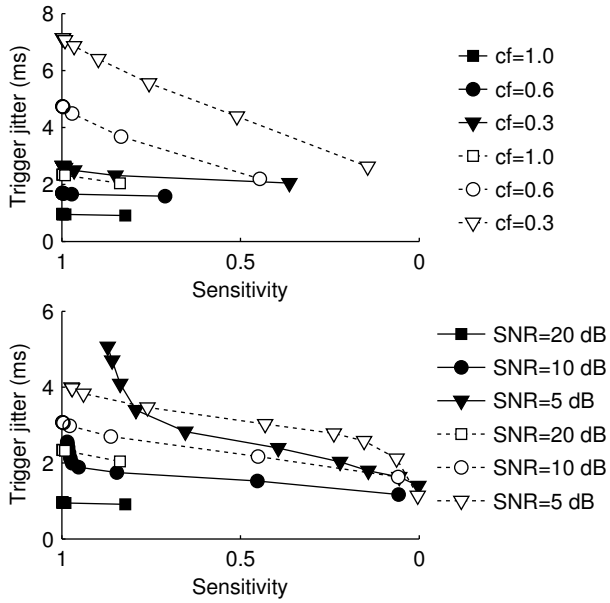


Figure 2.5: Trigger jitter versus sensitivity, obtained after detection with HT, with and without localization (black-filled and white-filled markers, respectively) and template matching, for different clipping factors (above), and SNR (below)

and to a consequent decrease in sensitivity, because besides false positives, a few true positives are also rejected. However, despite the fact that the sensitivity decrease is extremely small, the corresponding increase in PPV is substantial. Although this is true for both $\text{post}_{\text{slope}}$ and $\text{post}_{\text{none}}$, it is more pronounced for the case where post-processing is applied.

Figure 2.7 illustrates the PPV obtained for different f_i after rejecting peaks with correlations lower than thresholds chosen such that the sensitivity was higher than specific values. Note that even before locations are rejected, the PPV already increases after post-processing ($T = 0$). This happens because some false positive locations which are close to actual peaks in the ECG are merged to the same location on the first step of the localization algorithm (equation 2.2). With a sensitivity of 0.992 the PPV approaches 1 even for $f_i = 200\%$. With a sensitivity decrease of 0.005 (0.5%) and 0.008 (0.8%) the PPV increases by nearly 50% and 70% for $f_i = 100\%$ and $f_i = 200\%$ respectively, effectively rejecting almost all false positives added to the base locations.

2.4 Discussion and conclusions

Among the localization algorithms reported in literature, cross-correlation-based methods were found to be among the best performing. However, the localization precision of these algorithms is intrinsically linked to the sampling rate of the recording, which is not an issue for high-resolution

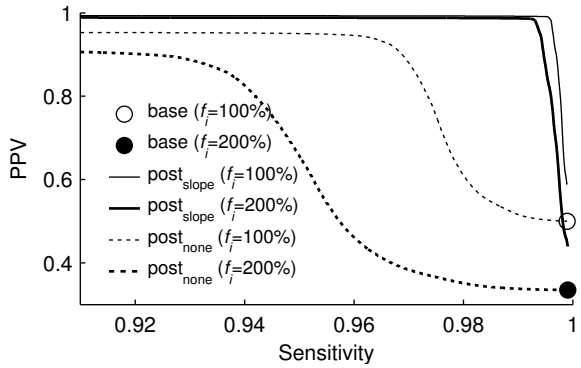


Figure 2.6: PPV versus sensitivity after template matching using different correlation thresholds, with ('slope') and without ('none') localization for two false positive fractions f_i; indicated are also the base sensivity/PPV before localization and template matching; note that the x- and y-axes do not start at 0.

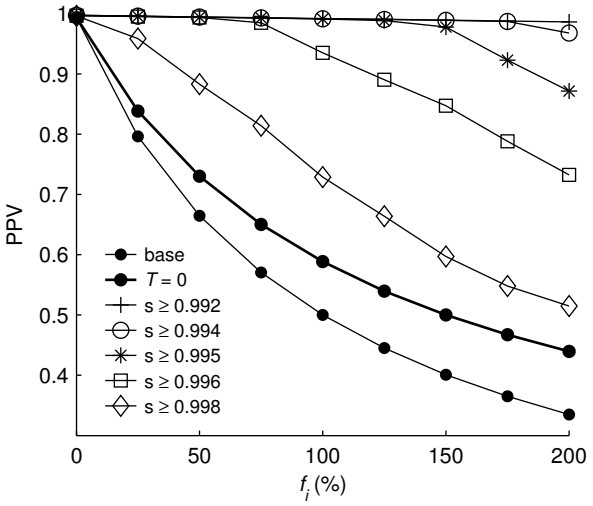


Figure 2.7: PPV per false positive fraction f_i after template matching with different thresholds; 'base' indicates the PPV without localization or template matching, $T = 0$ indicates the PPV after localization, without template matching, and each $s \geq X$ curve indicates the PPV after localization and template matching such that after rejecting locations with low correlation the resulting sensitivity is greater or equal than X.

ECG recordings, but is for recordings performed at lower sampling rates. In addition, these algorithms are computationally very expensive ($O(n^2)$, see Appendix B), which makes them less suited for low-power embedded devices; furthermore, because they typically require templates built from the entire recording, their applicability for real-time, online applications is limited. The

algorithm proposed in this chapter was developed with these requirements in mind. Despite its lower complexity ($O(n)$) it achieves comparable trigger jitter without requiring the data to first be upsampled. It also achieves high localization precision regardless of which QRS detector is used as input. Although it relies on the detection of just a few points in the ECG, it is extremely robust to noise, clipping and movement artifacts. In fact, post-processing with the *slope* algorithm was able to correct the original QRS locations of three different detectors under almost all noise conditions, yielding an average localization error below 2 ms, and a trigger jitter below 8 ms, even for recordings where the signal was severely degraded. Furthermore the increased localization precision improved template matching. After rejecting locations based on their (lower) correlation coefficient, it allowed the correct rejection of nearly all mock false positives added to a base list of detected QRS locations, with only a marginal decrease in sensitivity (0.5% for $f_i = 100\%$ and 0.8% for $f_i = 200\%$).

These results have important consequences for applications which require real-time ECG analysis. Given its low complexity, low localization error and robustness to signal clipping, which is common in sleep recordings, it can enable real-time detection of arousals during sleep. On the other hand, given its robustness to noise and artifacts, it can also enable real-time ECG morphology analysis even under challenging conditions such as during physical exercise. Finally, and as a byproduct of the detection process, the slopes and tangents actually have clinical value, for example in detecting and characterizing myocardial ischemia [136, 137].

2.5 Appendix A – Theoretical localization error

This appendix derives the theoretical localization error obtained using the *slope* algorithm on an ECG signal contaminated with white Gaussian noise.

Consider the ascending and descending slopes of y_A and y_D as defined by equations 2.6 and 2.9 and assume that the ratio β between them is given by

$$a_A = S \tag{2.17}$$

$$a_D = -S \cdot \beta. \tag{2.18}$$

Now consider that the ECG signal has been perturbed by a noise signal with amplitude E. This means that in case both the QRS complex and the noise component have a zero mean, the amplitude of each sample in the QRS complex may change by a maximum of $\pm e = E/2$. The worst possible localization error in this situation is illustrated in Figure 2.8.

The resulting tangents, y'_A and y'_D, are defined by

$$y'_A[n] = a'_A \cdot n + b'_A \tag{2.19}$$

$$y'_D[n] = a'_D \cdot n + b'_D, \tag{2.20}$$

with

$$a'_A = \frac{(s_2+e)-(s_1-e)}{\Delta} = S + \frac{2\cdot e}{\Delta} \tag{2.21}$$

$$a'_D = \frac{(s_4+e)-(s_3-e)}{\Delta} = -S\cdot\beta + \frac{2\cdot e}{\Delta} \tag{2.22}$$

$$b'_A = (s_1-e) - a'_A\cdot n_1 = s_1 - e - S\cdot n_1 - \frac{2\cdot e}{\Delta}\cdot n_1 \tag{2.23}$$

$$b'_D = (s_4+e) - a'_D\cdot n_4 = s_4 + e + S\cdot\beta\cdot n_4 - \frac{2\cdot e}{\Delta}\cdot n_4, \tag{2.24}$$

and intersect at a new location n'_I,

$$n'_I = \frac{b'_D - b'_A}{a'_A - a'_D}. \tag{2.25}$$

Using equations 2.13 and 2.24, the maximum localization error, e'_I is given by

$$e'_I = n_I - n'_I = \frac{b_D - b_A}{a_A - a_D} - \frac{b'_D - b'_A}{a'_A - a'_D}. \tag{2.26}$$

Noting that

$$a'_A - a'_D = S + \frac{2\cdot e}{\Delta} - \left(-S\cdot\beta + \frac{2\cdot e}{\Delta}\right) = (1+\beta)\cdot S, \tag{2.27}$$

we have that

$$a'_A - a'_D = a_A - a_D = (1+\beta)\cdot S. \tag{2.28}$$

Similarly,

$$(b_D - b_A) - (b'_D - b'_A) = -e + \frac{2\cdot e}{\Delta}\cdot n_4 - e - \frac{2\cdot e}{\Delta}\cdot n_1. \tag{2.29}$$

Using the relation $n_4 = n_1 + L$ we have

$$\begin{aligned}(b_D - b_A) - (b'_D - b'_A) &= -e + \frac{2\cdot e}{\Delta}\cdot(n_1+L) - e - \frac{2\cdot e}{\Delta}\cdot n_1 \\ &= 2\cdot e\cdot\left(\frac{L}{\Delta} - 1\right).\end{aligned} \tag{2.30}$$

Using equations 2.29 and 2.27 in 2.25, we finally have

$$e'_I = \frac{e}{S}\cdot\left(\frac{L}{\Delta} - 1\right)\cdot\frac{2}{1+\beta}. \tag{2.31}$$

As it can be seen, the localization error depends on the maximum amplitude of the noise component of the signal, the slopes of the QRS complex, the distance between the first and the last

sample and the distance between each pair of samples (differentiation step) used to compute the tangents. For simplicity consider that the ascending slope S of the QRS complex is constant and defined by the ratio between the amplitude of the ECG signal A and half of the length of the QRS complex, $Q/2$. The localization error is then given by

$$e'_I = \frac{E}{A} \cdot \frac{Q}{4} \cdot \left(\frac{L}{\Delta} - 1 \right) \cdot \frac{2}{1+\beta}. \tag{2.32}$$

Assuming that the noise component is mainly caused by electromyography (EMG) noise, it can be considered to follow a zero-mean Gaussian distribution with a standard deviation of 10% of the amplitude of the ECG signal [122]. Taking as slope ratio β the median value experimentally found in the recordings of the PTBD (1.13[4]), an average QRS duration of 0.08 sec [129], using a differentiation step Δ of 0.02 sec and choosing as L the length of the QRS complex Q (the largest possible value such that n_1 and n_4 are still part of the QRS), we can estimate the bounds of the standard deviation of the localization error (or trigger jitter),

$$\mathrm{sd}(e'_I) = 0.1 \cdot \frac{0.08\ \mathrm{s}}{4} \cdot \left(\frac{0.08\ \mathrm{s}}{0.02\ \mathrm{s}} - 1 \right) \cdot \frac{2}{1+1.13} \approx 5.63\ \mathrm{ms}. \tag{2.33}$$

2.6 Appendix B – Computational complexity

This appendix derives the computational complexity of the *xcorr* and *slope* algorithms and compares their performance using a computer simulation. Because for both algorithms peak localization takes place for every QRS complex, let n represent the average number of samples analyzed per complex. The *xcorr* algorithm performs the following steps:

1. Search for the local maximum, complexity $O(n)$.
2. Add the contribution of each sample to the template (consider that the template has the same number of samples n), $O(n)$.
3. Perform cross-correlation with the template for $2n+1$ positions around the center of the peak, $O(n \cdot (2n+1)) = O(n^2)$.
4. Find the position that yields the maximum cross-correlation, $O(n)$.
5. Repeat steps 2 to 4.

The overall complexity of the *xcorr* algorithm is $O(n^2)$. Note that the complexity of the up-sampling step, necessary to guarantee a high localization performance with this algorithm for

[4]Using the slope estimates given by the localization algorithm, the minimum and maximum slope ratios found in the PTBD were 0.68 and 1.84 respectively, confirming the expected asymmetry of QRS complexes described in literature [136, 137].

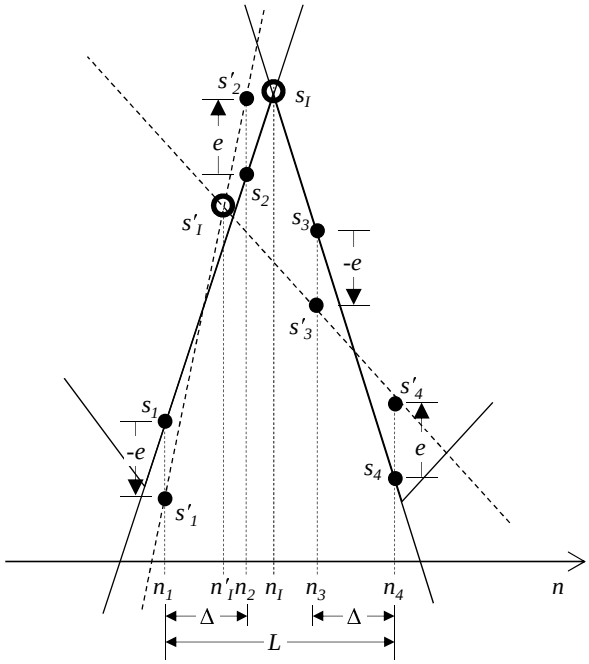


Figure 2.8: QRS complex with tangents to the slopes of the signal (solid lines) and on noisy signal in the worst case scenario (dashed lines) yielding the maximum localization error for a noise component with an amplitude of 2*e*

recordings with lower sampling rates was not included. Although that does not influence the overall complexity, it is an important factor to take into account when comparing the performance of both algorithms because it introduces an additional factor in the number of samples that need to be analyzed.

The *slope* algorithm performs the following steps:

1. Find the local maximum, $O(n)$.

2. Differentiate the signal, $O(n)$.

3. Find the maximum of the derivative before the local maximum, $O(n/2) = O(n)$.

4. Find the minimum of the derivative after the local maximum, $O(n/2) = O(n)$.

The overall complexity of the *slope* algorithm is $O(n)$.

Both algorithms were also compared in terms of running time, using a non-optimized implementation in Matlab R2012b (The MathWorks Inc., Massachusetts) running on a computer with an Intel Core i5-2540M at a clock speed of 2.60 GHz. Each algorithm was used to localize peaks

detected with the Hamilton-Tompkins detector on the 18 recordings of the MIT-BIH. In the case of the *xcorr* algorithm, the signal was first upsampled to 1000 Hz to guarantee a comparable localization performance, but the running time for this operation was not included. The average running time (per recording) of the *xcorr* algorithm was 210.63 ± 55.31 sec and of the *slope* algorithm, 23.08 ± 6.26 sec, nearly ten times faster.

Chapter 3

Estimating actigraphy from ECG and respiratory effort

Adapted from P. Fonseca, R. M. Aarts, X. Long, J. Rolink, and S. Leonhardt, Estimating actigraphy from motion artifacts in ECG and respiratory effort signals, *Physiological Measurement*, vol. 37, no. 1, pp. 67-82, 2016.

Abstract: Recent work in unobtrusive Sleep/Wake classification has shown that cardiac and respiratory features can help improve classification performance. Howbeit, actigraphy remains the single most discriminative modality for this task. Unfortunately, it requires the use of dedicated devices in addition to the sensors used to measure electrocardiography (ECG) or respiratory effort. This chapter proposes a method to estimate actigraphy from the body movement artifacts present in the ECG and respiratory inductance plethysmography (RIP) based on the time-frequency analysis of those signals. Using a continuous wavelet transform to analyze RIP, and ECG and RIP combined, it provides a surrogate measure of actigraphy with moderate correlation (for ECG+RIP, $\rho = 0.74$, $p < .001$) and agreement (mean bias ratio of 0.94 and 95% agreement ratios of 0.11 and 8.45) with reference actigraphy. More important, it can be used as a replacement of actigraphy in Sleep/Wake classification: after cross-validation with a data set comprising polysomnography (PSG) recordings of 15 healthy subjects and 25 insomniacs annotated by an external sleep technician, it achieves a statistically non-inferior classification performance when used together with respiratory features (average κ of 0.64 for 15 healthy subjects, and 0.50 for a data set with 40 healthy and insomniac subjects), and when used together with respiratory and cardiac features (average κ of 0.66 for 15 healthy subjects, and 0.56 for 40 healthy and insomniac subjects). Because this method eliminates the need for a dedicated actigraphy device, it reduces the number of sensors needed for Sleep/Wake classification to a single sensor when using respiratory features, and to two sensors when using respiratory and cardiac features without any loss in performance. It offers a major benefit in terms of comfort for long-term home monitoring and is immediately applicable for legacy ECG and RIP

monitoring devices already used in clinical practice and which do not have an accelerometer built-in.

3.1 Introduction

Although recent years have seen significant advances on unobtrusive sleep measurements, overnight polysomnography (PSG) recordings by expert technicians in dedicated laboratories remain the gold standard for sleep medicine [11]. However, PSG still suffers from severe drawbacks. Laboratory facilities, dedicated equipment and qualified personnel are very expensive; in addition, PSG is uncomfortable and can have a negative impact on normal sleep, being basically impossible to perform on long-term beyond one or two consecutive nights. All these factors motivated research in the area of unobtrusive sleep monitoring which can be used in a home setting. Actigraphy, in particular, has been quite popular in the assessment of sleep-wake disturbances [56]. It consists of the measurement of gross body movements and is usually performed with accelerometers mounted on the wrist, or on other limbs. The amount of body movements is quantified into so-called "activity counts" in epochs of fixed size (typically 30 seconds), which are then used to estimate whether the subject was awake or asleep during that period [55]. It has been indicated by the American Academy of Sleep Medicine (AASM) as a valid auxiliary method to evaluate patients with circadian disorders and sleep-wake disturbances, and also to evaluate their response to treatments of insomnia and circadian disorders [57].

Despite its growing popularity, the performance of actigraphy-based Sleep/Wake classifiers has been shown to be limited, and inferior to PSG [58, 138]. In an effort to improve this, additional modalities such as cardiac and respiratory activity have been also analyzed in the context of sleep and shown to improve actigraphy-based Sleep/Wake classification [139]. However, they require the inclusion of sensors which can measure those modalities, such as electrodes for electrocardiography (ECG) or respiratory inductance plethysmography (RIP) belts for respiratory effort. The addition of these sensors to the traditional actigraphy means an increase in set-up and data analysis complexity because the signals all need to be synchronized. Additionally, it also means a decrease in patient comfort because more sensors need to be worn.

Given the value of cardiac and respiratory modalities in ambulatory monitoring, and the fact that these characteristics cannot be captured with simple actigraphy sensors, the goal of this work is to eliminate the need for dedicated actigraphy sensors when cardiac and/or respiratory sensors are already used for sleep monitoring at home.

3.1.1 Body movement artifacts on ECG signals

The effect of body movements on ambulatory ECG signals has been thoroughly explored, especially in the context of signal improvement prior to clinical diagnosis of cardiac disorders. In particular, the last few years saw an increasing number of studies on methods to compensate [140], elimi-

nate [141], and otherwise detect [142] body movement artifacts (BMAs) on ECG signals [143]. However, very few of those studies have focused on the topic of quantifying these artifacts. The exception was the work of Pawar et al. [144] where the type of movements was classified based on the artifacts measured with ambulatory ECG.

One of the most commonly observed characteristics in literature is that portions of ECG affected by BMAs have different spectral properties than unaffected portions of the signal. Figure 3.1a shows the normalized power spectral density (PSD) plots of the ECG signal during periods with high (activity counts above 10) and with low body movement activity (activity counts equaling 0) for the synchronized ECG and actigraphy (with a Philips Actiwatch) recordings of 9 healthy subjects during the night. Although there is a significant overlap in the spectral content of the two signals, it is clear that periods with higher activity correspond to a higher power below 5 Hz, and also have a richer spectral content in the bands between 30 and 60 Hz.

3.1.2 Body movement artifacts on RIP signals

In contrast with the ECG, the effect of BMAs on RIP signals has been much less studied, the exceptions being the work of Motto et al. [145], Keenan and Wilhelm [146], and Aoude et al. [147]. What seems clear is that RIP-based measures of respiratory effort are particularly sensitive to body movements during periods of physical activity and even during sleep [146]. Although this makes respiratory monitoring more challenging, this modality seems like a good candidate for the estimation of actigraphy during the night. Figure 3.1b illustrates the normalized PSD plots of the RIP signal during periods with high and low body movement activity, in the same conditions as indicated before for ECG. The spectral content of the signal in these two conditions is clearly different: the characteristic peak between 0 and 1 Hz under low activity conditions (which corresponds to the average breathing rate) is less clear in the presence of high activity; in addition, the spectral content of the signal for frequencies above 1 Hz is much higher in the presence of high activity, making it easier to distinguish from artifact-free segments of the signal.

3.1.3 Objective

The objective of this chapter is to propose and evaluate a technique for quantifying BMAs in ECG and RIP signals during sleep. This will be achieved by estimating, based on time-frequency analysis with a continuous wavelet transform (CWT), the local power of each signal at different frequencies. A selected set of CWT coefficients are used to estimate the intensity of BMAs in the signal recordings. The potential of using this estimation as a method to detect body movements and as a replacement of wrist-actigraphy for Sleep/Wake classification is evaluated.

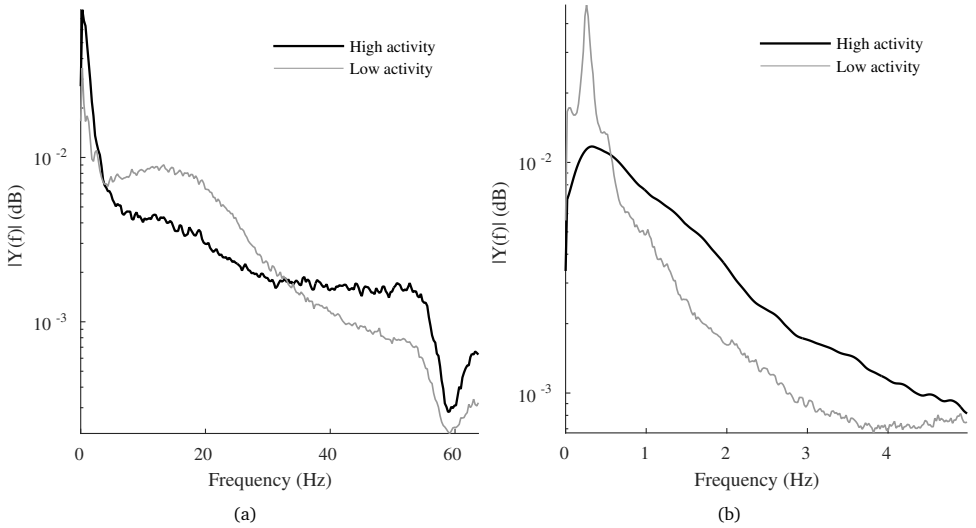


Figure 3.1: Normalized signal spectrum on periods with high and low body movement intensities for (a) ECG and (b) RIP signals. The low power of the ECG signal around 60 Hz is the result of a notch filter used during acquisition to filter out the mains hum from the power line.

3.2 Materials and methods

3.2.1 Data sets

The data set used in this study comprised full single-night PSG recordings of 40 subjects (21 females). 15 subjects (10 females) had a Pittsburgh Sleep Quality Index [148] of less than 6 and had no regular sleep complaints nor earlier diagnosis of sleep disorders (age 31.0 ± 10.4 years, sleep efficiency (SE) 87.8 ± 9.6% and total sleep time (TST) 6.67 ± 1.16 hour). The remaining 25 (11 female) subjects (age 45.6 ± 14.9 years, SE 73.6 ± 17.2% and TST 4.88 ± 1.41 hour) had a self-reported Insomnia Severity Index score above 14 and were classified as insomniacs [149]. The PSG data of six subjects was recorded at the Philips Experience Lab of the High Tech Campus, Eindhoven, The Netherlands, during 2010 (Vitaport 3 PSG, TEMEC). The PSG data of the remaining 34 subjects was recorded at the Sleep Health Center, Boston, United States of America (USA), during 2009 (Alice 5 PSG, Philips Electronics). All PSG montages had at least the recommended AASM set (including redundancy) for sleep scoring, namely EEG F4-M1, F3-M2, C4-M1, C3-M2, O2-M1 and O1-M2, left and right electrooculography (EOG), and submental chin electromyography (EMG) [11]. Both studies were approved by the local ethics committees and all participants signed an informed consent form. Sleep stages were scored by trained sleep technicians blind to the participants' condition (healthy or insomniac), in four classes (Wake, REM, N1, N2, N3) according to the AASM guidelines [11]. In the scope of this study, REM, N1, N2, and N3 were all merged into a

single Sleep class. Each PSG recording comprised, besides the standard signals required for sleep scoring, lead II ECG, chest RIP and synchronized actigraphy (Actiwatch Spectrum, Philips Electronics) worn on the wrist of the non-dominant arm and configured to log activity counts in epochs of 30 seconds. QRS complexes were detected and localized from ECG signals using a combination of a Hamilton-Tompkins detector [123, 124] and the localization algorithm described in Chapter 2 [150]. Prior to actigraphy estimation with CWT, all ECG signals were downsampled to a common sampling rate of 128 Hz. Before feature extraction and actigraphy estimation, RIP signals were downsampled to a common sampling rate of 10 Hz.

3.2.2 Continuous Wavelet Transform

The presence of movement artifacts changes the frequency content of ECG and RIP signals in different frequencies than those which normally contain information about actual cardiac and respiratory processes. Furthermore, the intensity and duration of the movements will cause the artifacts to have different characteristics: more intense, longer movements should lead to artifacts with a higher (overall) energy than less intense or shorter movements. Time-frequency analysis methods can be used to measure the energy of the signals over different frequency bands and over time, which in turn can be used to estimate these characteristics and provide an estimate of gross body movements.

One of the most popular methods for time-frequency analysis, the discrete short-time Fourier transform (STFT), uses the Discrete Fourier Transform together with a window of arbitrary but usually fixed length to calculate the frequency content of segments of a signal. The main disadvantage of STFT is that it relies on a window of constant size and although its length is arbitrary, there is a trade-off in its choice. A longer window will lead to an increase in frequency resolution, but the poorer time resolution will make it less adequate to capture the characteristics of shorter artifacts. A shorter window length improves time localization but will lead to a lower frequency resolution. Typical physiological signals include a combination of events such as spikes (e.g. ECG QRS complexes), transients (e.g. caused by folding and stretching of skin under ECG electrodes during body movements) and oscillations (e.g. respiratory waveform during stable breathing). Although the STFT is very well suited for the analysis of the latter, the limited time resolution, unavoidable when a higher frequency resolution is desirable, makes it less suited for the analysis of events with short duration. When both types of signals are present in the data, and both are of interest, the Wavelet Transform offers a more flexible compromise between time and frequency resolution.

The CWT measures the similarity between a signal and an analyzing wavelet function $\psi(t)$ for different dilations of the wavelet, at different locations in the signal. It is defined as

$$T(a,b) = \frac{1}{\sqrt{a}} \int_{-\infty}^{+\infty} x(t)\, \psi^* \left(\frac{t-b}{a} \right) dt, \tag{3.1}$$

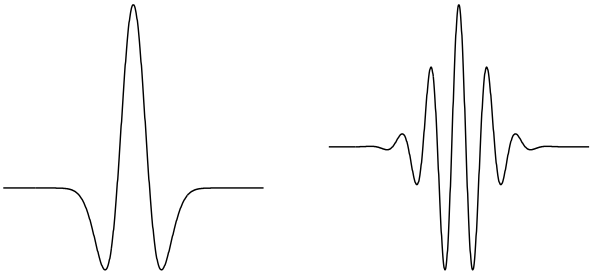


Figure 3.2: Wavelet functions used for CWT: (left) Mexican Hat, (right) Gaussian order 16.

where ψ^* is the complex conjugate of the wavelet function ψ, a is the scaling (dilation) parameter and b is the location parameter. The wavelet power spectrum at scale a and location b is given by

$$P(a,b) = |T(a,b)|^2 . \tag{3.2}$$

In contrast with the STFT, which is usually limited to sinusoidal functions with fixed scales, the CWT allows a wide range of wavelet functions to be used, and to be computed at arbitrary scales and locations. When applied to discrete signals, the CWT is computed over a discretized time-frequency grid. The integral is approximated by a summation over a number of samples for each time step location b.

A large number of wavelet functions has been proposed over the last decades, but one family, the Gaussian Wavelet [151], is widely used for biomedical signal analysis. It is given by

$$\psi^{(N)}(t) = \frac{d^N}{dt^N}\left[\pi^{-1/4} e^{-t^2/2}\right]. \tag{3.3}$$

The Mexican Hat [151] is a scaled version of the second derivative, up to a sign, of the Gaussian wavelet, and is given by

$$\psi(t) = \frac{2}{\sqrt{3}}\left(1-t^2\right) e^{-t^2/2}. \tag{3.4}$$

After experimentally comparing the performance of different wavelet function families and a different number and range of scales, the ECG signal was analyzed with a Mexican Hat wavelet function (Figure 3.2, left) with 320 scales chosen such that the length of the wavelets spanned between 1 and 60 seconds. The RIP signal was analyzed with a Gaussian wavelet of order 16 (Figure 3.2, right) with 150 scales chosen such that the length of the wavelets spanned between 1 and 30 seconds. Figure 3.3a and 3.3b illustrate examples of the CWT computed on short ECG and RIP segments with BMAs, respectively.

Since the wavelet function is shifted sample by sample over the entire signal, P in (3.2) effectively corresponds to a single value for each sample and for each scale. Since the goal is to estimate

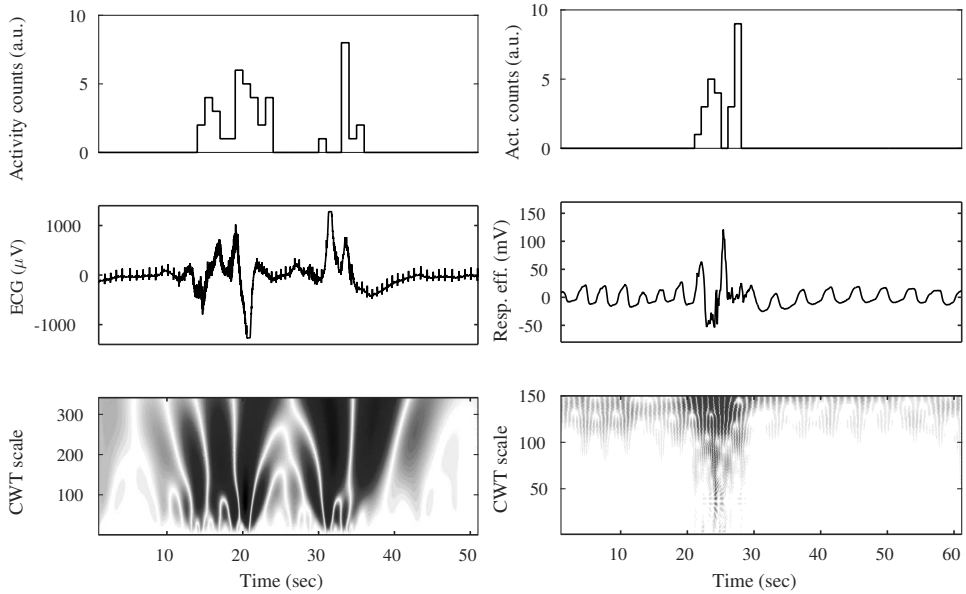


Figure 3.3: Example of the effect of a body movement artifact on (a) ECG and (b) RIP signals and corresponding CWT with a Mexican Hat and a Gaussian order 16 wavelet, respectively. The top plots of each figure indicate the corresponding activity counts measured in one-second epochs with actigraphy, for reference.

a value of actigraphy for each 30-second epoch, a feature vector is built by computing a number of sample statistics over the values of each epoch,

$$F[n] = \left\{P_{M(a_1)}[n], ..., P_{M(a_A)}[n], P_{m(a_1)}[n], ..., P_{m(a_A)}[n], \bar{P}_{(a_1)}[n], ..., \bar{P}_{(a_A)}[n]\right\}, \tag{3.5}$$

where $P_{M(a_i)}[n]$, $P_{m(a_i)}[n]$, and $\bar{P}_{(a_i)}[n]$ are the maximum, minimum and average wavelet power for scale a_i for the epoch starting at sample n, respectively,

$$P_{M(a_i)}[n] = \max(\{P[n, a_i], ..., P[n+N-1, a_i]\}) \tag{3.6}$$

$$P_{m(a_i)}[n] = \min(\{P[n, a_i], ..., P[n+N-1, a_i]\}) \tag{3.7}$$

$$\bar{P}_{(a_i)}[n] = \frac{1}{N}\sum_{b=n}^{n+N-1} P[b, a_i]. \tag{3.8}$$

This yields a total of 960 features for the ECG signal, and 450 features for the RIP signal.

The purpose of these sample statistics is to capture the power of different types of artifacts in the signal. BMAs which last for a large portion of the entire epoch should yield a high average

wavelet power for some scales. BMAs which last only a few seconds, or even a fraction of a second, should yield a high maximum wavelet power. When the minimum wavelet power is lower, on those same scales, it suggests that the epoch also contains portions of the signal which are not affected by BMAs. Note that this will allow the feature vector to represent artifacts with short and long durations, in contrast with the STFT, for which the average spectral density would only able to discriminate longer artifacts lasting a significant portion of the entire epoch.

The feature vectors are computed for the ECG signal, F_{ECG}, for the RIP signal, F_{RIP}, and for the combination of the two signals by concatenating the two previous vectors, $F_{ECG+RIP}$.

Besides the CWT method, an analogous analysis was performed after computing the STFT for 256 frequencies using Hanning windows of 60 seconds centered on each 30-second epoch of the ECG and the RIP signals. Since the results were inferior to the CWT and failed to yield a statistically non-inferior Sleep/Wake classification performance when compared with reference actigraphy, the results obtained with the STFT will not be described further and only the results obtained with the CWT will be presented and discussed.

3.2.3 Feature selection and regression

After the feature vectors are calculated with the CWT method for the ECG, RIP and combined ECG and RIP signals, an actigraphy value expressing the intensity of gross body movements is finally estimated.

This is achieved by finding the linear combination of the feature values calculated for each epoch which best quantifies those artifacts. The parameters of this linear combination can be obtained by multiple linear regression to a reference, in this case, simultaneously measured actigraphy.

This method generates a large number of coefficients (960 features for the ECG signal, 450 for RIP and 1410 for the combination of both), but not all are needed for the estimation procedure. In fact, there is a strong correlation between several pairs of these coefficients so it is actually advantageous to use only a subset for estimation. This allows the model to be strongly simplified, which in turn can help prevent phenomena such as overfitting.

The subset of features used in the linear model and the regression parameters for the selected features were automatically computed by means of stepwise regression [152] with the reference actigraphy. This procedure was performed separately for the features obtained for each modality (ECG, RIP, and for the combination of ECG and RIP). The stepwise regression search was performed with cross-validation, i.e., the subjects were first divided in two folds and the model was estimated on each fold while the fit of each coefficient to the data was evaluated on the other. The estimated model and the estimation error, per iteration, were obtained by averaging the results obtained for each validation fold. It has been shown that classical stepwise regression, where regression and evaluation are calculated on the same data set, can bias not only the parameter estimation, but also the tests used to determine whether variables should be included or excluded in the model.

Cross-validation alleviates these issues, and increases the robustness of the selection and regression procedure [153].

The result is, for each modality $m = \{ECG, RIP, ECG + RIP\}$, a list of N_m features included in the linear model, $S_m = \{s_1, ..., s_{N_m}\}$ and their regression coefficients, $C_m = \{\beta_1, ..., \beta_{N_m}\}$. These can be used to compute an estimate of surrogate actigraphy, X_m for each epoch n,

$$X_m[n] = \sum_{i=1}^{N_m} \beta_i F_m[n, s_i], \qquad s_i \in S_m, \beta_i \in C_m, \tag{3.9}$$

where $F_m[n, s_i]$ is the i-th selected coefficient in the feature vector of epoch n for modality m.

In order to evaluate this method, the estimated actigraphy was compared, for each modality, with the reference actigraphy measured with the Actiwatch using Pearson's correlation coefficient. For each subject, the correlation was computed between the regression estimate obtained with cross-validation on the final iteration of the stepwise regression procedure (the iteration yielding the final list of features) and the reference. The resulting values were then averaged over all subjects. In addition, the estimated and reference actigraphy signals were pooled across all subjects, and the corresponding correlation coefficients were computed. In the pooled case, the correlation coefficient obtained for each modality was tested against the null-hypothesis of no correlation. A Bland-Altman analysis was performed to evaluate the agreement between surrogate and reference actigraphy.

3.2.4 Sleep/Wake classification

Surrogate actigraphy was also evaluated in terms of its capacity to distinguish the presence or absence of movements. Note that this is particularly relevant in a Sleep/Wake classification scenario, where it is often more important to be able to identify whether the subject has moved rather than to quantify the amplitude of her movements. To compare the agreement in the detection of movements with reference and surrogate actigraphy, we used a simple criteria: if the actigraphy value is above a given threshold, the subject was considered to move. Traditional metrics of agreement such as sensitivity, positive predictive value (PPV) and Cohen's κ coefficient of agreement were computed for different actigraphy thresholds.

In order to evaluate whether surrogate actigraphy is at least as discriminative of Sleep and Wake states as reference actigraphy, the surrogate estimates were used together with a set of features shown in earlier work [139] to provide the best classification performance in this task. These feature sets comprise, besides actigraphy (`ac`), five respiratory features including the standard deviation of respiratory frequency over nine epochs (`sdf`), high frequency components (`hfc`) [98], sample entropy (`se`) [154], Dynamic Time Warping (`dtw`) and Dynamic Frequency Warping (`dfw`) [139] and one cardiac feature, mean heart rate (`mhr`). All features were extracted from ECG and RIP signals on non-overlapping epochs of 30 seconds.

The combination of features in the different feature sets is indicated in Table 3.1. The first

three sets were used to test surrogate actigraphy computed from the RIP signal alone. F_R, and F_{RA} correspond to the sets found by Long et al. [139] to give the best performance when respiratory effort and respiratory effort plus actigraphy were used, respectively. These served as benchmark for the performance obtained when surrogate actigraphy computed from RIP was used together with the respiratory features (F_{RX}) instead of reference actigraphy.

The last three feature sets were used to test surrogate actigraphy computed from combined ECG and RIP. F_{RCA} corresponds to the feature set found by Long et al. [139] to give the best performance when respiratory, cardiac and actigraphy features were used. F_{RC} corresponds to the same feature set but without actigraphy, and was used to establish the baseline performance improvement due to actigraphy. F_{RCX} comprises, besides the same respiratory and cardiac features, surrogate actigraphy computed from ECG and RIP.

A Bayesian linear discriminant classifier with time-varying prior probabilities similar to the one used by Long et al. [139] was also used in this work to classify Sleep and Wake. Despite its simplicity, this classifier has been shown to perform very well in this task [80, 139]. For each epoch, the selected class (Sleep or Wake) is the class ω_i that maximizes the posterior probability given an feature vector F [155],

$$\omega_i(F) = \arg\max_i [\mathrm{g_i}(F)] \tag{3.10}$$

where the discriminant function $\mathrm{g_i}$ for each class is given by

$$\mathrm{g_i}(F) = -\frac{1}{2}(F-\mu_i)^t \Sigma^{-1}(F-\mu_i) + \ln \mathrm{P}(\omega_i, t) \tag{3.11}$$

and where μ_i is the average feature vector for class i, Σ is the pooled covariance matrix for both classes, and $\mathrm{P}(\omega_i, t)$ is the prior probability for class i at time (since lights off) t. The time-varying prior probabilities are estimated during training by computing, for each epoch, the relative frequency of each class (according to the ground-truth) on all recordings on the training set [80].

Separate classifiers were trained and evaluated with each of the feature sets indicated in Table 3.1 using a leave-one-subject-out cross-validation procedure. The classification performance obtained with each feature set was evaluated using Cohen's κ coefficient of agreement [156], a measure of agreement between the result of Sleep/Wake classification and the reference annotations by the sleep technicians. This metric is particularly useful to evaluate performance in the presence of imbalanced classes, as is the case with Sleep and Wake. Since earlier work [139] only reports the performance in a set of healthy subjects, our results were analyzed separately for healthy subjects ($N = 15$) and for the entire data set ($N = 40$), allowing for a fair comparison with the results obtained in that work.

In order to evaluate whether the classifier with surrogate actigraphy provides an equivalent performance to the classifier with reference actigraphy, an equivalence test with a confidence interval (CI) of 90% was used [157]. Since the performance values (κ) do not follow a normal distribu-

Table 3.1: Feature sets used to train the different classifiers and compare performance with and without actigraphy, and with surrogate actigraphy

Feature set	Features (#)	Modality[a]	Feature set in [139]
F_R	`dtw`, `dfw` (2)	R	F_{DW}
F_{RA}	`ac`, `dtw`, `dfw` (3)	A, R	F_{AC-DW}
F_{RX}	X_{RIP}, `dtw`, `dfw` (3)	A†, R	n.a.
F_{RC}	`sdf`, `hfc`, `se`, `dtw`, `dfw`, `mhr` (6)	R, C	n.a.
F_{RCA}	`ac`, `sdf`, `hfc`, `se`, `dtw`, `dfw`, `mhr` (7)	A, R, C	F_{ARC2}
F_{RCX}	$X_{ECG+RIP}$, `sdf`, `hfc`, `se`, `dtw`, `dfw`, `mhr` (7)	A‡, R, C	n.a.

[a] A: actigraphy, R: respiratory effort, C: cardiac activity
† Surrogate actigraphy, estimated from RIP
‡ Surrogate actigraphy, estimated from ECG and RIP

tion, the CI was determined using a non-parametric method based on Wilcoxon's signed rank test [158]. Equivalence was statistically established if the 90% CI for difference in performance using surrogate actigraphy instead of reference actigraphy was fully contained in the interval established by a predefined margin δ, i.e., $(-\delta, \delta)$. The margin δ was chosen according to the recommendations of Walker and Nowacki [157], i.e., as the the lower bound of the 90% CI for difference in performance using actigraphy instead of only cardiac and/or respiratory features. This is the same as using a two one-sided test procedure, where equivalence is established if the null hypotheses

$$H_{01} : \text{med}(\kappa_a) - \text{med}(\kappa_b) < -\delta, \tag{3.12}$$

and

$$H_{02} : \text{med}(\kappa_a) - \text{med}(\kappa_b) > \delta \tag{3.13}$$

can be both rejected with $p < .05$. $\text{med}(\kappa)$ stands for the median κ obtained with a given condition for all subjets in the data set. In case only H_{01} (3.12) can be rejected, then non-inferiority (but not equivalence) is statistically established.

In addition, traditional performance metrics of precision, sensitivity, specificity, accuracy, and the area under the precision-recall curve were computed for the classification results obtained with each feature set.

3.3 Results

Table 3.2 indicates the correlation values obtained for each modality. The lowest pooled correlation is obtained with ECG, 0.62, followed by RIP, 0.65. The algorithm performs best when combining both modalities, achieving a pooled correlation coefficient of 0.74 and an average correlation coefficient of 0.67. All correlation coefficients were found to be significant with $p < .001$.

A visual inspection of the Bland-Altman plot between surrogate (EST) and reference (REF) actigraphy pooled across all subjects revelead heteroscedasticity, confirmed by a Spearman's corre-

Table 3.2: Correlation between estimated and reference actigraphy, and number of coefficients used from each signal

				Nr. ECG coeff.				Nr. RIP coeff.			
Signal(s)	ρ	95% CI	$\bar{\rho}$	P_m	P_M	$\bar{P}$	*tot*	P_m	P_M	$\bar{P}$	*tot*
ECG	0.62^{***}	(0.61, 0.62)	0.56	27	38	3	68				
RIP	0.65^{***}	(0.65, 0.66)	0.64					60	77	12	149
ECG, RIP	0.74^{***}	(0.73, 0.74)	0.67	15	28	1	44	57	73	5	135

*** $p < .001$

Note: P_m, P_M and $\bar{P}$ indicate the number of coefficients selected, for each modality, from the wavelet power statistics (minimum, maximum and average, respectively) in (3.5). *tot* indicates the total number of coefficients selected for each modality.

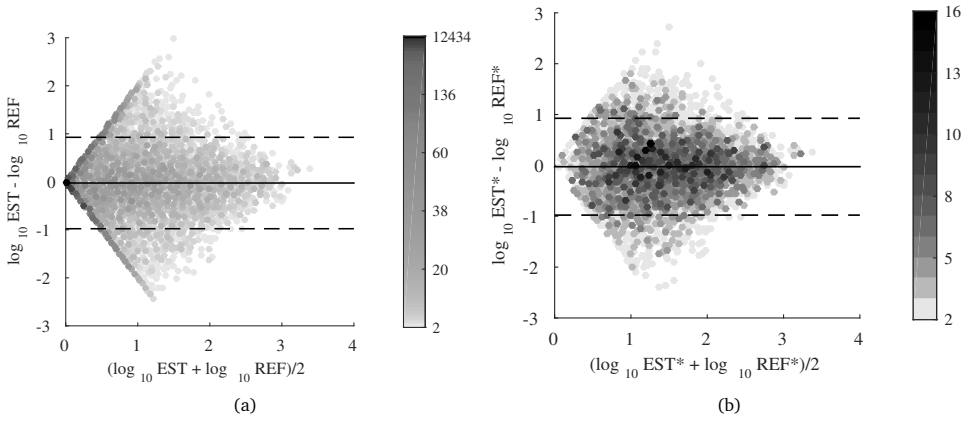


Figure 3.4: Bland-Altman plots (a) after log transformation, with mean bias ratio and 95% LoA indicated by the solid and dashed lines respectively, and (b) after restricting the plot to points where both methods agree regarding the presence of body movements, i.e., where both reference and surrogate actigraphy have a value greater or equal than 1.

lation of 0.81 ($p < .001$) between the absolute value of the differences and the average actigraphy values. Comparing log transformed data, as suggested by Bland and Altman [159], we obtained as mean difference ($\log_{10}\text{EST} - \log_{10}\text{REF}$) a value of -0.03 and 95% limits of agreement (LoA) $[-0.98, 0.93]$ (Figure 3.4a). Back-transforming (with antilog) the results we obtained values relating to the ratios between the two measurements. The pooled geometric mean bias ratio between EST and REF was 0.94 (average 1.06 ± 0.62 across all subjects) with 95% agreement ratio between 0.11 and 8.45 (0.15 ± 0.07 and 9.23 ± 10.87). Expressing the agreement ratios as function of the average actigraphy values [160] we obtain

$$-1.62\ \text{AVG} < \text{EST} - \text{REF} < 1.58\ \text{AVG}, \tag{3.14}$$

with $\text{AVG} = (\text{EST} + \text{REF})/2$.

