Journal of Neuroscience Methods 235 (2014) 130–137

Contents lists available at ScienceDirect

Journal of Neuroscience Methods

journal homepage: www.elsevier.com/locate/jneumeth

Clinical Neuroscience

Automatic sleep classification using a data-driven topic model reveals latent sleep states

Henriette Koch[a,*], Julie A.E. Christensen[a,b,c], Rune Frandsen[b,d], Marielle Zoetmulder[e], Lars Arvastson[c], Soren R. Christensen[c], Poul Jennum[b,d], Helge B.D. Sorensen[a]

[a] Technical University of Denmark, Department of Electrical Engineering, Ørsteds Plads, Building 349, 2800 Kgs. Lyngby, Denmark
[b] Danish Center for Sleep Medicine, Department of Clinical Neurophysiology, Glostrup Hospital, Nordre Ringvej 57, 2600 Glostrup, Denmark
[c] H. Lundbeck A/S, Ottiliavej 9, 2500 Valby, Denmark
[d] Center for Healthy Ageing, University of Copenhagen, Blegdamsvej 3B, 2200 Copenhagen N, Denmark
[e] Bispebjerg Hospital, Department of Neurology, Bispebjerg Bakke 23, 2400 Copenhagen NV, Denmark

HIGHLIGHTS

- We use a data-driven approach to develop a general and automatic sleep classifier.
- The model uses spectral EEG and EOG as well as eye correlation measures as input.
- Six latent sleep states with concordances to the golden standard are revealed.
- Sleep state transitions are expressed as continuous processes.
- We report good performance on subjects with and without neurodegenerative diseases.

ARTICLE INFO

Article history:
Received 25 January 2014
Received in revised form 29 June 2014
Accepted 1 July 2014
Available online 9 July 2014

Keywords:
Automatic sleep classification
Topic modeling
Electroencephalography (EEG)
Electrooculography (EOG)
Sleep state switching
Neurodegenerative diseases

ABSTRACT

Background: The golden standard for sleep classification uses manual scoring of polysomnography despite points of criticism such as oversimplification, low inter-rater reliability and the standard being designed on young and healthy subjects.
New method: To meet the criticism and reveal the latent sleep states, this study developed a general and automatic sleep classifier using a data-driven approach. Spectral EEG and EOG measures and eye correlation in 1 s windows were calculated and each sleep epoch was expressed as a mixture of probabilities of latent sleep states by using the topic model Latent Dirichlet Allocation. Model application was tested on control subjects and patients with periodic leg movements (PLM) representing a non-neurodegenerative group, and patients with idiopathic REM sleep behavior disorder (iRBD) and Parkinson's Disease (PD) representing a neurodegenerative group. The model was optimized using 50 subjects and validated on 76 subjects.
Results: The optimized sleep model used six topics, and the topic probabilities changed smoothly during transitions. According to the manual scorings, the model scored an overall subject-specific accuracy of 68.3 ± 7.44 (% $\mu \pm \sigma$) and group specific accuracies of 69.0 ± 4.62 (control), 70.1 ± 5.10 (PLM), 67.2 ± 8.30 (iRBD) and 67.7 ± 9.07 (PD).
Comparison with existing method: Statistics of the latent sleep state content showed accordances to the sleep stages defined in the golden standard. However, this study indicates that sleep contains six diverse latent sleep states and that state transitions are continuous processes.
Conclusions: The model is generally applicable and may contribute to the research in neurodegenerative diseases and sleep disorders.

1. Introduction

The golden standard for sleep analysis is defined by the American Academy of Sleep Medicine (AASM) (Iber et al., 2007) and uses manual scoring of polysomnography. AASM divides sleep

* Corresponding author. Tel.: +45 2513 6271/+1 650 716 8353.
E-mail address: henriette.koch@hotmail.com (H. Koch).

http://dx.doi.org/10.1016/j.jneumeth.2014.07.002
0165-0270/

into wakefulness (W), rapid-eye-movement (REM) and non-REM (NREM) sleep, where NREM is further divided into three stages N1–N3 according to the level of drowsiness and pattern of EEG changes. However, manual scoring is criticized for using the predefined epoch length of 30 s and oversimplification of the sleep structures by discretizing sleep stages into abruptly defined stages. Other weaknesses are the low inter-rater reliability of 67–91% (Ahmed and Tafreshi, 2009), large time consumption and the standard suited to fit young healthy subjects even though sleep changes with age (Crowley, 2011) and in patients with neurodegenerative diseases (Santamaria et al., 2011).