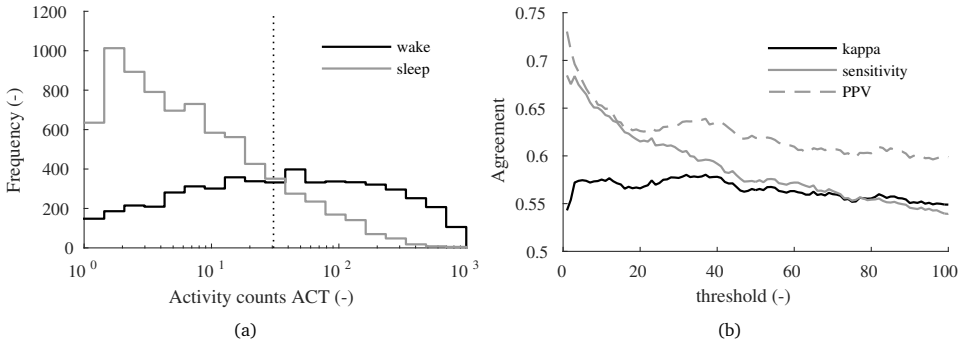


Figure 3.5: (a) histogram of actigraphy values for Sleep and Wake classes, with optimal separation indicated with a dotted line, and (b) agreement on presence of body movements detected with surrogate and reference actigraphy for thresholds below 100.

Correlation analysis was used to test the effects of age, body mass index, sleep efficiency and total sleep time on the correlation between surrogate and reference actigraphy, mean bias ratio and on the 95% limits of ratio agreement obtained for each subject, but no significant effects were found. Also, no significant differences between males and females were found after a t-test comparison of these performance metrics.

Figure 3.5a illustrates the histograms of (reference) actigraphy values for the classes Sleep and Wake for all subjects. Optimal separation (assuming equal prior probabilities) is achieved with a threshold of 30 (indicated with a dotted line in the figure). Using this threshold as average actigraphy in equation (3.14) we obtain as LoA $[-48.6, 47.4]$.

Figure 3.5b illustrates the agreement on presence of body movements (as defined by actigraphy above a given threshold) detected with surrogate and reference actigraphy for different metrics. For a threshold of 30, we obtain a κ of 0.58, a sensitivity of 0.61 and a PPV of 0.63. Overall, agreement is higher for lower thresholds, with sensitivity and PPV achieving a maximum of 0.68 and 0.73 respectively, for a threshold of 1.

Table 3.3 and 3.4 indicate the performance of Sleep/Wake classification for each feature set for the healthy subjects and for all subjects, respectively. Note that the results obtained with the same feature sets used by Long et al. [139] for healthy subjects are slightly different than those reported in that paper. The small differences are explained by the different handling of epochs where features could not be computed (e.g. no QRS complexes detected due to motion artifacts): Long et al. excluded those expochs while we linearly interpolated missing values in order to classify all epochs in each recording.

In case only respiratory features and actigraphy are used, surrogate actigraphy (F_{RX}) provides a comparable average performance across all metrics to the reference feature set with standard actigraphy (F_{RA}) both for healthy as well as for all subjects. The κ obtained was found to be significantly

Table 3.3: Mean and standard deviation of classification performance for healthy subjects ($N = 15$), per feature set

Feature set	Precision	Sensitivity	Specificity	Accuracy	AUC_{PR}	κ
F_R	0.61 (0.17)	0.67 (0.10)	0.96 (0.01)	0.92 (0.04)	0.66 (0.13)	0.58 (0.10)
F_{RA}	0.69 (0.16)	0.69 (0.13)	0.97 (0.01)	0.94 (0.02)	0.73 (0.12)	0.64 (0.11)***
F_{RX}	0.69 (0.15)	0.70 (0.11)	0.97 (0.01)	0.93 (0.04)	0.74 (0.13)	0.64 (0.09)***
F_{RC}	0.72 (0.19)	0.67 (0.16)	0.96 (0.04)	0.93 (0.04)	0.73 (0.12)	0.63 (0.14)
F_{RCA}	0.72 (0.18)	0.72 (0.15)	0.96 (0.04)	0.94 (0.04)	0.77 (0.12)	0.67 (0.14)***
F_{RCX}	0.69 (0.20)	0.76 (0.13)	0.95 (0.05)	0.93 (0.05)	0.76 (0.13)	0.66 (0.16)*

AUC_{PR}: area under the precision-recall curve
$^{*}p < .05$, $^{***}p < .001$, after a two-tailed Wilcoxon-signed rank test of whether the performance (κ) obtained after adding (surrogate) actigraphy to F_R and to F_{RC} was significantly different than without actigraphy.

Table 3.4: Mean and standard deviation of classification performance for all subjects ($N = 40$), per feature set

Feature set	Precision	Sensitivity	Specificity	Accuracy	AUC_{PR}	κ
F_R	0.60 (0.19)	0.58 (0.18)	0.93 (0.04)	0.83 (0.11)	0.65 (0.15)	0.45 (0.16)
F_{RA}	0.65 (0.19)	0.60 (0.19)	0.94 (0.04)	0.85 (0.11)	0.69 (0.16)	0.50 (0.17)***
F_{RX}	0.63 (0.20)	0.62 (0.21)	0.93 (0.07)	0.85 (0.11)	0.69 (0.17)	0.50 (0.18)**
F_{RC}	0.67 (0.20)	0.65 (0.21)	0.92 (0.10)	0.85 (0.11)	0.71 (0.15)	0.53 (0.17)
F_{RCA}	0.71 (0.19)	0.65 (0.21)	0.94 (0.08)	0.87 (0.10)	0.74 (0.15)	0.56 (0.17)***
F_{RCX}	0.70 (0.21)	0.67 (0.22)	0.93 (0.09)	0.87 (0.10)	0.73 (0.16)	0.56 (0.19)**

AUC_{PR}: area under the precision-recall curve
$^{**}p < .01$, $^{***}p < .001$, after a two-tailed Wilcoxon-signed rank test of whether the performance (κ) obtained after adding (surrogate) actigraphy to F_R and to F_{RC} was significantly different than without actigraphy.

higher than when respiratory features are used alone, without actigraphy (F_R). Furthermore, as illustrated in Figure 3.6a and 3.6b, the performance of the feature set with surrogate actigraphy (F_{RX}) is statistically non-inferior to the performance with reference actigraphy ($\kappa_{RX} - \kappa_{RA}$).

In case respiratory and cardiac features are used together with actigraphy, surrogate actigraphy (F_{RCX}) once again provides a comparable average performance to the reference feature set with standard actigraphy (F_{RCA}) both for healthy as well as for all subjects. Furthermore, surrogate actigraphy provides a statistically significantly higher κ when compared to the feature set without actigraphy (F_{RC}) and as illustrated in Figure 3.7a and 3.7b, the performance of that feature set ($\kappa_{RCX} - \kappa_{RCA}$) is statistically non-inferior to the performance with reference actigraphy.

3.4 Discussion

As indicated in Table 3.2, the CWT performs overall better when combining both modalities, achieving a correlation coefficient of 0.74 for ECG+RIP (significant with $p < .001$). The correlations obtained for ECG and RIP were similar, suggesting that this method is equally able to quantify body movements in both signals. The correlation increase obtained by combining ECG and RIP suggests that the BMA information in each signal is somewhat complementary, and the best estimate is obtained by combining both.

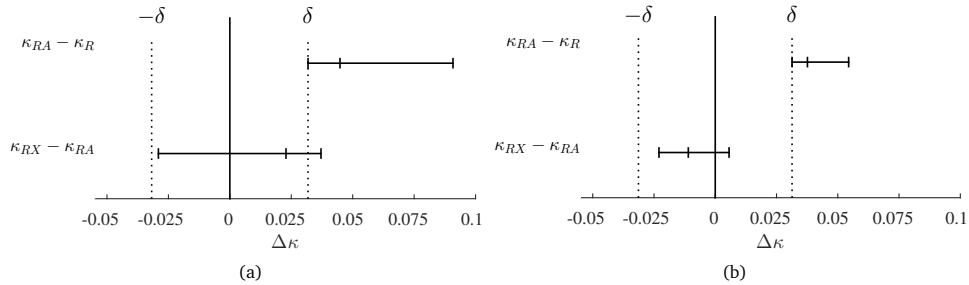


Figure 3.6: Confidence intervals (90%) for difference in κ obtained with different feature sets with (surrogate) actigraphy using respiratory features for (a) healthy and (b) all subjects. The lower bound of the CI for $\kappa_{RA} - \kappa_R$ is used to establish the equivalence margin δ for the comparison $\kappa_{RX} - \kappa_{RA}$. Equivalence for a given comparison is established if the corresponding CI is fully comprised between $(-\delta, \delta)$ (indicated with the vertical dashed lines). Non-inferiority is established if the lower bound of the CI is larger than $-\delta$. In this case, non-inferiority with surrogate actigraphy can be established for healthy subjects and equivalence can be established for all subjects.

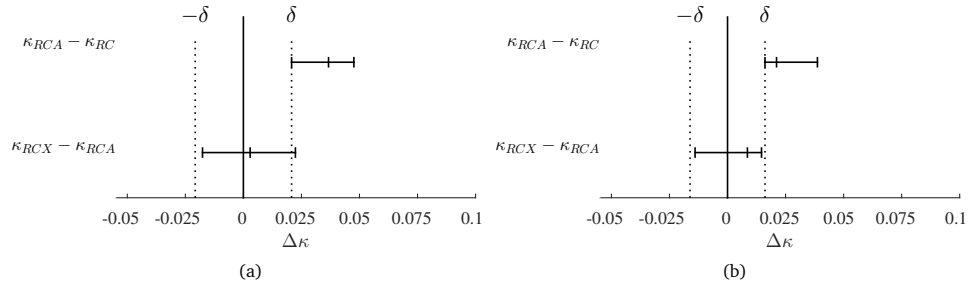


Figure 3.7: Confidence intervals (90%) for difference in κ obtained with different feature sets with (surrogate) actigraphy using respiratory and cardiac features for (a) healthy and (b) all subjects. The lower bound of the CI for $\kappa_{RCA} - \kappa_{RC}$ is used to establish the equivalence margin δ for the comparison $\kappa_{RCX} - \kappa_{RCA}$. Also in this case, non-inferiority with surrogate actigraphy can be established for healthy subjects and equivalence can be established for all subjects.

Regarding the Bland-Altman analysis (Figure 3.4), the heteroscedasticity found in the data suggests that this method is less suited for precisely quantifying moderate to large movements. This can be partly explained by the different location of the sensors: reference actigraphy was measured on the wrist, and surrogate actigraphy was calculated from sensors placed on the chest. Larger movements of one of the parts of the body may simply not have an equal correspondence in terms of amplitude on the other.

Furthermore, it is clear that a number of points in the Bland-Altman plot correspond to a value of zero from either surrogate or from reference actigraphy (3.4a). This can also be attributed to the different location of the sensors. For small movements of either the wrist or the chest, there isn't always a corresponding movement in the other body location. Overall, however, the method

achieves a reasonable agreement in the detection of small body movements. For a threshold of 1, it yields a sensitivity and PPV of 0.68 and 0.73 respectively, meaning that close to 70% of all wrist movements were detected with surrogate actigraphy on the chest, and vice versa. This sensitivity to movements of low amplitude is confirmed by the 95% LoA calculated as function of average actigraphy with (3.14) and which, for a value of 30 (which gives an optimal separation between Sleep and Wake) are moderate ($[-48.6, 47.4]$). Restricting the Bland-Altman plot to points where both methods agree regarding the presence of body movements (Figure 3.4b), the agreement between (log) actigraphy estimations is more clearly visible. From an analysis of the plot structure we note that for very large values of actigraphy (above 10^2) agreement is better than for intermediate values. A possible explanation for this is that very large, isolated movements, which likely correspond to movements of the entire body, affect both sensor readings in the two locations in a comparable way.

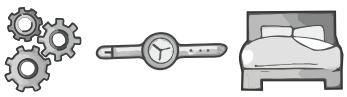
These results suggest that surrogate actigraphy is adequate for quantifying small body movements and, overall, to detect the presence of movements of any intensity and henceforth, to classify Sleep and Wake. This is confirmed by the results of Sleep/Wake classification, where surrogate actigraphy achieves statistically non-inferior performance to reference actigraphy when combined with RIP or with ECG and RIP signals, both for healthy as well as for all subjects.

3.5 Conclusions

Although recent work in unobtrusive Sleep/Wake classification has shown that cardiac and respiratory features are useful, actigraphy remains the single most discriminative modality for this task. Unfortunately it requires an additional dedicated sensor, typically a wrist-worn actigraphy device with an accelerometer. This chapter proposes a method to estimate actigraphy from the body movement artifacts present in the ECG and RIP signals based on the time-frequency analysis of those signals. It provides a surrogate measure of actigraphy with moderate correlation with reference actigraphy, and reasonable agreement on detection of body movements of low intensity. More importantly, it can be used as a replacement of actigraphy in Sleep/Wake classification without loss in performance. Although the best performance is obtained with ECG and RIP combined, it should be emphasized that the largest gains of adding actigraphy for Sleep/Wake classification were found using RIP alone. Importantly, RIP is one of the most used modalities in home sleep monitoring: together with pulse oximetry and respiratory flow, it is very often used for screening or diagnosis of sleep disordered breathing with home sleep tests.

The consequences are important: in case only respiratory effort and actigraphy are used for Sleep/Wake classification, this method eliminates the need to use an actigraphy device, relying only on the signals measured with a single (RIP) sensor. In case both respiratory effort and ECG are used together with actigraphy, it can also be used to eliminate the need for an actigraphy device, allowing the set-up to be restricted to two sensors (ECG and RIP). Although the economic impact of this benefit might seem small given the low price of accelerometers, it offers a major benefit in terms

of comfort, in particular for longer-term monitoring. Additionally, it is immediately applicable for legacy ECG and RIP monitoring devices already used in clinical practice and which do not have an accelerometer built-in. Finally, it allows the retrospective analysis of existing sleep data sets for which no actigraphy was collected but where body movements may be of clinical importance.



Part 3

Sleep stage classification based on unobtrusive modalities

Quero sentir o sono chegar como vida, e não como repouso.

I want to feel sleep come to me as life, not as rest.

- Bernardo Soares (heteronym), Livro do Desassossego
(*transl.* Richard Zenith, The Book of Disquiet)

Chapter 4

Sleep stage classification with ECG and respiratory effort

Adapted from P. Fonseca*, X. Long*, M. Radha, R. Haakma, R. M. Aarts, and J. Rolink, Sleep stage classification with ECG and respiratory effort, *Physiological Measurement*, vol. 36, no. 10, pp. 2027-40, 2015 (*joint first authorship).

Abstract: Automatic sleep stage classification with cardiorespiratory signals has attracted increasing attention. In contrast to the traditional manual scoring based on polysomnography (PSG), these signals can be measured using advanced unobtrusive techniques that are currently available, promising the application for personal and continuous home sleep monitoring. This chapter describes a methodology for classifying Wake, rapid eye movement (REM) sleep, and non-REM (NREM) light (N1 and N2) and deep (N3) sleep on a 30-second epoch basis. A total of 142 features were extracted from electrocardiography (ECG) and thoracic respiratory effort measured with respiratory inductance plethysmography (RIP). To improve the quality of these features, subject-specific Z-score normalization and spline smoothing were used to reduce between-subject and within-subject variability. A modified sequential forward search (SFS) feature selection procedure was applied, yielding 80 features while preventing the introduction of bias in the estimation of cross-validation performance. PSG data from 48 healthy adults were used to validate our methods. Using a Bayesian linear discriminant classifier and ten-fold cross-validation, we achieved a Cohen's κ coefficient of 0.49 and an accuracy of 69% in the four-class classification of Wake, REM, combined N1 and N2, and N3. These values increased to $\kappa = 0.56$ and accuracy = 80% when the classification problem was reduced to three classes, Wake, REM sleep, and NREM sleep.

4.1 Introduction

A problem with traditional sleep monitoring, known as polysomnography (PSG), is that a wide array of potentially sleep-disturbing sensors must be applied to the body and their measurements can only be interpreted by highly trained sleep technicians or scientists. Albeit invaluable in the diagnostic of sleep disorders, traditional PSG is rather ill-suited for regular, non-diagnostic monitoring of sleep and will only introduce more sleep disturbances when applied on a daily basis by untrained individuals. This scenario makes apparent a need for unobtrusive methods of sleep monitoring, preferably inexpensive and with no training required to operate them. Cardiorespiratory monitoring can be unobtrusive and the data can be analyzed by a computer, which makes this technology a promising candidate for personal, continuous and unobtrusive sleep monitoring.

Cardiorespiratory sleep staging or sleep stage classification is often based on heart rate variability (HRV) calculated from electrocardiography (ECG) and respiratory effort, often from respiratory inductance plethysmography (RIP). Usually cardiorespiratory information is combined with body movements from an accelerometer to more accurately distinguish Wake from Sleep. One of the earliest studies that presented a successful machine learning approach to cardiorespiratory sleep stage classification with these modalities was done by Redmond and Heneghan [98]. Using a set of HRV and respiration rate variability (RRV) features to characterize autonomic nervous activity activity, Redmond and colleagues showed the viability of a sleep stage classifier that can generate a simplified hypnogram for an entire night indicating, for each 30-second segment, a sleep stage classified as either Wake, rapid eye movement (REM) sleep, or non-REM (NREM), in a Wake/non-REM/REM (W/NR/R) classification for short. More recent research has shown that it is possible to obtain the same cardiorespiratory information from other sensors for sleep stage classification, such as from bed-mounted ballistocardiogram [93, 97] or contactless radio frequency [94]. Although these studies focused on distinction between Wake, REM sleep, and NREM sleep (without separating NREM sleep in other sleep stages) or between Wake and Sleep (merging REM and NREM sleep), these attempts promised that cardiorespiratory methods could one day be completely unobtrusive.

In previous work [161, 162] we proposed methods to simultaneously classify Wake, REM sleep, light sleep (NREM stage S1 and S2), and deep sleep or slow wave sleep (stage S3 and S4) using respiratory activity in order to estimate an overnight Wake/N1 and N2/N3/REM (W/N1-2/N3/R) hypnogram. In comparison with W/NR/R classification, achieving W/N1-2/N3/R classification would allow a more adequate assessment of sleep because N3 is thought by some researchers to be important in several cortical and physiological processes, such as energy conservation [163], cerebral restoration [164], and memory processing and consolidation (together with sleep spindles which occur during NREM stage 2) [165, 166]. In that work, we also reviewed the state-of-the-art in sleep stage classification with cardiac and/or respiratory activity. The methods presented there will be used to benchmark the method proposed in this chapter. In addition, two studies with comparable results have been proposed by Domingues et al. [99] and Willemen et al. [101].

However, these works only report results on a three-class task (W/NR/R classification) or use non-standard one-minute epochs for classification, respectively.

This chapter presents a methodology for automatic sleep stage classification based on machine learned models of the autonomic nervous system during sleep from ECG and RIP signals. Compared to previous studies, our methodology includes novel features, new feature post-processing methods, and a refined feature selection method which guarantees that no bias is introduced in the validation of the algorithm while avoiding the use of a hold-out validation set. These methods are applied for three-class (W/NR/R) and four-class (W/N1-2/N3/R) sleep stage classification of healthy subjects.

4.2 Materials and methods

4.2.1 Data sets

The data set was the same as used in earlier work [161, 162] and comprised full single-night PSG recordings of 48 subjects (27 females) acquired in the SIESTA project [167]. All subjects were healthy sleepers with a Pittsburgh Sleep Quality Index [148] of less than 6 and had no regular sleep complaints nor earlier diagnosis of sleep disorders. The subjects had an average age of 41.3 (±16.1) years at the time of the recording. Full subject demographics can be found in our earlier work [161]. Sleep stages were scored by trained sleep technicians in six classes according to the R&K rules [168]. In the scope of this study, S1 and S2 were merged in a single N1-2 class and S3 and S4 were merged in a single N3 class. Each PSG recording comprised, besides the standard signals required for sleep scoring, modified lead II ECG, and (thoracic) respiratory effort recorded with RIP. QRS complexes were detected and localized from ECG signals using a combination of a Hamilton-Tompkins detector [123, 124] and the post-processing localization algorithm described in Chapter 2 [150]. Prior to feature extraction, RIP signals were filtered with a 10th order Butterworth low-pass filter with a cut-off frequency of 0.6 Hz, after which baseline was removed by subtracting the median peak-to-through amplitude [161].

4.2.2 Feature extraction

We extracted a set of 142 features from cardiac and respiratory activity, and from cardiorespiratory interaction (CRI) using a sliding window centered on each 30-second epoch, guaranteeing sufficient data to capture the changes in autonomic activity [118]. Because some features are computed based on windows which exceed the epoch length, epochs at the start and end of each recording required a special handling: for each such feature, all epochs for which the window crosses the boundaries of the recording were marked as invalid; the feature values for these epochs were interpolated during post-processing using spline fitting (Section 4.2.2).

Cardiac features

Considering cardiac activity, 86 cardiac features were computed from the QRS complexes detected in the ECG signal. Time domain features, computed over nine epochs, include mean heart rate, mean heartbeat interval (detrended and non-detrended), standard deviation (SD) of heartbeat intervals, difference between maximal and minimal heartbeat intervals, root mean square and SD of successive heartbeat interval differences, and percentage of successive heartbeat intervals longer than >50 ms [80, 118]. We also computed the mean absolute difference and different percentiles (at 10%, 25%, 50%, 75%, and 90%) of detrended and non-detrended heart rates and heartbeat intervals [101, 103] as well as the mean, median, minimal, and maximal arousal likelihood ratios based on consecutive interbeat intervals (IBIs) [169]. In the frequency domain, the features included the logarithmic spectral powers in the very low frequency band (VLF) from 0.003 to 0.04 Hz, in the low frequency band (LF) from 0.04 to 0.15 Hz, in the high frequency band (HF) between 0.15 to 0.4 Hz, and the LF/HF ratio [170], where the power spectral densities were estimated over nine epochs. The spectral boundaries were adapted to the corresponding peak frequency, yielding their boundary-adapted versions [171]. We also computed the maximum module and phase of the HF pole [172] and the maximum power in the HF and its associated frequency, representing respiratory rate [80]. Features describing non-linear properties of heartbeat intervals were quantified with detrended fluctuation analysis (DFA) over 11 epochs [173] and its short-term (α_1), long-term (α_2), and all-time scaling exponents [174, 175], progressive DFA with non-overlapping segments of 64 heartbeats [176], windowed DFA over 11 epochs [177], and multi-scale sample entropy (MSE) over 17 epochs (length of 1 and 2 samples with scales of 1-10) [178]. The approximate entropy of the symbolic binary sequence that encodes the increase or decrease in successive heartbeat intervals over nine epochs was also calculated [179]. In addition, we propose new features based on a visibility graph (VG) and a difference visibility graph (DVG) method to characterize HRV time series in a two-dimensional complex network where samples are connected as nodes in terms of certain criteria [180, 181]. The network-based features, computed over seven epochs, comprised the mean, SD, and slope of node degrees and number of nodes in VG- and DVG-based networks with a small degree (≤ 3 for VG and ≤ 2 for DVG) and a large degree (≥ 10 for VG and ≥ 8 for DVG), and assortativity coefficient in the VG-based network [181, 182, 183].

Respiratory features

Concerning respiratory activity, 44 features were derived from the RIP signals. In the time domain, we estimated the variance of the respiratory effort signal, the respiratory frequency and its SD over 150, 210, and 270 s, the mean and SD of breath-by-breath correlations, and the SD of the breath lengths [80]. One of our previous studies [161] introduced respiratory amplitude features for sleep stage classification, including the standardized mean, standardized median, and sample entropy of respiratory peaks and troughs, median peak-to-trough difference, median volume and flow rate

for the complete breath cycle, inhalation, and exhalation, and inhalation-to-exhalation flow rate ratio. These features were adopted in this work. Besides, we also computed the similarity between the peaks and troughs by means of the envelope morphology using a dynamic time warping (DTW) metric [184]. From the respiratory spectrum, the respiratory frequency and its power, the logarithm of the spectral power in VLF (0.01-0.05 Hz), LF (0.05-0.15 Hz), and HF (0.15-0.5 Hz) bands, and the LF/HF ratio were estimated [98]. Respiratory regularity was measured by means of sample entropy over seven epochs [161, 185] and self-(dis)similarity based on DTW and dynamic frequency warping (DFW) [139] and uniform scaling [162] were derived. The same VG and a DVG network analysis features as for HRV were also computed for breath-to-breath intervals.

Cardiorespiratory interaction features

Numerous studies have shown that the interaction between cardiac and respiratory activity varies across sleep stages [181, 186, 187]. The power associated with respiratory-modulated heartbeat intervals was quantified over windows of nine epochs [186]. In addition, we also extracted the VG- and DVG-based features for CRI [181]. These resulted in a total of 12 CRI features in our feature set.

Feature post-processing

In order to reduce the impact of physiological differences and equipment-related variations from subject to subject, the features of each subject were first Z-score normalized by subtracting their mean and dividing by their SD. Further, it is known that the sleep pattern of healthy adults progresses with several cycles throughout the night [2]. For example, REM and NREM sleep alternate with 4-6 cycles of about 90-110 minutes with N3 usually dominating the NREM periods during the first half of the night. This suggests that the autonomic physiological response with its associated sleep stage is time-variant across the night for each subject. For this reason, we were motivated to smooth each feature for each subject by means of a cubic spline fitting method [188]. This is also expected to help reduce signal measurement noise and variability within subjects for each sleep stage conveyed by the feature values. The latter can be caused by body movements, conscious breathing control, internal physiological variations, or other external factors such as changes in environmental noise and temperature during bedtime sleep. Instead of other simpler low-pass filters, spline fitting was chosen because it can interpolate feature values which could not be computed, for example due to motion artifacts (about 10% of the epochs observed in our data set) or at the start and end of each recording for features computed with windows exceeding the epoch duration. This procedure allows all epochs in each recording to be classified.

Let $\mathbf{t}$ represent a sequence of feature values $\mathbf{v} = \{v_1, v_2, ..., v_n\}$ at their corresponding time (or epoch) indices $\mathbf{t} = \{t_1, t_2, ..., t_n\}$ (in 30 s), then a relation between them can be modeled by

$$v_i = h(t_i) + \varepsilon_i \quad (i = 1, 2, ..., n), \tag{4.1}$$

where h is a smoothing (spline) function, and ε_i are independent and identically distributed residuals. The smoothing function can be estimated by minimizing the objective function (penalized sum of squares) such that

$$\hat{h} = \arg\min_{h} \left[\sum_{i=1}^{n} [v_i - h(t_i)]^2 + \lambda \int_{t_1}^{t_n} h''(t)^2 dt \right], \tag{4.2}$$

where λ is a smoothing parameter that controls the trade-off between residual and local variation. The smoothing function can be expressed by cubic B-splines as basis functions and determined via least squares approximation [188, 189].

Each overnight recording with a total of m epochs is divided in s continuous segments ($s = \lceil m/n \rceil$), designated as smoothing splines. Each segment can then be modeled by the spline function, yielding a general spline fitting for the epochs over the entire recording. n represents the smoothing window size, and a larger n translates to a smoother fitting curve. In this work, a window size of nine epochs for modeling splines was experimentally found to be appropriate for the task of sleep stage classification.

4.2.3 Classifier

This work used a multi-class Bayesian linear discriminant with time-varying prior probabilities [80], similar to that used in previous work [161]. For each epoch, the selected class (N3, N1-2, R, or W) is the class ω_i that maximizes the posterior probability given an feature vector $\mathbf{x}$ [155],

$$\omega_i(\mathbf{x}) = \arg\max_{i} [\mathrm{g_i}(\mathbf{x})] \tag{4.3}$$

with the the discriminant function $\mathrm{g_i}$ for each class given by

$$\mathrm{g_i}(\mathbf{x}) = -\frac{1}{2} (\mathbf{x} - \mu_i)^t \, \Sigma^{-1} (\mathbf{x} - \mu_i) + \ln \mathrm{P}(\omega_i, t) \tag{4.4}$$

where μ_i is the average feature vector for class i, Σ is the pooled covariance matrix for all classes, and $\mathrm{P}(\omega_i, t)$ is the prior probability for class i at time (since lights off) t. All parameters were estimated during training.

4.2.4 Feature selection

To select the final list of features we used a wrapper feature selection method based on sequential forward search (SFS) [190] using as criterion the Cohen's κ coefficient of agreement [156] on the training set. This measure of agreement between the classification predictions and the ground-truth annotations is more adequate than traditional measures of accuracy for this problem because there is a strong imbalance between classes (N1-2 epochs, for example, account for more than 50% of all epochs in the data set) and this coefficient factors out chance agreement, compensating for

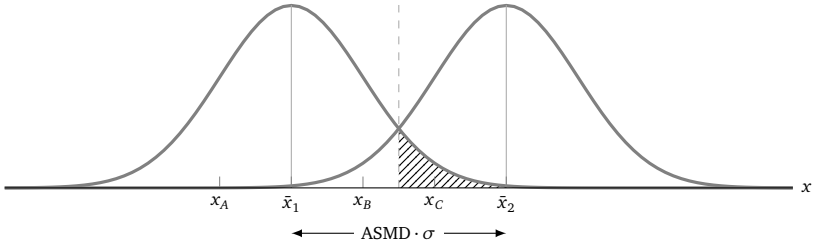


Figure 4.1: Relation between ASMD and separation between two normally distributed classes with the same standard deviation

class imbalance.

In many machine learning studies supervised feature selection is often applied on the entire data set, even if the training and validation are kept separate (for example using cross-validation). This common pitfall is known to introduce a bias in the evaluation of a classifier's performance, which will often be overestimated [191]. Although keeping a hold-out set for validation would solve this problem, the limited size of the data set would either mean that the model learning would be based on potentially insufficient examples, or that the classifier would be evaluated on a very small sample, possibly unrepresentative of the problem at hand. Instead, the feature selection procedure was executed by strictly separating, on an iterative procedure akin to cross-validation, the training and validation sets. For each iteration, unbound SFS was applied using as criteria the classification performance obtained in the training set of each iteration. The final number of features was chosen as the smallest number S that yield a certain percentage of the maximum training κ obtained across all iterations. The final list of selected features was chosen as the S features most often selected during the process.

The discriminative power of selected features was evaluated with the absolute standardized mean distance (ASMD) between the feature values of two classes, computed as

$$\text{ASMD} = \left| \frac{\bar{x}_1 - \bar{x}_2}{\sigma} \right| \tag{4.5}$$

where $\bar{x}_1$ and $\bar{x}_2$ are the sample means for class 1 and 2 and σ is the pooled sample SD.

To give an intuitive interpretation of the ASMD score, consider Figure 4.1 which illustrates the feature distribution for two normally distributed classes which share the same standard deviation. The probability of correctly assigning a class based only on the feature observation depends on the ASMD between the two classes. Assuming that a given observation of class 1 has a lower value than the critical point $\bar{x}_1 + 0.5 \cdot \text{ASMD} \cdot \sigma$, such as x_A and x_B in the figure, the criteria for class assignment to the class with highest probability would yield class 1 for these points. Above the critical point, such as for x_C, the same rule would yield, incorrectly, class 2. Generally speaking, and because both classes share the same standard deviation, the Z-score for normal distributions

can be used to determine the probability of correct class assignment based on the ASMD,

$$\begin{aligned} \mathrm{P}(x \leq \bar{x}_1 + 0.5 \cdot \mathrm{ASMD} \cdot \sigma) &= \mathrm{P}\left(Z \leq \frac{\bar{x}_1 + 0.5 \cdot \mathrm{ASMD} \cdot \sigma - \bar{x}_1}{\sigma}\right) \\ &= \mathrm{P}(Z \leq 0.5 \cdot \mathrm{ASMD}) \end{aligned} \tag{4.6}$$

For an ASMD of 0.5, 1, and 2, the probability of correct class assignment based on a feature value alone is approximately 60%, 69%, and 84%, respectively.

4.2.5 Validation and evaluation

After feature selection is performed and the set of features is chosen, the classification results per subject were evaluated using a ten-fold cross-validation procedure. The κ for all subjects in the data set as well as the average and pooled performance were then calculated. In addition to the κ, we also computed the traditional metric accuracy, i.e., the percentage of correctly identified epochs. For κ and accuracy, the results were computed both after pooling the predictions over all epochs of all subjects and after averaging the performance for each subject.

4.3 Results and discussion

4.3.1 Feature selection

Figure 4.2 indicates the κ obtained in the training set of each iteration of the feature selection procedure, for a varying number of features. As illustrated, the maximum average training performance is obtained for 105 features, with an average κ of 0.58. Also clear in the figure, is a plateau in performance between 70 and 100 features. This suggests that the number of features can be greatly decreased without affecting the training performance. A small feature set is often desirable to prevent over-fitting to the training data, as long as it is not so small that the model cannot learn the characteristics of the classes.

By choosing different operating points in Figure 4.2, we can choose a smaller feature set at the expense of a reduction in training performance. By allowing a reduction of 1% in training performance (from the maximum κ of 0.58 to 0.57), we can reduce the number of features from 150 to 80 features (a reduction of 23.8%). Below this number there is a statistically significant (after a Wilcoxon signed-rank test, with $p < .05$) decrease in training performance and further decreasing the number of features will likely lead to a decrease in classification performance after cross-validation. Using as criteria the smallest number of features which does not cause a statistically significant decrease below the maximum training performance, we chose a total of $S = 80$ features.

After ranking all features by the number of times they were selected during the iterative feature selection procedure and selecting the 80 features with the highest count, we found that all features in the final feature set were selected in at least 5 of the 10 iterations (with a mean count of 7.67)

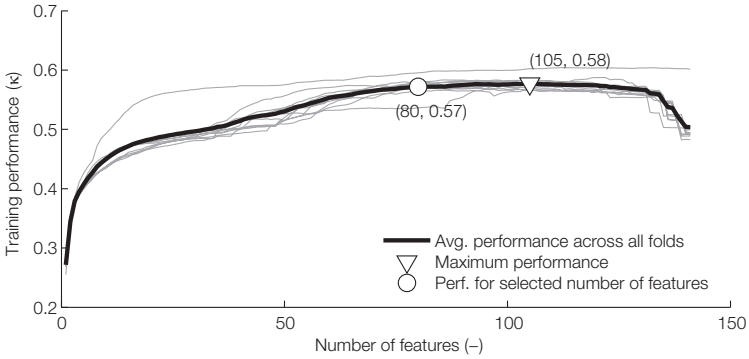


Figure 4.2: Training performance per iteration and average training performance during feature selection. Maximum performance and performance for the selected number of features are indicated with markers (in parenthesis, the number of features and the corresponding performance.)

with 14 features having been selected in all 10 iterations. This illustrates the robustness of the modified SFS method: despite their simplicity and computational efficiency, sequential selection algorithms are known to suffer from a so-called 'nesting effect', potentially leading to sub-optimal feature sets [192]. By iteratively performing several unbound SFS searches on different training sets and keeping only the features that are selected most often, this effect was reduced, as attested by the large number of iterations each feature in the final set was selected.

For brevity only the 14 features selected in all iterations will be discussed further. Table 4.1 indicates the discriminative power of each feature using the pooled ASMD. It was computed for each pair of classes after aggregating the feature values for all subjects and also the 90th percentile of the ASMD (in parenthesis) obtained for each feature, for all individual subjects. Pooled ASMD values below 0.5 were omitted and 90th percentile ASMD values below 1 were omitted.

The top features are clearly discriminative for different pairs of classes which helps explain the relatively large number of features selected. Additionally, it is interesting to observe that there is one feature (median likelihood ratio) which does not have a pooled ASMD above 0.5 for any class pair. However, its 90th percentile ASMD value is larger than 1 for the pairs N3/W and N1-2/W. This is a good example of a feature which is discriminative for only a subset of the subjects (at least 10%) but not for all subjects. The fact that it was selected in every single iteration using the wrapper method described in Section 4.2.4 suggests that it is complementary to other features for some subjects, helping raise the overall training performance.

A note should be made regarding long-term cardiac features such as α_2 or larger-scale MSE features. None of these features were part of the final set of features selected with our method. One possible explanation for this is related to the length of the time series used to compute them. The choice of window sizes (for MSE, 17 epochs, i.e., 8.5 minutes) represents a compromise between having as much data as possible to accurately calculate the features, while at the same time not

Table 4.1: Pooled and 90th percentile ASMD for features selected in all iterations

Feature	N3/N1-2	N3/R	N3/W	N1-2/R	N1-2/W	R/W
Respiratory features:						
VLF spectral power		0.56	1.02		0.86	0.68
		(1.34)	(1.69)		(1.52)	(1.26)
LF/HF spectral power ratio		0.56	0.85		0.95	0.70
		(1.36)	(1.62)	(1.10)	(1.62)	(1.30)
Frequency SD over 270 s	0.79	1.46	1.41	0.84	0.97	
	(1.20)	(1.82)	(1.87)	(1.38)	(1.67)	(1.09)
Mean breath-by-breath correlation	0.59	1.03	0.82			
	(1.27)	(1.78)	(1.76)	(1.61)	(1.51)	(1.46)
Sample entropy regularity		0.71		0.61	0.55	0.86
		(1.67)	(1.53)	(1.41)	(1.49)	(1.65)
DTW self-dissimilarity		0.59	0.86		0.86	0.56
		(1.62)	(1.58)		(1.68)	(1.39)
Standardized mean of troughs	0.82	1.21	0.97	0.56		
	(1.41)	(1.83)	(1.85)	(1.19)	(1.18)	(1.34)
DTW peak-to-trough similarity						0.55
		(1.06)	(1.34)		(1.04)	(1.38)
Uniform scaling self-dissimilarity	0.92	1.46	1.16	0.85	0.56	
	(1.47)	(1.87)	(1.85)	(1.50)	(1.47)	(1.22)
Cardiac (HRV) features:						
Mean likelihood ratio			0.86			
		(1.50)	(1.60)	(1.09)	(1.46)	(1.23)
Median likelihood ratio						
			(1.19)		(1.23)	
Adapted LF spectral power	0.65	0.88	0.70			
	(1.47)	(1.70)	(1.59)	(1.09)	(1.15)	(1.14)
Assortativity coefficient in VG		0.53				
		(1.34)	(1.11)	(1.22)		(1.44)
Number small-degree nodes in VG			0.59			
	(1.05)	(1.39)	(1.32)	(1.14)	(1.13)	(1.24)

The features are described in Section 4.2.2. The pooled ASMD was computed for each pair of classes after aggregating the feature values for all subjects (values below 0.5 were omitted); The 90th ASMD percentiles (in parentheses) were obtained after computing the ASMD of each feature, for each subject (values below 1 were omitted).

exceeding the average length of a given sleep stage (in our data set, the average length of N3 and REM sleep periods was found to be 5.1 and 8.7 minutes, respectively). Although theoretically the window sizes we used are sufficient to calculate these features (a window of 17 epochs corresponds, at an average heart rate of 60 bpm, to 510 samples), it has been suggested that the estimation of sample entropy is low in series shorter than 10^m (where m is the pattern length, in samples) [185, 193]. This means that for a pattern length of $m = 2$ and scales higher than 5, the coarse-grained time series used to calculate the sample entropy will have less data points than the suggested limit and the features might not be accurate, and therefore, not representative of the autonomic characteristics of different sleep stages.

Table 4.2: Cross-validation performance for three and four classes

	Pooled κ	Pooled accuracy	Average κ	Average accuracy
W/N1-2/N3/R	0.49	0.69	0.49 ± 0.13	0.69 ± 0.08
N3	0.51	0.89	0.50 ± 0.17	0.89 ± 0.04
N1-2	0.40	0.71	0.41 ± 0.14	0.71 ± 0.07
R	0.57	0.87	0.58 ± 0.19	0.87 ± 0.08
W	0.54	0.91	0.51 ± 0.18	0.91 ± 0.04
W/NR/R	0.56	0.80	0.56 ± 0.15	0.80 ± 0.08

The pooled performance was computed after aggregating all epochs of all subjects. The mean and SD were calculated based on the performance for each individual subject

Table 4.3: Confusion matrix after cross-validation

pred.↓ ref.→	N3		N1-2		R		W	
N3	**3431**	(7.6%)	1949	(4.3%)	5	(0.0%)	97	(0.2%)
N1-2	2969	(6.6%)	**19165**	(42.6%)	2947	(6.5%)	2302	(5.1%)
R	86	(0.2%)	2071	(4.6%)	**5383**	(12.0%)	404	(0.9%)
W	31	(0.1%)	952	(2.1%)	243	(0.5%)	**2996**	(6.7%)

4.3.2 Cross-validation

Table 4.2 indicates the overall classification performance obtained after 10-fold cross-validation using the selected set of 80 features. In addition, it indicates the performance per class, obtained by considering each class as the positive class and merging the remaining in a single negative class. The highest performance is obtained for R detection, followed by W. The lowest performance is obtained for N1-2. This is further confirmed by the confusion matrix of Table 4.3 which shows that the largest proportion of errors occurs when trying to distinguish N1-2 from the other classes. For all other classes, the percentage of misclassified epochs (relative to the total number of epochs) is below 1% except for N1-2.

In order to evaluate the performance of the classifier in a three-class task (W/NR/R), classes N3 and N1-2 were merged in a single NR class. Table 4.2 indicates the resulting performance. Analyzing the performance of the classifier we see that the classification performance rises substantially, to a κ of 0.56 and an accuracy of 80%. This was expected because a large number of classification errors occurred between N3 and N1-2, and in a W/NR/R task these two classes no longer need to be distinguished.

Figure 4.3 illustrates examples of predicted hypnograms, as compared with the reference, for three subjects in the data set: the subject with the worst performance (with $\kappa = 0.17$), with the median performance (with $\kappa = 0.50$) and with the best performance (with $\kappa = 0.68$). A possible explanation for the poor performance obtained for the worst subject is that the model trained with the characteristics of the general sample population does not fully capture this subject's cardiac and respiratory expression of different sleep stages. However, despite the low κ, the predicted

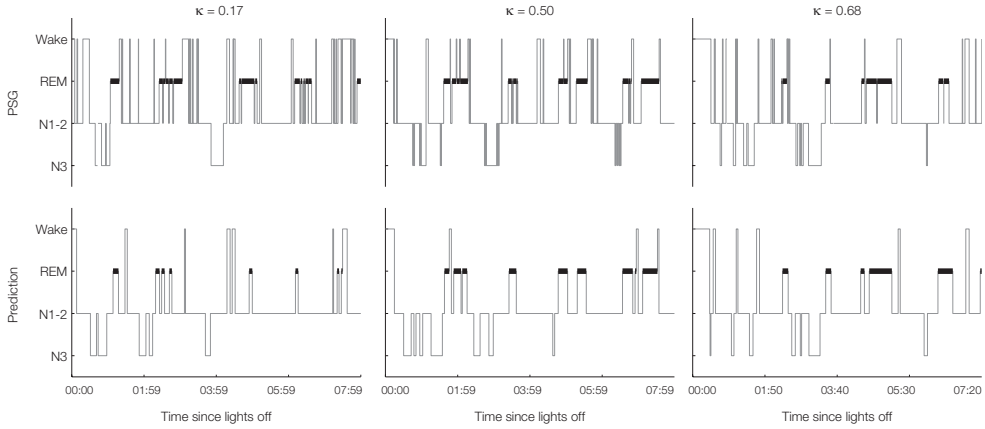


Figure 4.3: Example of sleep stage reference (top) and predictions (bottom) for the subject with the worst performance (left), with the median performance (middle) and with the best performance (right)

hypnogram still exhibits some correct features, namely, most REM intervals were detected, albeit with the incorrect length, and the two N3 periods were also detected. As the performance improves, we see that the predicted hypnograms match better the characteristics of the reference hypnogram, and in the best case the most obvious mistakes are in the missed detection of brief periods of Wake during the night while the rest of the sleep stages are correctly predicted. This is likely caused by the use of spline smoothing during feature post-processing, which is adequate to capture the slow-changing characteristics of most sleep stages, but penalizes short, abrupt changes such as brief periods of awakening.

4.3.3 Comparison with state-of-the-art

Table 4.4 compares the results of our work with other studies reported in literature. As indicated, only a few studies focused on W/N1-2/N3/R classification based on cardiac and/or respiratory signals and our results are amongst the best performing. The first observation is that the results of our previous work [162], which used only respiratory features, are worse than those produced in the present work, indicating that combining cardiac and respiratory activity can lead to an improved classification performance. The study of Hedner et al. [194] achieved similar results but they used more signal modalities including peripheral arterial tone, actigraphy, and pulse oximetry. The recent study by Willemen et al. [101] also achieved a good performance, although it was validated with a younger sample population, excluded 12% of the epochs from validation and used a basis of 60-second epochs instead of the standard scoring basis of 30 seconds which makes the results incomparable.