Sleep is primarily controlled by neurons in the brainstem and midbrain. State switching is determined by complex neuron pathways which can be described by two interacting flip–flip models (Dyken et al., 2012). These models suggest that only W, REM and NREM sleep are stable sleep states, states switches are continuous processes and that degeneration of the neurons involved will affect the overall sleep pattern.

In this study a general sleep classifier was designed using a data-driven approach to ensure that no biased definitions were used. The number of sleep states was predefined but this study's data-driven approach revealed the hidden and system immanent sleep states. Through a number of mathematical steps raw electroencephalography (EEG)/electrooculography (EOG) signals were converted into text strings and the topic model Latent Dirichlet Allocation (LDA) was applied to these text strings. LDA (Blei et al., 2003) analyzes the presence of words included in each text document in the text corpus. This is used to calculate the mixture of probabilities for each latent topic in each individual text document. In the context of sleep classification, raw EEG/EOG was analyzed in 1 s windows and converted into "words" with each 30 s epoch representing a document and topics representing latent sleep stages. The 1 s resolution ensures that micro-sleep events, which typically have a duration of 0.5–3 s, contribute to the epoch classification (Iber et al., 2007). The sleep model was trained on control subjects and applied to iRBD and PD patients, representing a neurodegenerative disease (NDD) group, and control subjects and patients with periodic leg movement (PLM), representing a non-NDD group. PLM patients were included to stress the model with consecutive limb movements and arousals but these patients show no signs of neurodegeneration.

2. Materials and methods

2.1. Data

Sleep recordings from 126 subjects were enrolled in this study. The Danish Center for Sleep Medicine in the Department of Clinical Neurophysiology, Glostrup University Hospital provided all data and carried out one manual sleep scoring according to the AASM standard for each polysomnography (PSG). Control subjects had no history of diagnosed sleep disorders, dream-enacting or movement disorder. Patients treated with anti-depressant drugs including hypnotics were excluded but dopaminergic treatments were continued. The dose of dopa may affect the vigilance states (Micallef et al., 2009) but it is clinical practice that patients continue the intake of dopa in parallel with the sleep recording. This study included clinical data to ensure a model suitable for clinical use was developed, and the dopaminergic treatments were therefore continued. The demographic data and leg movement (LM) index (number of leg movements per hour) are summarized in Table 1. Disease durations for the PD patients were 4.0 ± 4.0 years ($\mu \pm \sigma$) and 7.6 ± 4.0 years in the base term matrix and validation dataset respectively. One PD patient was later diagnosed with multipel system atrofi and one developed dementia with lewy bodies. Each dataset was designed to match age with no further knowledge about the subjects.

Table 1
Demographics and leg movement (LM) index.

Patient class	Subjects (F/M)	Age (years, $\mu \pm \sigma$)	LM-index ($\mu \pm \sigma$)
Train sleep model			
Control	10 (5/5)	57.2 ± 8.1	38.6 ± 41.3
Base term matrix			
Control	10 (5/5)	59.8 ± 8.0	21.3 ± 14.1
PLM	10 (6/4)	57.8 ± 9.3	48.0 ± 31.5
iRBD	10 (2/8)	59.0 ± 13.5	55.9 ± 34.2
PD	10 (4/6)	63.2 ± 6.0	52.8 ± 54.2
Validation			
Control	13 (11/2)	54.4 ± 9.1	15.5 ± 11.1
PLM	15 (6/9)	56.3 ± 12.6	62.2 ± 37.7
iRBD	22 (3/19)	63.6 ± 5.7	31.8 ± 23.3
PD	26 (8/18)	66.4 ± 6.7	31.2 ± 28.7

PSG was carried out as outpatient for control subjects, and for patients PSG was carried out as outpatient or inpatient according to the AASM standard (Iber et al., 2007). Recordings with artefacts such as electrode disconnections and continuous noise were excluded through visual inspection. Two or more PSGs were performed only if the prior recording did not meet the acceptable quality for clinical use. First night recordings may impact the sleep stage distribution but this was considered not to affect the micro-sleep distribution within the different sleep stages. Further, first night recordings were included in all subject groups in the model throughout the optimization and validation.