For W/NR/R classification we see that, to the best of our knowledge, our results also outperform those reported in almost all of the previous studies. In comparison with one of the best performing

Table 4.4: Performance comparison with state-of-the-art

Reference	Modalities[a]	N	Age (year)	Average κ	Average accuracy
W/N1-2/N3/R					
Our work	**RIP, ECG**	**48**	**41.3 ± 16.1**	**0.49**	**0.69**
Our previous work [162]	RIP	48	41.3 ± 16.1	0.41	0.65
Isa et al. (2011) [100]	ECG	16	[32 − 56]	0.26	0.60
Hedner et al. (2011) [194]	PAT, PO, ACT	227 OSA	49 ± 14	0.48	0.66
Willemen et al. (2014) [101][b]	RIP, ECG, ACT	85 (36 subj.)	22.1 ± 3.2	0.56	0.69
W/NR/R					
Our work	**RIP, ECG**	**48**	**41.3 ± 16.1**	**0.56**	**0.80**
Our previous work [162]	RIP	48	41.3 ± 16.1	0.48	0.77
Redmond and Heneghan (2006) [98]	RIP, ECG	37 OSA	46.7 ± 10.4	0.32	0.67
Redmond et al. (2007) [80]	RIP, ECG	31	42.0 ± 7.4	0.45	0.76
Mendez, Matteucci et al (2010)[172]	BCG	22 (11 subj.)	n.a.	0.42	0.72
Kortelainen et al. (2010) [93]	BCG	18 (9 subj.)	[20 − 54]	0.44	0.79
Migliorini et al. (2010) [195]	BCG	22 (11 subj.)	n.a.	0.55	0.77
Kurihara and Watanabe (2012) [105]	BCG	20	22.2	0.48	0.78
Xiao et al. (2013) [196]	ECG	45	[16 − 61]	0.46	0.73
Domingues et al. (2014) [99]	RIP, ECG, ACT	20	42.1 ± 9	0.58	0.78
Willemen et al. (2014) [101][b]	RIP, ECG, ACT	85 (36 subj.)	22.1 ± 3.2	0.62	0.81

[a] ECG: electrocardiography, RIP: respiratory inductance plethysmography, PAT: peripheral arterial tone, PO: pulse oximetry, ACT: actigraphy, BCG: ballistocardiography
[b] 60-second epochs; 12% of all epochs excluded from validation.
Note: All data sets comprise healthy subjects unless indicated. There are other studies in literature which present subject-dependent classification results. In this comparison only results obtained with subject-independent schemes were considered.

studies [99], we obtain a higher accuracy (albeit a slightly smaller κ) but require one less modality (actigraphy). Regarding the work of Willemen et al. [101] it is again important to note that the results in that study were obtained on basis of 60-second epochs.

These results suggest that our choice for a Bayesian linear discriminant was appropriate for this task. Besides its simplicity, it offers the benefit of a probabilistic framework which allows, for instance, the direct use of time-varying prior probabilities to improve classification. In comparison with increasingly popular black-box approaches, this classifier has the additional advantage that it does not require the tuning of critical parameters such as kernels for support vector machines, number of nodes in classification trees or number of hidden layers in neural networks.

4.4 Conclusions

This chapter presents a method to identify overnight sleep stages using cardiorespiratory features extracted from ECG and RIP signals. These features were post-processed by means of subject-specific Z-score normalization and spline smoothing, which helps reduce the influence of signal noise, between-subject, or within-subject variability in autonomic physiology. Eighty features were selected from a set of 142 features using a modified SFS-based feature selector designed to avoid biasing the validation performance. Using a linear discriminant classifier in a ten-fold

cross-validation procedure, the classification results (for both the four-class W/N1-2/N3/R and three-class W/NR/R classification tasks) achieved in this work (Table 4.4) outperform most of the previous studies.

As future work it would be interesting to investigate whether this methodology would achieve a comparable performance in subjects with sleep disorders such as sleep apnea or insomnia. If successful, and given its potential for unobtrusive, prolonged use at home, it could represent a significant step towards lowering the costs and complexity of home tests and thus complement the current medical practices for diagnosis of sleep disorders.

Chapter 5

Validation of PPG-based sleep staging in healthy middle aged adults

Adapted from P. Fonseca, T. Weysen, M.S. Goelema, E.I.S. Møst, M. Radha, C. Lunsingh Scheurleer, L. van den Heuvel, R.M. Aarts, Validation of photoplethysmography-based sleep staging compared with polysomnography in healthy middle aged adults, *Sleep*, vol. 40, no. 7, 2017.

Study Objectives: To compare the accuracy of automatic sleep staging based on HRV measured from photoplethysmography (PPG) combined with body movements measured with an accelerometer, with PSG and actigraphy.

Methods: Using wrist-worn PPG to analyze HRV and an accelerometer to measure body movements, sleep stages and sleep statistics were automatically computed from overnight recordings. Sleep/Wake, four-class (Wake/N1 and N2/N3/REM) and three-class (Wake/non-REM/REM) classifiers were trained on 135 simultaneously recorded PSG and PPG recordings of 101 healthy participants, and validated on 80 recordings of 51 healthy middle-aged adults. Epoch-by-epoch agreement and sleep statistics were compared with actigraphy for a subset of the validation set.

Results: The Sleep/Wake classifier obtained an epoch-by-epoch Cohen's κ between PPG- and PSG-sleep stages of 0.55 ± 0.14, sensitivity to Wake of 57.4 ± 17.7%, and accuracy of 91.5 ± 5.1%. κ and sensitivity were significantly higher than with actigraphy (0.40 ± 0.15 and 45.5 ± 19.3%, respectively). The three-class classifier achieved a κ of 0.46 ± 0.15 and accuracy of 72.9 ± 8.3%, and the four-class classifier, a κ of 0.42 ± 0.12 and accuracy of 59.3 ± 8.5%.

Conclusions: The moderate epoch-by-epoch agreement, and in particular, the good agreement in terms of sleep statistics suggests that this technique is promising for long-term sleep monitoring, although more evidence is needed to understand whether it can complement PSG in clinical practice. It also offers an improvement in Sleep/Wake classification over actigraphy for healthy

individuals, although this must be confirmed on a larger, clinical population.

5.1 Introduction

Sleep polysomnography (PSG) is considered the gold standard for objectively evaluating sleep and is the preferred methodology to diagnose sleep disorders in clinical practice and often used in research trials. However, traditional in-lab PSG set-ups which comprise electroencephalography (EEG) have several limitations, such as their high cost, they are labor-intensive and they produce an impact to the patient negatively affecting sleep [197]. Yet, the major disadvantage of PSG is that the methodology is not well suited for long-term monitoring beyond one or two nights.

Actigraphy is a methodology more suited for long-term monitoring of sleep and is highly appropriate for examining sleep variability [198]. The American Academy of Sleep Medicine (AASM) indicated actigraphy as a suitable method to assist in the evaluation of patients with circadian disorders and sleep-wake disturbances, and also to assess response to therapy of circadian disorders and insomnia [57]. The relative simplicity of actigraphy, together with the maturity and low-cost of accelerometer sensors, have fueled a growing trend in the area of health self-tracking, with many consumer devices in this area offering sleep monitoring as a feature of their products [81, 82, 199]. The first group of consumer sleep tracker (CST) devices, such as the Jawbone Up (Jawbone, San Francisco CA, United States of America (USA)), or the Fitbit Ultra (Fitbit Inc., San Francisco CA, USA), uses actigraphy to estimate periods of Sleep and Wake. However, because it relies exclusively on the measurement of gross body movements, actigraphy lacks the ability to describe sleep architecture. Furthermore, studies which validated actigraphy against PSG have shown that it tends to overestimate sleep as it is not well equipped to detect wakefulness when lying quietly [58, 59, 200]. The few CST devices that have been validated against PSG, such as the Jawbone Up [83] and the Fitbit Ultra (Fitbit Inc., San Francisco CA, USA) [84, 85] were found to overestimate total sleep time (TST) and sleep efficiency (SE) as compared to PSG [83, 84], confirming the general drawback of actigraphy-based sleep measurements.

The second group of more modern CST devices include, besides accelerometers, heart rate monitors based on reflective photoplethysmography (PPG). Devices such as the Jawbone Up3 (Jawbone, San Francisco CA, USA), Fitbit Alta HR (Fitbit Inc., San Francisco CA, USA) or the now discontinued Basis Peak (Intel, San Francisco CA, USA) claim, besides Sleep and Wake tracking, the ability to depict sleep architecture. However, nearly no validation information for these devices is available, and questions with respect to the accuracy of modern CST devices remain unanswered [81, 82, 199].

Despite the lack of evidence for CST devices, the physiological link between cardiac activity, for example indirectly measured with PPG, and sleep is relatively well-understood. PPG is a technique whereby reflected or transmitted light shone on the skin is measured by a photosensor. Light is absorbed by the skin, by venous blood and also by arterial blood and besides a slow-varying DC component, the measured signal has a pulsatile component related to the arterial

blood volume changes due to cardiac activity. Transmissive PPG, typically mounted on the fingertip, has been extensively used in various clinical applications and has been commercially available in pulse oximeters, vascular diagnostic tools and beat-to-beat blood pressure devices for the past four decades [90]. When mounted on the wrist, reflective PPG (simply referred to as PPG in the remainder of the chapter) measures blood volume changes in the microvascular bed of tissue of that part of the body. These sensors have been shown to accurately measure average heart rate, typically calculated over windows of a few seconds [91]. They also offer, in theory, the possibility of measuring heart rate variability (HRV) throughout the night based on the analysis of the distance between consecutive heart beats detected in the PPG signal. The relation between heart rate variability, expressing characteristics of the autonomic nervous system (ANS) and its sympathetic nervous system (SNS) and parasympathetic nervous system (PNS) divisions, and sleep stages has been extensively described in literature. For example, as non-REM (NREM) sleep progresses from N1 to N3, there is an accompanying increase in cardiovagal drive and PNS activity [87, 201] and a reduction in cardiac and SNS activity, translating to a decrease in heart rate and an increase in the respiratory mediation of HRV, visible in the high frequency band (HF) band (0.15 to 0.4 Hz), as compared to Wake [87]. In contrast, rapid eye movement (REM) sleep is a state of autonomic instability where PNS and SNS activity fluctuate, producing abrupt changes in heart rate. The average heart rate and the power in the low frequency band (LF) (0.04 to 0.15 Hz) of HRV is higher during REM than during NREM sleep, and there is a shift of the LF/HF ratio towards sympathetic dominance [87, 202]. It has also been shown that these relations extend beyond just a different autonomic control for different sleep stages but that they are in fact continuously associated with different EEG characteristics during sleep. The power in the LF band and the LF/HF ratio, for instance, has been found to be correlated with delta EEG power [203], and a high coherency was found between the power in the HF band and the power in the delta, theta, alpha, sigma and beta EEG bands [204], with non-linear interactions between delta, theta, alpha and HF power [205]. Machine learning techniques have successfully exploited the relations between HRV and sleep stages. Several algorithms were shown to be able to automatically score sleep stages based on HRV, typically measured with electrocardiography (ECG), often in combination with respiratory effort [80, 99, 101, 206] achieving moderate performance. Heart rate monitors embedded in CST devices could, in theory, enable similar PPG-based HRV characteristics to be measured, and, in combination with body movements measured with the accelerometers typically available in those devices, augment Sleep/Wake detection - already feasible with actigraphy - with more detailed sleep architecture information.

The current study has two objectives:

- To compare an automatic four-class sleep staging method for Wake/N1 and N2/N3/REM (W/N1-2/N3/R) classification based on PPG-based HRV and body movements, with PSG.
- To evaluate whether the addition of PPG-based HRV to body movement information improves the estimation of Sleep/Wake statistics as compared with traditional wrist-worn actigraphy

Table 5.1: Participant demographics for the set used to train the algorithm (Training set) and for the two hold-out sets used to validate the algorithm (Validation set 1 and Validation set 2)

	Training set		Validation set 1		Validation set 2	
	Mean (SD)	Range	Mean (SD)	Range	Mean (SD)	Range
N	101 participants, 135 recordings		16 participants, 26 recordings		35 participants, 54 recordings	
Sex	57 female participants (56.4%), 76 female recordings (56.3%)		8 female participants (50.0%), 13 female recordings (50.0%)		20 female participants (57.1%), 34 female recordings (63.0%)	
Age	44.2 (17.2)	[20, 83]	51.2 (8.4)	[41, 66]	52.0 (6.9)	[41, 64]
BMI (kg/m^2)	23.9 (3.1)	[17.2, 31.3]	25.9 (2.8)	[20.90, 29.86]	26.2 (4.0)	[19.27, 36.23]

in the same task.

5.2 Materials and methods

The current study was based on three separately collected data sets, used to train the sleep staging algorithm ('Training set') and to validate it ('Validation set 1', 'Validation set 2').

5.2.1 Training data set

The training data set consisted of a subset of the data collected during the SIESTA project in the period from 1997 to 2000 [167] and which included 165 healthy participants with a mean ± standard deviation (SD) age of 46.7 ± 7.7 year. The SIESTA study was carried out in five different countries in seven different sleep laboratories. The study was approved by the local ethical committee of each research group. None of the participants were using, or had a history of drug and/or alcohol use, were working at night, or had been diagnosed with a medical (or mental) disorder interfering with the aim of the study. All participants had a Mini-Mental State Examination score [207] > 25, a Pittsburgh Sleep Quality Index (PSQI) score [148] of < 6 and a bed time between 22:00 and 00:00. The total duration of the data collection per participant in the SIESTA data set was 15 days. At day 7 and day 8 participants were invited to sleep in the sleep laboratory to collect overnight PSG. Two hours before their habitual bed time the participants were asked to perform a battery of psychometric tests and afterwards the PSG was applied. The next morning, the PSG set-up was removed and after washing, getting dressed and breakfast, the same psychometric test battery was completed by the participants. All participants gave informed consent before participation. The PSG recordings of all participants were scored by two trained somnologists from different sleep centers, and revised by a third expert who took the final decision in case of disagreement. More details regarding participants and study design were described by Klosh et al. [167].

For the purpose of our study, a total of 135 recordings of 101 healthy sleepers (57 females) were included, with an age range between 20 and 83 years. Additional demographics can be found in Table 5.1.

5.2.2 Validation data sets

The sleep staging method was validated with two hold-out data sets. None of the participants in the validation sets were part of the training set.

Participants

Validation set 1 was collected in 2014 and consists of 16 healthy participants (8 females) with a mean age of 51.2 ± 8.4 years. The study lasted five nights, three measured at home which only included actigraphy and two in a hotel which included overnight PSG and PPG.

Validation set 2 was collected in 2015 and consisted of 35 healthy participants (20 females) with a mean age of 52.0 ± 6.9 years. The study lasted for 16 days, including 2 weeks of home monitoring including actigraphy and PPG and two nights at the end of the study in a hotel which included overnight PSG, PPG, and actigraphy.

Both studies included participants with no primary history of neurological, cardiovascular, psychiatric, pulmonary, endocrinological or sleep disorders. In addition, none of the participants were using sleep, antidepressant or cardiovascular medication, recreational drugs, excessive amounts of alcohol, nor were they pregnant or working in shifts, nor crossing more than two time zones in the last 2 months. All participants of both studies had a body mass index lower than 40, and a PSQI < 6.

The two studies were approved by the Internal Committee of Biomedical Experiments of Philips Research and were conducted in accordance with the Declaration of Helsinki. All participants gave informed consent before participation. The hotel gave approval to conduct the experiment on their premises.

Only the hotel night recordings (the last two nights in each study) were used because they comprised a complete PSG, needed to validate the sleep staging algorithm.

Procedure

Participants arrived at the hotel at 8:00 p.m. and were informed about the procedure. The recording devices were applied and participants were free to choose how to spend the time until their habitual bed time. Bed and lights off times were registered by the researcher, as well as wake up time and lights on time. Participants were free to choose how to spend the time between the two recording nights. The researcher only monitored the beginning and the end of the recording, but not the entire night. The researcher was available at the hotel for the duration of both recording nights, and the participants could call the researcher during the night using the hotel room phone.

Recording devices

Bed and wake times were logged by participants using the consensus sleep diary [208]. PSG was recorded with the Philips Respironics Alice PDx system (Philips Respironics Inc., Murrysville PA, USA)

Table 5.2: PSG sleep statistics, and bias, 95% limits of agreement, and RMSE between PPG and PSG in 80 recordings of 51 middle-aged adults

	PSG		PPG - PSG		
	Mean (SD)	Range	Mean error (SD)	95% LoA	RMSE
SOL	15.53 (8.23)	[3.50, 39.50]	-6.80 (7.69)	[-21.88, 8.28]	10.23
WASO	33.76 (29.63)	[2.00, 164.50]	-3.19 (25.28)	[-52.74, 46.36]	25.32
TWT	55.32 (36.04)	[14.00, 218.00]	-13.40 (31.74)	[-75.60, 48.80]	7.37
TIB	462.42 (43.15)	[329.00, 531.50]	n/a	n/a	n/a
TST	407.10 (50.13)	[232.00, 504.50]	13.40 (31.74)	[-48.80, 75.60]	34.27
SE	88.10 (7.67)	[51.56, 96.93]	2.90 (6.82)	[-10.46, 16.26]	7.37
N1-2	208.91 (44.58)	[107.00, 310.50]	-28.33 (42.22)	[-111.09, 54.42]	50.63
N3	107.22 (34.48)	[39.50, 188.00]	0.26 (38.32)	[-74.85, 75.37]	38.08
REM	90.97 (28.87)	[28.00, 150.50]	41.47 (33.62)	[-24.43, 107.37]	53.25

All statistics are shown in minutes, except SE which is shown as a percentage.

using the standard 10 – 20 system electrode placement with a referential system montage, assisted by BraiNet (Jordan Neuroscience Inc., Redlands CA, USA). The sleep scoring montage included the minimum set of three EEG channels recommended by AASM for offline scoring [11] (F4-A1, C4-A1, O2-A1) plus a backup electrode (C3-A2), left electrooculogram (LOC-A2), right electrooculogram (ROC-A1), a bi-polar submental electromyogram (EMG1–EMG2), an ECG (modified lead II) and two respiratory inductance plethysmography (RIP) belts mounted around the thorax and the abdomen (RESP1 and RESP2).

In addition to PSG, a CE-marked logging device containing a PPG and three-axial accelerometer sensors (Royal Philips, Amsterdam, the Netherlands) was used during the same recording periods. The logging device was mounted on the non-dominant wrist of the participant, with the sensor facing the skin on the dorsal side of the hand, above the ulnar styloid process.

Finally, the recordings of validation set 2 included actigraphy measured with Actiwatch Spectrum (Philips Respironics Inc., Murrysville PA, USA) which uses a piezoelectric accelerometer to detect and log limb movements. It was worn on the dominant wrist of the participant, configured to measure activity counts in epochs of 30 seconds.

Analysis

The PSG data were analyzed by an external experienced somnologist (SleepVision, Nijmegen, the Netherlands) blind to the health condition of the participants and to the purpose of the study and the sleep stages were manually scored in 30-second epochs according to the AASM guidelines [11]. No scoring of respiratory events was performed and the PSG did not highlight any sign of sleep disorders with any of the 51 participants. Table 5.2 indicates the PSG sleep statistics from the 80 recordings of the 51 participants in the two validation sets.

5.2.3 PPG-based sleep stage classification

In Chapter 4 [206] we presented a machine learning approach to sleep staging based on HRV measured from ECG and respiratory effort measured from RIP. The system validated in the present chapter uses a similar approach and similar set of HRV features, albeit computed from interbeat intervals detected from PPG instead of ECG. In short, the HRV feature set consisted of a combination of time-domain features such as sample statistics of heart-beat interval durations and consecutive differences [118], arousal likelihood ratios based on consecutive interbeat intervals (IBIs) [169], multiscale sample entropy [154] and detrended, progressive and windowed fluctuation analysis of heart-beat intervals [175], and measures of cardiorespiratory interaction based on visibility graphs [181]. The feature set also comprises frequency-domain features such as the spectral powers in the very low, low and high frequency bands, with [171] and without [118] adapted spectral boundaries. HRV features were combined with features related to body movements, calculated based on the three-axial accelerometer signal. Because no respiratory information is readily available from the wrist-worn sensors used, respiratory features were left out.

The sleep staging algorithm was trained with the same machine learning techniques described in Chapter 4 [206], using the data of the training set (Table 5.1). Because no PPG data were available in the training set, the HRV features used to train the algorithm were estimated from ECG. Because the features are based on the distance between consecutive R-R intervals, they are essentially equivalent (in the absence of cardiovascular conditions) to the IBIs computed from consecutive pulses measured from PPG. Using PPG-derived HRV, and body movement features, the algorithm was validated with a hold-out validation procedure on both validation sets (Table 5.1). Additionally, and because no accelerometer data were available in the training set, a surrogate measure of activity counts was computed from the respiratory effort RIP signal available in the PSG for that set using the algorithm described in Chapter 3.

The sleep staging algorithm provides an estimation of the sleep stage in W/N1-2/N3/R of each 30 seconds epoch. Based on this automatic scoring, overall Sleep/Wake statistics such as sleep onset latency (SOL), wake after sleep onset (WASO), total wake time (TWT), TST, and SE as well as the percentage of each sleep stage were computed for each overnight recording.

5.2.4 Comparison with PSG

In order to validate the output of the sleep staging algorithm, the clocks of the PPG sensor and of the PSG needed to be synchronized. This was achieved by compensating the clock offset and drift based on an interpolated (at 4 Hz) time series comprised of the interbeat intervals detected from the PPG and another series from the interbeat intervals calculated from R-peaks detected from the ECG signal from the PSG. This guaranteed that the ground-truth reference and the output of the sleep staging algorithm were perfectly aligned to the start time and that there was no clock drift between the two recorders.

After synchronizing both signals, the hypnogram derived from PSG annotations (henceforth referred to as 'reference hypnogram') and the hypnogram estimated by the PPG algorithm (henceforth referred to as 'estimated hypnogram') were restricted to the period between lights off in the evening and lights on in the morning. The estimated hypnogram was compared with the reference hypnogram both in terms of epoch-by-epoch agreement, as well as in terms of sleep statistics calculated for each overnight recording.

Regarding the epoch-by-epoch agreement, two quality metrics were used: accuracy and Cohen's κ coefficient of agreement [156] (or κ, in short). Accuracy indicates the percentage of 30-second epochs with correctly classified sleep stages. κ indicates the agreement between the two hypnograms corrected for agreement by chance. Though harder to interpret, the latter is more informative because of the imbalance in the occurrence of different sleep stages. κ is usually interpreted with the following terms: below 0.20 'slight agreement', 0.21 to 0.4 'fair agreement', 0.41 to 0.60 'moderate agreement', 0.61 to 0.8 'substantial agreement', and above 0.81 'almost perfect agreement'.

Regarding the sleep statistics, each measure was evaluated by calculating the mean and SD of the error and the root mean square error (RMSE) between the statistics obtained with the estimated hypnogram and the statistics calculated from the reference hypnogram. In addition, a Bland-Altman analysis of each statistic was performed, by calculating and plotting the mean and SD of the differences and the corresponding 95% limits of agreement (LoA) as the mean difference ± 1.96 times the SD. A positive mean difference value indicates that the algorithm tends to overestimate a particular statistic in comparison with reference PSG, while a negative value means that statistic is underestimated.

Similar to related work [83, 209, 210], we defined a priori a satisfactory agreement if differences between the estimated and the PSG reference for WASO, TWT, and TST, were smaller than 30 minutes, and if the differences for SE were below 5%. Because the differences between PPG and PSG do not follow a normal distribution, the 99.9% confidence intervals (CIs) for these errors were calculated with a non-parametric method based on Wilcoxon's signed rank test [158]. Statistically significant overall satisfactory agreement was considered (at $p < .001$) if both boundaries of the 99.9% CIs were (in absolute terms) smaller than 30 minutes (for WASO, TWT, and TST) and 5% (for SE) [157].

It should be noted that the sleep statistic time in bed (TIB) was not computed because the analysis was restricted to the lights off period according to protocol, corresponding not exactly to the time spent in bed, but rather to the period of time when the participants had the intention to sleep.

5.2.5 Comparison with actigraphy

Actigraphy-based Sleep/Wake statistics (SOL, WASO, TWT, TST, and SE) and epoch-by-epoch Sleep/Wake classification were estimated using the Actiware software (Philips Respironics Inc.,

Murrysville PA, USA) with the default sensitivity settings (medium, 40).

The Sleep/Wake statistics were evaluated against PSG on the 53 recordings of validation set 2 for which actigraphy was available. The results were compared with the Sleep/Wake statistics obtained with PPG for the same recordings. A Wilcoxon signed-rank test was used to compare the estimation errors obtained with actigraphy and with PPG.

In order to evaluate the performance of epoch-by-epoch classification, the clocks of the actigraphy device and the PSG were synchronized. This was achieved by computing a surrogate measure of actigraphy from the artifacts in the respiratory effort signal recorded with PSG as described in Chapter 3 [211] and then finding the clock offset that maximized the correlation between the actigraphy and the surrogate actigraphy signals. All comparisons were manually reviewed to guarantee that the clocks were precisely synchronized. Because this method depended on the availability of a valid respiratory effort signal, it could only be applied on 49 recordings of validation set 2 and the comparison between PPG and actigraphy was restricted to that subset. The same metrics of κ, accuracy, specificity and sensitivity (to Wake) were used to evaluate the epoch-by-epoch classification agreement with ground-truth annotations from PSG. A Wilcoxon signed-rank test was used to compare the performance obtained with actigraphy and with PPG.

5.3 Results

5.3.1 PPG-based Sleep/Wake statistics

SOL (minutes), WASO (minutes), TWT (minutes), TST (minutes), and SE (%) were computed from both the human annotation of the recordings as well as from the automated scoring by the PPG-based algorithm. Table 5.2 summarizes the mean differences, the 95% LoA, and the RMSE for these statistics, Figures 5.1, 5.2 and 5.3a illustrate the Bland-Altman analyses of the differences for SOL, WASO, TWT, TST and SE. The algorithm underestimates, in average, SOL and WASO by less than 10 minutes and TWT by less than 15 minutes, and overestimates TST by less than 15 minutes and SE by less than 5% in comparison with PSG. The standard deviation of the error for SOL is below 10 minutes, and for SE, below 10%, but for the remaining Sleep/Wake statistics, WASO, TWT and TST, it is close to 30 minutes. Closer inspection of the Bland-Altman plots of SOL and WASO reveals that despite the relatively low bias, and for SOL, the SD of the error, it tends to underestimate wake time when SOL and WASO increase beyond 15 minutes.

The algorithm achieves a satisfactory agreement according to the a priori defined margin of error of 30 minutes for 77.5% of the recordings for TST (62 recordings), 77.5% for TWT (62 recordings), 81.2% for WASO (65 recordings), and according to the a priori margin of 5% for 71.2% of the recordings for SE (57 recordings). Both boundaries of the 99.9% CIs for TST ([-0.75, 21.00] minutes), TWT ([-21.00, 0.75] minutes), WASO ([-10.00, 8.00] minutes), were contained within the priori interval of [-30, 30] minutes and the a priori defined satisfactory agreement can be statistically established ($p < .001$) for all statistics. Both boundaries of the 99.9% CIs for SE

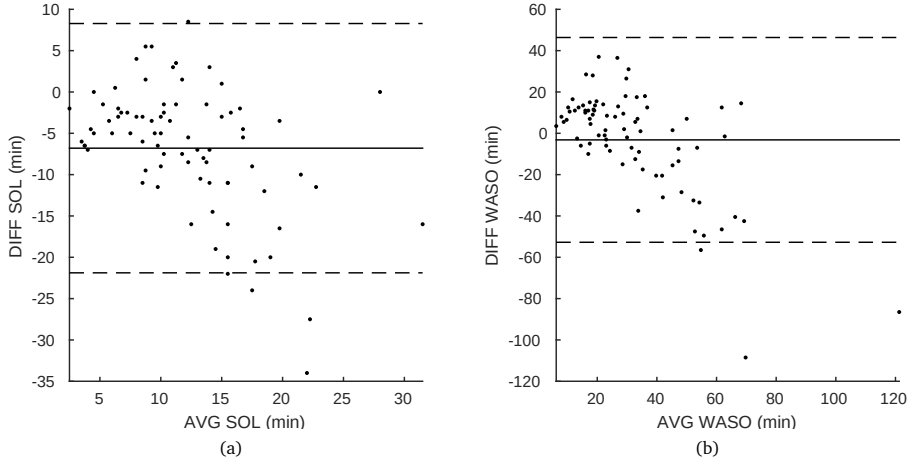


Figure 5.1: Bland-Altman plots for (a) SOL and (b) WASO. DIFF and AVG correspond to the difference and average between PPG and PSG, respectively.

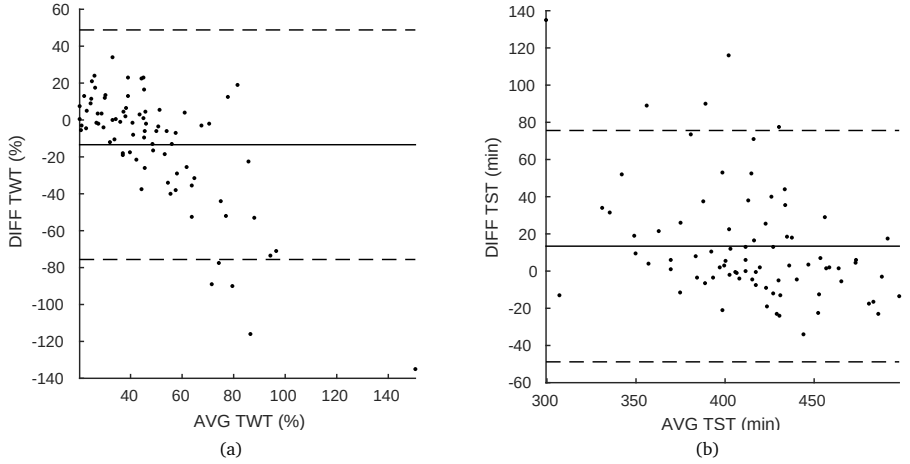


Figure 5.2: Bland-Altman plots for (a) TWT and (b) TST. DIFF and AVG correspond to the difference and average between PPG and PSG, respectively.

([-0.13, 4.61] %) were also contained within the a priori interval [5, 5] % and the a priori defined satisfactory agreement can also be statistically established ($p < .001$).

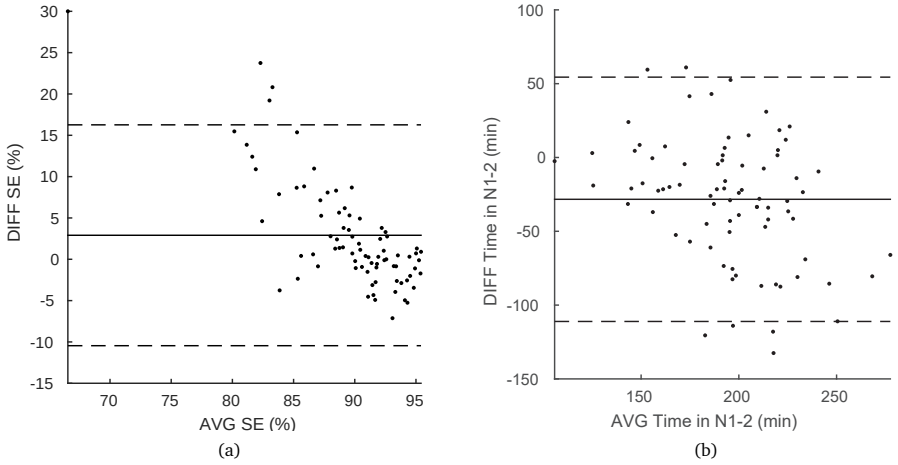


Figure 5.3: Bland-Altman plots for (a) SE and (b) time in N1-2. DIFF and AVG correspond to the difference and average between PPG and PSG, respectively.

5.3.2 PPG-based sleep staging performance

The mean differences, LoA and RMSE for the duration of sleep stages N1-2, N3 and REM are given in Table 5.2. Figures 5.3b and 5.4 illustrate the Bland-Altman analyses of the differences for these sleep stages. The best agreement was obtained for the estimation of time in N3, time in N1-2 was underestimated, in average, by nearly 30 minutes, and time in REM was overestimated by slightly more than 40 minutes.

The agreement between the automated system and the human annotation was also considered in terms of 30-second epoch-by-epoch agreement. Multiple cases were considered: first, the system was used for four-class classification (W/N1-2/N3/R, combining N1 and N2). In this case, the algorithm achieved a moderate agreement with a κ of 0.42 and accuracy of 59.3%. In the second case, the system was used for three-class classification (Wake/non-REM/REM (W/NR/R), combining N1, N2 and N3) and the κ increased to 0.46 and accuracy to 72.9%. Finally, the system was used for two-class classification, distinguishing only between Sleep and Wake (combining all sleep classes), achieving a κ of 0.55 and accuracy of 91.5%.

Table 5.3 shows the κ and accuracy for these three cases. In addition, the table also shows κ, accuracy, sensitivity and specificity in the detection of separate sleep stages. The worst single class detection performances were obtained for N1-2 followed by REM. This is in line with the Bland-Altman analysis of these two statistics, illustrated in Figures 5.3 and 5.4.

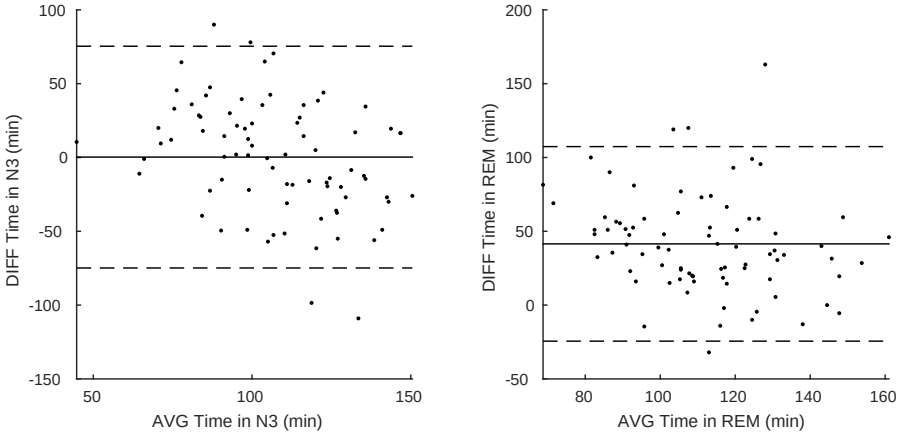


Figure 5.4: Bland-Altman plots for time in (a) N3 and (b) REM. DIFF and AVG correspond to the difference and average between PPG and PSG, respectively.

Table 5.3: Epoch-by-epoch agreement between PSG and PPG: mean (SD) of overall and per class κ and accuracy, per class and sensitivity and specificity.

	κ	Accuracy	Sensitivity	Specificity
Multiple classes[a]				
W/N1-2/N3/R	0.42 (0.12)	59.3 (8.5)	n/a	n/a
W/NR/R	0.46 (0.15)	72.9 (8.3)	n/a	n/a
Single class detection[b]				
Wake	0.55 (0.14)	91.5 (5.1)	58.2 (17.3)	96.9 (2.0)
N1-2	0.30 (0.13)	65.7 (7.1)	55.6 (9.2)	74.2 (8.2)
N3	0.50 (0.19)	82.5 (6.3)	63.9 (18.9)	89.0 (4.8)
NREM	0.46 (0.16)	75.3 (7.7)	77.5 (7.8)	71.1 (12.0)
REM	0.43 (0.19)	78.9 (7.2)	70.7 (18.2)	81.4 (6.9)

[a] Agreement for multiple classes computed based on the comparison between PPG epoch-based classification and PSG annotations; for the four-class task PSG N1 and N2 classes were merged into a single N1-2 class; for the three-class task, PPG N1-2 and N3 were merged into a single NREM class and PSG N1, N2 and N3 were merged into a single NREM class.
[b] For each class detection task, the remaining classes were merged in a single class, both for PPG and for PSG, and considered the negative class for purposes of sensitivity and specificity calculation.
REM, rapid eye movement; NREM non-REM.

5.3.3 Actigraphy-based Sleep/Wake classification performance

Regarding the error between the Sleep/Wake statistics obtained with actigraphy and with PSG, indicated in Table 5.4, SOL and WASO were underestimated by less than 10 minutes, TWT by less than 20 minutes, TST was overestimated by slightly more than 10 minutes and SE by less than 5%. Table 5.4 also indicates the error calculated between the Sleep/Wake statistics estimated with PPG and PSG for the same subset of validation set 2. Although there is a decrease in the bias for SOL,

Table 5.4: Bias, SD and RMSE in Sleep/Wake statistics between actigraphy and PSG, and between PPG and PSG in 53 recordings (subset of validation set 2)

	Actigraphy versus PSG		PPG versus PSG		Comparison between PPG and actigraphy[c]
	Mean error (SD)[a]	RMSE	Mean error (SD)[b]	RMSE	
SOL (minutes)	-8.59 (9.05)	12.42	-7.48 (6.64)	9.96	$p < .05$
WASO (minutes)	-8.32 (29.92)	30.79	-5.25 (27.89)	28.12	$p = .18$
TWT (minutes)	-17.94 (28.09)	33.10	-17.22 (34.19)	37.99	$p = .18$
TST (minutes)	10.60 (30.29)	31.83	15.66 (34.95)	38.00	$p = .22$
SE (%)	3.66 (5.85)	6.85	3.71 (7.34)	8.16	$p = .51$

[a] Actigraphy minus PSG, [b] PPG minus PSG.
[c] Wilcoxon signed-rank comparison of estimation error between actigraphy versus PSG and PPG versus PSG.

Table 5.5: Epoch-by-epoch Sleep/Wake classification agreement between actigraphy and PSG and between PPG and PSG in 49 recordings (subset of validation set 2)

	Actigraphy versus PSG, mean performance (SD)	PPG versus PSG, mean performance (SD)	Comparison between PPG and actigraphy[a]
κ (-)	0.40 (0.15)	0.55 (0.15)	$p < .001$
Accuracy (%)	91.8 (5.1)	91.3 (5.9)	$p = .58$
Sensitivity (%)	45.5 (19.3)	57.4 (17.7)	$p < .001$
Specificity (%)	97.1 (1.6)	97.4 (1.9)	$p = .21$

[c] Wilcoxon signed-rank comparison of estimation error between actigraphy versus PSG and PPG versus PSG.

WASO and TWT when estimating the Sleep/Wake statistics with PPG as compared with actigraphy, only the decrease in the SOL estimation error is significant ($p < .05$).

Regarding the epoch-by-epoch Sleep/Wake classification performance, Table 5.5 indicates the results obtained between actigraphy and PSG, and between PPG and PSG for a subset of validation set 2. The κ obtained with PPG (0.55) is significantly ($p < .001$) higher than obtained with actigraphy (0.40). Additionally there is a significant ($p < .001$) increase in sensitivity with PPG (57.4%) when compared with actigraphy (45.5%). There were no significant differences for accuracy and specificity, both above 90% for actigraphy and for PPG.

5.4 Discussion

This chapter describes the validation procedure and the results obtained with a PPG-based sleep staging algorithm when compared with human scored PSG and with wrist-worn actigraphy. Regarding the estimation of Sleep/Wake statistics the algorithm achieved a statistically significant ($p < .001$) satisfactory agreement for TST, TWT, WASO and SE for more than 70% of the recordings. Although it achieved a relatively small bias for SOL and WASO, inspection of the Bland-Altman plots revealed that it tends to underestimate wake time when SOL and WASO increase beyond 15 minutes. This is likely related to the behavior of participants who took longer to fall asleep, or who woke up during the night and had difficulties falling asleep again. Besides lying still, as verified by actigraphy, an analysis of the characteristics of the respiratory effort signal of

the PSG recordings suggested that some of these participants might have paced their breathing during these periods of wakefulness. This lowered the heart rate as a consequence, and probably also led to an increase in the HF and a decrease in the LF component of HRV when compared with irregular, uncontrolled breathing [212, 213] yielding HRV characteristics closer to those typically observed during N2 or even N3 [214]. Although the SD of the error for SOL and SE was found to be relatively low (below 10 minutes and 10% respectively), it reached nearly 30 minutes for the parameters WASO, TWT and TST. This highlighted the between-subject variability in the accuracy of the algorithm.

Furthermore, in comparison with the Sleep/Wake statistic estimations obtained with actigraphy, we observed a significant improvement in the estimation of SOL, but no significant improvement (nor deterioration) for any other statistic. Interestingly, the error in the actigraphy estimation of statistics such as sleep efficiency is lower than some of the values reported in literature [58, 59, 200]. This likely reflects the healthy nature of the individuals in our validation set and it would be interesting to investigate whether the results hold when estimating Wake in clinical populations or in subjects with fragmented sleep, and whether we would see a significant improvement in performance when using PPG to analyze Sleep and Wake in those cases.

In terms of epoch-by-epoch agreement with human annotated PSG, it is clear that the simpler the classification task, the better the overall performance is. This suggests that although the HRV features used in this work allow, to a certain extent, a reasonable partition of NREM classes into N1-2 and N3, these are not completely separable in that feature space. When N1, N2 and N3 are merged into a single NREM class, the problem of lack of separability between these disappears, and the performance naturally increases.

It is also interesting to remark that the performance of the algorithm for four- and three-class tasks is lower than that reported in Chapter 4 (κ of 0.49 and accuracy of 69% for the four-class task, and κ of 0.56 and accuracy of 80% for the three-class task). Furthermore, while in that work the best single class results were obtained for REM classification [206], the worst single class detection performances are now obtained for N1-2 followed by REM. One of the possible reasons for the deterioration in overall performance might be related to the highly susceptible nature of reflective PPG to subject and skin tissue motion, which are known to distort the PPG signal [215] and consequently decrease the precision of detected heart beats, interbeat intervals, and HRV features. Additionally, the absence of respiratory features in this study in comparison with Chapter 4 is likely to have a negative impact on performance. Although this alone is probably not sufficient to fully explain the decrease, other studies have shown the advantage of using features (besides body movements) in addition to HRV which help further express changes in sympathetic tone, important in the classification of REM. Herscovici et al. [216] for instance, combined comparable time- and frequency-domain HRV features (referred to by the authors as inter-pulse periods, or IIP) with features describing variations in the amplitude of peripheral arterial tone (PAT) measured at the fingertip. The PAT signal provides a sensitive surrogate measure of changes in sympathetic

tone and exhibits unique characteristics during REM, providing strong discriminative power for the detection of this sleep stage [216], and in general, overall automatic sleep staging based on autonomic features [194]. Although PAT is not directly available with the current sensor, the PPG signal actually contains respiratory-induced intensity variations (RIIV) which are related to ventilatory pressure and have been used, successfully, to estimate respiratory rate [217]. Although not explicitly related to RIIV, a preliminary study on ten sleep apnea patients by Uçar et al [218] has shown that adding morphological features of PPG to a standard set of HRV features increases the performance of sleep stage classification. It remains to be confirmed whether features extracted from RIIV, together with the addition of dedicated HRV features which may better reflect changes in sympathetic tone characteristic of REM help alleviate or overcome this degradation.

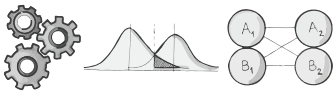
Finally, the best single-class performance is obtained for Sleep/Wake classification, with a κ of 0.55, accuracy of 91.5%, sensitivity to Wake of 58.2% and specificity to Wake of 96.9%. Comparing the performance with actigraphy-based epoch-by-epoch Sleep/Wake classification, we found that PPG led to a substantial and significant increase in κ (from 0.40 with actigraphy to 0.55 with PPG on the same recordings, $p < .001$) and in sensitivity to Wake (from 45.5% with actigraphy to 57.4% with PPG, $p < .001$). The results obtained with actigraphy in our work are, once more, slightly better than those reported in literature. Marino et al. [219] reported an accuracy of 86.3%, a sensitivity to sleep of 96.5% and a specificity to sleep of 32.9%. We calculated a κ of 0.37 based on their published data. Again, the differences can probably be explained by the exclusively healthy characteristics of the participants in our validation sets.

Notwithstanding, the increased performance in Sleep/Wake classification using PPG-HRV and body movements when compared with actigraphy is likely related to the presence of additional information captured by HRV characteristics which is better able to capture Wake states, especially when the subjects do not move. The reason why these improvements were less prominent in the computation of Sleep/Wake statistics is likely related to the high imbalance between Sleep and Wake classes (the participants in the validation sets had a mean sleep efficiency of 88.1%) and especially to the relative lack of variability in the Sleep/Wake statistics to start with (the SD of SOL – computed from PSG – was lower than 10 minutes). It is important to remark that the comparison with actigraphy is necessarily limited by the scope of our study. While actigraphy has been evaluated for many different clinical populations over the last decades, and its performance and limitations are well understood for different clinical settings, our results only allow tentative conclusions regarding a healthy group of middle-aged participants and cannot be automatically extended to other groups. Furthermore, this technology has the disadvantage that it assumes an intact autonomic function, and its performance will be probably impacted by disorders of the ANS which affect the regulation of blood pressure, heart rate and HRV. Besides, the accuracy of HRV features is intrinsically linked to the quality and availability of the PPG signal. Disorders such as peripheral artery disease may cause a restriction in the blood flow at peripheral arteries, limiting the quality of PPG or even rendering it unusable. By relying simply on the measurement of gross

body movements as an indicator of wakefulness, actigraphy might be more robust to the severity of such conditions.

Nevertheless, its relative low-cost, ease of use, and comfort for the subject, and its moderate epoch-by-epoch agreement, and good agreement in terms of sleep statistics suggests the adequacy of this technique for long-term sleep monitoring, although more evidence is needed to understand whether it can complement PSG in clinical practice. These results also suggest that besides being more sensitive to Wake, the agreement obtained for four sleep stages with this technology (with a κ of 0.42) is in fact slightly higher than the overall agreement for Sleep/Wake classification with traditional actigraphy (κ of 0.40) in our data set. Provided that the level of evidence for PPG-based sleep staging increases in the upcoming years and that these findings are confirmed in relevant clinical populations, this technology may have the potential to eventually replace actigraphy in clinical practice with the benefit of offering additional insights into the sleep architecture of the participants under investigation besides only their sleep-wake patterns. Regardless of its potential, however, it will only be usable when it has been properly validated on relevant clinical populations. To our knowledge, it has not yet been validated in individuals suffering from sleep disorders and therefore merits more research.

Long-term sleep monitoring was for a long time limited by the unavailability of small, comfortable sensors which could be used at home without assistance. A notable exception, accelerometers, enabled the development of actigraphy devices which albeit valuable in the assessment of some sleep disorders were limited to the analysis of sleep-wake patterns. Recent sensor developments have shown the feasibility of accurately monitoring other physiological parameters, with PPG, skin temperature and skin conductance sensors offering the promise, in the long run, to be able to complement PSG with enough accuracy to enable unobtrusive sleep monitoring in clinical practice.



Part 4

Probabilistic methods for sleep stage classification

Seria feliz se pudesse dormir. Esta opinião é deste momento, porque não durmo. A noite é um peso imenso por detrás do afogar-me com o cobertor mudo do que sonho. Tenho uma indigestão na alma.

I'd be happy if I could sleep. This is what I think now, because I'm not sleeping. The night is an enormous weight beyond the silent blanket of dreams under which I smother myself. I have indigestion of the soul.