The raw sleep data and hypnograms were extracted from Nervus (V5.5, Cephalon DK, Nørresundby, Denmark) using the built-in export data tool. Further analysis was carried out in MATLAB (R2011b, MathWorks, Inc., Natick, MA, USA) with a sampling frequency of 256 Hz.

2.2. Automatic sleep classification

The optimized sleep classifier used two EEG channels (C3-A2 and O1-A2) and two EOG channels (EOGL-A2 and EOGR-A1 placed one cm out/up and out/down, respectively, from the canthus) in between lights-off and lights-on. The steps involved in the classifier are shown in Fig. 1.

Initially, the signals were filtered forward and reversed in time with 4th order Butterworth filters (3 dB) using cut-off frequencies 0.3 and 35 Hz for EEG and 0.3 and 10 Hz for EOG. These filters were chosen due to narrow transition bands and no ripples. Double filtering ensured a linear phase-response and the squared magnitude was equalized later on. The cut off frequencies for EEG and EOG were chosen according to the AASM standard and to focus the signals on eye movements, respectively. No other removal of intermittent artefacts was performed because the study aimed at using a data-driven approach and should be able to handle such cases. The filtering was followed by word creation and word counting, which involves contiguous mathematical steps. The word count output describes the signal characteristics and is given as input to the LDA model. It is common to view data as text documents in the setting of LDA. This is no limitation to the actual type of data analyzed.

2.2.1. Signal characteristics

2.2.1.1. EEG words. The single-sided amplitude spectrum was calculated by fast Fourier transform in non-overlapping 1 s segments by using MATLAB's built-in *fft* function, a rectangular window and zero-padding. For EEG channels, the clinical frequency bands ($\delta < 4$, $4 \leq \theta < 8$, $8 \leq \alpha < 14$, $14 \leq \beta \leq 30$ Hz) were used, and within each frequency band the frequency content was summed. Symbolization was used to define the amplitude level in each 1 s segment and this

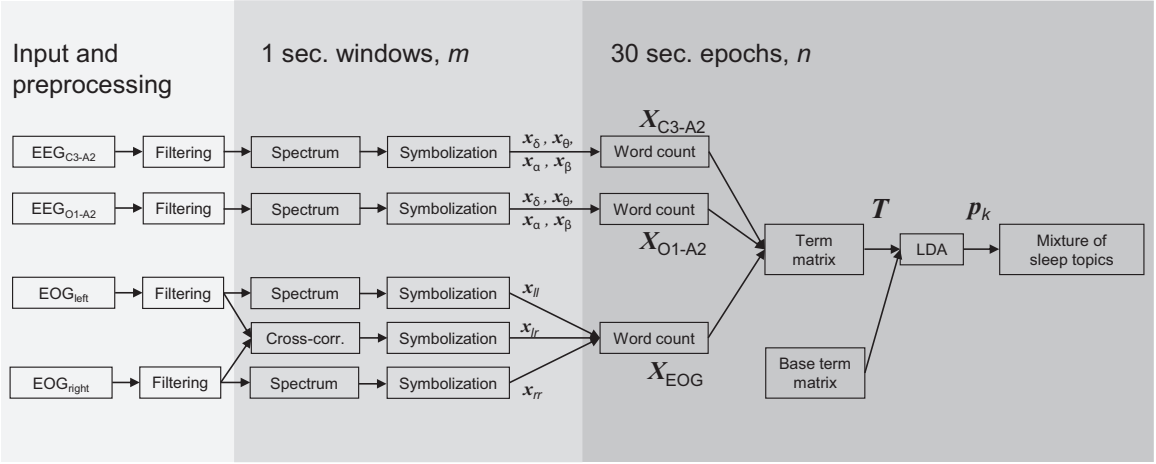

Fig. 1. Overview of the final sleep topic model applied to a test subject. Two EEG and two EOG channels were used as input to create a term matrix and the term matrix was merged with the base term matrix, which included the term matrix of 40 subjects (specified in Table 1). The merged term matrix was used as input in the topic model Latent Dirichlet Allocation (LDA) and the mixture of sleep topics was thereby obtained.

was carried out by symbolic aggregate approximation (SAX) (Lin et al., 2003): for each subject a histogram was produced for each frequency band and the four boundaries producing five equiprobable bins derived. The boundaries were calculated using the full night recording and not adjusted for eventual changes in electrode impedance or amplitude differences caused by the ultradian pattern. The boundaries were used to assign each 1 s segment a "symbol" between 1 and 5 according to the amplitude value (1 low amplitude, 5 high amplitude).

Each EEG 1 s segment was thereby expressed by four values, [x_δ, x_θ, x_α, x_β], and the full night recording expressed by four strings (one string for each frequency band). To maintain correlation with the AASM standard, all strings were divided into 30 s non-overlapping epochs, n. For each string, a word was defined as three contiguous symbols and the specific word length was found during the model optimization described in Section 2.2.3. A sliding window with length of three symbols and step-size one symbol was applied to each string to count the presence of the specific words within the individual frequency bands and within each 30 s epoch. Micro-sleep events with durations as short as 0.5–3 s may contain important information for sleep state classification and in order to capture these events, the step-size was predefined to 1 s. Thereby each frequency band contained 125 different words and all four frequency bands a total of 500 different words. A single EEG electrode was expressed as

$$\mathbf{X}_{EEG}(n) = \begin{bmatrix} X_\delta(n) \\ X_\theta(n) \\ X_\alpha(n) \\ X_\beta(n) \end{bmatrix}, \quad (1)$$

where $X(n)$ for each frequency band is a column vector containing word counts in each 30 s epoch $n = 1, 2, \ldots, N$. $\mathbf{X}_{EEG}$ has 500 rows (500 possible words) and N columns, where each column is a spectral fingerprint of the EEG content in epoch n.

Each 1 s window was described using the sum of amplitude values in the clinical frequency bands. This may not be the ideal way of describing a signal and an alternative approach was considered. The alternative method involved defining a couple of spectral templates describing specific micro-sleep events and calculating which template each 1 s window fits the most. The templates can be done fully automated or designed by sleep experts/scorers. This study used the clinical frequency bands because successful sleep classification has been achieved using these (Koch et al., 2013). However, the alternative approach may improve the method and can be included in future work.

2.2.1.2. EOG words. For the left and right EOG channel the single-sided amplitudes of frequencies ≤5 Hz were summed in each 1 s segment m, producing $x_{ll}(m)$ and $x_{rr}(m)$. Creation of words and word counting were performed similar to EEG but using four amplitude symbols (three SAX boundaries in contrast to four SAX boundaries used in EEG) and a word length of three symbols. The EOG characteristics in each epoch n were contained in $X_{ll}(n)$ and $X_{rr}(n)$. Further, the cross-correlation between the EOG channels in a 1 s segment was calculated as

$$X_{lr}(m) = \frac{\sigma_{lr}(m)}{\sqrt{\sigma_{ll}^2(m)\sigma_{rr}^2(m)}}, \quad (2)$$

where σ_{ll}^2 and σ_{rr}^2 denote the variance in the left and right EOG and σ_{lr} is the covariance. All values were aligned around zero by subtracting the subject-specific median from X_{lr}. EOG signals are anti-correlated during eye movements, which should induce negative cross-correlation during W, REM sleep and N1 (Iber et al., 2007). During deep sleep, high amplitude EEG artefacts in EOG could be present and would appear correlated and background EOG almost uncorrelated (Christensen et al., 2013). Four symbols were used for EOG cross-correlation and this dimension was chosen according to what was reported in (Christensen et al., 2013). The optimized SAX boundaries were [−0.7, 0, 0.7] and word counting in 30 s epochs was performed similar to EEG. The EOG fingerprint in epoch n was thereby expressed as

$$\mathbf{X}_{EOG}(n) = \begin{bmatrix} X_{ll}(n) \\ X_{rr}(n) \\ X_{lr}(n) \end{bmatrix}, \quad (3)$$

where $\mathbf{X}_{EOG}$ contains 192 possible words. Finally, the EEG and EOG words were combined in the term matrix

$$\boldsymbol{T}(n) = \begin{bmatrix} \boldsymbol{X}_{\mathrm{C3-A2}}(n) \\ \boldsymbol{X}_{\mathrm{O1-A2}}(n) \\ \boldsymbol{X}_{\mathrm{EOG}}(n) \end{bmatrix}, \quad (4)$$

which expresses the EEG and EOG characteristics in each epoch.