- Bernardo Soares (heteronym), Livro do Desassossego (*transl.* Richard Zenith, The Book of Disquiet)

Chapter 6

Cardiorespiratory sleep stage detection using conditional random fields

Adapted from P. Fonseca, N. den Teuling, X. Long, and R. M. Aarts, Cardiorespiratory sleep stage detection using conditional random fields, *IEEE Journal of Biomedical and Health Informatics*, vol. 21, no. 4, p. 956-66, 2017.

Abstract: This chapter explores the probabilistic properties of sleep stage sequences and transitions to improve the performance of sleep stage detection using cardiorespiratory features. A new classifier, based on Conditional Random Fields, is used in different sleep stage detection tasks (N3, REM, NREM, and Wake) in night-time recordings of ECG and RIP of healthy subjects. Using a data set of 342 PSG recordings of healthy subjects, among which 135 with regular sleep architecture, it outperforms hidden Markov models (HMM) and Bayesian linear discriminant (LD) classifiers in all tasks, achieving an average accuracy of 87.38% and κ of 0.41 (87.27% and 0.49 for regular subjects) for N3 detection, 78.71% and 0.55 (80.34% and 0.56 for regular subjects) for NREM detection, 88.49% and 0.51 (87.35% and 0.57 for regular subjects) for REM, and 85.69% and 0.51 (90.42% and 0.52 for regular subjects) for Wake. In comparison with the state-of-the-art, and having been tested on a much larger data set, the classifier was found to outperform most of the work reported in the literature for some of the tasks, in particular for subjects with regular sleep architecture. It achieves a comparable accuracy for N3, higher accuracy and κ for REM, and higher accuracy and comparable κ for NREM than the best performing classifiers described in literature.

6.1 Introduction

Manual sleep staging is a time-consuming task that requires the help of a sleep technician. Automatic sleep stage classification has been an active area of research for the past decades. It re-

moves the human element from sleep staging, allowing for new applications such as real-time sleep staging (useful for intervention studies), and remote monitoring (in-home sleep studies); and eliminates the problem of inter-scorer variability resulting from manual scoring [42]. Sleep stages are ordinarily measured from electroencephalography (EEG), but these sensors can disrupt sleep and often require care to apply correctly (i.e. requiring the help of an expert). Cardiorespiratory information provides a promising alternative to EEG for the purpose of sleep staging, with the benefit that it can be measured by unobtrusive methods. Cardiorespiratory-based sleep stage classification has been increasingly studied in recent years. Many studies have reported results on the classification of Wake, rapid eye movement (REM), 'light' (combined N1 and N2, or N1-2 for short) and 'deep' (N3) sleep stages [101, 103, 220], detecting REM sleep [80, 101, 172], or differentiating N1-2 and N3 stages with an ambulatory device [101, 221]. The results in the literature are promising, but not yet at the level required for reliable sleep staging. Classification is generally done using an extensive set of features. Many different classifiers have been tested over the years, such as Bayesian linear discriminant (LD) [80, 96, 195], hidden Markov models (HMM) [93], and support vector machines (SVM) [101] classifiers. For many classification tasks, the LD classifier was found to be amongst the best performing. The strength of LD lies in its underlying simplicity, providing a robust model of the features over the different sleep stages. However, LD classification is independent of time, whereas sleep is a structured process where the state and characteristics of each epoch are not independent from each other. Temporal classifiers can make use of this structure, improving classification. It has been shown that the process of sleep can be modeled using a Markov process [222]. The HMM classifier incorporates this Markov assumption, but the model does not handle correlations between features well. Furthermore, the model assumes that features are discriminative during the entire sleep stage. Some features could be indicative of a stage transition instead, but this information is not used by either HMM or LD.

This chapter investigates the use of a temporal classifier based on Markov networks, named conditional random fields (CRF) [223] as a new approach to automatic cardiorespiratory sleep stage classification. To the best of the authors' knowledge this method has not been used in the domain of sleep staging before, other than for a single paper on EEG sleep staging [224] which, given the difference in the sensing modalities, does not allow for a fair comparison. CRF is a generalization of HMM that conditions the model on the given observations. This allows for a more expressive model that can capture feature dependencies.

6.2 Materials and methods

6.2.1 Data sets

The data set used in this study comprises full polysomnography (PSG) data of 180 subjects from three different data sets. The first data set, with 327 recordings of 165 subjects (most subjects were monitored for two consecutive nights) was part of the data set created during the European

Union (EU) SIESTA project [167] between 1997 and 2000 in seven different sleep laboratories. The data set was restricted to subjects considered healthy (no sleep disorders, no shift work, no depression, usual bedtime before midnight), with a Pittsburgh Sleep Quality Index [148] of at most 5. Sleep stages were scored by consensus agreement between three sleep technicians blind to the condition of the participants in five classes ('Wake', 'REM', 'S1', 'S2', 'S3', 'S4') according to the R&K guidelines [168]. The second data set, comprising single-night recordings of six healthy subjects, was collected at the Philips Experience Lab of the High Tech Campus, Eindhoven, The Netherlands, during 2010 (Vitaport 3 PSG, TEMEC). The third data set, comprising nine healthy subjects, was collected at the Sleep Health Center, Boston, United States of America (USA), during 2009 (Alice 5 PSG, Philips Respironics). Electrocardiography (ECG) was recorded with a modified lead II derivation, sampled at 500 Hz (Boston data), 256 Hz (Eindhoven data) and 200 Hz (SIESTA data). Respiratory effort was recorded with thoracic respiratory inductance plethysmography (RIP) sampled at 10 Hz (Boston data), 256 Hz (Eindhoven data) and 200 Hz (SIESTA data). Subjects with diagnosed sleep disorders (sleep apnea, insomnia, restless legs syndrome, etc) were excluded from the three data sets. Sleep stages for subjects in the second and third data set were scored by individual sleep technicians, blind to the condition of the participants, in four classes ('Wake', 'REM', 'N1', 'N2', 'N3') according to the American Academy of Sleep Medicine (AASM) guidelines [11]. The protocols for the three data collection studies were reviewed and accepted by the local ethics committees where the experiments took place, and all participants signed informed consent forms. Although all subjects in the three data sets were considered healthy, it is reasonable to expect an impact of monitoring on the sleep quality of the subjects. This effect, also called 'first night effect' [43] actually often lasts more than the first night, and is variable from subject to subject. However, because it is likely to have an impact on the performance of sleep stage classifiers, two sets were created and analyzed separately: the first comprises only recordings which have at least a certain percentage of each sleep stage, representative of minimum sleep statistics of healthy adult sleep: at least 5% of N3, 15% of REM sleep, a sleep efficiency of at least 75%, and a minimum of 7 hours in bed [225] (set 'regular'). This resulted in a total of 135 recordings (101 subjects). The second set comprises all 342 recordings from the three data sets (set 'all'). Table 6.1 summarizes the subject demographics and sleep statistics of both sets.

6.2.2 Feature extraction

The RIP and ECG signals of all subjects were pre-processed before feature extraction. The RIP signal was first resampled to a common sampling rate of 10 Hz. It was then filtered with a 10th order Butterworth low-pass filter with a cut-off frequency of 0.6 Hz. Baseline was removed by subtracting the median amplitude calculated in a sliding window of 120 s. The baseline wander of the ECG signal was filtered with a linear phase high-pass filter using a Kaiser window of 1.016 sec, with a cut-off frequency of 0.8 Hz and a side-lobe attenuation of 30 dB [127]. QRS complexes were detected and localized from the ECG signals using a Hamilton-Tompkins QRS detector [123, 124]

Table 6.1: Demographics and sleep statistics of subjects in the two sets used in the study

	'regular'		'all'	
Parameter	Mean ± SD	Range	Mean ± SD	Range
N	101 subjects, 135 recordings		180 subjects, 342 recordings	
Sex	57 female subjects (56.44 %) 76 female recordings (56.30 %)		98 female subjects (54.44 %) 185 female recordings (54.09 %)	
Age (year)	42.8 ± 16.8	20 − 83	51.0 ± 19.6	20 − 95
BMI (kg/m^2)	23.8 ± 3.1	17.2 − 31.3	24.6 ± 3.5	17.0 − 35.3
TIB (hour)	8.0 ± 0.4	7.2 − 9.6	7.9 ± 0.5	5.3 − 9.6
SE (%)	88.9 ± 5.5	75.5 − 99.0	81.3 ± 11.9	21.7 − 99.0
REM (%)	21.7 ± 3.3	16.3 − 31.3	18.3 ± 5.7	0.0 − 34.8
N3 (%)	17.2 ± 5.7	6.0 − 36.2	15.2 ± 8.3	0.0 − 42.7

'regular' is a subset of the 'all' data set where only subjects with a minimum of 5% N3, 15% REM sleep and 75% sleep efficiency were considered.
REM and N3 percentages were calculated over the total sleep time for each recording.
SD: standard deviation, BMI: body mass index, TIB: time in bed, SE: sleep efficiency

Table 6.2: Respiratory features used in the study

Number	Feature name
1-3	Respiratory frequency calculated in time and frequency domain and its spectral power [80, 98]
4-7	PSD VLF, LF, and HF power and LF-to-HF ratio [98]
8-10	SD of respiratory frequency over 150, 210, and 270 s [80]
11-13	Mean and SD of breath-by-breath correlations, SD of breath lengths [98]
14-15	Sample entropy and variance of respiratory effort over 180 s [80, 185]
16-17	Respiratory dynamic time and frequency warping self-(dis)similarity [139]
18-21	Standardized mean and median of respiratory peaks and troughs [161]
22-23	Sample entropy of respiratory peaks and troughs [161]
24-25	Median peak-to-trough difference and dynamic time warping similarity [161]
26-31	Median volume and flow rate of breaths, inhalations, and exhalations [161]
32-33	Ratio of inhalation-to-exhalation volume and flow rate [161]

PSD, power spectral density; VLF, very low frequency; LF, low frequency; HF, high frequency; SD, standard deviation.
Features 1-7 and 11-17 were computed with windows of 30 s; features 18-33 were computed with windows of 150 s; for the remaining features the window length is indicated.

followed by the post-processing localization algorithm described in Chapter 2 [150]. The resulting interbeat interval (IBI) time series were resampled using linear interpolation at a sampling rate of 4 Hz.

Because the goal is to classify each 30 s epoch in each recording, all features were automatically extracted using sliding windows centered on each epoch. Tables 6.2, 6.3 and 6.4 give a list of all respiratory, cardiac and cardiorespiratory features used in this study, together with the window lengths used to compute them and a reference to the original work(s) where these features were first used for different sleep stage classification tasks.

During feature extraction an automatic process was used to determine whether or not portions of the ECG and RIP signals were adversely affected by noise or motion artifacts. Regarding the ECG signal, each 30-second epoch was separately evaluated in regard to the coverage of IBIs. The corresponding cardiac and cardiorespiratory features were only extracted for a given epoch if the sum of the length of all detected IBIs in the window used to extract each feature was equal or larger than 50% of the length of that window. Regarding the RIP signal, peaks and troughs were first

Table 6.3: Cardiac features used in the study

Number	Feature name
1-3	Mean HR, mean RR, and detrended mean RR [98, 118]
4-8	SDNN, RR range, pNN50, RMSSD, and SDSD [118]
9-12	RR logarithmic VLF, LF, and HF power and LF-to-HF ratio [118, 170]
13-16	RR mean respiratory frequency and power, max phase and module in HF pole [172]
17-36	Multiscale sample entropy of IBIs at length 1 and 2, scales 1-10 over 510 s [a] [154]
37-42	RR DFA, its short, long exponents and all scales, WDFA over 330 s and PDFA over non-overlapping segments of 64 heartbeats [173, 175, 176]
43-46	Mean absolute difference in HR and RR and in detrended HR and RR [103]
47-56	RR and HR percentiles (10%, 25%, 50%, 75%, and 90%) [103]
57-66	Detrended RR and HR percentiles (10%, 25%, 50%, 75%, and 90%) [103]
67	Sample entropy of symbolic binary changes in IBIs [179]
68-70	Power, 4th power and curve length of the ECG [104]
71-73	Nonlinear energy, Hjorth mobility and complexity of the ECG [104]
74-77	Peak power and corresponding frequency, mean and median ECG PSD [104]
78	Spectral entropy of the ECG [104]
79	Hurst exponent of the ECG [226]
80-81	Short- and long-range phase coordination of IBIs in patterns of up to 8 consecutive heartbeats [187, 227]

HR heart rate; RR R-R or interbeat interval; SDNN standard deviation of RR; pNN50 pecentage of successive RR differences >50 ms; RMSSD, root mean square of successive RR differences; SDSD, standard deviation of successive RR differences; VLF, very low frequency; LF, low frequency; HF, high frequency; DFA, detrended fluctuation analysis; PDFA, progressive DFA; WDFA, windowed DFA; PSD, power spectral density.
Features 1-16 and 43-67 were computed with windows of 270 s; features 68-79 were computed with windows of 30 s; for the remaining features the window length is indicated.
[a] The estimation accuracy of sample entropy is lower in series shorter than 10^m (where m is the pattern length, in samples) [185, 193]. In practice this means that this feature will be accurate for all scales with $m = 1$ and for scales below 6 with $m = 2$. The choice of window size was discussed in Section 4.3.1 of Chapter 4.

Table 6.4: Cardiorespiratory features used in the study

Number	Feature name
1	Co-power between RR and respiratory effort over 9 epochs [186]
2-3	Short- and long-range phase coordination between respiration and RR in patterns of up to 8 consecutive heartbeats [187, 227]

RR R-R or interbeat interval.

detected based on the sign change of the RIP signal slope, and then marked as false detections if (1) the sum of two successive peak-to-trough intervals was less than the median of all intervals or if (2) the peak-to-trough distance was less than 15% of the median of all intervals [161]. Corresponding respiratory features were only extracted for epochs where the window used to extract a feature did not have false peak/trough detections. After automatically rejecting epochs with artifacts, and after extracting each feature from the remaining epochs on each recording, the values of rejected epochs were estimated with linear interpolation between the values of neighboring (valid) epochs. Finally, in order to reduce physiological and equipment-related variations between subjects, all features were normalized in terms of mean and standard deviation (Z-score normalization). This allowed common decision thresholds to be used for all subjects.

6.2.3 Bayesian linear discriminant

The LD classifier is based on the Bayes decision rule for minimizing the probability of error, i.e. to choose the class that maximizes its posterior probability given an observation (feature vector) $\mathbf{x}$, or, in the case of two classes a and b, to choose class a if

$$g_a(\mathbf{x}) - g_b(\mathbf{x}) \geq D, \tag{6.1}$$

where D is a decision threshold and $g_a(\mathbf{x})$ is a so-called discriminant function, $g_a(\mathbf{x}) = \ln P(a|\mathbf{x})$. Using the Bayes rule, and under the assumption that the observations of each class are drawn from multivariate normal distributions and that the covariance matrices of all classes are identical, the discriminant function is given [155] by

$$g_a(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}-\mu_a)^T\Sigma^{-1}(\mathbf{x}-\mu_a) + \ln P(a), \tag{6.2}$$

where μ_a is the mean feature vector for class a, Σ is the pooled covariance matrix for all classes, and $P(a)$ is the prior probability of class a. Considering the prior probabilities of all classes as constant, the decision boundary D in (6.1) is given by

$$(\mathbf{x}-\mu_a)^T\Sigma^{-1}(\mathbf{x}-\mu_a) - (\mathbf{x}-\mu_b)^T\Sigma^{-1}(\mathbf{x}-\mu_b). \tag{6.3}$$

It is clear that the further apart the mean vectors μ_a and μ_b are in the feature space, i.e., the larger the factor $\Sigma^{-1}(\mu_a - \mu_b)$ is, the more separable the classes are.

6.2.4 Probabilistic graphical models

As explained, the decision rule for LD is based on the estimation of the posterior probabilities $P(a|\mathbf{x})$ of each class, which depends, for classification purposes, solely on the likelihood $P(\mathbf{x}|a)$ and on the prior probability $P(a)$. Sleep is a structured process, and when the body enters a certain sleep stage, it will stay in this stage for a while before transitioning to the next stage [228]. Provided that the body is in a certain stage, the chance of the body being in that stage in the next epoch is generally larger than switching to another stage. This suggests that a classifier can benefit from taking past information into account. Although the LD classifier is adequate for many problems, in the case of sleep, which is a structured process, it might not exploit the full potential of all information available. Consider for simplicity a process, illustrated in Figure 6.1, comprised of a sequence of states which can have one of two classes a and b. Now consider that the posterior probability of a class w_i for a given state also depends on the class w_j of the previous state and on the likelihood $P(\mathbf{x}_i|w_i)$ of observing $\mathbf{x}_i$ given the class w_i,

$$P(w_i|\mathbf{x}_i, w_j) = \frac{P(w_i|w_j)P(\mathbf{x}_i|w_i)}{P(\mathbf{x}_i)}. \tag{6.4}$$

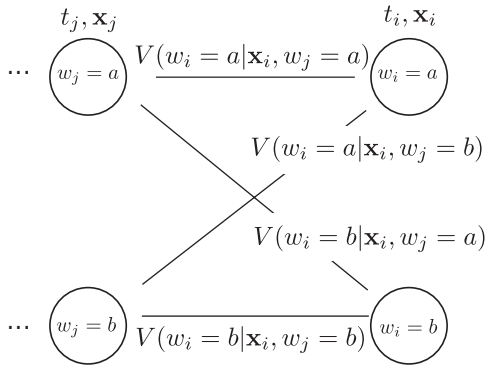


Figure 6.1: Trellis diagram for process with two classes

The class chosen for the current state, w_i, should correspond to the class that yields the largest posterior probability. Because the probability $P(\mathbf{x}_i)$ is irrelevant for the choice of the most likely class w_i, we can write

$$P(w_i|\mathbf{x}_i, w_j) \propto P(w_i|w_j)P(\mathbf{x}_i|w_i). \tag{6.5}$$

In the previous example it was assumed that w_j is known. However, when presented with a sequence of observations X, the previous class is not known because the likelihood of the model for a given sequence of states might not be optimal on the next observation. The choice of w_i therefore depends on the previous time point only, but the optimal w_j can only be chosen after the optimal path up to time point t_i is determined, which depends on the observations $\mathbf{x}_i$. Consider a state w_k that happens at time t_k, just before t_j. In this case, t_k is conditionally independent from w_i and can be fixed depending on the choice of w_j. Writing out the dependencies, we obtain the recursive relation

$$V(w_i|\mathbf{x}_i, w_j) = P(w_i|w_j)P(\mathbf{x}_i|w_i)\,V(w_j|\mathbf{x}_j, w_k). \tag{6.6}$$

In the following, for brevity, the choice of a class a for a state w_i will be represented by i_a and the conditions for choosing class a will be derived. The conditions for choosing class b can be derived in an analogous way.

Because the state w_j, corresponding to the optimal path until time t_i, is not fixed until $\mathbf{x}_i$ is observed, determining the class w_i requires an evaluation of the likelihood of four possible combinations, $V(i_a|\mathbf{x}_i, j_a)$, $V(i_b|\mathbf{x}_i, j_a)$, $V(i_a|\mathbf{x}_i, j_b)$, and $V(i_b|\mathbf{x}_i, j_b)$. The class at time t_i will be given by

$$\arg\max_{w_i} V(w_i|\mathbf{x}_i, w_j),\ \forall w_i, w_j = a, b. \tag{6.7}$$

Expanding the previous equation, class a is selected if

$$V(i_a|\mathbf{x}_i, j_a) \geq V(i_b|\mathbf{x}_i, j_a) \wedge\ V(i_a|\mathbf{x}_i, j_a) \geq V(i_b|\mathbf{x}_i, j_b), \tag{6.8}$$

which corresponds to the case where in the optimal path, $j = a$, or

$$V(i_a|\mathbf{x}_i, j_b) \geq V(i_b|\mathbf{x}_i, j_a) \wedge\ V(i_a|\mathbf{x}_i, j_b) \geq V(i_b|\mathbf{x}_i, j_b), \tag{6.9}$$

where in the optimal path, $j = b$. Using (6.6) in the previous equations, the conditions become

$$\frac{P(\mathbf{x}_i|i_a)}{P(\mathbf{x}_i|i_b)} \geq \frac{P(i_b|j_a)}{P(i_a|j_a)} \wedge \frac{P(\mathbf{x}_i|i_a)}{P(\mathbf{x}_i|i_b)} \geq \frac{P(i_b|j_b)V(j_b|\mathbf{x}_j, w_k)}{P(i_a|j_a)V(j_a|\mathbf{x}_j, w_k)} \tag{6.10}$$

or

$$\frac{P(\mathbf{x}_i|i_a)}{P(\mathbf{x}_i|i_b)} \geq \frac{P(i_b|j_a)V(j_a|\mathbf{x}_j, w_k)}{P(i_a|j_b)V(j_b|\mathbf{x}_j, w_k)} \wedge \frac{P(\mathbf{x}_i|i_a)}{P(\mathbf{x}_i|i_b)} \geq \frac{P(i_b|j_b)}{P(i_a|j_b)}. \tag{6.11}$$

Transitional features

One of the important consequences of using state-dependent posterior probabilities for classification is that the estimation will also benefit from features that are only discriminative at state transitions. Consider the (univariate) toy example of Figure 6.2a, where $P(\mathbf{x}_i|a) \approx P(\mathbf{x}_i|b)$ during the majority of intervals. However, after each class transition, the feature is extremely discriminative, with $P(\mathbf{x}_i|a) >> P(\mathbf{x}_i|b)$ or vice versa. At time point t_1 in the figure, and supposing that at that point $V(j_b|\mathbf{x}_j, w_k) > V(j_a|\mathbf{x}_j, w_k)$, the rules (6.10) and (6.11) can be applied to determine whether class a is chosen. Assuming that the probability of changing states is smaller than the probability of remaining in the same state and that the probability of staying in the same state is approximately the same for both classes, the first conditions of (6.10) and of (6.11) are met and the conditions can be simplified as

$$\frac{P(\mathbf{x}_i|i_a)}{P(\mathbf{x}_i|i_b)} \geq \frac{V(j_b|\mathbf{x}_j, w_k)}{V(j_a|\mathbf{x}_j, w_k)} \vee \frac{P(\mathbf{x}_i|i_a)}{P(\mathbf{x}_i|i_b)} \geq \frac{P(i_b|j_b)}{P(i_a|j_b)}. \tag{6.12}$$

As long as one of these conditions is satisfied, time point t_1 will be classified with a, and consequently, $V(i_a|\mathbf{x}_i, w_j) > V(i_b|\mathbf{x}_i, w_j)$. This condition will persist after the transition until the time point t_2 in the figure, i.e., as long as $P(\mathbf{x}_i|a) \approx P(\mathbf{x}_i|b)$, (6.8) will favor the choice of class a.

Non-separable features

Another situation where state-dependent posterior probabilities will help classification is in the case of features which are not easy to (linearly) separate. Consider the toy example of Figure 6.2b,

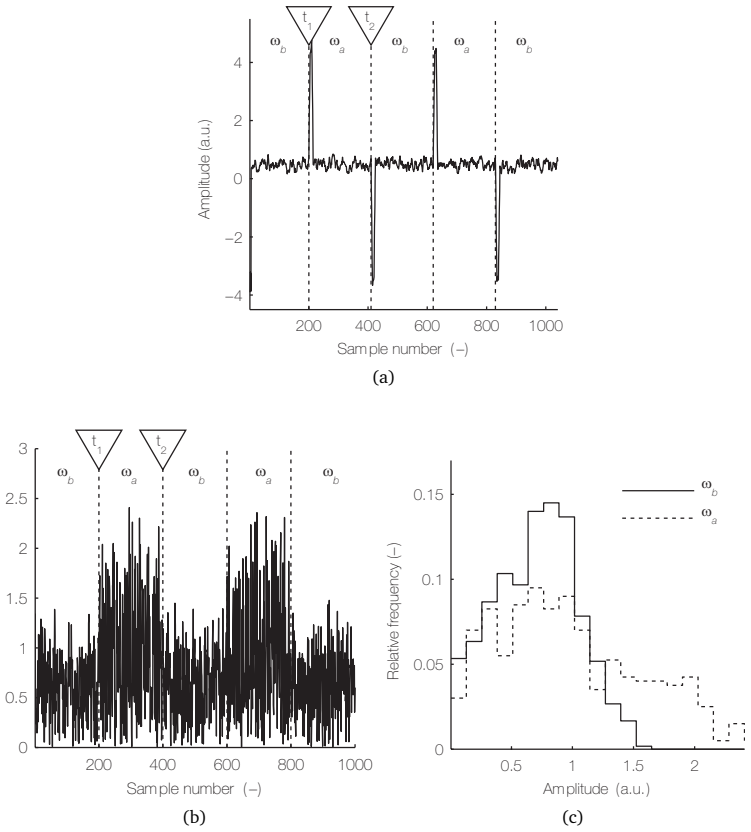


Figure 6.2: Toy example of a feature which is (a) only discriminative at the transition between states and (b) difficult to separate linearly. (c) Histogram of the feature in (b).

and the corresponding histogram for the values of each class in Figure 6.2c. Because there is a large overlap between the histograms of both classes, the classification error with the linear discriminant of (6.3) will be correspondingly large. However, during the intervals of class a, for example between t_1 and t_2, the feature occasionally has a higher value, uncharacteristic of b. These points would be correctly classified using a linear discriminant because here, $P(\mathbf{x}|a) > P(\mathbf{x}|b)$. However, while a linear discriminant would make an error as soon as the feature value would be lower again, these points will have an important effect on state-dependent factors $V(w_i|\mathbf{x}_i, w_j)$. With the same assumptions as before, class a would be chosen if one of the conditions in (6.12) is satisfied. However, unlike the previous case, the likelihood $P(\mathbf{x}_i|i_a)$ will never be substantially higher than $P(\mathbf{x}_i|i_b)$, meaning that the right condition of (6.12) will likely not be satisfied. Each time $P(\mathbf{x}_i|i_a)$ is slightly larger than $P(\mathbf{x}_i|i_b)$ the ratio $V(j_b|\mathbf{x}_j, w_k)/\,V(j_a|\mathbf{x}_j, w_k)$ will become smaller, eventually allowing the left condition of (6.12) to be satisfied and class a to be chosen.

Hidden Markov Models

The framework introduced by the LD is not the most appropriate to model dependency on previous states. Bayesian networks, represented by directed acyclic graphs, however, are very well suited for this purpose. These graphs represent dependencies between random variables using directed edges. The joint probability of the variables $V = \{v_1, .., v_N\}$ in a Bayesian network is given by

$$P(V) = \prod_{i=1}^{N} P\left(v_i | \, pa(v_i)\right), \tag{6.13}$$

where $pa(v)$ denotes the parent nodes of v. In the problem of sleep stage detection, these dependencies can be simplified by allowing the current state to depend on the previous state and on the observations on the current state. In this case the network corresponds to a HMM where the class of each state, w_t, remains hidden, and the only observable variables are the feature vectors, $\mathbf{x}_t$, which depend on them [229]. The joint probability of the variables in this network is given by

$$P(W,X) = P(w_1)P(\mathbf{x}_1|w_1)\prod_{t=2}^{T} P(w_t|w_{t-1})P(\mathbf{x}_t|w_t), \tag{6.14}$$

where $P(w_t|w_{t-1})$ represents the probability of transitioning from one state to the following, $P(\mathbf{x}_t|w_t)$ the probability that observation $\mathbf{x}_t$ was generated by w_t, and $P(w_1)$ is the initial state probability, i.e., the initial belief about the starting state. For a given sequence of observations $X = \{\mathbf{x}_1, .., \mathbf{x}_T\}$, the corresponding sequence of states $W = \{w_1, .., w_T\}$ can be estimated as the sequence of states $\hat{w}_t$ that maximizes the conditional probability $P(w|X)$,

$$\hat{w}_t = \arg\max_{w_t} P(w_t|X). \tag{6.15}$$

Because $\max P(w_t|X) \propto \max P(w_t, X)$, the states can be estimated as the sequence which maximizes the joint probability of the whole chain. This can be efficiently computed with the Viterbi algorithm [230] where the most likely sequence of states is progressively determined as the sequence that maximizes the joint probability up to each state,

$$V_t(w_t) = \max_{w_{1,..,t-1}} P\left(w_{1,..,t}, X_{1,..,t}\right), \text{ for } 2 \leq t \leq T, \tag{6.16}$$

which can be factorized and recursively computed as

$$V_t(w_t) = \max_{w_{t-1}} [P(X_t|w_t)P(w_t|w_{t-1})V_{t-1}(w_{t-1})], \tag{6.17}$$

with the tail function $V_1(w_t)$ given by

$$V_1(w_1) = \max_{w_1} [P(X_1|w_1)P(w_1)]. \tag{6.18}$$

The most probable class in state t, $\hat{w}_t$, will finally be given by

$$\hat{w}_t = \arg\max_{w_t} V_t(w_t). \tag{6.19}$$

Note that the classification rule in (6.19) is equivalent to the rule in (6.7). A HMM, decoded with the Viterbi algorithm, will therefore benefit from the properties described in the previous sections.

Conditional Random Fields

In generative models such as HMM classifiers, the parameters are learned by maximizing the joint probability distribution $P(W,X)$, which in turn requires the distribution of the observations, $P(X)$, to be modeled or somehow learned from the data. When the features of the observed variable X are not independent, the joint distribution may be extremely difficult to model, requiring either large amounts of training data, or strong assumptions about the variables. Discriminative models, such as CRF [223] avoid this problem by computing the probability $P(y|\mathbf{x})$ of a possible output $y = (y_1, y_2, \cdots, y_n)$ given an observation $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n)$, avoiding the explicit modeling of the marginal $P(X)$. By simplifying the modeling problem and not requiring any assumption about the independence of the features (only about the states), discriminative models make better use of correlated, interdependent features, which are common in the case of sleep stage detection using cardiorespiratory features. CRF are a special case of undirected graphs which are globally conditioned on the observation X. Parameter learning and inference is usually performed by means of factor graphs [231], a type of model which describes the probability distribution of the network using non-negative factors to express interaction between random variables. The joint distribution can be factorized over all maximal cliques,

$$P(V) = \frac{1}{Z} \prod_{c \in C} \Psi_c(v_c), \tag{6.20}$$

where the potential functions Ψ_c are the factor potentials of a variable v_c in a clique c, and Z is a normalization function, needed because the potential functions can be any arbitrary function,

$$Z = \sum_{v} \prod_{c \in C} \Psi_c(v_c). \tag{6.21}$$

Because they are strictly positive, the joint distribution can be described by a log-linear model [232],

$$P(V) = \frac{1}{Z} \exp\left(\sum_{c \in C} \lambda_c f_c(v_c) \right), \tag{6.22}$$

$$Z = \sum_{v \in V} \exp\left(\sum_{c \in C} \lambda_c f_c(v_c) \right), \tag{6.23}$$

where the set of parameters $\lambda_c \in \mathbb{R}_0^+$ is selected to maximize the model fit. Defining a set of K (with $K = S^2 + S$ and S as the number of classes or states) parameters $\theta = \{\lambda_{s,r}, \lambda_v\}$ and feature functions $f_k(y_t, y_{t-1}, \mathbf{x}_t)$ incorporating state-transition functions and state-observation functions,

$$f_k(y_t, y_{t-1}, \mathbf{x}_t) = \begin{cases} 1_{\{y_t=s\}} 1_{\{y_{t-1}=r\}}, & \text{for each } (s,r) \\ 1_{\{y_t=s\}} \mathbf{x}_t, & \text{for each } (s,x) \end{cases} \tag{6.24}$$

where the notation $1_{\{\text{condition}\}}$ has a value of 1 if the condition is true, and 0 otherwise, the joint probability in (6.22) can be simplified

$$P(X,Y) = \frac{1}{Z} \exp\left(\sum_t^T \sum_k^K \lambda_k f_k(y_t, y_{t-1}, \mathbf{x}_t) \right), \tag{6.25}$$

which corresponds to the general log-linear model of (6.22) and (6.23). Using Bayes' rule, the conditional probability $P(Y|X)$ can be obtained,

$$P(Y|X) = \frac{\frac{1}{Z} \exp\left(\sum_t^T \sum_k^K \lambda_k f_k(y_t, y_{t-1}, \mathbf{x}_t) \right)}{\frac{1}{Z} \sum_y \exp\left(\sum_t^T \sum_k^K \lambda_k f_k(y_t, y_{t-1}, \mathbf{x}_t) \right)}, \tag{6.26}$$

which is equivalent to

$$P(Y|X) = \frac{1}{Z(X)} \prod_t^T \exp\left(\sum_k^K \lambda_k f_k(y_t, y_{t-1}, \mathbf{x}_t) \right), \tag{6.27}$$

with

$$Z(X) = \sum_y \prod_t^T \exp\left(\sum_k^K \lambda_k f_k(y_t, y_{t-1}, \mathbf{x}_t) \right). \tag{6.28}$$

These equations correspond to the general form of a CRF, (6.20) and (6.21), with the potentials Ψ_t given by the function

$$\Psi_t = \exp\left(\sum_k^K \lambda_k f_k(y_t, y_{t-1}, \mathbf{x}_t) \right). \tag{6.29}$$

It should be mentioned that the set of observations used at time point t and denoted by $\mathbf{x}_t$, can actually contain observations from different points in time as well. In this study, and to allow a fair comparison with LD where the posterior probabilities of each class at a given time are determined by the feature vector (observations) at that time, only features observed at time point t were used. The parameters $\theta = \{\lambda_{s,r}, \lambda_v\}$ need to be estimated such that the model has a good fit with the training data set. Using N input sequences (corresponding to state and observation sequences from

N different subjects), the model fit is expressed using a conditional log-likelihood [233],

$$L(\theta) = \sum_{i=1}^{N} \log P\left(Y^i|X^i\right) = \sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{k=1}^{K} \lambda_k f_k\left(y_t^i, y_{t-1}^i, \mathbf{x}_t^i\right) - \sum_{i=1}^{N} \log Z\left(X^i\right), \tag{6.30}$$

where the superscript i indicates the i-th state-observation sequence pair. The set of parameters θ that maximizes $L(\theta)$ can be found using numerical optimization techniques such as the Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm [234]. The decoding of a linear-chain CRF is done in a similar way as HMM, using a modified version of the Viterbi algorithm [233]. The goal is still to estimate the sequence of states that maximizes the conditional probability $P(w|X)$, as in (6.15). The relation $\max\ P(w|X) \propto \max\ P(w,X)$ still holds, and the estimation is equivalent to estimating the maximum $V_t(w_t)$,

$$\begin{aligned} V_t(w_t) &= \max_{w_{1,..,t-1}} P\left(w_{1,..,t}, \mathbf{x}_{1,..,t}\right) \\ &= \max_{w_{1,..,t-1}} \exp\left(\sum_{i}^{t}\sum_{k}^{K} \lambda_k f_k(w_i, w_{i-1}, \mathbf{x}_t)\right) \\ &= \max_{w_{1,..,t-1}} \prod_{i}^{t} \Psi_t(w_i, w_{i-1}, \mathbf{x}_i). \end{aligned} \tag{6.31}$$

Factorizing the equation and rewriting it recursively, we get

$$V_t(w_t) = \max_{w_{t-1}} \Psi_t(w_t, w_{t-1}, \mathbf{x}_t)\, V_{t-1}(w_{t-1}), \tag{6.32}$$

with the tail function $V_1(w_1)$ given by

$$V_1(w_1) = \max_{w_1} \Psi_t(w_1, \mathbf{x}_1). \tag{6.33}$$

CRF was used to classify the toy example of Figure 6.2a and Figure 6.3a illustrates the results. It is clear that the score obtained in the case of a feature with transitional properties matches almost perfectly the class annotations. This illustrates how such features can be exploited by CRF, but not by LD. Regarding the CRF classification of the toy example of Figure 6.2b, illustrated in Figure 6.3b, the score obtained in the case of a feature which is not easily separable also behaves as expected. Despite the large number epochs with ambiguous values, which would cause LD to erroneously classify many instances, as more and more unambiguous feature values are found, the classification converges to the correct class. Apart from a delay in the transition between detected classes, the classification remains correct until the end of that interval. Besides being able to respond to transitional features, CRF is thus also adequate to classify sequences where states remain stable for a certain amount of time, such as sleep.

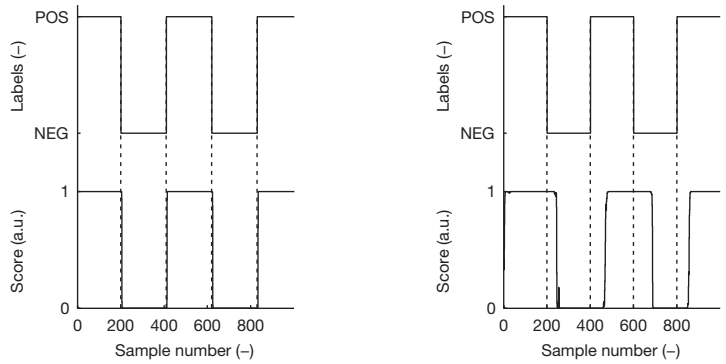


Figure 6.3: CRF scores and corresponding classes for toy example of (a) Figure 6.2a and (b) Figure 6.2b. In both cases, the positive label corresponds to class w_b and the negative label to class w_a. A score closer to 1 indicates a higher posterior probability for class w_b. 'a.u.' stands for 'arbitrary units'.

6.2.5 Training and evaluation

In order to compare the performance of the LD, HMM and CRF classifiers, four separate, non complementary detection tasks were considered: N3, non-REM (NREM), REM, and Wake. For the N3 detection task, S3 and S4 from the SIESTA data were merged into a single N3 class; for the NREM detection task, S1, S2, S3 and S4 from the SIESTA data and N1, N2 and N3 from the Boston and Eindhoven sets were merged into a single NREM class; for the REM and Wake tasks, the corresponding classes in each data set were used. Each of these classes was considered in a 'one versus rest' setting and for each task a 10-fold cross-validation scheme was used. Furthermore, to guarantee that the validation gives, as much as possible, an unbiased estimate of subject-independent classification, care was taken to guarantee that two recordings of the same subject were part of the same fold. To allow a paired comparison, the same folds were used to validate each classifier.

LD Classifier

There is a large redundancy in the extensive set of cardiac and respiratory features used in this study. Because the LD classifier is particularly sensitive to the presence of redundant, and more importantly, non-discriminative features, the correlation feature selection algorithm [235] with sequential forward search was used to restrict the features to a set which maximizes their discriminative power, while minimizing redundancy between them. Feature selection was performed on the training set of each iteration of the cross-validation procedure to avoid biasing the validation performance. The classification score is obtained as the difference between the discriminant functions of the positive class and the remaining classes (the left-hand side of (6.1)).

HMM Classifier

A discrete HMM classifier [229] was used for sleep stage detection. In order to estimate the emission probabilities, a mapping between the feature space and a discrete set of symbols was first defined. This was achieved with k-means clustering, where the number of clusters and the centroid of each cluster were determined during the training step of each cross-validation iteration. The number of clusters was automatically determined for each classification task by computing, after k-means clustering for a varying value of k, the value that maximizes the normalized mutual information (NMI) between the labels (class) of each feature vector and the corresponding, closest clusters [236],

$$\arg\max_{k} \mathrm{NMI}\left(\lambda^{w}, \lambda_{K}^{k}\right), \tag{6.34}$$

with

$$\mathrm{NMI}\left(\lambda^{w}, \lambda_{K}^{k}\right) = \frac{\sum_{i=1}^{W}\sum_{j=1}^{K} n_{i,j} \log\left(\frac{n \cdot n_{i,j}}{n_i^w n_j^k}\right)}{\sqrt{\left(\sum_{i=1}^{W} n_i^w \log\left(\frac{n_i^w}{n}\right)\right)\left(\sum_{j=1}^{K} n_j^k \log\left(\frac{n_j^k}{n}\right)\right)}}, \tag{6.35}$$

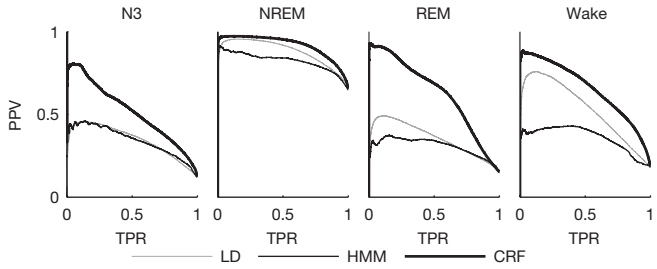
and where λ^w is the mapping between each feature vector and its class, and λ_K^k is the mapping between each feature vector and the closest cluster after k-means clustering with K clusters, n_i^w is the number of feature vectors belonging to class i according to λ^w, n_j^k is the number of feature vectors assigned to cluster j according to λ_K^k, and $n_{i,j}$ is the number of feature vectors with class i and in cluster j. During classification, the cluster index of each feature vector is determined by finding the closest cluster centroid. The index of this cluster is used as input to the Viterbi algorithm to obtain a score expressing the posterior probability of the positive class according to (6.17).

CRF Classifier

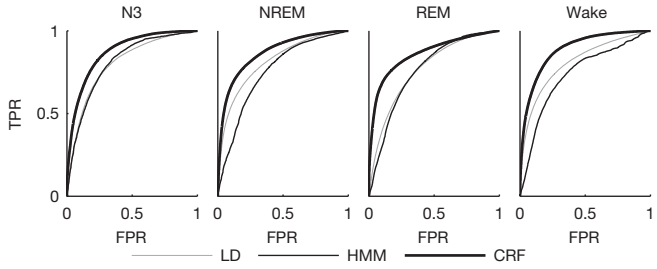
The CRF classifier was trained with the BFGS algorithm, and classification was performed with the modified version of the Viterbi algorithm, yielding a score given by (6.32) for each epoch which can be interpreted as the posterior probability of the positive class in that epoch.

Classification Performance

In order to compare the classification performance of each classifier, the scores obtained for each test subject in each iteration of the cross-validation procedure were collected and aggregated (pooled). The precision-recall (PR) curve - which plots the positive predictive value (PPV) against the true positive rate (TPR) - and the receiver operating characteristic (ROC) - which plots the TPR versus the false positive rate (FPR) - were computed for a varying threshold on the scores output by each classifier. Although the PR curve depends on the priors of each class, it is most useful in the



(a)



(b)

Figure 6.4: Pooled (a) PR and (b) ROC curves per classifier and classification task, for data set 'all'

case of heavily imbalanced classes, such as when detecting Wake (in average, 18.7% of all epochs in our 'all' data set), N3 (12.4% of all epochs) or REM (15.2% of all epochs). In these cases this curve allows us to easily evaluate whether an increase in true positives comes at a disproportionate cost in false positives. The threshold leading to the maximum pooled Cohen's κ coefficient of agreement [156] was then computed. Based on this threshold, the κ for each subject was computed. Note that because this threshold was selected based on the pooled κ, it will not necessarily correspond to the maximum κ for each subject. Significance was tested with a two-tailed Wilcoxon signed-rank test [237] for each evaluation metric.

6.3 Results and discussion

Figures 6.4a and 6.4b compare the pooled PR and ROC curves obtained with each classifier, for each detection task in the 'all' data set, respectively. In all detection tasks the CRF classifier outperforms the other classifiers over the entire solution space.

Figures 6.5a and 6.5b compare the average κ and accuracy obtained with each classifier for the different classification tasks in both data sets, respectively. The performance of the CRF classifier is significantly higher than both the HMM and the LD classifiers in all tasks. The performance in the 'regular' data set is also higher than in the 'all' data set, reflecting the more regular sleep structure

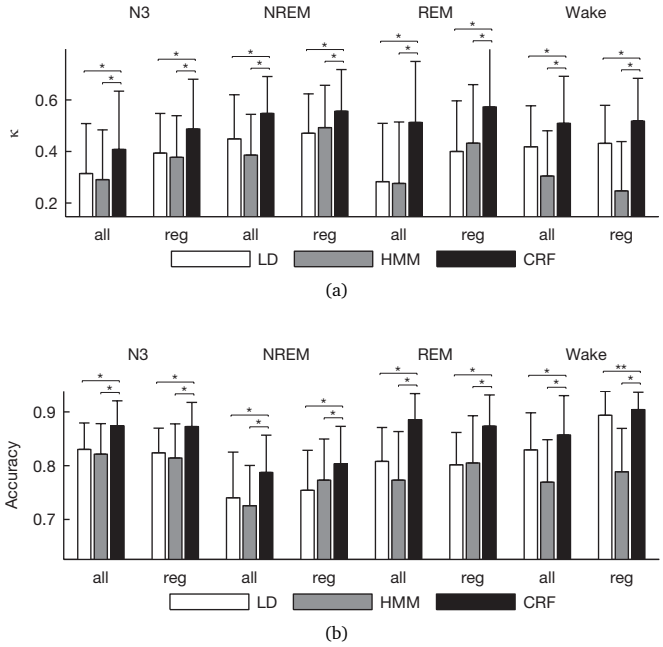


Figure 6.5: Mean (a) κ and (b) accuracy, per classifier, per detection task for both data sets. * and ** indicate significantly higher performance ($p < .001$ and $p < .01$ respectively), after a two-tailed Wilcoxon signed-rank test [237].

of those subjects.

As illustrated in Figure 6.6, the area under the curve (AUC) for the PR and ROC curves is also highest with the CRF classifier. Using a Wilcoxon rank sum test we found no significant differences in CRF performance for any task between the subjects of each of the three data sets comprising the 'all' data set. Regarding the differences in performance improvement, particularly apparent in the PR curves, it is important to analyze both PR and ROC curves in context: while it seems that the improvement in PPV for NREM is much smaller than for N3 and REM, it is worth reminding that the prior probability of NREM is much higher than of the remaining classes and that in this case the PR curve will always seem optimistic. In this case it should be noticed that the improvement in AUC for the ROC curve of NREM is in line with the improvements obtained for the other classes.