2.2.2. Training and applying a topic model

The topic model Latent Dirichlet Allocation (LDA) optimized by variational inference (Blei et al., 2003) was used to calculate the mixture of latent sleep states, $p_k(n)$, in each epoch. LDA is a three-level hierarchical Bayesian model and allows documents to be associated with multiple topics. Further, LDA assumes exchangeability for the words in a document as well as documents in a text corpus and this ensures that no prior information is needed when classifying an epoch.

The MATLAB LDA toolbox implemented by Verbeek (2006) was used in this study. The sleep model was trained on 10 control subjects by using the manually scored hypnograms to randomly extract 60 epochs from each AASM sleep stage in each subject. All epochs were included if a smaller number of epochs in a sleep stage were present. The extracted epochs were combined in a term matrix and fed into LDA. The LDA parameters describing the word probabilities, β_{jk} (word j, topic k) and the topic Dirichlet, α_k, were trained and saved to allow equal scoring terms and thereby direct comparison between subjects.

The training only included control subjects due to the hypothesis that control subjects express the "true" sleep stages. Further, control subjects move less during sleep, which enhances the data-driven approach to reveal EEG/EOG characteristics in contrast to describing movement artefacts. Epochs included in the building process were randomly chosen across the entire night to eliminate ultradian differences in EEG/EOG characteristics. AASM stages were used to ensure equal representation of all stages during the building process and elimination of the effect of overrepresentation of certain sleep stages, especially N2. Visual evaluation of the number of epochs used to build the model showed a tendency of increased performance with increased training epochs. There were a limited number of wake and N1 epochs present, which lead to a decreased wake detection and more detailed description of deep sleep if an unbalanced amount of building data was used (building data dominated by REM and slow wave sleep). The balanced sleep stage distribution used to generate the model may be interpret as conflicting with the data-driven approach. However, equal AASM sleep stage representation was used because certain states e.g. wakefulness are vital for a sleep classifier.

The model was stabilized and correct topic assignments obtained by introducing a base term matrix. The base term matrix included the term matrix defined in Eq. (4) derived for each subject in the base term matrix group (defined in Table 1). Applying LDA to a test subject was done by merging the specific test subject's term matrix with the base term matrix and using the trained topic and word Dirichlet distributions in LDA. This produced the mixture of sleep topics.

LDA is a complex topic model and other and simpler models were considered. Probabilistic latent semantic indexing was considered but discarded due to the uncertainty of how to apply the model to test data (Blei et al., 2003). Non-negative matrix factorization was investigated because it is simple to apply to test data. However, non-negative matrix factorization described sleep in less detail and was therefore found not fit for use.

2.2.3. Optimizing the sleep model parameters

The parameters used and stated above were found through testing a variety of settings. To minimize the deviances caused by the approximation in LDA, 10 models were generated at each setting using the 10 training control subjects from Table 1. Multiclass RBF kernel support vector machines (mSVM) using a one-versus-all approach were applied to compare the model with the manual scoring. The algorithm was implemented in MATLAB using the built-in functions *svmtrain* and *svmclassify*. Sequential minimal optimization was used and the kernel parameters tuned using accuracy as the optimization parameter. mSVM enables comparing the five AASM stages with a different number of sleep topics. Testing was carried out on a per-subject basis to reduce the effect of sleep differences between subjects and the low hypnogram interrater-reliability.

Table 2
Parameters tested and selected parameters.

Parameter	Tested	Initial	Selected
FFT window (s)	0.5, 1, 2	1	1
Number of SAX boundaries	2, 3, 4, 5	4	4
Word length	1, 2, 3, 4	3	3
Number of topics	3, 4, 5, 6, 7	5	6
EEG electrodes	C3, F3, O1, C3/F3, C3/O1, F3/O1, C3/F3/O1	C3	C3/O1
EOG boundaries	No EOG, 0.7, 0.8, 0.9	No EOG	0.7