To give some examples of CRF classification, Figures 6.7 and 6.8 illustrate the reference hypnogram and corresponding results obtained for the subject closest to the average κ performance (across all tasks) and for the subject with the best average performance, respectively. Regarding Wake detection, we can see in Figure 6.7 that some of the false positives occur during REM periods, most notably during the second half of the night. This is likely related to the increase in sympathetic activity during phasic REM [238], which shares some characteristics with a Wake

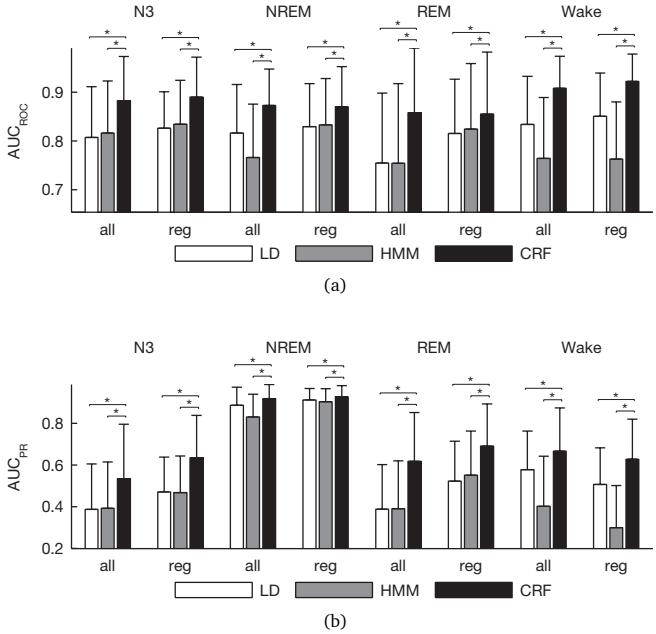


Figure 6.6: Mean area under the (a) PR and (b) ROC curves, per classifier, per detection task for both data sets. * indicates significantly higher performance ($p < .001$), after a two-tailed Wilcoxon signed-rank test [237].

state. Regarding N3 detection, it is interesting to note, also in Figure 6.7, that the first (shorter) periods of N3 are incorrectly classified. Paying closer attention to the posterior probabilities for this class we see a gradual increase which comes close to the detection threshold; although the periods were misclassified, this suggests that if the periods were longer, even for such a difficult detection of N3 the classification would probably eventually converge to the correct class, as happened for the correctly detected period after epoch 200. Finally, regarding REM detection in Figure 6.7, it is interesting to note the missed REM detections between epochs 550 and 600: looking at the posterior probabilities of NREM in these epochs it is clear that the classifier is confusing many epochs of this REM period with NREM. This is likely due to the presence of intervals of tonic REM, which are parasympathetically driven [238] and therefore more difficult to distinguish from NREM.

A direct comparison between the results obtained with the CRF classifier proposed in this chapter and previous work in this area is not easy. With the exception of Wake and N3 detection, almost no literature was found on binary classification. An additional factor which makes the comparison difficult is related to the large between-subject variations, where small data sets might not be representative enough of the problem at hand. Nevertheless, in order to compare the performance of the CRF classifier with the state-of-the-art in this area, the problem was restricted to binary classification using the same modalities described in the literature (ECG and/or respiratory effort). In

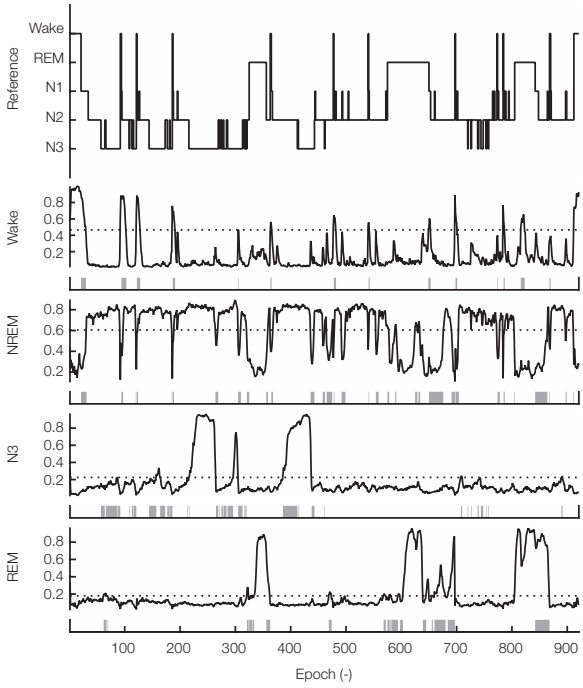


Figure 6.7: Example reference hypnogram (top) and classification results (remaining plots) for the subject closest to the average performance (averaged over all classification tasks). For each classification task, each plot illustrates the posterior probability obtained with CRF (solid line), the classification threshold (dashed line), and epochs for which there was a classification error (gray bars). The κ values obtained for this subject were 0.55 (Wake), 0.58 (NREM), 0.45 (N3), and 0.56 (REM).

cases where no 'one versus rest' classifiers were found, binary classifiers were chosen which had the target class as one of the classes. For example when literature described NREM versus REM, we analyzed separately the performance of NREM and REM detections. Based on this selection, classifiers with the best reported κ and with the best accuracy were chosen for comparison and are summarized in Table 6.5.

Regarding N3 detection, the CRF classifier achieves a lower κ but comparable accuracy than the LD classifier described in our earlier work [239] for both data sets. It is relevant to note that in that work we used the same set of cardiorespiratory features and a similar LD classifier on a comparable data set. However, that work proposed the use of spline fitting to reduce the effect of body motion artifacts and within-subject variability and more importantly, the use of a time-delay of 5 minutes in the classification. It is interesting to note that even without these two critical elements, the CRF classifier still achieves a comparable accuracy in the same task. To provide a fair comparison, we also compared the CRF classifier with the the results reported in that work [239] when spline fitting and time-delay were not used. In that case it is clear that CRF achieves a higher

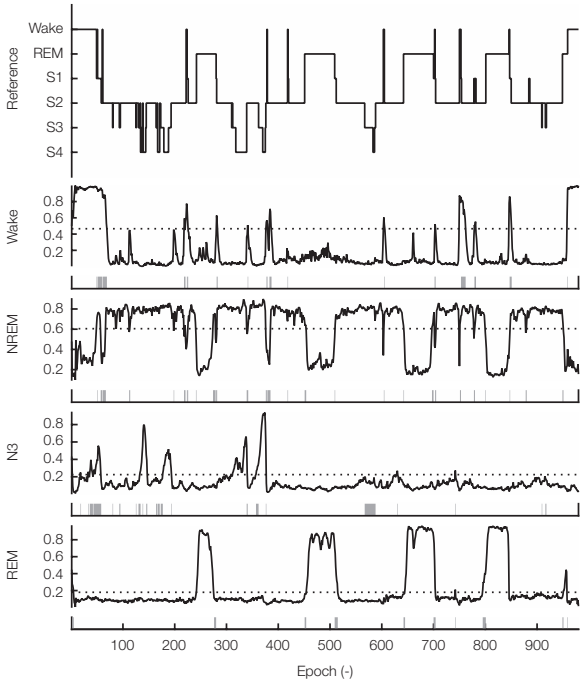


Figure 6.8: Example reference hypnogram (top) and classification results (remaining plots) for the subject with the best average results (across all tasks). The κ values for this subject were 0.72 (Wake), 0.86 (NREM), 0.58 (N3), 0.90 (REM).

accuracy and κ for the 'regular' data set, suggesting its superiority over LD for the same task, using the same base features. Regarding NREM detection, the CRF classifier achieves a comparable κ and accuracy as the HMM classifier of Mendez et al. [172] although they use a much smaller data set limited to a narrow age range between 40 and 50 years. Regarding REM detection, CRF achieves a comparable κ for the 'regular' data set, and a lower coefficient for the 'all' data set. In both cases, it achieves a higher accuracy. Finally, regarding Wake detection, the classifier achieves a lower κ but higher accuracy (for the 'regular' data set) than the classifier of Redmond et al. [80], and a lower κ and accuracy than the classifier of Long et al. [139]. Regarding the latter, it is interesting to note that the set of features used was the same as in this chapter and that the data set used in that work was also a subset of the 'regular' data set described here. Because the LD classifier used in this chapter is in all similar to the one used in that work, and that CRF outperformed LD for this data set, we speculate that the decrease in performance could come from the characteristics of the remaining subjects, arguably more difficult from a classification point of view. It should be noted that for Sleep/Wake classification there are studies which report higher performance [96]. However, they were excluded from this comparison because they use actigraphy, which remains the single most discriminative feature for this task.

Table 6.5: Comparison with classifiers reported in the literature

	Author	N	Modalities	Accuracy (%)	κ (-)
N3	Long [239][a†]	325	ECG RE	88.9	0.51
	Long [239][*]	257	ECG RE	86.1	0.45
	CRF ('regular')	135	ECG RE	87.27 (4.49)	0.49 (0.19)
	CRF ('all')	342	ECG RE	87.38 (4.64)	0.41 (0.23)
NREM	Mendez [172][b†]	12	ECG	79.3	0.56
	CRF ('regular')	135	ECG RE	80.34 (6.97)	0.56 (0.16)
	CRF ('all')	342	ECG RE	78.71 (6.93)	0.55 (0.14)
REM	Mendez [172][b†]	12	ECG	79.3	0.56
	CRF ('regular')	135	ECG RE	87.35 (5.79)	0.57 (0.22)
	CRF ('all')	342	ECG RE	88.49 (4.91)	0.51 (0.24)
Wake	Redmond [80] [c‡]	31	ECG RE	89	0.6
	Long [139] [c§]	15	RE	94.2	0.58
	CRF ('regular')	135	ECG RE	90.42 (3.23)	0.52 (0.17)
	CRF ('all')	342	ECG RE	85.69 (7.32)	0.51 (0.18)

Comparison with best studies reported in the literature, restricted to classifiers which use the same modalities. The indicated accuracy and κ correspond to the average, and between parenthesis (when available), the standard deviation.
Original classification task: [a] N3/other, [b] NREM/REM, [c] Sleep/Wake.
Classifier in the literature with the highest [†] κ and accuracy, [‡] κ, [§]accuracy.
[*] Same classifier as the best reported for N3 classification [239], but features are not smoothed with spline fitting and no time delay is used. Reported results restricted to subjects with more than 30 minutes of N3.

Finally, it is important to highlight some of the limitations of this study. First, and despite the use of a large data set with a wide range of ages, none of the subjects suffered from sleep disorders. This has obvious limitations in a clinical context for which disordered patients are of particular interest. Second, the classification tasks evaluated in this study were binary. Although it is clear from the examples in Figures 6.7 and 6.8 that the output of each binary task is somewhat complementary, this technique will be most useful when the classifier is extended to a multi-class task. An additional note should be made in regard to the use of a discrete HMM classifier. Given the transitional nature of sleep stages (in particular during NREM sleep), it is possible that continuous HMM, modeled for example using Gaussian Mixture Models, could be better suited for the task at hand. Nevertheless, despite these limitations, this chapter demonstrates the superiority of CRF for sleep stage classification over the more popular LD and discrete HMM classifiers. Because the CRF framework presented is not limited to binary classification, the extension to multiple stage classification will be addressed in future work.

6.4 Conclusions

The probabilistic properties of sleep stage sequences and transitions are explored to improve the performance of sleep stage detection using cardiorespiratory features. A new classifier, based on Conditional Random Fields, is compared with a classifier based on Hidden Markov Models and a Bayesian Linear Discriminant for different sleep stage detection tasks: N3, REM, NREM and Wake.

The CRF classifier was evaluated using cross-validation on a large data set comprising 342 PSG recordings of healthy subjects. It was found to significantly outperform the other classifiers in all performance metrics, suggesting the adequacy of this classifier for the problem of sleep stage detection. In addition, it was tested on a larger data set than most state-of-the-art work reported in the literature and therefore with a wider range of population characteristics. It achieves, in particular for subjects with regular sleep characteristics, comparable accuracy for N3, higher accuracy and κ for REM, and higher accuracy and comparable κ for NREM. Future work will explore the post-processing of cardiorespiratory features used in classification, which has proven useful in our earlier work in N3 classification and the use of the CRF classifier in multi-stage classification problems and in recordings of subjects with sleep disorders such as insomnia and sleep-disordered breathing.

Chapter 7

A comparison of probabilistic classifiers for sleep stage classification

Adapted from P. Fonseca, N. den Teuling, X. Long, and R. M. Aarts, A comparison of probabilistic classifiers for sleep stage classification using cardiorespiratory features, *submittted to Physiological Measurement*.

Abstract: In this chapter we investigate sleep stage classification in a home sleep monitoring context restricted to cardiorespiratory (ECG and RIP) inputs and evaluate the feasibility of sleep staging on obstructive sleep apnea (OSA) patients. Considering previous use of CRF in this area, we compare CRF to HMM and LD classifiers on a five-class sleep staging task (Wake/N1/N2/N3/REM) and explore the benefit of incorporating time information in the classification. Evaluating the six classifiers on a data set of 443 night-time PSG recordings from 231 subjects (100 subjects with healthy 'regular' sleep architecture, and 51 subjects previously diagnosed with OSA), we conclude that CRF with time information (CRFt) outperforms all other classifiers on all tasks, achieving a median accuracy and Cohen's κ for all subjects of 62.8% and 0.44 for five classes, 68.8% and 0.47 for four classes (W/N1 and N2/N3/REM), and 77.6% and 0.55 for three classes (Wake/non-REM/REM). No advantage was found in training classifiers dedicated specifically to three- or four-class tasks. However, an advantage was found in training classifiers specifically for 'regular' and 'OSA' subjects, achieving an improvement in classification performance for these groups. For 'regular' subjects, CRFt achieved a median accuracy and Cohen's κ of 67.0% and 0.51, 70.8% and 0.53 and 81.3% and 0.62 for five-, four- and three-classes respectively, and for 'OSA' patients, of 60.0% and 0.40, 70.0% and 0.45, and 75.8% and 0.51 for five-, four- and three-classes respectively. Overall, the results suggest that CRFt is not only better at learning and predicting more complex and irregular sleep architectures, but it also performs reasonably well in five-class classification, the standard for sleep scoring used in clinical PSG. Additionally, and albeit with a decrease in per-

formance when compared with healthy subjects, sleep stage classification in OSA patients using cardiorespiratory features and CRFt seems feasible with reasonable accuracy.

7.1 Introduction

Despite substantial advances in the last decades in simpler and less obtrusive techniques for sleep monitoring, overnight polysomnography (PSG) remains the gold standard for sleep assessment and diagnosis of sleep disorders. However valuable, it is a costly, highly specialized procedure, with substantial burden to the subjects under investigation and often limited to one or two consecutive nights. Because PSG is not practical for long-term sleep monitoring, significant research effort has been put in the development of sensors and techniques which allow sleep to be measured with less expensive sensors that are easier to set up and which can be used at home [80, 93, 101, 194, 206]. Given the limitations of home sleep monitoring, however, many sensors commonly used in PSG such as electroencephalography (EEG), electrooculography (EOG) or electromyography (EMG), are essentially excluded. The only physiological characteristics that can conveniently and reliably monitored are body movements, for example by means of actigraphy [198], cardiac activity, for instance with electrocardiography (ECG), photoplethysmography (PPG) [240, 241], or ballistocardiography (BCG) [92], and surrogates of respiratory effort based on respiratory movements measured with respiratory inductance plethysmography (RIP) belts around the thorax, BCG, or even Doppler radars [94]. Although these modalities can be comfortably used in a home setting, and even set up by the subjects themselves, the relation between these physiological characteristics and sleep stages cannot be trivially annotated using visual rules as is done with standard PSG. Exploiting the varying autonomic characteristics during different stages of sleep, algorithms have been described which automate this process, typically using cardiac characteristics computed from heart rate variability and respiratory features from respiratory effort to classify epochs in states of Wake, 'light sleep' (typically combining N1 and N2, or N1-2 for short), 'deep sleep' or N3, and rapid eye movement (REM) [101, 206, 241].

In Chapter 6 we explained how transitional features and non-separable features are better handled by conditional random fields (CRF) than by probabilistic classifiers traditionally used in sleep stage classification literature, such as hidden Markov models (HMM) or Bayesian linear discriminant (LD). However, that chapter only explored binary classification problems and did not address the more complex problem of multiple-class classification. Here we investigate the use of CRF for the simultaneous classification of multiple sleep stages using the same cardiorespiratory features as in Chapter 6, and compare it with HMM and LD for the same task. In addition, we will evaluate the extension of the multiple-class sleep stage classification problem to the five classes defined by the American Academy of Sleep Medicine (AASM) [11] to separately classify Wake/N1/N2/N3/REM (W/N1/N2/N3/R). This is in contrast with most work done in the area of cardiorespiratory sleep staging where classification is usually limited to either three classes in Wake/non-REM/REM (W/NR/R) [80, 99], or four classes in Wake/N1 and N2/N3/REM (W/N1-

2/N3/R) [101, 206]. Finally, we will evaluate the feasibility of sleep stage classification of obstructive sleep apnea (OSA) patients using CRF and the same set of cardiorespiratory features. Earlier work reported in literature has shown that the presence of this condition negatively impacts the performance of cardiorespiratory classifiers, achieving modest performance on three-class classification of Wake, REM and non-REM (NREM) with accuracy and Cohen's κ of 67% and 0.32 [98] and 70% and 0.41 [102].

7.2 Materials and methods

7.2.1 Data sets

The data sets used in this project were the same as used in Chapter 6, further extended with subjects previously diagnosed with OSA. They comprise a total of 443 recordings of 231 subjects from three databases recorded in different locations during different time periods. The first data set consists of 428 recordings of 216 subjects, and was collected as part of the European Union (EU) SIESTA project [167] between 1997 and 2000. The second data set comprises six subjects, recorded at the Philips Experience Lab of the High Tech Campus, Eindhoven, The Netherlands, during 2010 and the third data set comprises nine subjects, recorded at the Sleep Health Center, Boston, United States of America (USA), during 2009. The protocols for the three data collection studies were reviewed by the local ethics committees and all participants signed informed consent. The healthy subjects in the three data sets were not suffering from depression nor other mental or sleep disorders, and had a Pittsburgh Sleep Quality Index [148] of at most 5. The apnea patients in the first data set had been diagnosed according to the International Statistical Classification of Diseases and Related Health Problems, 10th Ed. (ICD-10) (G47.3) and at the date of investigation were free of psychopharmacotherapeutic drugs for at least five times the half-life of the medication, and in a steady state for on-going non-psychopharmacotherapeutic medication for at least two months. All sleep technicians were blind to the condition of the participants and scored sleep stages on the second and third data sets in five classes (Wake, REM, N1, N2, and N3) according to the AASM guidelines [11], and on the first data set, in six classes (Wake, REM, S1, S2, S3, S4) according to the R&K guidelines [168]. To use a common list of ground-truth classes, S1 was mapped to N1, S2 to N2, and S3 and S4 were merged in a single N3 class. These sleep stage annotations were used as ground-truth for training and validation in our study. Besides all modalities needed for offline visual scoring of sleep stages according to the corresponding guidelines, all recordings comprise ECG recorded with a modified lead II derivation, and respiratory effort recorded with thoracic RIP. Although all subjects in the data set were healthy, we expect an effect of monitoring in the actual characteristics of sleep. The well-known 'first night effect' is known to have an impact on sleep architecture, namely with a reduction in sleep efficiency, a delay in the onset of N3 and REM, and increased number of stage transitions [43]. In order to analyze the impact of these changes, and similar to our previous study, a subset of the complete data set was analyzed separately, com-

Table 7.1: Demographics and sleep statistics of subjects in the two subsets used in the study

	'regular'		'OSA'		'all'	
Parameter	Mean (SD)	Range	Mean (SD)	Range	Mean (SD)	Range
N	100 subjs., 134 recs.		51 subjs., 102 recs.		231 subjs., 443 recs.	
Sex	56 female subj. (56.0 %) 75 recs. females (56.0 %)		7 female subj. (13.7 %) 14 female recs. (13.7 %)		105 female subj. (45.5 %) 198 female recs. (44.7 %)	
Age (year)	44.3 (17.2)	20 – 83	51.4 (9.8)	29 – 73	50.4 (17.95)	20 – 95
BMI (kg/m^2)	24.0 (3.0)	19.0 – 31.3	30.7 (5.1)	22.9 – 43.3	25.9 (4.67)	17.0 – 43.3
TIB (hour)	8.0 (0.4)	7.2 – 9.7	8.3 (0.3)	7.8 – 9.6	8.0 (0.50)	5.8 – 9.7
SE (%)	88.8 (5.6)	75.5 – 99.0	82.9 (11.1)	25.3 – 97.7	81.6 (11.73)	21.7 – 99.0
N1 (%)	8.5 (3.9),	1.6 – 21.8	18.3 (10.6)	4.0 – 63.9	13.0 (8.09)	1.6 – 63.9
N2 (%)	51.8 (6.4)	30.6 – 67.1	53.0 (10.8)	14.0 – 74.2	53.8 (9.10)	14.0 – 76.2
N3 (%)	17.1 (5.7)	6.0 – 36.2	9.4 (7.7)	0.0 – 44.2	13.8 (8.50)	0.0 – 44.2
REM (%)	21.7 (3.3)	16.3 – 31.3	18.1 (5.6)	0.0 – 30.0	18.3 (5.63)	0.0 – 34.8
AHI (evnt/hour)	n.a.		38.3 (23.7)	1.7 – 97.1	n.a.	

'regular' is a subset of the healthy subjects in the 'all' data set where only subjects with a minimum of 5% N3, 15% REM sleep and 75% sleep efficiency were considered. 'OSA' is a subset of the subjects in 'all' data set that had a previous diagnosis of sleep apnea.
N1, N2, N3, and REM percentages were calculated over the total sleep time for each recording.
SD: standard deviation, BMI: body mass index, TIB: time in bed, SE: sleep efficiency, AHI: apnea/hypopnea index, calculated as the number of all apnea and hypopnea events divided by the total sleep time.

prising 134 recordings (100 subjects) with a minimum percentage of each sleep stage considered representative of sleep statistics of healthy adults: at least 5% of N3, 15% of REM sleep, a sleep efficiency of at least 75%, and a minimum of 7 hours in bed [225] (set 'regular'). The second set, also analyzed separately, comprised all 102 recordings of the 51 OSA patients (set 'OSA'). The third set comprised all 443 recordings from all three data sets (set 'all'). Table 7.1 summarizes all subject demographics and sleep statistics of the three subsets.

7.2.2 Feature extraction

The classifiers used in this study use the same set of cardiac and respiratory features described in Chapter 6 and detailed in Section 6.2.2 and Tables 6.2, 6.3 and 6.4. The process and resulting features are summarized in the following.

The baseline wander of the ECG signal was first removed with a high-pass filter using a Kaiser window of 1.016 sec, a cut-off frequency of 0.8 Hz and a side-lobe attenuation of 30 dB. QRS complexes were detected with a Hamilton-Tompkins QRS detector [123, 124] and further localized with the post-processing algorithm described in Chapter 2 [150]. RIP signals were first resampled to a common sampling rate of 10 Hz, and then low-pass filtered with a 10th order Butterworth filter with a cut-off frequency of 0.6 Hz. Baseline was removed by subtracting the median value calculated in a sliding window of 120 s. After the signals were processed, a total of 81 features from ECG, 33 from RIP and 3 from the simultaneous analysis of both signals, were extracted using windows centered on non-overlapping epochs of 30 s, aligned with the ground-truth annotations from PSG. The length of the extraction windows varied from feature to feature as described in Section 6.2.2. An automatic artifact rejection procedure was employed to determine whether seg-

ments of the ECG and RIP signals were too affected by noise or motion artifacts. For each feature extracted from ECG, a given epoch was marked as invalid if the sum of the length of the inter-beat intervals (IBIs) in the window used to compute that feature amounted to less than 50% of the length of that window. For all features extracted from RIP, respiratory peaks and troughs on each epoch were first detected, and a criteria was used to determine whether they had a plausible length [161]. Epochs with one or more implausible peaks or troughs were marked as invalid. After feature extraction and artifact rejection, the values of features for epochs marked as invalid were estimated with linear interpolation between the values of neighboring (valid) epochs. Finally, to reduce between-subject physiological and equipment-related variations [242], all features were normalized to have zero mean and unit standard deviation.

7.2.3 Multiple-sleep stage classification

A detailed explanation of Bayesian linear discriminants, Bayesian networks, hidden Markov models and conditional random fields, and how they behave in the presence of transitional and non-separable features was given in Chapter 6. This section will give only a brief summary of the three classifiers.

Applied in Chapter 4 for multiple-sleep stage classification, the LD classifier uses Bayes' rule for minimizing the probability of error by selecting, separately for each epoch, the most likely class, i.e., the class that maximizes the discriminant function

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}-\mu_i)^t \Sigma^{-1}(\mathbf{x}-\mu_i) + \ln P(\omega_i) \tag{7.1}$$

where μ_i is the average feature vector for class i, Σ is the pooled covariance matrix for all classes, and $P(\omega_i)$ is usually a fixed term with the prior probability of each class. It has been observed that the probability of different sleep stages changes with sleep time [80]. In order to exploit this characteristic, and in addition to the standard LD of (7.1), a variation of this classifier with time-varying prior probabilities (LDtvp) is used [80, 206]. In this case, the last term of the discriminant equation corresponds to $P(\omega_i, t)$, representing the prior probability of class i at time (since lights off) t, with $t = 1, .., T$ where T corresponds to the duration of the recording. All parameters μ_i, Σ, and $P(\omega_i, t)$ are determined during training.

Unlike LD classifiers, where the most likely class is considered separately for each epoch, HMM classifiers make use of Bayesian networks where the dependency between consecutive states can be modeled by means of directed edges. HMM are a special case of Bayesian networks where the class of each state, w_t, remains hidden, and the only observable variables are the feature vectors, $\mathbf{x}_t$, which depend on them [229]. Allowing the current state to depend on the previous state and on the current observations, the joint probability of the variables in this network is given by

$$P(W,X) = P(w_1)P(\mathbf{x}_1|w_1)\prod_{t=2}^{T} P(w_t|w_{t-1})P(\mathbf{x}_t|w_t). \tag{7.2}$$

Using the Viterbi algorithm [230], the most likely sequence of states is progressively determined as the sequence that maximizes the joint probability up to each state,

$$V_t(w_t) = \max_{w_{t-1}} [P(X_t|w_t)P(w_t|w_{t-1})V_{t-1}(w_{t-1})], \tag{7.3}$$

with the tail function for $t = 1$,

$$V_1(w_1) = \max_{w_1} [P(X_1|w_1)P(w_1)]. \tag{7.4}$$

Emission probabilities expressing the likelihood of an observation given a class, $P(X_t|w_t)$, and transition probabilities expressing the likelihood of transitioning to a state w_t given the previous state w_{t-1}, i.e., $P(w_t|w_{t-1})$ are learned during training. Similar to Chapter 6, in this study we used a discrete HMM classifier [229] whereby the estimation of emission probabilities is obtained by mapping the feature space to a discrete set of symbols, or clusters. The number of clusters k was automatically determined during training as the value of k that maximized the normalized mutual information between the labels (class) of each feature vector and the corresponding closest clusters obtained with k-means clustering [236]. During classification, the cluster index of each feature vector was determined by finding the closest cluster centroid. The index of this cluster was used as input to the Viterbi algorithm to obtain the posterior probability of each class and the most likely class for each state.

HMM, like other generative models, require the distribution of the observations to be modeled, or somehow learned from the training data. When the features are not independent this requires large amounts of training data or strong assumptions about the distribution and the relation between features. In the case of our study, this was achieved by mapping the feature space to a limited, discrete, number of clusters. Discriminative models such as CRF [223] do not have this requirement, because they compute the probability of a class based on an observation, $P(w|X)$, directly from data. Because discriminative models do not require any assumptions about the independence of the observed variables, they make better use of interdependent, correlated features, very common in the case of cardiorespiratory sleep stage classification. Parameter learning with CRF is usually done with factor graphs [231] and the conditional probability $P(W|X)$ can be obtained from

$$P(W|X) = \frac{1}{Z(X)} \prod_t^T \exp\left(\sum_k^K \lambda_k f_k(w_t, w_{t-1}, \mathbf{x}_t)\right), \tag{7.5}$$

with

$$Z(X) = \sum_w \prod_t^T \exp\left(\sum_k^K \lambda_k f_k(w_t, w_{t-1}, \mathbf{x}_t)\right), \tag{7.6}$$

where λ_k are the model parameters and f_k the feature functions, in this study specified as the set

of state-transition functions $f_t(w_t, w_{t-1}, \mathbf{x}_t)$ defined for all possible state-transition pairs (s, r), and state-observation functions $f_e(w_t, w_{t-1}, \mathbf{x}_t)$ defined for all state-observation pairs (s, x),

$$f_t(w_t, w_{t-1}, \mathbf{x}_t) = 1_{\{w_t=s\}} 1_{\{w_{t-1}=r\}}, \quad \text{for each } (s, r), \tag{7.7}$$

and

$$f_e(w_t, w_{t-1}, \mathbf{x}_t) = 1_{\{w_t=s\}} \mathbf{x}_t, \quad \text{for each } (s, x), \tag{7.8}$$

where the notation $1_{\{\text{condition}\}}$ has a value of 1 if the condition is true, and 0 otherwise. A total of $K = S^2 + S$ parameters $\theta = \{\lambda_{s,r}, \lambda_{s,x}\}$, with S as the number of classes, are estimated during training using N input sequences (corresponding to state and observation sequences from N different recordings), by maximizing the conditional log-likelihood [233],

$$L(\theta) = \sum_{i=1}^{N} \log P\left(W^i | X^i\right) = \sum_{i=1}^{N} \sum_{t=1}^{T} \sum_{k=1}^{K} \lambda_k f_k\left(w_t^i, w_{t-1}^i, \mathbf{x}_t^i\right) - \sum_{i=1}^{N} \log Z\left(X^i\right), \tag{7.9}$$

where the superscript i indicates the i-th state-observation sequence pair. The set of parameters θ that maximizes $L(\theta)$ can be found using numerical optimization techniques such as the Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm [234]. Decoding of a linear-chain CRF is done in a similar way as HMM, using a modified version of the Viterbi algorithm [233] where the most likely sequence of states is progressively determined as the sequence of classes that maximizes the recursion

$$V_t(w_t) = \max_{w_{t-1}} \exp\left(\sum_{k}^{K} \lambda_k f_k\left(w_t, w_{t-1}, \mathbf{x}_t\right)\right) V_{t-1}\left(w_{t-1}\right). \tag{7.10}$$

In a manner comparable to the use of time-varying prior probabilities for LD, a trivial way of encoding time information in Bayesian chains is to simply add a time feature to the feature set. By defining the lights-off and lights-on instants as time $t = 0$ and $t = 1$ respectively, and by linearly interpolating the values for all epochs, a comparable (and normalized) time feature can be used for each recording. This technique was used for both the HMM and CRF classifiers, and will be further specified as HMMt and CRFt to explicitly indicate the use of time information.

7.2.4 Evaluation

To compare the performance of the three classifiers (and their corresponding time-variations) in the multiple-stage classification tasks, a 10-fold cross-validation procedure was used. Furthermore, to eliminate the possibility of subject-dependent bias between training and validation, when a subject had more than one recording they were both made part of the same fold for either training or validation. The same folds were used for all classifiers to allow for paired comparisons.

The performance of each of the six classifiers (LD, LDtvp, HMM, HMMt, CRF and CRFt) was

evaluated in terms of epoch-per-epoch agreement between the predicted classes on a five-class classification task (W/N1/N2/N3/R) in terms of accuracy (percentage of epochs where the predicted and ground-truth sleep stage agreed) and Cohen's κ coefficient of agreement, an agreement measure that compensates for the chance of random agreement [156]. In addition, the performance of all classifiers was also evaluated on four-class (W/N1-2/N3/R) and three-class (W/NR/R) tasks. These results were obtained by merging both the predicted and ground-truth classes in the following way: for the four-class task, N1 and N2 were merged into a single N1-2 class. For three-class classification, N1, N2 and N3 were merged into a single NR class. A two-tailed Wilcoxon signed-rank test [237] was used to test for significant differences between the best and the remaining classifiers for each task. This evaluation procedure was performed for the 'regular', the 'OSA' and the 'all' data sets separately.

In order to evaluate whether the class merging process is an adequate procedure when a higher performance is desired at the cost of simplifying the classification problem, the performance of each classifier was compared with that obtained with dedicated classifiers, trained specifically for the four- and three-class problem. This was done by merging the ground-truth annotation classes before training. Two-tailed Wilcoxon signed-rank tests were used to test for significant differences between the class-merging and the dedicated-training approaches.

Finally, and given the difference in complexity between all classifiers and to heuristically check for the presence of over-fitting or generalization problems, the training and validation performance of all classifiers was computed using cross-validation for a varying number of recordings and so-called learning curves were plotted.

7.3 Results

Tables 7.6, 7.7 and 7.8 compare the classification performance of each classifier for the 'regular', 'OSA' and 'all' data sets, for each classification task, in terms of κ and accuracy. For three- and four-class classification tasks, the results were obtained by merging the classes resulting from five-class classification. The subset comparisons of Table 7.8 were obtained with cross-validation on the 'all' data set but analyzed separately for the recordings that were part of the 'regular', 'OSA' and 'irregular' (healthy subjects not part of the 'regular' set) data sets.

An unpaired Wilcoxon rank sum test [237] was used to compare, for the CRFt classifier, the performance obtained between the subsets when performing five-class classification. The performance was not significantly different between the 'regular' and the 'irregular' sets, neither for accuracy ($p = .64$) nor for κ ($p = .46$). It was, however, significantly higher for 'regular' than for 'OSA', ($p < .001$ and $p < .001$), and for 'irregular' than for 'OSA' ($p < .01$ and $p < .001$).

Table 7.2 indicates the results for LDtvp and CRFt obtained with a dedicated classifier trained specifically with three and four classes and compares them with the results obtained by merging the classes detected with the five-class classifier for each set.

Table 7.2: Classification performance with classifiers trained with three and four classes

	W/NR/R		W/N1-2/N3/R	
	κ (-)	accuracy (%)	κ (-)	accuracy (%)
'regular' (N=102)				
LDtvp	0.55 (0.15)	80.4 (6.4)	0.46* (0.12)	68.0 (7.0)
CRFt	0.59 (0.15)	**81.8*** (6.1)	**0.51** (0.14)	**71.0** (8.4)
'OSA' (N=102)				
LDtvp	0.40*** (0.12)	74.2 (7.7)	0.39 (0.11)	**68.2** (8.9)
CRFt	**0.50** (0.14)	**77.0** (7.6)	**0.45** (0.13)	**70.6** (8.7)
'all' (N=443)				
LDtvp	0.47*** (0.15)	76.2 (8.0)	0.41*** (0.13)	66.1 (8.4)
CRFt	0.53 (0.15)	**78.0**** (7.8)	**0.46** (0.14)	**68.8*** (9.1)

The table indicates for each classifier and classification task the mean (standard deviation) κ and accuracy.
Indicated in **bold** are the results for which the dedicated classifiers obtained a higher performance than that achieved by merging the results of the five-class classifier. Significant differences are indicated (*** $p < .001$, ** $p < .01$, * $p < .05$).

Table 7.3: Pooled confusion matrix for CRFt and LDtvp classification of 'all' subjects

	pred.↓ ref.→	N1	N2	N3	REM	Wake	PPV (%)
CRFt	N1	1625	2187	51	450	1771	**26.7**
	N2	21071	152450	30925	17899	16142	**63.9**
	N3	539	6340	14968	111	904	**65.5**
	REM	6507	8435	242	37896	6411	**63.7**
	Wake	14053	15875	753	7136	49975	56.9
	Sens. (%)	3.7	**82.3**	31.9	**59.7**	**66.5**	
LDtvp	N1	1607	2032	103	1170	1523	25.0
	N2	26648	150789	25877	20638	22609	61.2
	N3	567	13462	18946	328	1323	54.7
	REM	7054	10837	721	35658	8427	56.9
	Wake	7919	8167	1292	5698	41321	**64.2**
	Sens. (%)	3.7	81.4	**40.4**	56.2	54.9	

The confusion matrices compare the predicted and the reference sleep stages for all epochs of all subjects in the data set after cross-validation with CRFt (above) and LDtvp (below). **Bold** indicates the classifier (between CRFt and LDtvp) which obtained the best sensitivity and PPV for each class.

Tables 7.3, 7.4, and 7.5 indicate the confusion matrix for five-class classification with the CRFt classifier for the three data sets. In addition, the table indicates the sensitivity and the positive predictive value (PPV) for each class and, for comparison, the performance obtained with LDtvp. For Tables 7.4 and 7.5 the performance obtained after evaluating the results of each subset after cross-validation with CRFt on 'all' data set are also indicated.

Finally, Figures 7.1a, 7.1b and 7.2 illustrate the so-called 'learning curves' for the LDtvp, HMM, and CRFt classifiers for the three data sets. For each classifier, the curves illustrate the training and validation performance for five-class classification, averaged after cross-validation, for a varying number of recordings.

Table 7.4: Pooled confusion matrix for CRFt classification of 'regular' subjects

	pred.↓ ref.→	N1	N2	N3	REM	Wake	PPV (%)		
							CRFt	CRFt 'all'	LDtvp
	N1	691	1200	20	155	520	**26.7**	24.7	22.4
	N2	3750	47869	10884	5939	2542	**67.4**	66.2	64.9
	N3	186	2473	8131	24	339	72.9	**78.9**	60.0
	REM	2006	3770	111	17256	1585	69.8	**74.0**	65.5
	Wake	2829	3076	230	1103	8696	54.6	44.1	**65.2**
	CRFt	**7.3**	**82.0**	42.0	**70.5**	63.6			
Sens. (%)	CRFt 'all'	6.7	80.4	31.1	64.1	**74.4**			
	LDtvp	5.7	78.4	**50.3**	66.5	54.1			

See notes to Table 7.3. For comparison, the sensitivity and PPV obtained with LDtvp for the same data set, and with CRFt evaluated for the 'regular' set after cross-validation on the 'all' data set are also given.

Table 7.5: Pooled confusion matrix for CRFt classification of 'OSA' subjects

	pred.↓ ref.→	N1	N2	N3	REM	Wake	PPV (%)		
							CRFt	CRFt 'all'	LDtvp
	N1	1496	1405	16	486	1049	33.6	30.7	**34.2**
	N2	7996	36832	5210	3991	4020	**63.4**	62.5	60.4
	N3	159	1298	2050	50	272	53.5	**54.8**	49.6
	REM	2470	2198	93	9069	2154	56.7	**58.9**	53.0
	Wake	3079	2796	193	1641	9590	55.4	49.5	**60.8**
	CRFt	9.8	**82.7**	27.1	**59.5**	56.1			
Sens. (%)	CRFt 'all'	2.1	81.9	**44.8**	51.8	**59.8**			
	LDtvp	**13.1**	81.8	37.6	49.8	47.8			

See notes to Table 7.4.

7.4 Discussion

As indicated in Table 7.6, the CRF classifier with normalized time (CRFt) outperforms all other classifiers for the 'regular' data set for all classification tasks. The differences are significant for all performance metrics, except for accuracy in the three-class task. The same can be observed for the 'all' data set in Table 7.8, where the differences are significant for both κ and accuracy, for all classification tasks. It is clear that the addition of normalized time information as a feature improves the classification performance for healthy subjects in comparison with the standard CRF. Regarding the 'OSA' data set, as indicated in Table 7.7, CRFt achieves a significantly higher κ and accuracy for the five-class task, accuracy, than all other classifiers except CRF. Interestingly, CRF achieves a better accuracy for the three- and four-class task, albeit not significantly in comparison with CRFt and LDtvp. In contrast with the classification for healthy subjects, it is clear that for the 'OSA' data set the addition of time information leads to smaller improvements over standard CRF. OSA is known to cause increased sleep fragmentation due to the arousals which often follow apnea and hypopnea events [243], and also a decrease of N3 [244, 245] and REM [245]. Likewise, it is reasonable to expect that, at least for patients with a higher OSA severity, the occurrence of apnea events possibly dictates more changes in sleep architecture than the natural progression of

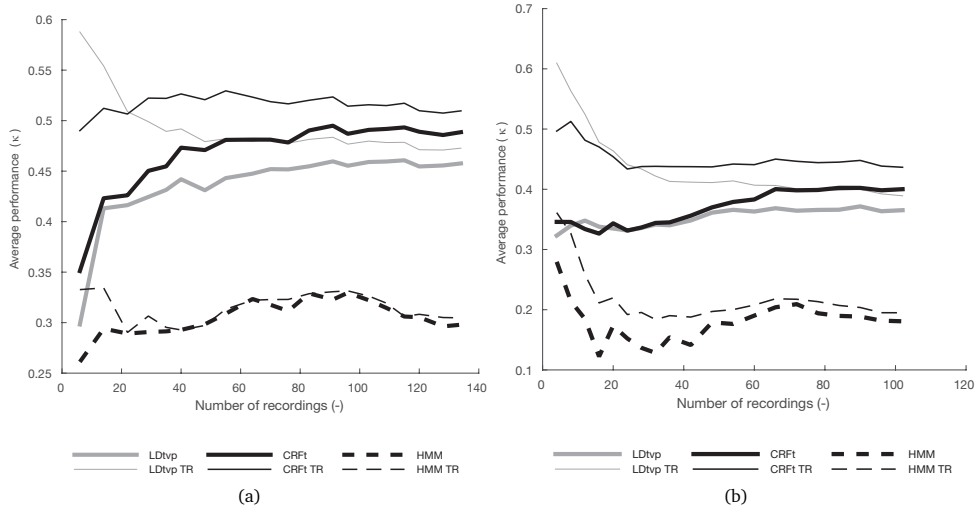


Figure 7.1: Average validation and training (suffix TR) performance κ for a varying number of recordings in the (a) 'regular' and (b) 'OSA' data sets, for the LDtvp, HMM and CRFt five-class classifiers

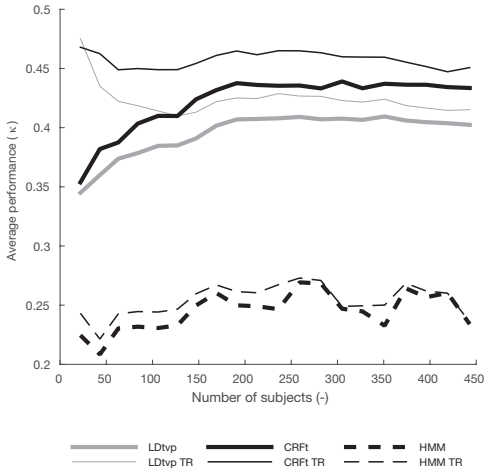


Figure 7.2: Average validation and training (suffix TR) performance κ for a varying number of recordings in the 'all' data set, for the LDtvp, HMM and CRFt five-class classifiers

sleep over time. While for healthy subjects, for instance, there is a strong association between the occurrence of N3 and REM with sleeping time, the presence and progression of these sleep stages for an apnea patient may depend more on the disordered breathing events which can occur during the entire night.

Comparing Tables 7.6, 7.7 and 7.8 it is obvious that the overall performance for 'all' subjects is lower than for 'regular' subjects but higher than for 'OSA' subjects. Accordingly, healthy sleep architectures seem easier to model and classify with the current set of cardiorespiratory features, in contrast with the more complex characteristics of healthy subjects who do not have regular sleep architectures and even more so with OSA patients for which the architecture may be disrupted by their condition. Interestingly, as illustrated by the results for the subsets 'regular' and 'OSA' of the 'all' data set (also indicated in Table 7.8), the performance for those subjects is overall lower than that what was obtained when the classifier was trained separately with examples of only 'regular' and only 'OSA' subjects (Tables 7.6 and 7.7). This suggests that by attempting to model the common characteristics of all subjects, including those with regular, irregular and disrupted sleep architectures, the classifier compromises on the ability to learn characteristics that were specific of individual groups of subjects to be able to better model overall sleep architectures. A prior stratification of the sleep architecture (at least between regular and irregular subjects) and screening for a sleep disorder such as apnea would allow specific classifiers to be used instead of a globally trained one and improve overall classification performance.

Regarding the comparison between the performance of CRFt and LDtvp for the subsets of the 'all' data set, it can be observed that the differences depends on the characteristics of the subset. Although there are no significant differences in performance for the 'regular' subset, the differences are significant for all tasks for the 'irregular' subset and for the five-class task for the 'OSA' subset. This suggests that a CRFt classifier trained with examples of all subjects is better able to learn complex sleep architectures than LDtvp, in particular for the most complex case of five sleep stages.

Regarding Figures 7.1a and 7.1b it is noticeable that the training and validation performances of the CRFt classifier converge at approximately the same number of recordings (around 70-80) as the LDtvp classifier and that both achieve a better training and classification performance than HMM, suggesting that the latter is simply not able to learn the characteristics of the sleep stages for either data set. The same can be seen in Figure 7.2 although, because the complexity of the problem increases with the addition of subjects with irregular sleep architectures, both classifiers require a larger number of subjects for the performances to converge (around 200 recordings). What is clear from the three figures is that for the 'regular', 'OSA' and 'all' sets, the number of recordings in the data sets seems to be sufficient for an adequate training and validation. There is no evidence of over-fitting nor large problems generalizing to new data, in which cases the training performance would seem substantially higher than the validation performance when using the complete data sets.

Table 7.2 shows that there is, generally speaking, no clear advantage to training classifiers dedicated specifically to the three- and four-class tasks as compared with the results obtained by appropriately merging the classes resulting from five-class classification. For CRFt, the only significant improvement is in terms of accuracy for the three-class task in the 'regular' data set and also in terms of accuracy for both tasks in the 'all' set. κ actually decreases (albeit not significantly)

when compared with the results obtained by merging the five-class classification results. For LDtvp there are no significant improvements for any task, and the results are significantly worse in terms of κ for the four-class task for 'regular' and 'all', and for the three-class task for the 'regular' and 'all' tasks. This is likely due to substantial differences between the classes: when merged together, the resulting class is likely very heterogeneous and a classifier trained with fewer classes still needs to learn - with a more complex model - the characteristics of the original classes, not being able to generalize better than a classifier trained with all classes to start with.

Table 7.3 indicates the confusion matrices obtained after cross-validation with CRFt and LDtvp for 'all' subjects. It is clear that the CRFt classifier achieves a superior performance both in terms of sensitivity and PPV for almost all classes, with the exception of a lower sensitivity to N3 and a lower PPV to Wake. Furthermore, it is interesting to note that the LDtvp classifier confuses more often classes which are, in principle, more disparate, than CRFt does. For instance, notice the much larger number of predictions of N1 in epochs of REM (1170 versus 450 with CRFt), or when predicting REM or Wake in epochs of N3 (721 and 1292 against 242 and 753 with CRFt respectively). This suggests that not only does the CRFt do an overall better job of sleep stage classification, but also that the output is overall more plausible than that with LDtvp. Regarding the confusion matrix for the 'regular' data set, Table 7.4 indicates once more that CRFt, either trained with 'regular' subjects, or with 'all' subjects, yields better sensitivity and PPV than LDtvp for the same classes as in Table 7.3, with all subjects. Interestingly, the same is not observed for 'OSA' subjects, as indicated in Table 7.5. In this case, LDtvp achieves the best PPV for Wake, but not the best sensitivity for N3. Where LDtvp outperforms CRFt seems to be in the detection of N1, both in terms of sensitivity as well as PPV. A note should be made about the apparent superior performance of CRFt trained with 'all' subjects versus CRFt trained with 'OSA' for some classes, for example achieving a much higher sensitivity to N3. Although Tables 7.7 and 7.8 show that a dedicated classifier trained with OSA subjects outperforms, overall, a classifier trained with all subjects and then applied to this subset, the results in the confusion matrix appear to contradict that observation. This apparent contradiction stems from the fact that the confusion matrices are based on all epochs aggregated over all subjects in that data set. As we had seen in Table 7.1, these OSA patients have, in average, less REM and especially less N3 than the average for 'regular' subjects and in the complete set. The confusion matrix thus hides the much lower performance obtained by CRFt trained with 'all' subject, for example on subjects with nearly no N3 nor REM sleep.

To illustrate some examples of the classification output with the CRFt classifier, Figures 7.3 and 7.4 illustrate a comparison between the ground-truth reference hypnogram and the classification output for the subject with the best performance and the subject with the performance closest to the median (both in terms of κ), in the 'regular' data set. As it can be clearly seen in Figure 7.3, there is a substantial match between the reference and the prediction for the best subject, with most mistakes occurring in fast transitions between classes, such as between epoch 50 and 100

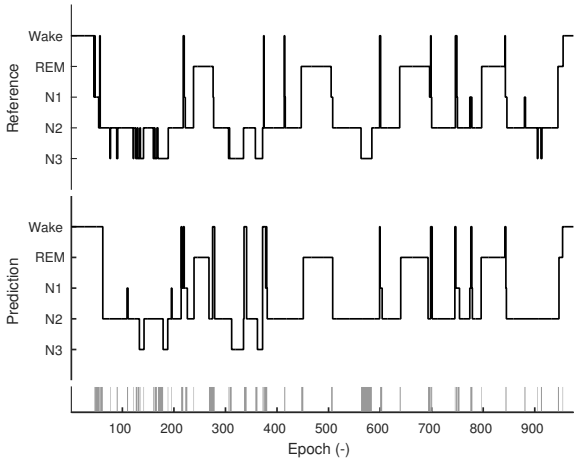


Figure 7.3: Classification example for the subject with the highest κ of 'regular' subjects (κ of 0.74 and an accuracy of 84.2%). From top to bottom: reference hypnogram, result of classification with CRFt classifier, epochs for which there is a mismatch between reference and classification.

where a sequence of transitions between N3 and N2 lead to most mistakes in the first half of that recording. Regarding the subject with the performance closest to the median, in Figure 7.4, it is clear that many classification errors occur once more during periods with faster transitions between stages, such as between epochs 550 and 600 where there are successive transitions between REM and N2, or between epochs 700 and 800 and at the end of the recording where there are successive transitions between Wake, REM and N1. Similar observations are made for 'OSA' subjects, where the best and the subject with performance closest to the median are illustrated in Figures 7.5 and 7.6. Although the cause for these successive transitions is more likely related to the occurrence of apnea events, the classifier is once more not being able to fully cope with fragmented sleep architectures.

Regarding the classification of subjects of the 'all' data set, Figures 7.7 and 7.8 illustrate a comparison between the ground-truth reference hypnogram and the classification output for the subject with the best performance and the performance closest to the median in the 'all' data set. The first remarkable aspect in Fi gure 7.7 is the extremely low sleep efficiency of this subject, who spent a large part of the night awake. The classifier is correctly able to detect this period, and most mistakes occur once more during short periods of N3 and N1 which alternate with N2. Regarding Figure 7.8 we again see that many mistakes occur during fast transitions to and from short periods of N3 although in this particular subject we see a rather low sensitivity to N3, likely related to a weaker expression of some of the hallmark cardiac and respiratory characteristics of N3 seen in most subjects in the training set.

In Chapter 4 we used a linear discriminant similar to the LDtvp classifier used in this chapter and achieved a κ of 0.49, on par with the best classifiers reported in literature at the time [206].

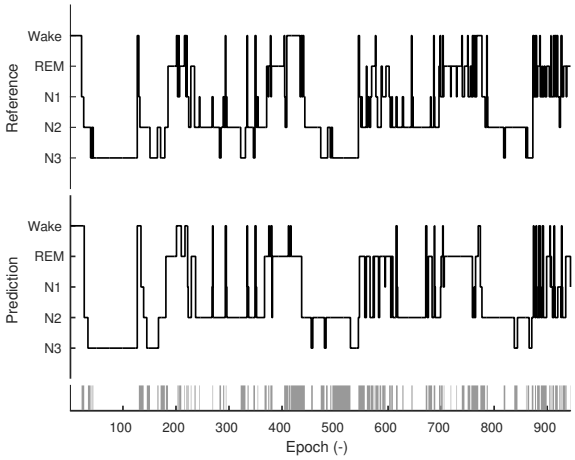


Figure 7.4: Classification example for the subject closest to the median κ of 'regular' subjects (κ of 0.51 and an accuracy of 64.3%). See caption to Figure 7.3.

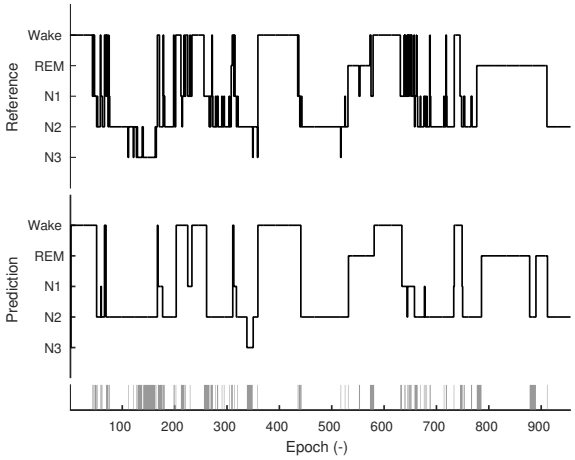


Figure 7.5: Classification example for the subject with the highest κ of 'OSA' subjects (κ of 0.71 and an accuracy of 79.9%). See caption to Figure 7.3.

In that work we used a smaller data set, comprised exclusively of healthy subjects with regular sleep architecture. In comparison with the present work, we are now able to achieve nearly the same performance (κ of 0.48) but for a five-class classification task using the CRFt classifier. In more recent work on four-class sleep stage classification of healthy adults, Beattie et al. [241] used an accelerometer and a PPG sensor to extract movement, breathing and heart rate variability features, achieving a κ of 0.52 and accuracy of 69% on healthy adults. Albeit on a different, larger data set, and when trained exclusively with 'regular' subjects with cardiac and respiratory (but not

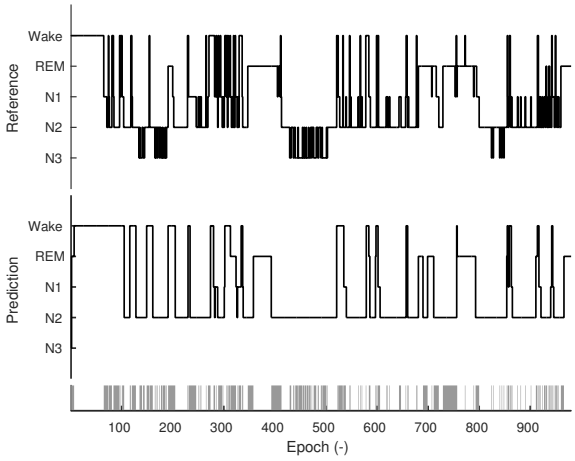


Figure 7.6: Classification example for the subject closest to the median κ of 'OSA' subjects (κ of 0.40 and an accuracy of 61.1%). See caption to Figure 7.3.

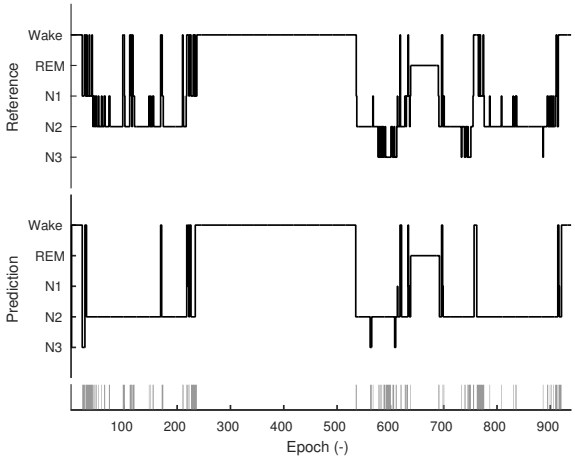


Figure 7.7: Classification example for the subject with the highest κ of 'all' subjects (κ of 0.74 and an accuracy of 83.8%). See caption to Figure 7.3.

movement) features, our CRFt classifier obtained comparable results with an average and median κ of 0.50 and 0.53 and accuracy of 69.6% and 70.8% respectively.

Additionally, this classifier seems better able to handle subjects with irregular sleep architectures and with sleep disorders, achieving consistently better results than LDtvp both with a generally trained classifier evaluated on 'irregular' and 'OSA' subsets, as well as with a dedicated classifier trained only on 'OSA' patients. In this regard, it is interesting to compare the results reported in literature for sleep stage classification of OSA patients. Redmond and Heneghan reported an ac-

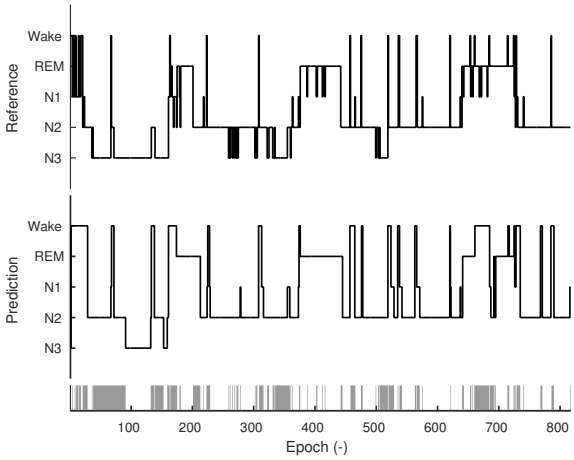


Figure 7.8: Classification example for the subject closest to the median κ of 'all' subjects (κ of 0.44 and an accuracy of 63.4%). See caption to Figure 7.3.

curacy and κ on a three-class task of 67% and 0.32 [98] for 37 recordings, and Willemen et al. reported, also for a three-class task, accuracy and κ of 70% and 0.41 [102] for 25 recordings. In comparison, the CRFt classifier in our work achieved an accuracy and κ of 75.4% and 0.50 respectively, for the same three-class problem, representing a substantial improvement, albeit on a different data set.

Overall, these results suggest that the CRF classifier is better at modeling the transitional characteristics between sleep stages, while offering an improved separation between sleep stages with overlapping cardiorespiratory characteristics. These advantages allow it to not only better learn and predict more complex and irregular sleep architectures, but also to make the first steps towards five-class classification, the standard for sleep scoring used in clinical PSG and, importantly, to achieve first promising results in the sleep stage classification of patients with sleep disorders.

Finally, we should highlight the limitations of this technology. As discussed, the classification performance on OSA patients is higher when using a dedicated classifier, than when using a classifier trained with healthy and disordered subjects. In a real-world application, this would imply knowing, a priori, whether a patient suffers from a sleep disorder (OSA in this case) or not, before the appropriate classifier can be used. This may of course limit the immediate applicability of this technology in a clinical context, where the analysis of sleep architecture might be relevant for the complete understanding of a given condition in the first place. Alternative methods for automatic detection of, in this case, respiratory events, would be necessary in order to obtain a first segmentation of the subjects between those with OSA and those without. Furthermore, OSA is only one of many sleep disorders, and the applicability of these methods to patients suffering from other pathologies remains to be demonstrated.

Regarding the HMM classifier, it clearly underperforms when compared with the LD and the CRF classifiers. A possible explanation for this is related to the limitations imposed by the nature of the HMM classifier used, where the discretization into a limited number of symbols - needed to compute the emission and transition probabilities - possibly limits the learning capabilities in this complex task. Although this problem was not so evident in the results of Chapter 6, where HMM was superior to LD for certain binary sleep classification tasks, this simplification is clearly not appropriate in the more complex problem of multi-class classification. The applicability of continuous HMM models remains to be evaluated.

7.5 Conclusions

The performance of a CRF classifier was evaluated in a home sleep monitoring context with cardiorespiratory features, and compared to HMM and LD classifiers. In Chapter 6 we have shown how CRF can exploit transitional features and better handle non-separable features, but the evaluation was limited to binary classification in 'one class versus the rest' tasks. In this chapter we show that CRF is superior to the other classifiers also in the more useful multiple-stage classification problem. Furthermore, we explored the advantage of incorporating time as a feature for classification.

A total of six classifiers (CRF, HMM and LD, with and without time information) were evaluated on five-class (W/N1/N2/N3/R), four-class (W/N1-2/N3/R), and three-class (W/NR/R) sleep stage classification tasks using cross-validation on 443 night-time PSG records from 231 subjects. The CRF classifier with normalized time (CRFt) was found to outperform all other classifiers for all classification tasks, and was found to yield more plausible outputs than the more popular LDtvp.

The results also demonstrated that five-class classification is feasible, and that dedicated three- and four-class classifiers do not outperform a five-class classifier. Finally, it was shown that CRF classifiers also achieve moderate performance in the sleep stage classification of OSA patients, outperforming earlier work reported in literature for this population.

More research is needed to improve the sensitivity to N1 stage, and to evaluate this technology in subjects suffering from other sleep disorders.

Table 7.6: Mean (SD), and median κ and accuracy after sleep staging of 'regular' subjects (N=134)

	W/NR/R		W/N1-2/N3/R		W/N1/N2/N3/R	
	κ (-)	accuracy (%)	κ (-)	accuracy (%)	κ (-)	accuracy (%)
LD	0.55 (0.14), 0.56	78.91 (6.83), 79.95	0.44 (0.11), 0.45	62.66 (7.25), 63.25	0.39 (0.09), 0.40	53.33 (7.49), 54.54
LDtvp	0.57 (0.14), 0.59	80.52 (6.48), 81.32	0.49 (0.12), 0.50	68.41 (7.14), 69.99	0.45 (0.11), 0.47	63.61 (7.79), 64.95
HMM	0.40 (0.15), 0.40	68.76 (8.35), 70.35	0.31 (0.12), 0.32	50.75 (8.76), 51.47	0.29 (0.11), 0.31	46.98 (8.56), 47.95
HMMt	0.41 (0.14), 0.41	69.29 (7.96), 70.57	0.33 (0.11), 0.34	52.17 (8.05), 53.08	0.31 (0.10), 0.32	48.38 (8.12), 48.55
CRF	0.58 (0.15), 0.61	80.54 (6.49), 81.64	0.49 (0.14), 0.51	68.76 (8.69), 69.46	0.47 (0.13), 0.49	64.66 (8.86), 65.66
CRFt	**0.59 (0.15), 0.62***	**80.67 (6.93), 81.29**†	**0.50 (0.15), 0.53***	**69.59 (9.03), 70.79***	**0.48 (0.13), 0.51***	**65.66 (9.20), 67.02***

* CRFt performs significantly better than all others ($p < .001$). † CRFt performs significantly better than LD, HMM, and HMMt ($p < .001$), but not LDtvp ($p = .37$) nor CRF ($p = .11$).

Table 7.7: Mean (SD), and median κ and accuracy after sleep staging of 'OSA' subjects (N=102)

	W/NR/R		W/N1-2/N3/R		W/N1/N2/N3/R	
	κ (-)	accuracy (%)	κ (-)	accuracy (%)	κ (-)	accuracy (%)
LD	0.44 (0.12), 0.45	71.86 (6.85), 72.95	0.38 (0.10), 0.38	60.72 (7.30), 61.71	0.32 (0.09), 0.33	47.91 (6.88), 48.15
LDtvp	0.46 (0.12), 0.48	74.80 (7.07), 76.26	0.42 (0.11), 0.43	67.98 (8.05), 69.90	0.36 (0.10), 0.37	57.27 (9.30), 57.62
HMM	0.21 (0.17), 0.20	58.88 (12.08), 61.06	0.19 (0.12), 0.20	46.28 (11.07), 46.63	0.17 (0.11), 0.17	38.27 (10.26), 38.53
HMMt	0.21 (0.16), 0.22	58.72 (11.25), 60.04	0.19 (0.12), 0.20	46.71 (10.69), 47.12	0.17 (0.11), 0.17	38.98 (10.38), 39.99
CRF	0.49 (0.13), 0.50	**75.52 (6.70), 76.05**$^{\dagger 1}$	0.44 (0.13), 0.44	**68.85 (8.38), 69.66**$^{\dagger 2}$	0.39 (0.12), 0.40	59.26 (10.27), 60.19
CRFt	**0.50, (0.13), 0.51**$^{\ddagger 1}$	75.40 (6.83), 75.77	**0.44 (0.13), 0.45**$^{\ddagger 2}$	68.67 (8.31), 69.67	**0.40 (0.12), 0.40**$^{\ddagger 3}$	**59.30 (10.20), 59.92**$^{\ddagger 4}$

$^{\dagger 1,2}$ CRF performs significantly better than LD, HMM and HMMt ($p < .001$), but not LDtvp ($^{\dagger 1}p = .15$, $^{\dagger 2}p = .17$) nor CRFt ($^{\dagger 1}p = .61$, $^{\dagger 2}p = .64$).
$^{\ddagger 1,2,3,4}$ CRFt performs significantly better than LD, HMM and HMMt ($p < .001$), LDtvp ($^{\ddagger 1,3,4}p < .001$, $^{\ddagger 2}p < .01$), but not always CRF ($^{\ddagger 1}p = .28$, $^{\ddagger 2}p = .32$, $^{\ddagger 3}p < .05$, $^{\ddagger 4}p = .55$).

Table 7.8: Mean (SD), and median κ and accuracy after sleep staging of 'all' subjects (N=443)

	W/NR/R		W/N1-2/N3/R		W/N1/N2/N3/R	
	κ (-)	accuracy (%)	κ (-)	accuracy (%)	κ (-)	accuracy (%)
LD	0.49 (0.14), 0.50	74.00 (8.10), 75.46	0.39 (0.12), 0.40	59.16 (7.80), 59.98	0.34 (0.10), 0.35	48.76 (8.08), 49.59
LDtvp	0.51 (0.14), 0.53	76.46 (7.89), 78.07	0.44 (0.13), 0.45	66.67 (8.46), 67.78	0.40 (0.12), 0.42	59.83 (9.62), 61.37
HMM	0.32 (0.16), 0.34	59.63 (11.96), 61.72	0.26 (0.13), 0.26	46.05 (10.32), 46.74	0.25 (0.12), 0.26	42.92 (10.68), 43.46
HMMt	0.32 (0.16), 0.34	59.64 (11.95), 61.70	0.25 (0.13), 0.26	45.18 (10.03), 45.73	0.25 (0.12), 0.25	42.10 (10.41), 42.62
CRF	0.53 (0.14), 0.54	76.31 (7.39), 76.90	0.44 (0.14), 0.45	66.54 (8.97), 67.37	0.42 (0.13), 0.43	61.10 (9.77), 62.18
CRFt	**0.54 (0.14), 0.55**^{*1}	**76.70 (7.43), 77.62**†	**0.46 (0.14), 0.47**^{*1}	**67.36 (8.93), 68.77**^{*2}	**0.43 (0.13), 0.44**^{*1}	**61.83 (9.86), 62.80**^{*1}
subset 'regular' (N=134)						
LDtvp	0.55 (0.14), 0.57	**78.94 (6.06), 79.50**$^{\S 1}$	**0.46 (0.12), 0.49**$^{\S 2}$	**67.11 (7.48), 67.76**$^{\S 3}$	0.44 (0.11), 0.46	62.65 (7.75), 63.47
CRFt	**0.56 (0.13), 0.58**$^{\S 4}$	78.27 (6.60), 79.24	0.46 (0.14), 0.47	66.51 (8.97), 68.26	**0.45 (0.12), 0.46**$^{\S 5}$	**63.11 (8.94), 63.82**$^{\S 6}$
subset 'irregular' (N=207)						
LDtvp	0.50 (0.15), 0.51	75.39 (8.35), 76.46	0.43 (0.13), 0.44	66.47 (8.85), 67.78	0.38 (0.12), 0.39	58.61 (10.10), 59.45
CRFt	**0.53 (0.15), 0.54**$^{\|\|1}$	**76.02 (7.67), 77.08**$^{\|\|2}$	**0.45 (0.14), 0.47**$^{\|\|1}$	**67.73 (8.91), 69.04**$^{\|\|1}$	**0.42 (0.13), 0.43**$^{\|\|1}$	**61.27 (10.20), 62.04**$^{\|\|1}$
subset 'OSA' (N=102)						
LDtvp	0.46 (0.14), 0.47	**74.08 (8.00), 75.50**$^{\P 1}$	0.41 (0.12), 0.43	66.51 (8.42), 68.41	0.35 (0.12), 0.35	56.43 (10.47), 57.53
CRFt	**0.46 (0.14), 0.48**$^{\P 2}$	73.62 (6.92), 74.49	**0.43 (0.13), 0.44**$^{\P 3}$	**67.18 (8.07), 68.96**$^{\P 4}$	**0.38 (0.13), 0.39**$^{\#}$	**58.52 (10.51), 59.51**$^{\#}$

^{*1,2} CRFt performs significantly better than LD, HMM, HMMt and CRF ($p < .001$) and LDtvp ($^{*1}p < .001$, $^{*2}p < .01$).
† CRFt performs significantly better than LD, HMM, HMMt and CRF ($p < .001$) but not LDtvp ($p = .45$).
For the 'regular' subset, there is no significant difference in performance between LDtvp and CRFt ($^{\S 1}p = .08$, $^{\S 2}p = .22$, $^{\S 3}p = .06$, $^{\S 4}p = .06$, $^{\S 5}p = .14$, $^{\S 6}p = .40$).
For the 'irregular' subset, CRFt performs significantly better than LDtvp ($^{||1}p < .001$, $^{||2}p < .05$).
For the 'OSA' subset, CRFt performs significantly better than LDtvp for five classes ($^{\#}p < .001$) but there isn't always a significant difference between CRFt and LDtvp for the remaining tasks ($^{\P 1}p = .14$, $^{\P 2}p = .47$, $^{\P 3}p < .05$, $^{\P 4}p = .23$).

Part 5

Discussion

Dormir, ser longínquo sem o saber, estar deitado, esquecer com o próprio corpo; ter a liberdade de ser inconsciente, um refúgio do lago esquecido, estagnado entre frondes árvores, nos vastos afastamentos das florestas.

To sleep, to be far away, remote without knowing it, to forget with one's very body, to have the freedom of unconsciousness like a refuge on a forgotten lake, stagnating among thick foliage in the hidden depths of forests.

- Bernardo Soares (heteronym), Livro do Desassossego (*transl.* Richard Zenith, The Book of Disquiet)

Chapter 8

General discussion

This thesis set out with three related objectives. The first was to develop methods to increase the performance of automatic cardiorespiratory-based sleep stage classification and bring it closer, for healthy subjects, to the gold standard polysomnography (PSG). Logically organized in a traditional machine learning approach, the thesis started with a contribution to the improvement of physiological signal processing, with **Chapter 2** describing a method to improve the precision in QRS localization. This is an essential step in electrocardiography (ECG) signal pre-processing, and eventual heart rate variability (HRV) feature extraction. Continuing with the step of feature extraction, **Chapter 3** described a method which allows body movements to be estimated from artifacts in ECG and respiratory inductance plethysmography (RIP) signals. This method alleviates the need to wear an additional actigraphy sensor in home registrations and, perhaps more importantly in the scope of the thesis, it allowed retrospective data sets for which actigraphy was not registered to be used in subsequent chapters. Continuing on with the steps of feature selection and classification, **Chapter 4** described an algorithm which used HRV and respiration rate variability (RRV) features extracted from ECG and RIP to automatically perform four-class sleep stage classification (Wake/N1 and N2/N3/REM) of healthy subjects. Importantly, this chapter showed that automatic sleep stage classification was feasible, with reasonable performance, without resorting to central nervous system (CNS) characteristics traditionally used to score sleep. Further exploring classification approaches to this problem, **Chapter 6** described how Markov chains in general, and conditional random fields (CRF) in particular, could exploit the structured nature of sleep and use features discriminative not only of different sleep stages, but also of stage transitions and sleep stage continuity, achieving reasonable performance not only for the common four-class task mostly described in literature, but also in the more complex task of five-class classification (Wake/N1/N2/N3/REM), the standard for sleep scoring in PSG.

The second objective of the thesis was to prove that these methods could work with sensors

that are less obtrusive for long-term home sleep monitoring. **Chapter 5** validated a sleep stage classifier which used HRV features derived from a photoplethysmography (PPG) sensor, common in modern consumer sleep tracker (CST) devices. The classifier achieved a moderate performance for four-class classification of healthy middle-aged adults and showed better performance than actigraphy for Sleep/Wake classification in this target group.

Finally, the third and last objective of the thesis was to evaluate whether these methods could be applied in subjects suffering from a sleep disorder. This would provide first evidence of whether this technology would have the potential to ever be used in clinical practice. **Chapter 7** explored the performance of CRF using HRV and RRV features on patients diagnosed with obstructive sleep apnea, achieving a reasonable performance for four-class classification in this group.

Having been the object of study for decades, it is only in recent years that unobtrusive cardiorespiratory sleep staging is starting to reach reasonable performance, highlighting an increasing maturity of not only of cardiac and respiratory sensors, but also of the methodology used to interpret changes in autonomic characteristics during sleep. Although this technology holds the promise of complementing PSG for long-term home monitoring and perhaps even replacing other tools already used in clinical practice such as actigraphy, there are important limitations in the state of the art which are relevant to highlight.

The first is related to the relatively small size of the data sets used in the most promising work reported in literature. In their seminal work in 2006 and 2007, Redmond and colleagues reported their findings on data sets comprising 37 apneic [98] and 31 healthy subjects [80]. Later, Mendez et al. [172] and Kortelainen et al. [93] reported improvements in performance, but again in limited data sets of 24 and 18 healthy subjects, respectively. In 2013, Willemen et al. [101] used a larger data set of 85 healthy subjects, reporting further improvements, followed in 2014 by Domingues et al. [99] but again with a smaller data set of 20 subjects. Finally, Willemen et al. [102] reported their findings on a data set of 25 apneic subjects. Although the performance of sleep stage classification has steadily increased and not yet showing signs of plateauing, questions must be raised regarding whether the described techniques generalize well to new data. As shown throughout the thesis, and in particular, in later chapters, the performance of the algorithms varies substantially between different subjects, depending for example on whether they have a 'regular' sleep architecture, or whether they deviate, even in the absence of sleep disorders, from population averages. By repeatedly reporting new techniques on small data sets of just a few dozen subjects, it remains difficult to assess whether they will ever have the potential to be used across a wider population.

The second limitation, and arguably the biggest obstacle to a faster adoption of this technology in clinical practice, remains the relative lack of validation studies on clinical populations suffering from different disorders. With the notable exception of a few articles [98, 102], nearly no publications can be found on the subject. It is unclear whether this is related to a negative bias in literature. Articles reporting poor results will have a lower chance of being accepted by editors

and peer reviewers of notable journals. Alternatively, it could simply reflect the lack of research in this area, perhaps due to the increased difficulty in acquiring large data sets with sleep disordered patients. In comparison, actigraphy has deserved substantial research attention over decades now, delivering a large body of evidence which eventually lead not only to its endorsement by associations such as the American Academy of Sleep Medicine (AASM), but perhaps equally important, allowing practitioners to understand its limitations and use that knowledge when interpreting its results.

Although it could be argued that the performance of cardiorespiratory sleep staging has now reached a maturity and level of evidence which is sufficient for healthy consumers looking to learn more about their own sleep, caution should be exercised by manufacturers of CST devices offering this technology to unwary consumers. And this highlights a third limitation of this technology. Giving otherwise healthy subjects negative - but non-actionable - information about their sleep might cause stress and lead to actual sleep problems. On the other hand, giving positive feedback about sleep while missing out on clinically relevant conditions might give a false sense of reassurance and detract users from actually visiting a doctor. This could lead to a serious adverse consequence and even a liability case to a manufacturer of CST devices: a sleep apnea patient who did not visit a doctor because his device gave him positive feedback about his sleep could dismiss his excessive daytime sleepiness as a consequence of other factors such as excessive stress at work, with potentially dangerous consequences ensuing.

The next sections will analyze and discuss some important results and insights regarding the work reported in this thesis, and suggest aspects which should be further investigated to increase the performance and maturity of cardiorespiratory sleep staging, and its extension to clinically relevant populations.

8.1 Feature analysis

A substantial portion of the work reported in this thesis, in particular in later chapters, focused on the use of advanced probabilistic methods to exploit the structured nature of sleep and thereby increase classification performance. How adequate these methods proved to be, they eventually depended on the power of the available set of cardiac and respiratory features to express changes in autonomic activity related to different stages of sleep.

Table 8.1 gives an overview of the three features with the highest average absolute standardized mean distance (ASMD) per sleep stage across all healthy subjects in our data set. In addition, and in order to understand whether the discriminative power of each of these features is affected by the characteristics of different subjects, the correlation between the ASMD per subject and the sleep statistics for that subject (based on ground-truth annotations) were computed. The three statistics with the highest correlation are indicated in the table, together with the level of statistical significance after a Wilcoxon signed-rank test. The first apparent observation is that only for Wake and N3 do we find features with an ASMD above 1. As explained in Chapter 4, under the

assumption of normal feature distributions, an ASMD of 1 corresponds to a 69% probability of correctly identifying a class based on the feature scores alone. However, that is not the case for N1, N2 and REM. For N1 and N2, the ASMD is between 0.5 and 0.6 (probability of correctly identifying a class between 60% and 62% based on feature scores) and for REM, around 0.75 (around 64%). This suggests that some classes are more easily separable than others based on cardiorespiratory features alone and that more features are needed to provide adequate separation between them.

Another observation is that (surrogate) actigraphy is the single most discriminative feature for Wake classification, followed by a respiratory feature (dynamic time warping (DTW) distance), and a statistic (90th percentile) of the mean heart rate. Noting that actigraphy is actually not available in the data set used in the final chapters, this highlights the importance of the surrogate actigraphy estimation technique proposed and described in Chapter 3. Furthermore, it confirms the intuitive observation that, in the studied healthy population, Wake mostly coincides with periods of more intense (or more prolonged) movement. This is not always the case, however, as the significant positive correlations between the ASMD of actigraphy for Wake with sleep efficiency (SE) (0.58) and total sleep time (TST) (0.50) show: the longer the subjects remain awake (and the lower their sleep efficiency is), the lower the ASMD becomes. These are the highest amongst the correlations between the ASMD of any feature and all sleep statistics, and they are easy to interpret: when subjects are awake for longer periods of time, they might lie still, waiting to fall sleep. In this case actigraphy will yield low values, but the subjects are still awake. Interestingly, the ASMD of the 90th percentile of the mean heart rate feature correlates less with these sleep statistics (never higher than 0.23) and given its low correlation with actigraphy (Pearson's $\rho = 0.30$ with $p < .001$) it captures complementary information that improves Wake classification.

Finally, it is also interesting to notice the difference in modalities in the features found to be the most discriminative for each stage. While for Wake the three modalities are used, N3 sees the most discriminative features being derived from RIP while rapid eye movement (REM), N1 and N2, from interbeat intervals (IBIs) computed from ECG. This suggests that the two modalities are not only complementary, but important to distinguish different sleep stages accurately.

To have a simple visual impression of the collinearity between the many HRV and RRV features used in this thesis, we calculated the absolute Pearson's correlation between each pair of features extracted for all subjects in the data set described in Chapter 7 (Figure 8.1). Although many features are highly correlated, it is clear that there many uncorrelated features, suggesting that they offer complementary information for these classification tasks.

While no single feature alone is discriminative of all sleep stages, nor individually discriminative of any single sleep stage, it is clear that a combination of features provides the necessary information to separate the classes. However, this highlights one of the issues with cardiorespiratory-based sleep analysis: while with traditional PSG a set of well-defined rules based on a small set of channels (electroencephalography (EEG), electrooculography (EOG) and electromyography (EMG)) allows humans to visually score sleep stages, a large number of cardiorespiratory features would

Table 8.1: Correlation between the most discriminative cardiorespiratory features per class and sleep statistics

Mod.	Variable	Class	ASMD	N3T	N3Ti	N3TpC	N3pC	N3pCsd	NC	REMTi	REMpC	SE	SOL	TRT	TST	WASO	WKNi
A	Surrogate activity counts from resp.	W	1.22 ± 0.59								0.28^{***}	0.58^{***}			0.50^{***}		
R	Constrained DTW distance	W	1.08 ± 0.64								0.24^{***}	0.51^{***}			0.43^{***}		
C	90th percentile of the mean heart rate	W	1.01 ± 0.51				0.17^{**}					0.23^{***}			0.20^{***}		
C	Windowed detrended fluctuation analysis (DFA)	N1	0.56 ± 0.29						0.27^{***}			0.43^{***}			0.38^{***}		
C	Log-SD of the normalized IBIs	N1	0.53 ± 0.31					0.25^{***}				0.38^{***}			0.32^{***}		
C	Mean absolute deviation of the detrended IBIs	N1	0.52 ± 0.33								0.28^{***}	0.37^{***}			0.29^{***}		
C	Scaling exponent using DFA for all time scales	N2	0.61 ± 0.29		0.39^{***}											0.32^{***}	0.21^{***}
C	Mean of the normalized IBIs	N2	0.61 ± 0.31										0.035	0.039		0.15^{**}	
C	25th percentile of the IBIs	N2	0.61 ± 0.29										0.053	0.063		0.16^{**}	
R	Standardized median trough value	N3	1.08 ± 0.57		0.059											0.14^{*}	0.16^{**}
R	Standardized mean trough value	N3	1.06 ± 0.58		0.076											0.14^{**}	0.16^{**}
R	Standardized median peak value	N3	1.03 ± 0.57										0.057			0.18^{***}	0.17^{**}
C	Approximate entropy of sign changes of IBIs	R	0.77 ± 0.42				0.13^{*}	0.17^{**}				0.16^{**}					
C	Scaling exponent of DFA for all time scales	R	0.76 ± 0.44							0.21^{***}						0.17^{**}	0.16^{**}
C	Sample entropy of detrended IBIs (length 2, scale 1)	R	0.75 ± 0.46	0.14^{**}		0.15^{**}				0.16^{**}							

$^{***}p < .001$, $^{**}p < .01$, $^{*}p < .05$

Modality: A: actigraphy, R: respiratory, C: cardiac

N3T/N3Ti/N3TpC: number of transitions from N3 - absolute/per hour of N3/per sleep cycle, N3pC/N3pCsd: mean/standard deviation of N3 duration per sleep cycle, NC: number of sleep cycles, REMTi: number of transitions from REM per hour of REM, REMpC: mean REM duration per sleep cycle, SE: sleep efficiency, SOL: sleep onset latency, TRT: total recording time, TST: total sleep time, WASO: wake after sleep onset, WKNi: number of transitions to Wake per hour of sleep

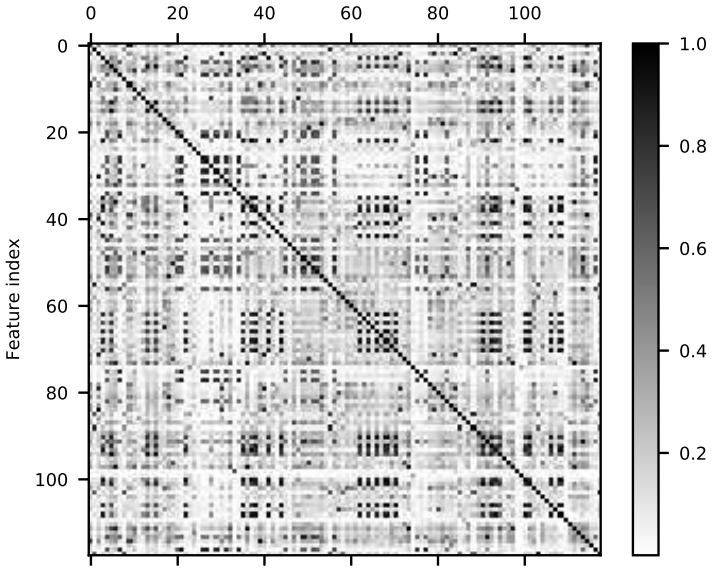


Figure 8.1: Absolute Pearson's correlation between each pair of cardiorespiratory features used in Chapter 7

be needed to allow such a process. Furthermore, the fact that most classes are not separable with single features and the high degree of between-subject variability in these characteristics makes it nearly impossible to define rules for visual scoring.

Regardless, the reasonable level of agreement obtained between automatic sleep staging with cardiorespiratory characteristics and human annotations suggests a hidden process, not directly observable, of which cardiorespiratory features only partially reflect or express. This is supported by the improved performance when Markov chains are used instead of single epoch classification, as shown in Chapter 6 and Chapter 7.

8.2 Subject characteristics and sleep disorders

Although in Chapter 7 we showed how probabilistic graphical models also perform reasonably well in patients suffering from obstructive sleep apnea (OSA), it must be emphasized that most research described in this thesis was focused on healthy, normal sleep, in particular in the development of features describing characteristics of the autonomic nervous system (ANS) of healthy adults. However, it is clear that the macro-structure of sleep already differs substantially between healthy individuals, depending, amongst other factors, on their age [225]. Figure 8.2 illustrates

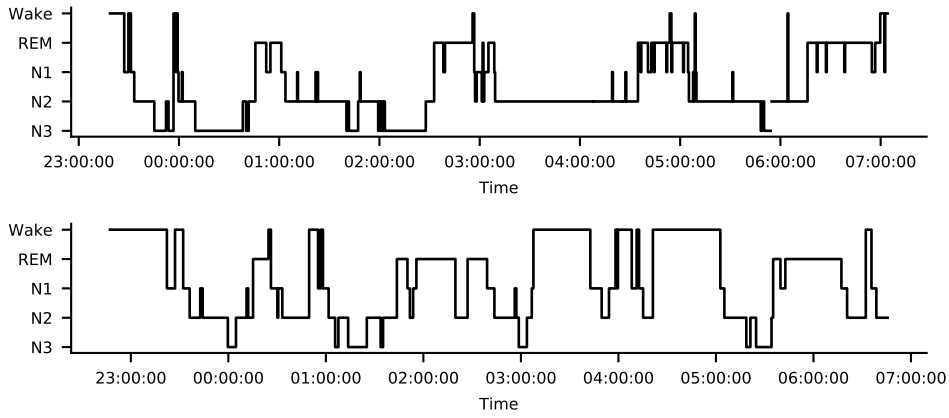


Figure 8.2: Hypnogram of the youngest (above) and oldest (below) healthy subjects in our data set

the hypnograms of the youngest (20 years old) and the oldest (95 years old) healthy subjects in our data sets. As expected, the oldest subject exhibits less slow wave sleep (SWS), a higher number of awakenings, and a longer wake after sleep onset (WASO).

Although the impact of different sleep characteristics in the performance of the classifier was already clear from the results in Chapter 6 and Chapter 7 where higher classification performance was achieved for recordings of regular sleep architectures than for irregular ones, it is interesting to further understand how sleep characteristics might limit the performance of the models developed in this thesis. To this end, we calculated a number of sleep statistics for the recordings of healthy subjects used to validate the five-class classifier of Chapter 7. These statistics were all computed based on the ground-truth hypnogram used as reference in that work, annotated in consensus scoring by independent sleep technicians. Besides the traditional sleep statistics described in most work, we calculated additional statistics describing the number and characteristics of sleep cycles and the number of transitions from N3 and REM. Table 8.2 indicates the Pearson's correlation between the performance (κ and accuracy), and subject demographics (age, sex and BMI) and each sleep statistic.

The five statistics most correlated with κ were age, number of awakenings per hour of sleep (WKNi), number of transitions from REM per hour of REM (REMTi), percentage of N3 and finally, the sex of the participant. Although all individual statistics were weakly correlated, it appears that besides the two demographics (age, sex), two statistics are related to some measure of sleep fragmentation (number of awakenings, and number of transitions from REM) and the last one is related to the macro-structure of sleep. To better understand the interplay between these sleep characteristics and the classifier performance, we performed a stepwise ordinary least squares (OLS) regression with forward selection, using as selection criteria the R-squared of the model.

Table 8.2: Correlation between sleep statistics and five-class staging performance κ and accuracy

Variable	ρ_κ	$\rho_{accuracy}$	Variable	ρ_κ	$\rho_{accuracy}$
Age	-0.232^{***}	-0.211^{***}	TRT	0.114^{*}	0.132^{*}
Sex	0.163^{**}	0.143^{**}	NC	0.066	0.087
BMI	−0.065	−0.060	N3pC	-0.147^{**}	-0.246^{***}
SOL	0.020	0.014	N3pCsd	−0.007	−0.067
WASO	−0.086	-0.145^{**}	N3T	−0.033	0.002
WKNi	-0.218^{***}	-0.252^{***}	N3Ti	0.023	0.088
N3%	-0.165^{**}	-0.284^{***}	N3TpC	−0.040	−0.021
REM%	0.075	0.014	REMpC	0.088	0.048
SE	0.084	0.146^{**}	REMpCsd	0.066	0.051
TST	0.121^{*}	0.189^{***}	REMT	−0.082	−0.078
WUT	0.102	0.131^{*}	REMTi	-0.194^{***}	-0.157^{**}
ST	0.007	−0.026	REMTpC	-0.113^{*}	-0.118^{*}

$^{***}p < .001$, $^{**}p < .01$, $^{*}p < .05$

BMI: Body max index, SOL: sleep onset latency, WASO: wake after sleep onset, WKNi: number of awakenings per hour of sleep, SE: sleep efficiency, TST: total sleep time, WUT: wake-up time, ST: snooze time, TRT: total recording time, NC: number of sleep cycles, {N3,REM}pC: average duration of {N3,REM} per sleep cycle, {N3,REM}pCsd: standard deviation of {N3,REM} duration per sleep cycle, {N3,REM}T: number of transitions from {N3,REM}, {N3,REM}Ti: number of transitions from {N3,REM} per hour of {N3,REM}, {N3,REM}TpC: average number of transitions from {N3,REM} per sleep cycle.

Table 8.3 indicates the variables selected for the final OLS model, together with their regression coefficients and the standard error. It should be noted that all variables were normalized (with zero-mean and unit standard deviation) before fitting the model and the loading coefficients reflect this. Figure 8.3 illustrates a scatter plot of the OLS model predictions for κ and accuracy.

Although the model only explains about 25% and 36% of the variance of κ and accuracy, respectively, these results suggest that these characteristics are likely having a limiting effect on the performance of the classifier.

In view of this, future research in this area should focus on developing features or classification models that are more robust to differences in subject and sleep characteristics. An age-specific classifier would probably lead to overall better results. Alternatively, different normalization schemes could help mitigate the effect of age and allow for a universal classifier which performs equally well for all age groups. Furthermore, is is clear that fragmentation is deteriorating the performance. This is perhaps unsurprising, considering that most features use windows of several minutes and are therefore better suited to capture longer, more stable states. In the presence of quickly changing sleep stages they might be losing some of their discriminative potential. Besides developing new features which are better able to handle these aspects, another approach would be to analyze, with finer granularity, the micro-structure of sleep to detect events such as arousals or sleep stage transitions to better handle this aspect.

When sleep disorders come into the picture these differences are further exacerbated. One of the hallmarks of insomnia, for instance, is an extended wake time and shortened TST [246], sometimes accompanied, in a sub-group of patients, by a decrease in SWS [246, 247, 248]. Figure 8.4 illustrates the hypnograms of the subject with the longest sleep onset latency and the subject with

Table 8.3: Coefficient (standard error) for all variables selected with stepwise regression for κ and accuracy

Feature name	κ	Accuracy
Intercept	0.445 (0.006)***	0.629 (0.004)***
Age	−0.036 (0.007)***	−0.027 (0.005)***
N3T	−0.024 (0.018)	−0.015 (0.013)
N3Ti		−0.006 (0.006)
N3TpC	0.035 (0.020)	0.030 (0.015)*
N3%	−0.001 (0.024)	−0.016 (0.023)
N3pC	−0.066 (0.027)*	−0.053 (0.024)*
N3pCsd	0.023 (0.010)*	0.016 (0.007)*
NC		0.021 (0.015)
REM%	−0.012 (0.008)	−0.044 (0.015)**
REMTi	−0.023 (0.007)**	−0.017 (0.005)***
REMpC		0.026 (0.015)
Sex (Female)	0.024 (0.006)***	0.019 (0.004)***
TRT	0.011 (0.007)	
TST		0.001 (0.012)
WASO	0.020 (0.010)	0.018 (0.007)*
WKNi	−0.033 (0.009)***	−0.028 (0.006)***
R-squared	0.249	0.355
Nr. observations	340	340

$^{***}p < .001$, $^{**}p < .01$, $^{*}p < .05$
See acronyms defined in Table 8.2.

the highest WASO from the insomnia data set used in Chapter 3.

OSA patients, on the other hand, have an increased sleep fragmentation caused by the arousals which often follow apnea and hypopnea events [243], and can have a decreased SWS [244, 245] and, especially in more severe cases, decreased REM duration [245]. Figure 8.5 illustrates the hypnograms of the subject with the apnea-hypopnea index (AHI) closest to 15, and the subject with the highest AHI. The sleep fragmentation is clearly visible for both subjects, but especially for the subject with highest OSA severity, with a very large number of transitions to N1 (probably following apnea-induced arousals), reduced time in N3 and in REM. It should be noted that this picture of high sleep fragmentation is not exclusive of OSA. Patients suffering from paradoxical insomnia with sleep misperception have a higher arousal index [249], as do patients suffering from periodic limb movement disorder (PLMD) [250]. Narcoleptics, on the other hand, are also known to have a higher number of awakenings, and higher REM sleep fragmentation [251, 252].