The testing strategy resembled cross-validation. To make up for the very unequal sleep stage distribution an equal number of epochs from each AASM stage was extracted from each subject (the least dominating sleep stage determined the number of epochs) and randomly split into two training datasets. Two mSVMs were trained, one on each training dataset, and applied to the remaining epochs for that subject. The number of epochs, where the mSVM prediction was in accordance with the manually scored hypnogram, was counted and divided with the total number of epochs to compare accuracies. For each subject the accuracy was calculated as the mean of the two mSVMs (mean and standard deviation of the inaccuracies between the two mSVMs for the 10 training control subjects were $2.6 \pm 2.9\%$). Mean accuracy across subjects was calculated for each parameter setting and the parameters were compared by one-way analysis of variance at a 95% significance level. If performing equally, the simpler model was selected. One parameter was tested at a time and the testing initialized using EEG channel C3-A2, no EOG channel, single-sided amplitude, four SAX boundaries and word length three symbols. The parameters and order of testing are shown in Table 2 (testing was carried out from top to bottom, starting with the initial parameters listed).

2.2.4. Interpretation of sleep topics

Interpretation of sleep topics was carried out by analyzing the word probabilities contained in β_{kj} in LDA. The marginal distribution of a word, $p(w_j)$, describes the prevalence of word j in the sleep recording and can be derived by

$$p(w_j) = \sum_{k=1}^{K} p(w_{jk}|z_k), \tag{5}$$

where $k = 1, \ldots, K$ is the sleep topic and z is the latent topic variable. A word may frequently appear across all topics, making it not powerful for a certain topic. To evaluate a word's impact on a topic, WI_{jk}, the ratio between the conditioned and marginal distribution can be expressed as

$$WI_{jk} = \log_2 \left[\frac{p(w_{jk}|z_k)}{p(w_j)} \right]. \tag{6}$$

By ranking WI_{jk} in descending order and calculating the cumulative sum, C_k, the complexity of each topic can be investigated. This is derived by

$$C_k(i) = \sum_{i=1}^{V} \sum_{r=1}^{i} R_{rk}, \tag{7}$$

where $i = 1, 2, \ldots, V$ is the number of included words and R_{rk} is WI_{jk} ranked in descending order for each topic.