In addition to the changes in sleep architecture which, as we saw, may hamper the performance of sleep staging, sleep disorders are also known to have an impact on cardiorespiratory activity and ANS regulation. Night-time HRV is known to be different in some insomnia sub-types [107] with an increase in the LF component, and a decrease of the HF component [108], reflecting a constant sympathetic overactivity [253] and reduction of vagal modulation [108, 254]. Apnea events in OSA, on the other hand, are accompanied by bursts in sympathetic activity [27, 107], with a characteristic change in heart rate and HRV [107]. Furthermore, the repetitive occurrence of apnea events has been linked with a global change in autonomic balance, with a shift towards sympa-

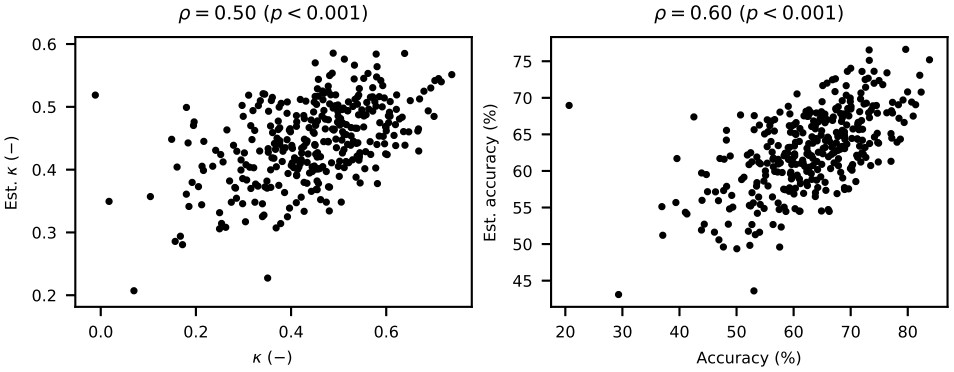


Figure 8.3: Scatter plot of (left) estimated κ and (right) accuracy based on demographics and sleep statistics

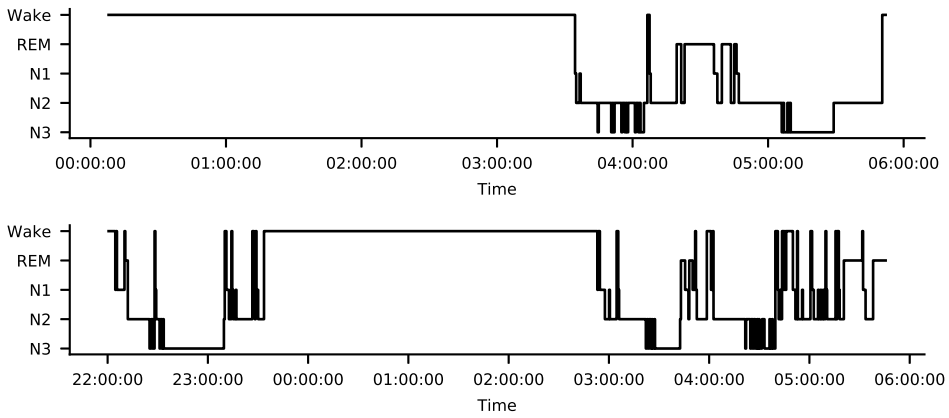


Figure 8.4: Hypnograms for the subject with the longest sleep onset latency (SOL) (above) and the highest WASO (below) from the insomnia data set used in Chapter 3

thetic dominance and blunted parasympathetic control [107, 255]. Clearly, any cardiorespiratory features and classification models developed for, and trained with the characteristics of healthy individuals, are likely to under-perform in the presence of a sleep disorder.

Furthermore, and as shown in Chapter 7, the performance of sleep staging on OSA patients is lower than that for healthy subjects with regular sleep. This is likely due to the nature of the features used for classification, which are clearly better suited to model the ANS characteristics of sleep in the absence of events that disrupt its normal progression. As Figure 8.6 (left) illustrates, the performance of the classifier described in Chapter 7, when trained with examples of healthy subjects and OSA patients is significantly negatively (albeit weakly) correlated with AHI. With a dedicated classifier to either healthy or OSA, on the other hand, this association disappears

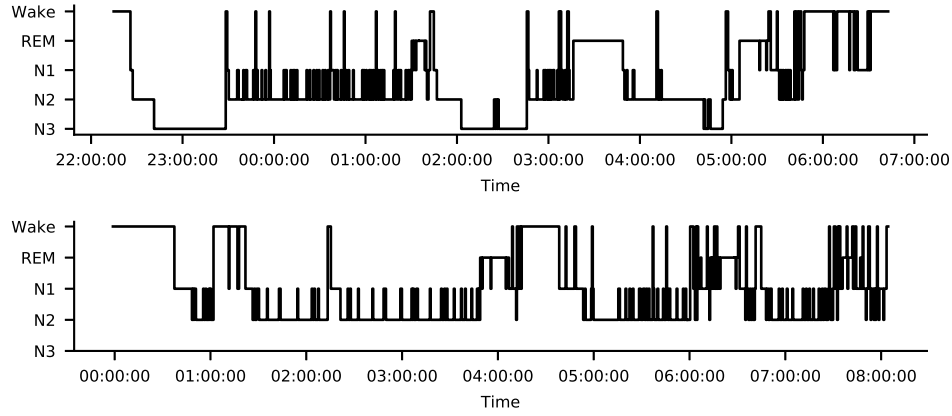


Figure 8.5: Hypnograms for the subject with AHI closest to 15, and the subject with the highest AHI in the OSA data set used in Chapter 7

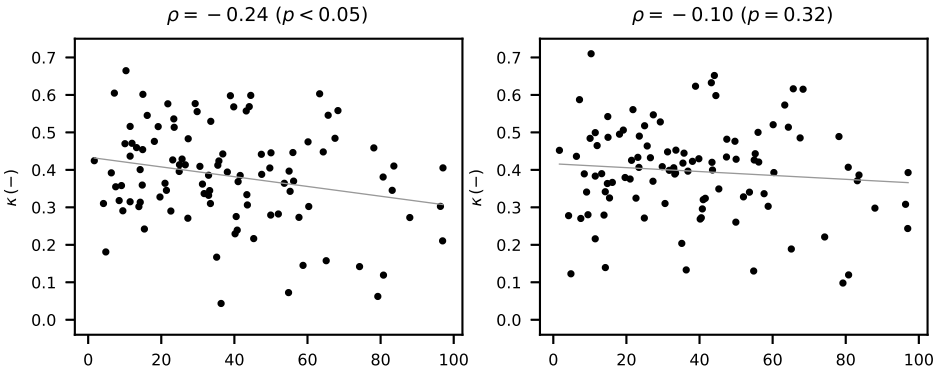


Figure 8.6: Performance κ versus AHI for (left) classifier trained with all subjects, and (right) dedicated classifier separately trained for each subject group

(Figure 8.6, right).

One of the possible reasons for the lower performance on OSA patients is likely related to the factors indicated above, such as the global change in ANS characteristics, the increased frequency of arousals and associated bursts of sympathetic activity and consequent changes in HRV, and increased sleep fragmentation. As explained in Chapter 6, most features are extracted in windows of at least a few minutes and may simply not be able to handle these successive, abrupt changes in HRV and sleep architecture. It remains to be studied whether the addition of dedicated features, possibly designed to detect the presence of these events and transitions, or calculated with shorter windows, help improve performance.

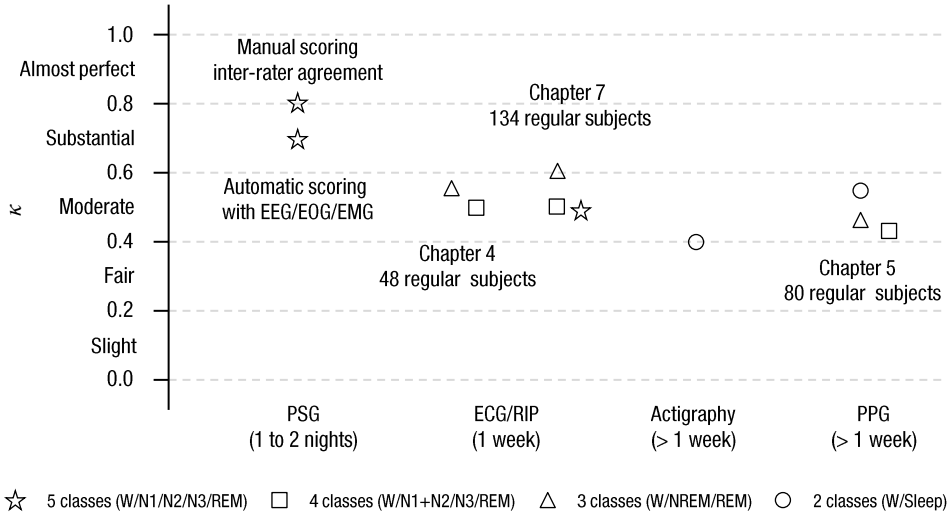


Figure 8.7: Classification performance (in κ) for each modality

8.3 Sensing modalities

Albeit not yet on par with the performance of automatic sleep scoring based on PSG signals such as EEG, or EMG [256], cardiorespiratory sleep monitoring offers the promise of long-term, unassisted sleep monitoring at home, with the potential to complement in-lab PSG for the assessment of sleep. Although recent progress in this area suggests that this might become a reality in the near future, it is also clear that any physiological sensors used in such systems must also be compatible with such a long-term monitoring practice, because they must be set up and used without the involvement of medical professionals, and must also be comfortable enough to be used for consecutive nights without disrupting sleep. Although this thesis focused on the use of physiological characteristics such as HRV or respiratory activity which are, in principle, sensor-agnostic, the algorithms described in different chapters were evaluated with different sensor modalities, for different classification tasks. Figure 8.7 illustrates the performance obtained for the healthy subjects in each validation set, as described in the different chapters of this thesis, together with the sensors used and the number of evaluated subjects. One of the most striking observations is that there is a inversely proportional relation between how unobtrusive - and therefore how adequate for long-term monitoring - a sensor is, and the sleep stage classification performance it obtains.

Notwithstanding that the performance of cardiorespiratory sleep staging is still lower than human scoring, or even automatic scoring using PSG, it is clear that the best performance is still obtained when ECG and RIP are used to compute cardiorespiratory features. However, these sensing modalities are still relatively obtrusive, and although they can, arguably, be used for longer

time periods than PSG, it is not conceivable that are used for a period longer than a few days, or a week. On the other side of the spectrum, we have actigraphy, an attractive monitoring modality which due to its relative comfort and ease of use, has been adopted and used in clinical practice. However, it is also clear that due to the use of a single sensing modality, i.e., the measurement of gross body movements, its performance is limited to monitoring sleep-wake periods, with only a fair to moderate performance. The bridge between the comfort offered by a wrist-worn sensor such as actigraphy, and the richness of cardiorespiratory characteristics needed to achieve a better performance in sleep staging, may very well lie in the advent of modern CST devices, such as those based on wrist-worn PPG sensors. However, two main challenges remain before this technology can be fully adopted in clinical practice. The first, already addressed earlier, lies in the lack of validation studies describing its performance in clinically relevant populations. The second is related to its still inferior performance, especially when compared with more unobtrusive sensing modalities. As discussed in Chapter 5, PPG has been mainly used to extract HRV features which, together with body movements measured with the accelerometer often included in these devices, only account for a fraction of the set of cardiorespiratory features used, for example, in Chapter 4 or Chapter 7. Although the absence of respiratory information alone might not fully explain the decrease in performance, it is also clear that respiratory features are important for sleep stage classification. Consider, for instance, that the three most discriminative features for N3 classification (Table 8.1) are all computed from a RIP signal. Interestingly, the PPG signal actually contains respiratory-induced intensity variations which are related to ventilatory pressure and have been used to estimate respiratory rate [217]. Future work should explore the presence of respiratory information in these signals, and the use of dedicated features, or otherwise, features adapted from earlier work with RIP to further drive the performance of sleep stage classification with these types of sensors.

8.4 General conclusions and future perspective

This thesis set out to develop methods which would increase the performance of cardiorespiratory sleep staging and bring it closer to the gold standard PSG. Although this field has received increasing attention over the last few years, the methods proposed in the later chapters are among the best reported in literature, and represent a substantial improvement over the performance reported in earlier years of research in this area. As a second objective, this thesis aimed to prove that these methods could work in truly unobtrusive sensors, such as wrist-worn PPG included in modern CST devices. The slight decrease in classification performance when compared with ECG and RIP-based classifiers is arguably offset by the convenience of PPG, which in turn achieves performances higher, in the tested subject group, than actigraphy, with the added benefit of providing richer information about sleep.

Despite advances in this area, many challenges remain. Further increasing the performance of these techniques to bring them closer to PSG is not only desirable, but given the trend in this

area, probably achievable in the long run. In this regard, it is relevant to raise the question of what is the maximum achievable performance? Although sleep, and sleep stages in particular, are based on the (visual) analysis of EEG activity, which reflects cortical and therefore CNS function, cardiorespiratory-based systems are based on ANS characteristics which might be correlated with, but are not equivalent to, CNS states. Furthermore, if the agreement between human scorers analyzing the same sleep recording with the same set of rules is, as explained, limited, what constitutes an actual gold standard ground-truth reference which can be used to train models and classifiers, and against which they can be evaluated?

But perhaps more important and arguably more challenging than these points, is to further develop and validate this technology beyond healthy subjects and towards more relevant clinical populations. This thesis took a first meaningful step in this direction by achieving one of the highest performances in sleep staging with patients suffering from sleep apnea using ECG and RIP. This must now be extended to further patient populations and eventually translated to truly unobtrusive sensors such as PPG. But also here, another question arises: what constitutes a 'good enough' performance? To facilitate acceptance and a quicker adoption by the medical community, the target performance should be a comparable performance to the de facto standard in this area, PSG. However, as actigraphy has shown over the last decades, even with a modest performance in the detection of Sleep and Wake states it found its place in clinical practice, becoming a valuable tool in the assessment, diagnosis and follow-up of certain diseases. With proper validation on relevant clinical populations, it is conceivable that unobtrusive sleep monitoring devices follow the same route and even if they never perform as well as PSG, find their place in clinical practice.

With a veritable explosion in the offer and adoption of wearable monitors, many of which emphasize sleep, and sleep quality as one of the key health factors, together with an increasing awareness that sleep is a fundamental element in balanced living, it is inevitable that this technology will reach widespread use in the next few years, perhaps before it is even adopted and endorsed by (medical) sleep associations. The research agenda in this area should thus focus not on dismissing the value of this technology, but rather on validating and further improving it on a wider patient population and also understanding its fundamental limitations. Furthermore, it should be redirected from a simple monitoring and reporting function to a tool that can give users in general, and disordered patients in particular, actionable suggestions. Correspondingly, it may be of less relevance to provide users with the results of a sleep staging algorithm, reliable as it may ever be. Instead, accurate, validated sleep monitoring technology should help people follow a healthier sleep hygiene, track trends, and highlight or flag potential problems as they occur, but before they lead to more severe consequences.

As unobtrusive sleep monitoring technology continues to improve, allied with the relative low cost and consequent ubiquity of wearable sensors in day-to-day life, it is conceivable that it might, one day, deliver a window into the physiology and characteristics of many sleeping users, and provide the means to help the millions of people suffering from undiagnosed or unsuccessfully

treated sleeping disorders.

Bibliography

[1] K. Adam and I. Oswald, "Sleep helps healing," *British Medical Journal (Clinical research ed.)*, vol. 289, no. 6456, pp. 1400–1, 1984.

[2] M. A. Carskadon and W. C. Dement, "Normal Human Sleep," in *Principles and Practice of Sleep Medicine*, 5th ed., W. C. Dement and M. Kryger, Eds. St. Louis: Elsevier, 2011, ch. 2, pp. 16–26.

[3] R. Wolk, A. S. Gami, A. Garcia-Touchard, and V. K. Somers, "Sleep and cardiovascular disease," *Current Problems in Cardiology*, vol. 30, no. 12, pp. 625–62, 2012.

[4] K. Spiegel, R. Leproult, and E. Van Cauter, "Impact of sleep debt on metabolic and endocrine function," *Lancet*, vol. 354, no. 9188, pp. 1435–9, 1999.

[5] K. L. Knutson, K. Spiegel, P. Penev, and E. Van Cauter, "The metabolic consequences of sleep deprivation," *Sleep Medicine Reviews*, vol. 11, no. 3, pp. 163–78, 2007.

[6] P. A. Bryant, J. Trinder, and N. Curtis, "Sick and tired: Does sleep have a vital role in the immune system?" *Nature Reviews. Immunology*, vol. 4, no. 6, pp. 457–67, 2004.

[7] S. Lautenbacher, B. Kundermann, and J. C. Krieg, "Sleep deprivation and pain perception," *Sleep Medicine Reviews*, vol. 10, no. 5, pp. 357–69, 2006.

[8] P. Meerlo, R. E. Mistlberger, B. L. Jacobs, H. Craig Heller, and D. McGinty, "New neurons in the adult brain: The role of sleep and consequences of sleep loss," *Sleep Medicine Reviews*, vol. 13, no. 3, pp. 187–94, 2009.

[9] G. Curcio, M. Ferrara, and L. De Gennaro, "Sleep loss, learning capacity and academic performance," *Sleep Medicine Reviews*, vol. 10, no. 5, pp. 323–37, 2006.

[10] V. K. Somers, M. E. Dyken, A. L. Mark, and F. M. Abboud, "Sympathetic-nerve activity during sleep in normal subjects," *New England Journal of Medicine*, vol. 328, no. 5, pp. 303–7, 1993.

[11] C. Iber, S. Ancoli-Israel, A. Chesson, and S. F. Quan, *The AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology and Technical Specifications*, 1st ed. Westchester, IL: American Academy of Sleep Medicine, 2007.

[12] F. Amzica and M. Steriade, "Electrophysiological correlates of sleep delta waves," *Electroencephalography and Clinical Neurophysiology*, vol. 107, no. 2, pp. 69–83, 1998.

[13] S. R. Pandi-Perumal, L. K. Seils, L. Kayumov, M. R. Ralph, A. Lowe, H. Moller, and D. F. Swaab, "Senescence, sleep, and circadian rhythms," *Ageing Research Reviews*, vol. 1, no. 1, pp. 559–604, 2002.

[14] T. Young, M. Palta, J. Dempsey, P. E. Peppard, F. J. Nieto, and K. M. Hla, "Burden of sleep apnea: rationale, design, and major findings of the Wisconsin Sleep Cohort study," *WMJ*, vol. 108, no. 5, pp. 246–9, aug 2009.

[15] K. L. Knutson, E. Van Cauter, P. J. Rathouz, T. DeLeire, and D. S. Lauderdale, "Trends in the prevalence of short sleepers in the USA: 1975-2006," *Sleep*, vol. 33, no. 1, pp. 37–45, jan 2010.

[16] M. Calem, J. Bisla, A. Begum, M. Dewey, P. E. Bebbington, T. Brugha, C. Cooper, R. Jenkins, J. Lindesay, S. McManus, H. Meltzer, N. Spiers, S. Weich, and R. Stewart, "Increased prevalence of insomnia and

changes in hypnotics use in England over 15 years: Analysis of the 1993, 2000, and 2007 National Psychiatric Morbidity surveys," *Sleep*, vol. 35, no. 3, pp. 377–84, 2012.

[17] E. S. Ford, A. G. Wheaton, T. J. Cunningham, W. H. Giles, D. P. Chapman, and J. B. Croft, "Trends in outpatient visits for insomnia, sleep apnea, and prescriptions for sleep medications among US adults: Findings from the National Ambulatory Medical Care survey 1999-2010," *Sleep*, vol. 37, no. 8, pp. 1283–93, aug 2014.

[18] Centers for Disease Control and Prevention, "Percentage of adults who reported an average of less than 6 hours of sleep per 24-hour period, by sex and age group - United States, 1985 and 2004," *Morbidity and Mortality Weekly Report*, vol. 54, no. 37, p. 933, 2005.

[19] G. A. Kerkhof, "Epidemiology of Sleep and Sleep Disorders in The Netherlands," *Sleep Medicine*, vol. 30, pp. 229–39, 2017.

[20] H. R. Colten and B. M. Altevogt, "Extent and Health Consequences of Chronic Sleep Loss and Sleep Disorders," in *Sleep Disorders and Sleep Deprivation: an Unmet Public Health Problem*, H. R. Colten and B. M. Altevogt, Eds. National Academies Press, 2006, ch. 3, pp. 55–136.

[21] K. L. Lichstein, D. J. Taylor, C. S. McCrae, and M. E. Petrov, "Insomnia: Epidemiology and Risk Factors," in *Principles and Practice of Sleep Medicine*, 6th ed., M. Kryger, T. Roth, and W. C. Dement, Eds. Philadelphia: Elsevier, 2016, ch. 81, pp. 761–8.

[22] M. M. Ohayon and C. F. Reynolds III, "Epidemiological and clinical relevance of insomnia diagnosis algorithms according to the DSM-IV and the International Classification of Sleep Disorders," *Sleep Medicine*, vol. 10, no. 9, pp. 952–60, 2009.

[23] T. Roth, C. Coulouvrat, G. Hajak, M. D. Lakoma, N. A. Sampson, V. Shahly, A. C. Shillington, J. J. Stephenson, J. K. Walsh, and R. C. Kessler, "Prevalence and perceived health associated with insomnia based on DSM-IV-TR; ICD-10; and ICSD-2 criteria: Results from the America Insomnia Survey," *Biological Psychiatry*, vol. 69, no. 6, pp. 592–600, 2011.

[24] T. Roth, "Prevalence, associated risks, and treatment patterns of insomnia," *Journal of Clinical Psychiatry*, vol. 66, no. suppl 9, pp. 10–3, 2005.

[25] D. Leger and B. Poursain, "An international survey of insomnia: under-recognition and under-treatment of a polysymptomatic condition," *Current Medical Research and Opinion*, vol. 21, no. 11, pp. 1785–92, 2005.

[26] M. H. Hall, J. Fernandez-Mendoza, C. E. Kline, and A. N. Vgontzas, "Insomnia and Health," in *Principles and Practice of Sleep Medicine*, 6th ed., M. Kryger, T. Roth, and W. C. Dement, Eds. Philadelphia: Elsevier, 2016, ch. 84, pp. 794–803.

[27] H. Greenberg, V. Lakticova, and S. M. Scharf, "Obstructive Sleep Apnea: Clinical Features, Evaluation, and Principles of Management," in *Principles and Practice of Sleep Medicine*, 6th ed., M. Kryger, T. Roth, and W. C. Dement, Eds. Philadelphia: Elsevier, 2016, ch. 114, pp. 1110–24.

[28] N. M. Punjabi, K. Bandeen-Roche, J. J. Marx, D. N. Neubauer, P. L. Smith, and A. R. Schwartz, "The association between daytime sleepiness and sleep-disordered breathing in NREM and REM sleep," *Sleep*, vol. 25, no. 3, pp. 307–14, 2002.

[29] V. Somers, D. White, R. Amin, W. Abraham, F. Costa, A. Culebras, S. Daniels, J. Floras, C. Hunt, L. Olson, T. Pickering, R. Russell, M. Woo, and T. Young, "Sleep apnea and cardiovascular disease," *Circulation*, vol. 118, no. 14, pp. 1497–518, 2008.

[30] D. J. Durgan and R. M. Bryan, "Cerebrovascular consequences of obstructive sleep apnea," *Journal of the American Heart Association*, vol. 1, no. 4, p. e000091, 2012.

[31] S. Horstmann, C. W. Hess, C. Bassetti, M. Gugger, and J. Mathis, "Sleepiness-related accidents in sleep apnea patients," *Sleep*, vol. 23, no. 3, pp. 383–9, 2000.

[32] J. Ulfberg, N. Carter, M. Talbäck, and C. Edling, "Excessive daytime sleepiness at work and subjective work performance in the general population and among heavy snorers and patients with obstructive sleep apnea," *Chest*, vol. 110, no. 3, pp. 659–63, 1996.

[33] T. Young, M. Palta, J. Dempsey, J. Skatrud, S. Weber, and S. Badr, "The occurrence of sleep-disordered breathing among middle-aged adults," *New England Journal of Medicine*, vol. 328, no. 17, pp. 1230–5, 1993.

[34] P. E. Peppard, T. Young, J. H. Barnet, M. Palta, E. W. Hagen, and K. M. Hla, "Increased prevalence of sleep-disordered breathing in adults," *American Journal of Epidemiology*, vol. 177, no. 9, pp. 1006–14, 2013.

[35] T. Young, L. Evans, L. Finn, and M. Palta, "Estimation of the clinically diagnosed proportion of sleep apnea syndrome in middle-aged men and women," *Sleep*, vol. 20, no. 9, pp. 705–6, 1997.

[36] V. Kapur, K. P. Strohl, S. Redline, C. Iber, G. O'Connor, and J. Nieto, "Underdiagnosis of sleep apnea syndrome in U.S. communities," *Sleep and Breathing*, vol. 6, no. 2, pp. 49–54, 2002.

[37] M. Kryger, T. Roth, and W. C. Dement, *Principles and Practice of Sleep Medicine*, 6th ed., M. Kryger, T. Roth, and W. C. Dement, Eds. Philadelphia: Elsevier, 2016.

[38] Philips Sleep Diagnostics, http://www.usa.philips.com/healthcare/solutions/sleep-and-respiratory-care/sleep-diagnostic, (accessed November, 2016).

[39] C. W. Whitney, D. J. Gottlieb, S. Redline, R. G. Norman, R. R. Dodge, E. Shahar, S. Surovec, and F. J. Nieto, "Reliability of scoring respiratory disturbance indices and sleep staging," *Sleep*, vol. 21, no. 7, pp. 749–57, 1998.

[40] S. Redline, R. Budhiraja, V. Kapur, C. L. Marcus, J. H. Mateika, R. Mehra, S. Parthasarthy, V. K. Somers, K. P. Strohl, L. G. Sulit, D. Gozal, M. S. Wise, and S. F. Quan, "The scoring of respiratory events in sleep: Reliability and validity," *Journal of Clinical Sleep Medicine*, vol. 3, no. 2, pp. 169–200, 2007.

[41] G. Aarab, F. Lobbezoo, H. L. Hamburger, and M. Naeije, "Variability in the apnea-hypopnea index and its consequences for diagnosis and therapy evaluation," *Respiration*, vol. 77, no. 1, pp. 32–7, 2009.

[42] N. Punjabi, N. Shifa, G. Doffner, S. Patil, G. Pien, and R. Aurora, "Computer-assisted automated scoring of polysomnograms using the Somnolyzer system," *Sleep*, vol. 38, no. 10, pp. 1555–66, 2015.

[43] H. W. Agnew, W. B. Webb, and R. L. Williams, "The first night effect: An EEG study of sleep," *Psychophysiology*, vol. 2, no. 3, pp. 263–6, 1966.

[44] O. Le Bon, L. Staner, G. Hoffmann, M. Dramaix, I. San Sebastian, J. R. Murphy, M. Kentos, I. Pelc, and P. Linkowski, "The first-night effect may last more than one night," *Journal of Psychiatric Research*, vol. 35, no. 3, pp. 165–72, 2001.

[45] W. W. Flemons, N. J. Douglas, S. T. Kuna, D. O. Rodenstein, and J. Wheatley, "Access to diagnosis and treatment of patients with suspected sleep apnea," *American Journal of Respiratory and Critical Care Medicine*, vol. 169, no. 6, pp. 668–72, 2004.

[46] L. R. A. Bittencourt, D. Suchecki, S. Tufik, C. Peres, S. M. Togeiro, M. D. C. Bagnato, and L. E. Nery, "The variability of the apnoea-hypopnoea index," *Journal of Sleep Research*, vol. 10, no. 3, pp. 245–51, 2001.

[47] S. S. Mosko, M. J. Dickel, and J. Ashurst, "Night-to-night variability in sleep apnea and sleep-related periodic leg movements in the elderly," *Sleep*, vol. 11, no. 4, pp. 340–8, 1988.

[48] M. Hornyak, M. Kopasz, B. Feige, D. Riemann, and U. Voderholzer, "Variability of periodic leg movements in various sleep disorders: Implications for clinical and pathophysiologic studies," *Sleep*, vol. 28, no. 3, pp. 331–5, 2005.

[49] J. Newell, O. Mairesse, P. Verbanck, and D. Neu, "Is a one-night stay in the lab really enough to

conclude? First-night effect and night-to-night variability in polysomnographic recordings among different clinical population samples," *Psychiatry Research*, vol. 200, no. 2-3, pp. 795–801, 2012.

[50] J. D. Edinger, A. I. Fins, R. J. Sullivan, G. R. Marsh, D. S. Dailey, T. V. Hope, M. Young, E. Shaw, D. Carlson, and D. Vasil, "Sleep in the laboratory and sleep at home: Comparisons of older insomniacs and normal sleepers," *Sleep*, vol. 20, no. 12, pp. 1119–26, 1997.

[51] J. D. Edinger, D. M. Glenn, L. A. Bastian, G. R. Marsh, D. Daile, T. V. Hope, M. Young, E. Shaw, and G. Meeks, "Sleep in the laboratory and sleep at home II: Comparisons of middle-aged insomnia sufferers and normal sleepers," *Sleep*, vol. 24, no. 7, pp. 761–70, 2001.

[52] M. E. Shayeb, L. A. Topfer, T. Stafinski, L. Pawluk, and D. Menon, "Diagnostic accuracy of level 3 portable sleep tests versus level 1 polysomnography for sleep-disordered breathing: A systematic review and meta-analysis," *CMAJ*, vol. 186, no. 1, pp. 25–51, 2014.

[53] J. F. Masa, J. Corral, R. Pereira, J. Duran-Cantolla, M. Cabello, L. Hernández-Blasco, C. Monasterio, A. Alonso, E. Chiner, M. Rubio, E. Garcia-Ledesma, L. Cacelo, R. Carpizo, L. Sacristan, N. Salord, M. Carrera, J. N. Sancho-Chust, C. Embid, F.-J. Vázquez-Polo, M. A. Negrín, and J. M. Montserrat, "Effectiveness of home respiratory polygraphy for the diagnosis of sleep apnoea and hypopnoea syndrome." *Thorax*, vol. 66, no. 7, pp. 567–73, 2011.

[54] R. Ferber, R. Millman, M. Coppola, J. Fleetham, C. F. Murray, C. Iber, V. McCall, G. Nino-Murcia, M. Pressman, and M. Sanders, "Portable recording in the assessment of obstructive sleep apnea. ASDA standards of practice." *Sleep*, vol. 17, no. 4, pp. 378–92, 1994.

[55] R. J. Cole, D. F. Kripke, W. Gruen, D. J. Mullaney, and J. C. Gillin, "Automatic sleep/wake identification from wrist activity," *Sleep*, vol. 15, no. 5, pp. 461–9, 1992.

[56] A. Sadeh and C. Acebo, "The role of actigraphy in sleep medicine," *Sleep Medicine Reviews*, vol. 6, no. 2, pp. 113–24, 2002.

[57] T. Morgenthaler, C. Alessi, L. Friedman, J. Owens, V. Kapur, B. Boehlecke, T. Brown, A. Chesson, J. Coleman, T. Lee-Chiong, J. Pancer, T. J. Swick, and A. Chesson Jr, "Practice parameters for the use of actigraphy in the assessment of sleep and sleep disorders: An update for 2007," *Sleep*, vol. 30, no. 4, pp. 519–29, 2007.

[58] J. Paquet, A. Kawinska, and J. Carrier, "Wake detection capacity of actigraphy during sleep," *Sleep*, vol. 30, no. 10, pp. 1362–9, 2007.

[59] K. Lichstein, K. Stone, J. Donaldson, S. Nau, J. Soeffing, D. Murray, K. Lester, and R. Aguillard, "Actigraphy validation with insomnia," *Sleep*, vol. 29, no. 2, pp. 232–9, 2006.

[60] M. M. Sánchez-Ortuño, J. D. Edinger, M. K. Means, and D. Almirall, "Home is where sleep is: An ecological approach to test the validity of actigraphy for the assessment of insomnia," *Journal of Clinical Sleep Medicine*, vol. 6, no. 1, pp. 21–9, 2010.

[61] J. L. Martin and A. D. Hakim, "Wrist actigraph," *Chest*, vol. 139, no. 6, pp. 1514–27, jun 2011.

[62] A. Vallières and C. M. Morin, "Actigraphy in the assessment of insomnia," *Sleep*, vol. 26, no. 7, pp. 902–6, oct 2003.

[63] B. Sivertsen, S. Omvik, O. E. Havik, S. Pallesen, B. Bjorvatn, G. H. Nielsen, S. Straume, and I. H. Nordhus, "A comparison of actigraphy and polysomnography in older adults treated for chronic primary insomnia," *Sleep*, vol. 29, no. 10, pp. 1353–8, oct 2006.

[64] A. Vallières, H. Ivers, C. H. Bastien, S. Beualieu-Bonneau, and C. M. Morin, "Variability and predictability in sleep patterns of chronic insomniacs," *Journal of Sleep Research*, vol. 14, no. 4, pp. 447–53, dec 2005.

[65] M. Means, "Accuracy of sleep perceptions among insomnia sufferers and normal sleepers," *Sleep Medicine*, vol. 4, no. 4, pp. 285–96, jul 2003.

[66] Sleep Cycle Alarm Clock, https://www.sleepcycle.com, (accessed November, 2016).

[67] SleepBot, https://www.mysleepbot.com, (accessed November, 2016).

[68] Sleep as Android, http://sleep.urbandroid.org, (accessed November, 2016).

[69] FitBit, https://www.fitbit.com, (accessed November, 2016).

[70] JawBone, https://jawbone.com, (accessed November, 2016).

[71] Misfit, https://misfit.com, (accessed November, 2016).

[72] Philips Health Watch, http://www.usa.philips.com/c-m-hs/health-programs/health-watch, (accessed November, 2016).

[73] FitBit Ultra, https://help.fitbit.com/?p=ultra, (accessed November, 2016).

[74] JawBone Up, https://jawbone.com/support/up, (accessed November, 2016).

[75] Misfit Shine, https://misfit.com/products/misfit-shine, (accessed November, 2016).

[76] ResMed S+, http://www.resmed.com/us/en/consumer/s-plus.html, (accessed November, 2016).

[77] Withings Aura Sleep Sensor, https://www.withings.com/ca/en/products/aura/sleep-sensor-accessory, (accessed November, 2016).

[78] Beddit Sleep Tracker, https://www.withings.com/ca/en/products/aura/sleep-sensor-accessory, (accessed November, 2016).

[79] Sense with Sleep Pill, https://hello.is/, (accessed November, 2016).

[80] S. J. Redmond, P. de Chazal, C. O'Brien, S. Ryan, W. T. McNicholas, and C. Heneghan, "Sleep staging using cardiorespiratory signals," *Somnologie-Schlafforschung und Schlafmedizin*, vol. 11, no. 4, pp. 245–56, 2007.

[81] K. Russo, B. Goparaju, and M. T. Bianchi, "Consumer sleep monitors: Is there a baby in the bathwater?" *Nature and Science of Sleep*, vol. 7, pp. 147–57, 2015.

[82] P. R. T. Ko, J. A. Kientz, E. K. Choe, M. Kay, C. A. Landis, and N. F. Watson, "Consumer sleep technologies: A review of the landscape," *Journal of Clinical Sleep Medicine*, vol. 11, no. 12, pp. 1455–61, 2015.

[83] M. de Zambotti, F. C. Baker, and I. M. Colrain, "Validation of sleep-tracking technology compared with polysomnography in adolescents," *Sleep*, vol. 38, no. 9, pp. 1461–8, 2015.

[84] H. E. Montgomery-Downs, S. P. Insana, and J. A. Bond, "Movement toward a novel activity monitoring device," *Sleep and Breathing*, vol. 16, no. 3, pp. 913–7, 2012.

[85] L. J. Meltzer, L. S. Hiruma, K. Avis, H. Montgomery-Downs, and J. Valentin, "Comparison of a commercial accelerometer with polysomnography and actigraphy in children and adolescents," *Sleep*, vol. 38, no. 8, pp. 1323–30, 2015.

[86] N. R. Carlson, *Physiology of Behavior*, 8th ed., N. R. Carlson, Ed. Allyn & Bacon, 2003.

[87] P. A. Lanfranchi, J.-L. Pépin, and V. K. Somers, "Cardiovascular Physiology: Autonomic Control in Health and in Sleep Disorders," in *Principles and Practice of Sleep Medicine*, 6th ed., M. Kryger, T. Roth, and W. C. Dement, Eds. Philadelphia: Elsevier, 2016, ch. 14, pp. 142–54.

[88] D. J. Eckert and J. E. Butler, "Respiratory Physiology: Understanding the Control of Ventilation," in *Principles and Practice of Sleep Medicine*, 6th ed., M. Kryger, T. Roth, and W. C. Dement, Eds. Philadelphia: Elsevier, 2016, ch. 16, pp. 167–73.

[89] R. Heinzer and F. Sériès, "Physiology of Upper and Lower Airways," in *Principles and Practice of Sleep Medicine*, 6th ed., M. Kryger, T. Roth, and W. C. Dement, Eds. Philadelphia: Elsevier, 2016, ch. 17, pp. 174–81.

[90] J. Allen, "Photoplethysmography and its application in clinical physiological measurement," *Physiological Measurement*, vol. 28, no. 3, pp. R1–R39, 2007.

[91] D. K. Spierer, Z. Rosen, L. L. Litman, and K. Fujii, "Validation of photoplethysmography as a method to detect heart rate during rest and exercise," *Journal of Medical Engineering & Technology*, vol. 39, no. 5, pp. 264–71, 2015.

[92] D. C. Mack, J. T. Patrie, P. M. Suratt, R. A. Felder, and M. Alwan, "Development and preliminary validation of heart rate and breathing rate detection using a passive, ballistocardiography-based sleep monitoring system," *IEEE Transactions on Information Technology in Biomedicine*, vol. 13, no. 1, pp. 111–20, 2009.

[93] J. M. Kortelainen, M. O. Mendez, A. M. Bianchi, M. Matteucci, and S. Cerutti, "Sleep staging based on signals acquired through bed sensor," *IEEE Transactions on Information Technology in Biomedicine*, vol. 14, no. 3, pp. 776–85, 2010.

[94] P. de Chazal, N. Fox, E. O'Hare, C. Heneghan, A. Zaffaroni, P. Boyle, S. Smith, C. O'Connell, and W. T. McNicholas, "Sleep/wake measurement using a non-contact biomotion sensor," *Journal of Sleep Research*, vol. 20, no. 2, pp. 356–66, 2011.

[95] W. Karlen, C. Mattiussi, and D. Floreano, "Sleep and wake classification with ECG and respiratory effort signals," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 3, no. 2, pp. 71–8, 2009.

[96] S. Devot, R. Dratwa, and E. Naujokat, "Sleep/wake detection based on cardiorespiratory signals and actigraphy," in *Proceedings of the 32nd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, vol. 2010. IEEE, jan 2010, pp. 5089–92.

[97] T. Watanabe and K. Watanabe, "Noncontact method for sleep stage estimation," *IEEE Transactions on Biomedical Engineering*, vol. 51, no. 10, pp. 1735–48, 2004.

[98] J. Redmond and C. Heneghan, "Cardiorespiratory-based sleep staging in subjects with obstructive sleep apnea," *IEEE Transactions on Biomedical Engineering*, vol. 53, no. 3, pp. 485–96, 2006.

[99] A. Domingues, T. Paiva, and J. M. Sanches, "Hypnogram and sleep parameter computation from activity and cardiovascular data," *IEEE Transactions on Biomedical Engineering*, vol. 61, no. 6, pp. 1711–9, 2014.

[100] S. M. Isa, I. Wasito, and A. M. Arymurthy, "Kernel dimensionality reduction on sleep stage classification using ECG signal," *International Journal of Computer Science Issues*, vol. 8, no. 4, pp. 115–23, 2011.

[101] T. Willemen, D. Van Deun, V. Verhaert, M. Vandekerckhove, V. Exadaktylos, J. Verbraecken, S. V. Huffel, B. Haex, and J. Vander Sloten, "An evaluation of cardio-respiratory and movement features with respect to sleep stage classification," *IEEE Journal of Biomedical and Health Informatics*, vol. 18, no. 2, pp. 661–9, 2014.

[102] T. Willemen, C. Varon, A. Caicedo Dorado, B. Haex, J. Vander Sloten, and S. Van Huffel, "Probabilistic cardiac and respiratory based classification of sleep and apneic events in subjects with sleep apnea," *Physiological Measurement*, vol. 36, no. 10, pp. 2103–18, 2015.

[103] B. Yılmaz, M. H. Asyalı, E. Arıkan, S. Yetkin, and F. Özgen, "Sleep stage and obstructive apneaic epoch classification using single-lead ECG," *Biomedical Engineering Online*, vol. 9, no. 39, 2010.

[104] A. Noviyanto, S. M. Isa, I. Wasito, and A. M. Arymurthy, "Selecting features of single lead ECG signal for automatic sleep stages classification using correlation-based feature subset selection," *International Journal of Computer Science Issues*, vol. 8, no. 5, pp. 139–48, 2011.

[105] Y. Kurihara and K. Watanabe, "Sleep-stage decision algorithm by using heartbeat and body-movement signals," *IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans*, vol. 42, no. 6, pp. 1450–9, 2012.

[106] M. Macrea, E. S. Katz, and A. Malhotra, "Central Sleep Apnea: Definitions, Pathophysiology, Genetics and Epidemiology," in *Principles and Practice of Sleep Medicine*, 6th ed., M. Kryger, T. Roth, and W. C. Dement, Eds. Philadelphia: Elsevier, 2016, ch. 109, pp. 1049–58.

[107] E. Tobaldini, L. Nobili, S. Strada, K. R. Casali, A. Braghiroli, and N. Montano, "Heart rate variability in normal and pathological sleep," *Frontiers in Physiology*, vol. 4, no. 294, pp. 1–11, oct 2013.

[108] M. H. Bonnet and D. L. Arand, "Heart rate variability in insomniacs and matched normal sleepers," *Psychosomatic Medicine*, vol. 60, no. 5, pp. 610–5, 1998.

[109] B. U. Köhler, C. Hennig, and R. Orglmeister, "The principles of software QRS detection," *IEEE Engineering in Medicine and Biology*, vol. 21, no. 1, pp. 42–57, 2002.

[110] Association for the Advancement of Medical Instrumentation Electrocardiograph Committee, *Testing and Reporting Performance Results of Cardiac Rhythm and ST Segment Measurement Algorithms*. Arlington, Virginia, USA: ANSI/AASM/ISO, 2008, no. EC57:1998/(R)2008.

[111] P. G. Catcheside, S. C. Chiong, R. S. Orr, J. Mercer, N. A. Saunders, and R. D. McEvoy, "Acute cardiovascular responses to arousal from non-REM sleep during normoxia and hypoxia," *Sleep*, vol. 24, no. 8, pp. 895–902, 2001.

[112] M. F. Bangash, A. Xie, J. B. Skatrud, K. J. Reichmuth, S. R. Barczi, and B. J. Morgan, "Cerebrovascular response to arousal from NREM and REM sleep," *Sleep*, vol. 31, no. 3, pp. 321–7, 2008.

[113] G. Breithardt, M. E. Cain, N. El-Sherif, N. C. Flowers, V. Hombach, M. Janse, M. B. Simson, and G. Steinbeck, "Standards for analysis of ventricular late potentials using high-resolution or signal-averaged electrocardiography," *Journal of the American College of Cardiology*, vol. 17, no. 5, pp. 999–1006, 1991.

[114] R. Jané, H. Rix, P. Caminal, P. Laguna, and R. Jane, "Alignment methods for averaging of high-resolution cardiac signals: A comparative study of performance," *IEEE Transactions on Biomedical Engineering*, vol. 38, no. 6, pp. 571–9, 1991.

[115] J. F. Thayer, S. S. Yamamoto, and J. F. Brosschot, "The relationship of autonomic imbalance, heart rate variability and cardiovascular disease risk factors," *International Journal of Cardiology*, vol. 141, no. 2, pp. 122–31, 2010.

[116] M. Volterrani, "Decreased heart rate variability in patients with chronic obstructive pulmonary disease," *Chest*, vol. 106, no. 5, pp. 1432–7, 1994.

[117] P. K. Stein and Y. Pu, "Heart rate variability, sleep and sleep disorders," *Sleep Medicine Reviews*, vol. 16, no. 1, pp. 47–66, 2012.

[118] Task Force of the European Society of Cardiology and the North American Society of Pacing and Electrophysiology, "Heart rate variability: Standards of measurement, physiologic interpretation, and clinical use," *European Heart Journal*, vol. 17, no. 3, pp. 354–81, 1996.

[119] T. Bragge, M. P. Tarvainen, P. O. Ranta-aho, and P. A. Karjalainen, "High-resolution QRS fiducial point corrections in sparsely sampled ECG recordings," *Physiological Measurement*, vol. 26, no. 5, pp. 743–51, 2005.

[120] L. Sörnmo, "Vectorcardiographic loop alignment and morphologic beat-to-beat variability," *IEEE Transactions on Biomedical Engineering*, vol. 45, no. 12, pp. 1401–13, 1998.

[121] J. L. Ferreira, P. J. M. Cluitmans, and R. M. Aarts, "Detection of sharp wave activity in biological signals using differentiation between consecutive samples," in *Proceedings of the 6th International Conference on Bio-inspired Systems and Signal Processing (BIOSIGNALS)*. Barcelona, Spain: SciTePress, feb 2013, pp. 327–32.

[122] G. M. Friesen, T. C. Jannett, M. A. Jadallah, S. L. Yates, S. R. Quint, and H. T. Nagle, "A comparison of the noise sensitivity of nine QRS detection algorithms," *IEEE Transactions on Biomedical Engineering*, vol. 37, no. 1, pp. 85–98, 1990.

[123] P. Hamilton, "Open source ECG analysis," in *Computers in Cardiology*. IEEE, sep 2002, pp. 101–4.

[124] P. S. Hamilton and W. J. Tompkins, "Quantitative investigation of QRS detection rules using the

MIT/BIH arrhythmia database," *IEEE Transactions on Biomedical Engineering*, vol. 33, no. 12, pp. 1157–65, 1986.

[125] M. E. Nygårds and L. Sörnmo, "Delineation of the QRS complex using the envelope of the ECG," *Medical and Biological Engineering and Computing*, vol. 21, no. 5, pp. 538–47, 1983.

[126] V. X. Afonso, W. J. Tompkins, T. Q. Nguyen, and S. Luo, "ECG beat detection using filter banks," *IEEE Transactions on Biomedical Engineering*, vol. 46, no. 2, pp. 192–202, 1999.

[127] J. A. van Alsté, W. van Eck, and O. E. Herrmann, "ECG baseline wander reduction using linear phase filters," *Computers and Biomedical Research*, vol. 19, no. 5, pp. 417–27, 1986.

[128] J. Pan and W. J. Tompkins, "A real-time QRS detection algorithm," *IEEE Transactions on Biomedical Engineering*, vol. 32, no. 3, pp. 230–36, 1985.

[129] R. M. Rangayyan, *Biomedical Signal Analysis: A Case-Study Approach*. Wiley-IEEE Press, 2001.

[130] A. L. Goldberger, L. A. N. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C. K. Peng, and H. E. Stanley, "PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals," *Circulation*, vol. 101, no. 23, pp. e215–e220, 2000.

[131] G. B. Moody and R. G. Mark, "The impact of the MIT-BIH arrhythmia database," *IEEE Engineering in Medicine and Biology*, vol. 20, no. 3, pp. 45–50, 2001.

[132] R. Bousseljot, D. Kreiseler, and A. Schnabel, "Nutzung der EKG-Signaldatenbank CARDIODAT der PTB über das Internet," *Biomedizinische Technik/Biomedical Engineering*, vol. 40, no. s1, pp. 317–8, 1995.

[133] G. R. Shaw and P. Savard, "On the detection of QRS variations in the ECG," *IEEE Transactions on Biomedical Engineering*, vol. 42, no. 7, pp. 736–41, 1995.

[134] K. L. Venkatachalam, J. E. Herbrandson, and S. J. Asirvatham, "Signals and signal processing for the electrophysiologist: Part II: Signal processing and artifact," *Circulation. Arrhythmia and Electrophysiology*, vol. 4, no. 6, pp. 974–81, dec 2011.

[135] N. V. Thakor, J. G. Webster, and W. J. Tompkins, "Estimation of QRS complex power spectra for design of a QRS filter," *IEEE Transactions on Biomedical Engineering*, vol. 31, no. 11, pp. 702–6, 1984.

[136] E. Pueyo, L. Sornmo, and P. Laguna, "QRS slopes for detection and characterization of myocardial ischemia," *IEEE Transactions on Biomedical Engineering*, vol. 55, no. 2, pp. 468–77, 2008.

[137] D. Romero, M. Ringborn, P. Laguna, and E. Pueyo, "Detection and quantification of acute myocardial ischemia by morphologic evaluation of QRS changes by an angle-based method," *Journal of Electrocardiology*, vol. 46, no. 3, pp. 204–14, 2013.

[138] P. Fonseca, X. Long, J. Foussier, and R. M. Aarts, "On the impact of arousals on the performance of sleep and wake classification using actigraphy," in *Proceedings of the 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, vol. 2013. Osaka - Japan: IEEE, jan 2013, pp. 6760–3.

[139] X. Long, P. Fonseca, J. Foussier, R. Haakma, and R. M. Aarts, "Sleep and wake classification with actigraphy and respiratory effort using dynamic warping," *IEEE Journal of Biomedical and Health Informatics*, vol. 18, no. 4, pp. 1272–84, 2014.

[140] I. Romero, "PCA-based noise reduction in ambulatory ECGs," in *Computing in Cardiology*. IEEE, 2010, pp. 677–80.

[141] S.-H. Liu, "Motion artifact reduction in electrocardiogram using adaptive filter," *Journal of Medical and Biological Engineering*, vol. 31, no. 1, pp. 67–72, 2010.

[142] J. Lee, D. D. McManus, S. Merchant, and K. H. Chon, "Automatic motion and noise artifact detection in Holter ECG data using empirical mode decomposition and statistical approaches," *IEEE Transactions on Biomedical Engineering*, vol. 59, no. 6, pp. 1499–506, 2012.

[143] G. D. Clifford and G. B. Moody, "Editorial: signal quality in cardiorespiratory monitoring," *Physiological Measurement*, vol. 33, no. 9, p. Article number E01, 2012.

[144] T. Pawar, S. Chaudhuri, and S. P. Duttagupta, "Body movement activity recognition for ambulatory cardiac monitoring," *IEEE Transactions on Biomedical Engineering*, vol. 54, no. 5, pp. 874–82, 2007.

[145] A. L. Motto, H. L. Galiana, K. A. Brown, and R. E. Kearney, "Detection of movement artifacts in respiratory inductance plethysmography: performance analysis of a Neyman-Pearson energy-based detector," in *Proceedings of the 26th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, vol. 1. IEEE, jan 2004, pp. 49–52.

[146] D. B. Keenan and F. H. Wilhelm, "Adaptive and wavelet filtering methods for improving accuracy of respiratory measurement," *Biomedical Sciences Instrumentation*, vol. 41, pp. 37–42, 2005.

[147] A. A. Aoude, A. L. Motto, H. L. Galiana, K. A. Brown, and R. E. Kearney, "Power-based segmentation of respiratory signals using forward-backward bank filtering," in *Proceedings of the 28th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, vol. 1. IEEE, jan 2006, pp. 4631–4.

[148] D. J. Buysse, C. F. Reynolds, T. H. Monk, S. R. Berman, and D. J. Kupfer, "The Pittsburgh Sleep Quality Index: A new instrument for psychiatric practice and research," *Psychiatry Research*, vol. 28, no. 2, pp. 193–213, 1989.

[149] C. H. Bastien, A. Vallières, and C. M. Morin, "Validation of the Insomnia Severity Index as an outcome measure for insomnia research," *Sleep Medicine*, vol. 2, no. 4, pp. 297–307, 2001.

[150] P. Fonseca, R. M. Aarts, J. Foussier, and X. Long, "A novel low-complexity post-processing algorithm for precise QRS localization," *SpringerPlus*, vol. 3, pp. 376–1/13, 2014.

[151] I. Daubechies, "The wavelet transform, time-frequency localization and signal analysis," *IEEE Transactions on Information Theory*, vol. 36, no. 5, pp. 961–1005, 1990.

[152] N. R. Draper and H. Smith, *Applied Regression Analysis*, 3rd ed. Wiley-Interscience, 1998.

[153] J. Fox, *Regression Diagnostics: an Introduction*. SAGE, 1991.

[154] M. Costa, A. Goldberger, and C.-K. Peng, "Multiscale entropy analysis of complex physiologic time series," *Physical Review Letters*, vol. 89, no. 6, p. 068102, 2002.

[155] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed. Wiley-Interscience, 2000.

[156] J. Cohen, "A Coefficient of agreement for nominal scales," *Educational and Psychological Measurement*, vol. 20, no. 1, pp. 37–46, 1960.

[157] E. Walker and A. S. Nowacki, "Understanding equivalence and noninferiority testing," *Journal of General Internal Medicine*, vol. 26, no. 2, pp. 192–6, 2011.

[158] M. Hollander, D. A. Wolfe, and E. Chicken, *Nonparametric Statistical Methods*, 3rd ed. Wiley, 2013.

[159] J. M. Bland and D. G. Altman, "Measuring agreement in method comparison studies," *Statistical Methods in Medical Research*, vol. 8, no. 2, pp. 135–60, 1999.

[160] A. M. Euser, F. W. Dekker, and S. le Cessie, "A practical approach to Bland-Altman plots and variation coefficients for log transformed variables," *Journal of Clinical Epidemiology*, vol. 61, no. 10, pp. 978–82, 2008.

[161] X. Long, J. Foussier, P. Fonseca, R. Haakma, and R. M. Aarts, "Analyzing respiratory effort amplitude for automated sleep stage classification," *Biomedical Signal Processing and Control*, vol. 14, pp. 197–205, 2014.

[162] X. Long, J. Yang, T. Weysen, R. Haakma, J. Foussier, P. Fonseca, and R. M. Aarts, "Measuring dissimilarity between respiratory effort signals based on uniform scaling for sleep staging," *Physiological Measurement*, vol. 35, no. 12, pp. 2529–42, 2014.

[163] R. J. Berger and N. H. Phillips, "Energy conservation and sleep," *Behavioural Brain Research*, vol. 69, no. 1-2, pp. 65–73, 1995.

[164] J. H. Benington and H. C. Heller, "Restoration of brain energy metabolism as the function of sleep," *Progress in Neurobiology*, vol. 45, no. 4, pp. 347–60, 1995.

[165] R. Stickgold, "Sleep-dependent memory consolidation," *Nature*, vol. 437, no. 7063, pp. 1272–8, 2005.

[166] M. P. Walker, "The role of slow wave sleep in memory processing," *Journal of Clinical Sleep Medicine*, vol. 5, no. 2 Suppl, pp. S20–6, 2009.

[167] G. Klosch, B. Kemp, T. Penzel, A. Schlogl, P. Rappelsberger, E. Trenker, G. Gruber, J. Zeitlhofer, S. L. Himanen, D. Kunz, M. J. Barbanoj, J. Roschke, A. Varri, and G. Dorffner, "The SIESTA project polygraphic and clinical database," *IEEE Engineering in Medicine and Biology*, vol. 20, no. 3, pp. 51–7, 2001.

[168] A. Rechtschaffen and A. Kales, *A Manual of Standardized Terminology, Techniques and Scoring System for Sleep Stages of Human Subjects*. Washington, DC: Public Health Service, U.S. Government Printing Office, 1968.

[169] M. Basner, B. Griefahn, U. Müller, G. Plath, and A. Samel, "An ECG-based algorithm for the automatic identification of autonomic activations associated with cortical arousal," *Sleep*, vol. 30, no. 10, pp. 1349–61, 2007.

[170] P. Bušek, J. Vaňková, J. Opavský, J. Salinger, and S. Nevšímalová, "Spectral analysis of the heart rate variability in sleep," *Physiological Research*, vol. 54, no. 4, pp. 369–76, 2005.

[171] X. Long, P. Fonseca, R. Haakma, R. M. Aarts, and J. Foussier, "Spectral boundary adaptation on heart rate variability for sleep and wake classification," *International Journal on Artificial Intelligence Tools*, vol. 23, no. 3, p. 1460002, 2014.

[172] M. O. Mendez, M. Matteucci, V. Castronovo, L. Ferini-Strambi, S. Cerutti, and A. M. Bianchi, "Sleep staging from heart rate variability: Time-varying spectral features and hidden Markov models," *International Journal of Biomedical Engineering and Technology*, vol. 3, no. 3, pp. 246–63, 2010.

[173] J. W. Kantelhardt, E. Koscielny-Bunde, H. H. A. Rego, S. Havlin, and A. Bunde, "Detecting long-range correlations with detrended fluctuation analysis," *Physica A: Statistical Mechanics and its Applications*, vol. 295, no. 3, pp. 441–54, 2001.

[174] N. Iyengar, C. K. Peng, R. Morin, A. L. Goldberger, and L. A. Lipsitz, "Age-related alterations in the fractal scaling of cardiac interbeat interval dynamics," *The American Journal of Physiology*, vol. 271, pp. R1078–84, 1996.

[175] T. Penzel, J. W. Kantelhardt, L. Grote, J.-H. H. Peter, and A. Bunde, "Comparison of detrended fluctuation analysis and spectral analysis for heart rate variability in sleep and sleep apnea," *IEEE Transactions on Biomedical Engineering*, vol. 50, no. 10, pp. 1143–51, 2003.

[176] S. Telser, M. Staudacher, Y. Ploner, A. Amann, H. Hinterhuber, and M. Ritsch-Marte, "Can one detect sleep stage transitions for on-line sleep scoring by monitoring the heart rate variability?" *Somnologie*, vol. 8, no. 2, pp. 33–41, 2004.

[177] M. Adnane, Z. Jiang, and Z. Yan, "Sleep-wake stages classification and sleep efficiency estimation using single-lead electrocardiogram," *Expert Systems with Applications*, vol. 39, no. 1, pp. 1401–13, 2012.

[178] M. Costa, A. L. Goldberger, and C.-K. Peng, "Multiscale entropy analysis of biological signals," *Physical Review E*, vol. 71, pp. 1–18, 2005.

[179] D. Cysarz, H. Bettermann, and P. van Leeuwen, "Entropies of short binary sequences in heart period dynamics," *American Journal of Physiology - Heart and Circulatory Physiology*, vol. 278, no. 6, pp. 2163–72, 2000.

[180] L. Lacasa, B. Luque, F. Ballesteros, J. Luque, and J. C. Nuño, "From time series to complex networks: The visibility graph," *Proceedings of the National Academy of Sciences*, vol. 105, no. 13, pp. 4972–5, 2008.

[181] X. Long, P. Fonseca, R. M. Aarts, R. Haakma, and J. Foussier, "Modeling cardiorespiratory interaction during human sleep with complex networks," *Applied Physics Letters*, vol. 105, no. 203701, pp. 1–4, 2014.

[182] Z. G. Shao, "Network analysis of human heartbeat dynamics," *Applied Physics Letters*, vol. 96, no. 7, p. 073703, 2010.

[183] G. Zhu, Y. Li, and P. P. Wen, "Analysis and classification of sleep stages based on difference visibility graphs from a single channel EEG signal," *IEEE Journal of Biomedical and Health Informatics*, vol. 18, no. 6, pp. 1813–21, 2014.

[184] D. Berndt and J. Clifford, "Using dynamic time warping to find patterns in time series," *Workshop on Knowledge Discovery in Databases*, vol. 10, no. 16, pp. 359–70, 1994.

[185] J. S. Richman and J. R. Moorman, "Physiological time-series analysis using approximate entropy and sample entropy," *American Journal of Physiology - Heart and Circulatory Physiology*, vol. 278, no. 6, pp. H2039–49, 2000.

[186] Y. Ichimaru, K. P. Clark, J. Ringler, and W. J. Weiss, "Effect of sleep stage on the relationship between respiration and heart rate variability," in *Computers in Cardiology*, 1990, pp. 657–60.

[187] D. Cysarz, H. Bettermann, S. Lange, D. Geue, and P. van Leeuwen, "A quantitative comparison of different methods to detect cardiorespiratory coordination during night-time sleep," *Biomedical Engineering Online*, vol. 3, no. 1, p. 44, 2004.

[188] C. De Boor, *A Practical Guide to Splines*. New York: Springer, 2001.

[189] M. Unser, "Splines: A perfect fit for signal and image processing," *IEEE Signal Processing Magazine*, vol. 16, no. 6, pp. 22–38, 1999.

[190] A. W. Whitney, "A direct method of nonparametric measurement selection," *IEEE Transactions on Computers*, vol. 100, no. 9, pp. 1100–3, 1971.

[191] P. Smialowski, D. Frishman, and S. Kramer, "Pitfalls of supervised feature selection," *Bioinformatics*, vol. 26, no. 3, pp. 440–3, 2010.

[192] P. Pudil, J. Novovičová, and J. Kittler, "Floating search methods in feature selection," *Pattern Recognition Letters*, vol. 15, no. 11, pp. 1119–25, 1994.

[193] J. M. Yentes, N. Hunt, K. K. Schmid, J. P. Kaipust, D. McGrath, and N. Stergiou, "The appropriate use of approximate entropy and sample entropy with short data sets," *Annals of Biomedical Engineering*, vol. 41, no. 2, pp. 349–65, 2013.

[194] J. Hedner, D. P. White, A. Malhotra, S. Herscovici, S. D. Pittman, D. Zou, L. Grote, and G. Pillar, "Sleep staging based on autonomic signals: A multi-center validation study," *Journal of Clinical Sleep Medicine*, vol. 7, no. 3, pp. 301–6, 2011.

[195] M. Migliorini, A. M. Bianchi, D. Nisticò, J. Kortelainen, E. Arce-Santana, S. Cerutti, and M. O. Mendez, "Automatic sleep staging based on ballistocardiographic signals recorded through bed sensors," in *Proceedings of the 32nd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, vol. 2010. IEEE, jan 2010, pp. 3273–6.

[196] M. Xiao, H. Yan, J. Song, Y. Yang, and X. Yang, "Sleep stages classification based on heart rate variability and random forest," *Biomedical Signal Processing and Control*, vol. 8, no. 6, pp. 624–33, 2013.

[197] M. Bruyneel, C. Sanida, G. Art, W. Libert, L. Cuvelier, M. Paesmans, R. Sergysels, and V. Ninane, "Sleep efficiency during sleep studies: Results of a prospective study comparing home-based and in-hospital polysomnography," *Journal of Sleep Research*, vol. 20, no. 1 Pt 2, pp. 201–6, 2011.

[198] S. Ancoli-Israel, R. Cole, C. Alessi, M. Chambers, W. Moorcroft, and C. P. Pollak, "The role of actigraphy in the study of sleep and circadian rhythms," *Sleep*, vol. 26, no. 3, pp. 342–92, 2003.

[199] J. Lee and J. Finkelstein, "Consumer sleep tracking devices: A critical review," *Digital Healthcare Empowering Europeans*, vol. 210, pp. 458–60, 2015.

[200] C. A. Kushida, A. Chang, C. Gadkary, C. Guilleminault, O. Carrillo, and W. C. Dement, "Comparison of actigraphic, polysmonographic, and subjective assessment of sleep parameters in sleep-disordered patients," *Sleep medicine*, vol. 2, no. 5, pp. 389–96, 2001.

[201] J. Trinder, J. Kleiman, M. Carrington, S. Smith, S. Breen, N. Tan, and Y. Kim, "Autonomic activity during human sleep as a function of time and sleep stage," *Journal of Sleep Research*, vol. 10, no. 4, pp. 253–64, 2001.

[202] M. H. Bonnet and D. L. Arand, "Heart rate variability: Sleep stage, time of night, and arousal influences," *Electroencephalography and clinical neurophysiology*, vol. 102, no. 5, pp. 390–6, 1997.

[203] M. Ako, T. Kawara, S. Uchida, S. Miyazaki, K. Nishihara, J. Mukai, K. Hirao, J. Ako, and Y. Okubo, "Correlation between electroencephalography and heart rate variability during sleep," *Psychiatry and Clinical Neurosciences*, vol. 57, no. 1, pp. 59–65, 2003.

[204] F. Jurysta, P. Van De Borne, P. F. Migeotte, M. Dumont, J. P. Lanquart, J. P. Degaute, P. Linkowski, and P. V. D. Borne, "A study of the dynamic interactions between sleep EEG and heart rate variability in healthy young men," *Clinical Neurophysiology*, vol. 114, no. 11, pp. 2146–55, 2003.

[205] M. Dumont, F. Jurysta, J. P. Lanquart, P. F. Migeotte, P. Van De Borne, and P. Linkowski, "Interdependency between heart rate variability and sleep EEG: Linear/non-linear?" *Clinical Neurophysiology*, vol. 115, no. 9, pp. 2031–40, 2004.

[206] P. Fonseca, X. Long, M. Radha, R. Haakma, R. M. Aarts, and J. Rolink, "Sleep stage classification with ECG and respiratory effort," *Physiological Measurement*, vol. 36, no. 10, pp. 2027–40, 2015.

[207] M. F. Folstein, S. E. Folstein, and P. R. McHugh, ""Mini-mental state": A practical method for grading the cognitive state of patients for the clinician," *Journal of Psychiatric Research*, vol. 12, no. 3, pp. 189–98, 1975.

[208] C. E. Carney, D. J. Buysse, S. Ancoli-Israel, J. D. Edinger, A. D. Krystal, K. L. Lichstein, and C. M. Morin, "The consensus sleep diary: Standardizing prospective sleep self-monitoring," *Sleep*, vol. 35, no. 2, pp. 287–302, 2012.

[209] H. Werner, L. Molinari, C. Guyer, and O. G. Jenni, "Agreement rates between actigraphy, diary, and questionnaire for children's sleep patterns," *Archives of Pediatrics & Adolescent Medicine*, vol. 162, no. 4, pp. 350–8, 2008.

[210] L. J. Meltzer, C. M. Walsh, J. Traylor, and A. M. L. Westin, "Direct comparison of two new actigraphs and polysomnography in children and adolescents," *Sleep*, vol. 35, no. 1, pp. 159–66, 2012.

[211] P. Fonseca, R. M. Aarts, X. Long, J. Rolink, and S. Leonhardt, "Estimating actigraphy from motion artifacts in ECG and respiratory effort signals," *Physiological Measurement*, vol. 37, no. 1, pp. 67–82, 2016.

[212] V. Novak, P. Novak, J. de Champlain, A. Le Blanc, R. Martin, and R. Nadeau, "Influence of respiration on heart rate and blood pressure fluctuations," *Journal of Applied Physiology*, vol. 74, no. 2, pp. 617–26, 1993.

[213] M. V. Pitzalis, F. Mastropasqua, F. Massari, A. Passantino, R. Colombo, A. Mannarini, C. Forleo, and P. Rizzon, "Effect of respiratory rate on the relationships between RR interval and systolic blood pressure fluctuations: a frequency-dependent phenomenon," *Cardiovascular Research*, vol. 38, no. 2, pp. 332–9, 1998.

[214] S. Elsenbruch, M. J. Harnish, and W. C. Orr, "Heart rate variability during waking and sleep in healthy males and females," *Sleep*, vol. 22, no. 8, pp. 1067–71, 1999.

[215] R. W. C. G. R. Wijshoff, M. Mischi, J. Veen, A. M. van der Lee, and R. M. Aarts, "Reducing motion

artifacts in photoplethysmograms by using relative sensor motion: phantom study," *Journal of Biomedical Optics*, vol. 17, no. 11, pp. 117 007–1–15, 2012.

[216] S. Herscovici, A. Pe'er, S. Papyan, and P. Lavie, "Detecting REM sleep from the finger: an automatic REM sleep algorithm based on peripheral arterial tone (PAT) and actigraphy," *Physiological Measurement*, vol. 28, no. 2, pp. 129–40, 2007.

[217] L. M. Nilsson, "Respiration signals from photoplethysmography," *Anesthesia & Analgesia*, vol. 117, no. 4, pp. 859–65, 2013.

[218] M. K. Uçar, M. R. Bozkurt, C. Bilgin, and K. Polat, "Automatic sleep staging in obstructive sleep apnea patients using photoplethysmography, heart rate variability signal and machine learning techniques," *Neural Computing and Applications*, pp. 1–16, 2016.

[219] M. Marino, Y. Li, M. N. Rueschman, J. W. Winkelman, J. M. Ellenbogen, J. M. Solet, H. Dulin, L. F. Berkman, and O. M. Buxton, "Measuring sleep: Accuracy, sensitivity, and specificity of wrist actigraphy compared to polysomnography," *Sleep*, vol. 36, no. 11, pp. 1747–55, 2013.

[220] A. Noviyanto and A. M. Arymurthy, "Sleep stages classification based on temporal pattern recognition in neural network approach," in *Proceedings of the 2012 International Joint Conference on Neural Networks (IJCNN)*, 2012, pp. 1–6.

[221] M. Bresler, K. Sheffy, G. Pillar, M. Preiszler, and S. Herscovici, "Differentiating between light and deep sleep stages using an ambulatory device based on peripheral arterial tonometry," *Physiological Measurement*, vol. 29, no. 5, pp. 571–84, 2008.

[222] J. W. Kim, J.-S. S. Lee, P. A. Robinson, and D.-U. U. Jeong, "Markov analysis of sleep dynamics," *Physical Review Letters*, vol. 102, no. 17, p. 178104, 2009.

[223] J. Lafferty, A. Mccallum, and F. C. N. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proceedings of the 18th International Conference on Machine Learning (ICML)*, vol. 2001, 2001, pp. 282–9.

[224] G. Luo and W. Min, "Subject-adaptive real-time sleep stage classification based on conditional random field," in *Proceedings of the AMIA Annual Symposium*, vol. 2007. American Medical Informatics Association, jan 2007, p. 488.

[225] M. M. Ohayon, M. A. Carskadon, C. Guilleminault, and M. V. Vitiello, "Meta-analysis of quantitative sleep parameters from childhood to old age in healthy individuals: Developing normative sleep values across the human lifespan," *Sleep*, vol. 27, no. 7, pp. 1255–73, 2004.

[226] U. R. Acharya, E. C. Chua, O. Faust, T. C. Lim, and L. F. Lim, "Automated detection of sleep apnea from electrocardiogram signals using nonlinear parameters," *Physiological Measurement*, vol. 32, no. 3, pp. 287–303, 2011.

[227] H. Bettermann, D. Cysarz, and P. Van Leeuwen, "Detecting cardiorespiratory coordination by respiratory pattern analysis of heart period dynamics - the musical rhythm approach," *International Journal of Bifurcation and Chaos*, vol. 10, no. 10, pp. 2349–60, 2000.

[228] X. Long, J. B. Arends, R. M. Aarts, R. Haakma, P. Fonseca, and J. Rolink, "Time delay between cardiac and brain activity during sleep transitions," *Applied Physics Letters*, vol. 106, no. 14, pp. 143 702–1/4, 2015.

[229] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–86, 1989.

[230] A. J. Viterbi, "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm," *IEEE Transactions on Information Theory*, vol. 13, no. 2, pp. 260–9, 1967.

[231] F. R. Kschischang, B. J. Frey, and H.-A. Loeliger, "Factor graphs and the sum-product algorithm," *IEEE Transactions on Information Theory*, vol. 47, no. 2, pp. 498–519, 2001.

[232] J. M. Hammersley and P. Clifford, "Markov fields on finite graphs and lattices," *unpublished*, 1971.

[233] C. Sutton and A. McCallum, "An Introduction to Conditional Random Fields for Relational Learning," in *Introduction to Statistical Relational Learning*. MIT press, 2007, vol. 93, pp. 142–6.

[234] J. E. Dennis, Jr and J. J. More, "Quasi-Newton methods, motivation and theory," *SIAM Review*, vol. 19, no. 1, pp. 46–89, 1997.

[235] M. A. Hall, "Correlation-based Feature Selection for Machine Learning," Ph.D. dissertation, The University of Waikato, 1999.

[236] A. Strehl and J. Ghosh, "Cluster ensembles - a knowledge reuse framework for combining multiple partitions," *The Journal of Machine Learning Research*, vol. 3, pp. 583–617, 2002.

[237] F. Wilcoxon, "Individual comparisons by ranking methods," *Biometrics Bulletin*, vol. 1, no. 6, pp. 80–3, 1945.

[238] R. L. Verrier and R. M. Harper, "Cardiovascular Physiology: Central and Autonomic Regulation," in *Principles and Practice of Sleep Medicine*, 5th ed., M. H. Kryger, T. Roth, and W. C. Dement, Eds. Elsevier Saunders, 2011, pp. 215–25.

[239] X. Long, P. Fonseca, R. M. Aarts, R. Haakma, J. Rolink, and S. Leonhardt, "Detection of nocturnal slow wave sleep based on cardiorespiratory activity in healthy adults," *IEEE Journal of Biomedical and Health Informatics*, vol. 21, no. 1, pp. 123–33, 2017.

[240] C. Yu, Z. Liu, T. McKenna, A. T. Reisner, and J. Reifman, "A method for automatic identification of reliable heart rates calculated from ECG and PPG waveforms," *Journal of the American Medical Informatics Association*, vol. 13, no. 3, pp. 309–20, 2006.

[241] Z. Beattie, Y. Oyang, A. Statan, A. Ghoreyshi, A. Pantelopoulos, A. Russell, and C. Heneghan, "Estimation of sleep stages in a healthy adult population from optical plethysmography and accelerometer signals," *Physiological Measurement*, vol. 38, pp. 1968–79, 2017.

[242] X. Long, R. Haakma, T. R. M. Leufkens, P. Fonseca, and R. M. Aarts, "Effects of between-and within-subject variability on autonomic cardiorespiratory activity during sleep and their limitations on sleep staging: a multilevel analysis," *Computational Intelligence and Neuroscience*, vol. 2015, pp. 583 820–1/17, 2015.

[243] T. Roehrs, F. Zorick, R. Wittig, W. Conway, and T. Roth, "Predictors of objective level of daytime sleepiness in patients with sleep-related breathing disorders," *Chest*, vol. 95, no. 6, pp. 1202–6, jun 1989.

[244] R. Ratnavadivel, N. Chau, D. Stadler, A. Yeo, R. D. McEvoy, and P. G. Catcheside, "Marked reduction in obstructive sleep apnea severity in slow wave sleep," *Journal of Clinical Sleep Medicine*, vol. 5, no. 6, pp. 519–24, dec 2009.

[245] M. T. Bianchi, S. S. Cash, J. Mietus, C.-K. Peng, and R. Thomas, "Obstructive sleep apnea alters sleep stage transition dynamics," *PLoS ONE*, vol. 5, no. 6, p. e11356, jun 2010.

[246] A. N. Vgontzas, J. Fernandez-Mendoza, D. Liao, and E. O. Bixler, "Insomnia with objective short sleep duration: the most biologically severe phenotype of the disorder," *Sleep medicine reviews*, vol. 17, no. 4, pp. 241–54, aug 2013.

[247] J.-M. Gaillard, "Chronic primary insomnia: Possible physiopathological involvement of slow wave sleep deficiency," *Sleep*, vol. 1, no. 2, pp. 133–47, jun 1978.

[248] W. R. Pigeon and M. L. Perlis, "Sleep homeostasis in primary insomnia," *Sleep Medicine Reviews*, vol. 10, no. 4, pp. 247–54, 2006.

[249] L. Parrino, G. Milioli, F. De Paolis, A. Grassi, and M. G. Terzano, "Paradoxical insomnia: The role of CAP and arousals in sleep misperception," *Sleep Medicine*, vol. 10, no. 10, pp. 1139–45, 2009.

[250] R. Ferri, B. B. Koo, D. L. Picchietti, and S. Fulda, "Periodic leg movements during sleep: Phenotype, neurophysiology, and clinical significance," *Sleep Medicine*, vol. 31, pp. 29–38, 2017.

[251] J. Montplaisir and R. Godbout, "Nocturnal sleep of narcoleptic patients: Revisited," *Sleep*, vol. 9, no. 1, pp. 159–61, 1986.

[252] S. Overeem, E. Mignot, J. G. van Dijk, and G. J. Lammers, "Narcolepsy: Clinical features, new pathophysiologic insights, and future perspectives," *Journal of Clinical Neurophysiology*, vol. 18, no. 2, pp. 78–105, 2001.

[253] M. de Zambotti, N. Covassin, M. Sarlo, G. De Min Tona, J. Trinder, and L. Stegagno, "Nighttime cardiac sympathetic hyper-activation in young primary insomniacs," *Clinical Autonomic Research*, vol. 23, no. 1, pp. 49–56, feb 2013.

[254] K. Spiegelhalder, L. Fuchs, J. Ladwig, S. D. Kyle, C. Nissen, U. Voderholzer, B. Feige, and D. Riemann, "Heart rate and heart rate variability in subjectively reported insomnia," *Journal of Sleep Research*, vol. 20, no. 1 Pt 2, pp. 137–45, 2011.

[255] K. Narkiewicz, C. A. Pesek, M. Kato, B. G. Phillips, D. E. Davison, and V. K. Somers, "Baroreflex control of sympathetic nerve activity and heart rate in obstructive sleep apnea," *Hypertension*, vol. 32, no. 6, pp. 1039–43, 1998.

[256] P. Anderer, G. Gruber, S. Parapatics, and G. Dorffner, "Automatic sleep classification according to Rechtschaffen and Kales," in *Proceedings of the 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, vol. 2007. IEEE, jan 2007, pp. 3994–7.

List of Publications

Journals

- X. Long, **P. Fonseca**, J. Foussier, R. Haakma, and R. M. Aarts, 'Sleep and wake classification with actigraphy and respiratory effort using dynamic warping,' IEEE Journal of Biomedical and Health Informatics, vol. 18, no. 4, pp. 1272–84, 2014.
- **P. Fonseca**, R. M. Aarts, J. Foussier, and X. Long, 'A novel low-complexity post-processing algorithm for precise QRS localization, SpringerPlus, vol. 3, no. 1, p. 376, 2014.
- X. Long, **P. Fonseca**, R. M. Aarts, R. Haakma, and J. Foussier, 'Modeling cardiorespiratory interaction during human sleep with complex networks,' Applied Physics Letters, vol. 105, no. 20, p. 203701, 2014.
- X. Long, **P. Fonseca**, R. Haakma, R. M. Aarts, and J. Foussier, 'Spectral boundary adaptation on heart rate variability for sleep and wake classification,' International Journal on Artificial Intelligence Tools, vol. 23, no. 3, p. 1460002, 2014.
- X. Long, J. Foussier, **P. Fonseca**, R. Haakma, and R. M. Aarts, 'Analyzing respiratory effort amplitude for automated sleep stage classification,' Biomedical Signal Processing and Control, vol. 14, pp. 197–205, 2014.
- X. Long, J. Yang, T. Weysen, R. Haakma, J. Foussier, **P. Fonseca**, and R. M. Aarts, 'Measuring dissimilarity between respiratory effort signals based on uniform scaling for sleep staging,' Physiological Measurement, vol. 35, no. 12, pp. 2529–42, 2014.
- X. Long, J. B. Arends, R. M. Aarts, R. Haakma, **P. Fonseca**, and J. Rolink, 'Time delay between cardiac and brain activity during sleep transitions,' Applied Physics Letters, vol. 106, no. 14, p. 143702, 2015.
- X. Long, R. Haakma, T. R. M. Leufkens, **P. Fonseca**, and R. M. Aarts, 'Effects of between-and within-subject variability on autonomic cardiorespiratory activity during sleep and their limitations on sleep staging: a multilevel analysis,' Computational Intelligence and Neuroscience, vol. 2015, p. 583820, 2015.
- **P. Fonseca***, X. Long*, M. Radha, R. Haakma, R. M. Aarts, and J. Rolink, 'Sleep stage classification with ECG and respiratory effort,' Physiological Measurement, vol. 36, no. 10, pp. 2027–40, 2015. *Joint first authorship.
- **P. Fonseca**, R. M. Aarts, X. Long, J. Rolink, and S. Leonhardt, 'Estimating actigraphy from motion artifacts in ECG and respiratory effort signals,' Physiological Measurement, vol. 37, no. 1, pp. 67–82, 2016.
- **P. Fonseca**, N. den Teuling, X. Long, and R. M. Aarts, 'Cardiorespiratory sleep stage detection using conditional random fields,' IEEE Journal of Biomedical and Health Informatics, vol. 21, no. 4, pp. 956–66, 2017.

- X. Long, **P. Fonseca**, R. M. Aarts, R. Haakma, J. Rolink, and S. Leonhardt, 'Detection of nocturnal slow wave sleep based on cardiorespiratory activity in healthy adults,' IEEE Journal of Biomedical and Health Informatics, vol. 21, no. 1, pp. 123–33, 2017.
- **P. Fonseca**, T. Weysen, M. S. Goelema, E. I. S. Møst, M. Radha, C. Lunsingh Scheurleer, L. van den Heuvel, and R. M. Aarts, 'Validation of photoplethysmography-based sleep staging compared with polysomnography in healthy middle aged adults,' Sleep, vol. 40, no. 7, 2017.
- M.-M. Nano, **P. Fonseca**, R. Vullings, and R. M. Aarts, 'Measures of cardiovascular autonomic activity in insomnia disorder: a systematic review,' PLoS ONE, vol. 12, no. 10, p. e0186716, 2017.
- **P. Fonseca**, N. den Teuling, X. Long, R. M. Aarts, 'A comparison of probabilistic classifiers for sleep stage classification,' *submitted to Physiological Measurement*, 2018.

Conference Proceedings

- X. Long, **P. Fonseca**, J. Foussier, R. Haakma, and R. M. Aarts, 'Using dynamic time warping for sleep and wake discrimination,' in Proceedings of the 2012 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI), 2012, vol. 25, pp. 886–9.
- X. Long, **P. Fonseca**, J. Foussier, R. Haakma, and R. M. Aarts, 'Time-frequency analysis of heart rate variability for sleep and wake classification,' in Proceedings of the IEEE 12th International Conference on Bioinformatics & Bioengineering (BIBE), 2012, pp. 85–90.
- **P. Fonseca**, X. Long, J. Foussier, and R. M. Aarts, 'On the impact of arousals on the performance of sleep and wake classification using actigraphy,' in Proceedings of the 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 2013, vol. 2013, pp. 6760–3.
- J. Foussier, **P. Fonseca**, X. Long, and S. Leonhardt, 'Automatic feature selection for sleep/wake classification with small data sets,' in Proceedings of the 6th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2013), 2013, pp. 178–84.
- J. Foussier, **P. Fonseca**, X. Long, B. Misgeld, and S. Leonhardt, 'Combining HRV features for automatic arousal detection,' in Proceedings of the IFMBE International Conference on Health Informatics (ICHI), 2013.
- X. Long, J. Foussier, **P. Fonseca**, R. Haakma, and R. M. Aarts, 'Respiration amplitude analysis for REM and NREM sleep classification,' in Proceedings of the 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 2013, vol. 2013, pp. 5017–20.
- X. Long, **P. Fonseca**, R. Haakma, J. Foussier, and R. M. Aarts, 'Automatic detection of overnight deep sleep based on heart rate variability: a preliminary study,' in Proceedings of the 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 2014, pp. 50–3.
- X. Long, R. Haakma, R. M. Aarts, **P. Fonseca**, and J. Foussier, 'Between-laboratory and demographic effects on heart rate and its variability during sleep,' in Workshop on Smart Healthcare and Healing Environments in conjunction with AMI'14, European Conference on Ambient Intelligence, 2014.
- X. Long, R. Haakma, J. Rolink, **P. Fonseca**, and R. M. Aarts, 'Improving sleep/wake detection via boundary adaptation for respiratory spectral features,' in Proceedings of the 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 2015, pp. 374–7.
- X. Long, **P. Fonseca**, R. Haakma, and R. M. Aarts, 'Actigraphy-based sleep/wake detection for insomniacs,' in Proceedings of the IEEE 14th International Conference on Wearable and Implantable Body Sensor Networks (BSN), 2017, pp. 9–12.

- G. B. Papini, **P. Fonseca**, X. L. Aubert, S. Overeem, J. W. M. Bergmans, and R. Vullings, 'Photoplethysmography beat detection and pulse morphology quality assessment for signal reliability estimation,' in Proceedings of the 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 2017.

Patents

List of filed patent applications in the scope of this thesis, sorted by publication date:

- **P. Fonseca**, A. Heinrich, I. Berezhnoy, R. Haakma, and R. de Bruijn, 'Separating cardiac signal and respiratory signal from vital signs,' WO2013179189 (A1), 2013.
- I. Berezhnoy, E. Schuijers, and **P. Fonseca**, 'Device and method for sleep monitoring,' US2015351693 (A1), 2015.
- I. Berezhnoy, **P. Fonseca**, A. Heinrich, and R. Haakma, 'Electronic switch for controlling a device in dependency on a sleep stage,' US2015302722 (A1); US9754471 (B2), 2015.
- **P. Fonseca**, G. N. Garcia Molina, and R. Haakma, 'System and method for enhancing sleep slow wave activity based on cardiac characteristics or respiratory characterics,' US2016310696 (A1), 2016.
- R. Haakma, **P. Fonseca**, and M. Radha, 'Device and method for monitoring a physiological state of a subject,' WO2016206921 (A1), 2016.
- **P. Fonseca**, R. Haakma, R. M. Aarts, and X. Long, 'Actigraphy methods and apparatuses,' US2016128641 (A1), 2016.
- A. Heinrich and **P. Fonseca**, 'Baby sleep monitor,' EP3232924 (A1), 2017.
- X. Long, R. Haakma, **P. Fonseca**, and R. M. Aarts, 'System and method for determining spectral boundaries for sleep stage classification,' EP3229661 (A1), 2017.
- **P. Fonseca**, X. Long, N. den Teuling, R. Haakma, and R. M. Aarts, 'System and method for slow wave sleep detection,' EP3232913 (A1), 2017.
- **P. Fonseca**, X. Long, N. den Teuling, R. Haakma, and R. M. Aarts, 'System and method for cardiorespiratory sleep stage classification,' EP3232912 (A1), 2017.

Acknowledgments

When I think back of the journey, I am humbled by the fact that it would have been impossible without the help of so many incredible, inspiring people. Those who made every step worth it, bad days bearable, good days memorable. This list will always be incomplete. To everyone that was a part of my life and my work in the last five years, a heartfelt thank you.

To my promotor, Prof. Ronald Aarts. Ronald, your patience, critical attitude, and valuable advice showed me which turns to take, and which path to follow. And you taught me a valuable lesson in science: that even when we think we found an answer, we must keep on questioning it. You have helped me become a better scientist, and for that I will always be grateful.

To my co-promotor, Prof. Sebastiaan Overeem. Sebastiaan, your drive, passion, knowledge and vision of the future are inspiring, and I am thankful I could count on your vast expertise to help improve my dissertation. I look forward to our exciting cooperations, and to make home sleep monitoring a reality.

To Prof. Raymond van Ee, Prof. Dirk Pevernagie, Prof. Loe Feijs, thank you for accepting to be part of the Doctorate Committee and for reviewing my thesis, and to Prof. Jan Bergmans for offering me the opportunity to pursue my PhD in the Signal Processing Systems group of the TU/e.

To my colleagues at Philips, not only for the years of doctoral work, but for all years before that. Because of you I drive to work every morning with the same enthusiasm as when I started, almost fourteen years ago.

To Jörg Habetha. Thanks in great part to you, I chose to pursue a PhD. I am grateful that you gave me not only the opportunity, but also the support, throughout the years, to work on it next to my full time job at Philips Research.

To Xi Long, you are one of the most prolific scientific writers I have met, and were always an inspiration and a pleasure to work with. You gave me the best example I could have of how to start, work on, and finish a PhD in this area.

To Reinder Haakma. Pursuing a PhD next to a full time job was not easy, but it would have been a whole lot more difficult without a research project to back it up and for that, I am grateful. The fact that the topic is very much alive six years after we started it together at Philips Research is a testament to your efforts as a project leader. But also to the great researchers and fantastic colleagues we worked with, in and outside our project. So a big thank you to Maaike, Adrienne, Igor, Jenny, Koen, Marija, Jos, Alberto, Guido, Helma, Fons, Ton, Ad, Lieke and to David, Paul, Els, Sander, Stefan, Tsvetomira, Henning, Leonie, Charlotte, Erik, Gary and of course, Sandrine, who started this all.

Despite having worked with so many, some colleagues in particular did and continue to impress me. A special thanks to Niek den Teuling, whose creativity and hard work during his internship paved the way to the first sleep stage classifier with good performance. And to Mustafa Radha, whose genuine excitement about cutting-edge AI and machine learning is helping me continue to develop myself, and our home sleep monitoring technology.

To Tim Weysen and Tim Leufkens, an endless source of creativity and knowledge about sleep, thank you for always being ready with valuable advice. I look forward to seeing some of our many ideas making it to the outside world.

To Marina Nano and Gabriele Papini, who are well under way on their own doctorate road, I hope I can help you as others have helped me. Thank you for your energy, creativity and hard work.

To my current and former colleagues in the Innovations for Sleep Apnea project, in particular, Mareike Klee, Angela Grassi and Xavier Aubert in Eindhoven, Marco Ross, Peter Anderer, Arnaud Moreau, and Andreas Cerny in Philips Vienna, and Edmund Shaw, Charles Kosinski, Greg Matthews, Mike Kane, Ray Vasko, Dave Smith and Rich Sofranko in Philips North America. Mareike, you are truly an example of dedication to your job and to Philips, and I admire how you manage such a vast portfolio of research activities without ever losing sight of the well-being of your project members. Angela, you are taking your first steps in this area, but with the energy and passion you keep showing, I am sure you will succeed in whatever it is that you set as a goal. Xavier, while everyone seems increasingly and exclusively concerned with the 'big picture', I learned and took great inspiration from your critical attention to detail.

To my colleagues in the IC of the Future research project, in particular Thomas Falck, and Esther van der Heide. Thank you for giving me the opportunity to work in such a challenging, exciting topic in a great research team.

To my colleagues at Kempenhaeghe Sleep Medicine Center, TU/e and KU Leuven, Merel van Gilst, Lieke Hermans, Bruno Arsenali, Hans van Dijk, Prof. Sabine Van Huffel, Carolina Varon, for the outstanding collaboration in our common projects.

Ao Prof. Paulo Novais e aos alunos de Mestrado da Universidade do Minho, Marco Prata, Pedro Mateus, Cármina Azevedo, António Rebelo e Guilherme Barbosa que, como eu, se aventuraram pelos Países-Baixos no que, espero, tenha sido uma enriquecedora experiência.

Ao Prof. Fernando Pereira, do Instituto Superior Técnico em Lisboa, um agradecimento especial, por ter despertado em mim, há tantos anos atrás, o interesse pela investigação científica, e pelos exemplos de rigor e dedicação nos quais me inspiro até hoje.

Aan mijn vrienden, Bert, Kathleen, Annelies, Antony, Liv, Jelle, Björn, Hendrik, Ellen, Nico, Jan, Kim, Elke, Jan, Guy, Kathy, Matthias, Elien, Iris, Amber. Jullie hebben de deuren van jullie land, jullie huis and jullie hart voor mij geopend. Dankzij jullie voel me ik thuis. Bedankt.

Aan Francine, bedankt voor uw steun, niet alleen tijdens mijn doctoraat, maar ook daarnaast. U staat altijd klaar voor mij, een betere schoonmoeder kan ik me niet wensen en ik ben dankbaar dat u deel uitmaakt van ons gezin.

Mãe, pai, sei que a distância é um grande preço a pagar pela felicidade e realização profissional que constantemente procuro. E nunca vos posso agradecer o suficiente pelos sacrifícios que fizeram para que eu chegasse aqui e pelos valores que me deram. Pelos exemplos de honestidade, justiça, profissionalismo, e bondade com os que nos rodeiam, que sempre tentei copiar. Espero que tenham orgulho da pessoa em que me tornei.

And to my sweet Ellen. I know, when I started, how apprehensive you were, afraid that I was going to suffer all you went through during your PhD. But despite all of that, you supported me during this journey,

so often encouraging me when I really needed a push. Not everyone can count themselves lucky to have, as a wife, their best friend. You are not only my wife and my best friend, but also my diving buddy, my travel companion, my inspiration, my super-hero, and my love. If all experiences we had over the last thirteen years are any example, I can hardly wait for what we have yet to live together. I cannot ever thank you enough. This is dedicated to you.

Curriculum Vitae

Pedro Fonseca was born in Lisbon, Portugal, in 1979. In 2002 he received the Dipl.-Ing. degree in Electrical Engineering and in 2004, the M.Sc. degree in Electrical and Computer Engineering from Instituto Superior Técnico, Technical University of Lisbon, Portugal. He joined Philips Research Eindhoven in 2004 and has worked in the areas of image and video analysis and digital content management and, since 2010, physiological sensing, ambulatory monitoring, and sleep. In 2013 he started his PhD project in the Signal Processing Systems Group of the Eindhoven University of Technology. He is currently a Senior Scientist in the Personal Health department of Philips Research, where in parallel to his PhD work, he contributes to several research projects in the area of unobtrusive physiological sensing, home sleep monitoring, and sleep diagnostics.

The TU/e Sleep Series

1. Maaike Goelema. Perceived sleep quality in a personal health monitoring context.
2. Pedro Fonseca. Home sleep monitoring.