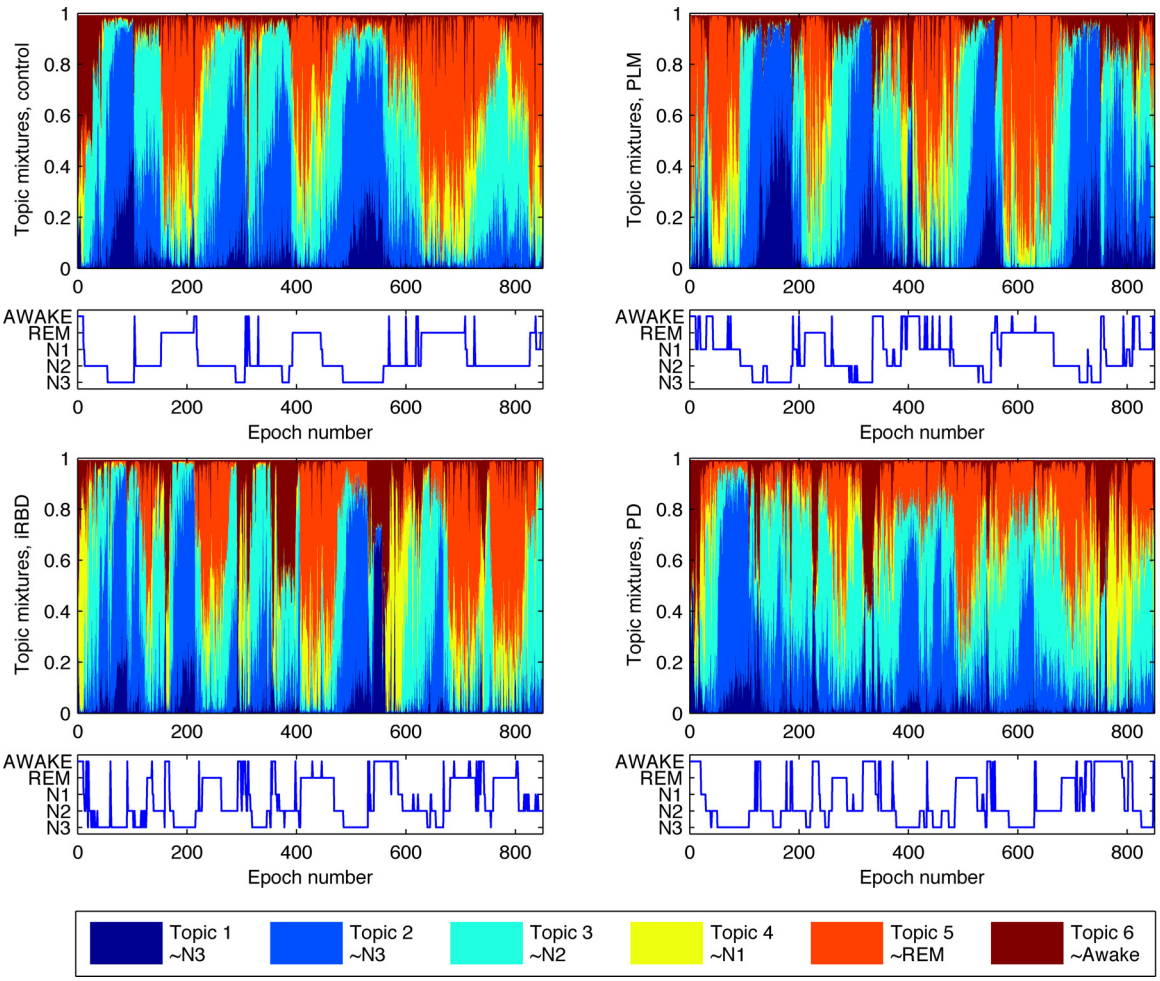

Fig. 2. Examples of sleep topic diagrams and corresponding manually scored hypnograms. Top row: control subject and PLM patient (non-NDD subjects), bottom row: iRBD and PD patient (NDD patients). An epoch is 30 s and represented as a colored vertical bin where the amount of each color is the individual sleep topic probabilities $p_k(n)$.

3. Results

The mixture of sleep topics gives a detailed description of sleep, and optimization of the model showed that sleep contains six topics. Parameter testing showed that inclusion of EOG significantly improved the sleep classification, especially classification of REM sleep epochs. Inclusion of F3-A2 did not improve the model performance.

The model shows that the probability for each sleep topic changes smoothly during state transitions and clear visual similarities between the sleep topic diagrams and manually scored hypnograms are present. Fig. 2 shows examples of the model applied to a subject from each group and the AASM stage and sleep topic concordances are summarized in Table 3. Interpretation of sleep topics showed similarities with AASM stages (Iber et al., 2007), which in general contain:

W: EEG 8–13 Hz, EOG irregular/REMs
REM sleep: EEG 2–6 Hz, EOG REMs
N1: EEG 4–7 Hz, EOG slow eye movements (SEMs)
N2: 11–16 Hz, EOG: SEMs/abscent
N3: 0.25–2 Hz, EOG: SEMs/abscent

The impacting EEG/EOG characteristics in each sleep topic were evaluated using the word Dirichlet distribution, β_{jk}, obtained during training. Evaluating the topic complexity by C_k showed that topic 3 is the most complex sleep topic followed by 2, 5, 4, 6 and 1 in descending order, see Fig. 3. This indicates that N2 spans most EEG/EOG characteristics. This is in accordance with the AASM standard, where N2 contains many micro-sleep events and is the most prevalent sleep stage. As seen in Fig. 3 the initial gradient is similar for all topics, but the simpler topics reach plateau earlier and have less impacting words (more "background" words).

The word-topic analysis illustrates the complexity of the model and shows that not a single but a large amount of words define each sleep topic. Epochs with one highly dominating topic were

Table 3
Concordances between AASM sleep stages and sleep topics.

Topic no./color (≈ AASM)	EEG: clinical frequency band content EOG: power and eye-correlation
6 / red (≈W)	Low δ. Scattered θ. High α (β minor impact). Scattered power, high anti-correlation.
5 / orange (≈REM)	Low: δ, θ, α, β (β most impact). Scattered power, high anti-correlation.
4 / yellow (≈N1)	Low: δ, θ. Scattered α, high β. Low power, medium anti-correlation.
3 / turquoise (≈N2)	Medium δ. Scattered: α, β. Low power, medium correlation.
2 / light blue (≈N3)	High: δ, θ. Sporadic: α, β. High power, high correlation.
1 / dark blue (≈N3)	High: δ, θ. High correlation.

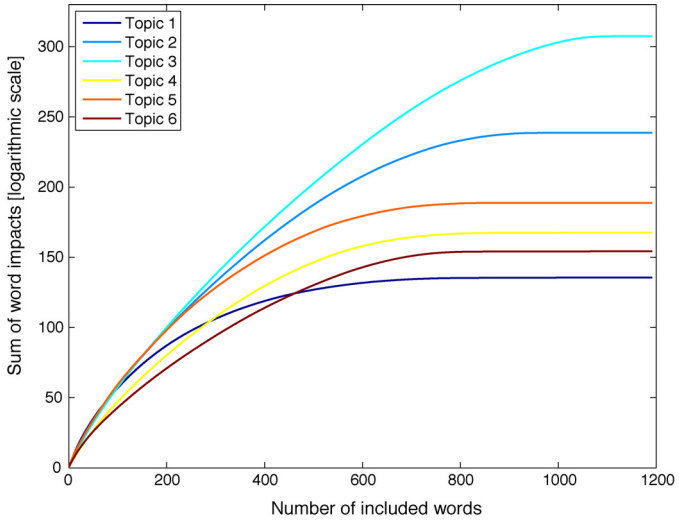

Fig. 3. Cumulative sum of word impacts transformed by 2^{C_k}. Each graph illustrates the sum of word impacts when including more words, starting with the most powerful word. Topic 1 is the simplest topic with few impacting words and a small total impact. Topic 3 is the most complex topic with many impacting words and a large total impact.

investigated visually, and Fig. 4 shows examples of the four PSG signals in these epochs. The figure supports the concordances between topics and AASM sleep stages stated in Table 3.

The final model was validated on a per-subject basis using mSVM (explained in Section 2.2.3) and 20-fold cross validation. The manual scoring was used to ensure representation of all AASM sleep stages in all folds. Accuracy was calculated for each subject on a by-epoch basis and Cohen's kappa was derived to take the agreement occurring by chance into account. The mean accuracy and Cohen's kappa were calculated for all subjects in the validation and base term matrix data, for each subject group and overall. The validation data reached an overall accuracy of 68.3 ± 7.44 (% $\mu \pm \sigma$) and Cohen's kappa of 0.67 ± 0.08 ($\mu \pm \sigma$) and the base term matrix reached an overall accuracy of 68.8 ± 9.02 and Cohen's kappa of 0.67 ± 0.09. For comparison, well performing automatic sleep stage classifiers have accuracies 71–91% (Ahmed and Tafreshi, 2009). Table 4 contains the group specific accuracies for the individual AASM sleep stages and all NREM stages combined. Visually the topic diagrams showed that no specific combination of topics predict wake periods, and the validation supports this observation by low wake prediction accuracy. Further, the validation shows that N3 and REM sleep epochs are called more accurately compared to wakefulness and N1.

It should be mentioned that the accuracy is highly dependent on the manually scored hypnogram. Further, the accuracy will decrease if the transition inflection points differ inconsistently between the manual scoring and sleep topic model. The document exchangeability in LDA and the fact that manual scoring takes prior epochs into account may introduce more differences.

4. Discussion

4.1. Automatic sleep classifier

The proposed model carries out automatic generalized sleep classification and reveals latent sleep states using EEG and EOG. A detailed description of sleep is given with continuous state switching using the clinical frequency bands as the only biased knowledge. In order to evaluate the model against a gold standard, the sleep model was optimized and compared using the manual scored hypnograms. The optimized model used six sleep topics which indicates that sleep has six latent states. The modelling approach may be interpreted as partially data-driven because equal representaion of the AASM sleep stages was used for optimization. However, this was done to benchmark the novel sleep classification with the golden standard's sleep staging and to ensure representation of topics describing certain states e.g. wakefulness (visual evaluation of models built on all-night data showed a decreased wake description and more detailed description of deep sleep). Sleep topic 1 and 2 corresponds to N3 in the AASM standard but the content of topic 1 is in accordance with the Rechtschaffen and Kales manual scoring standard, which defines

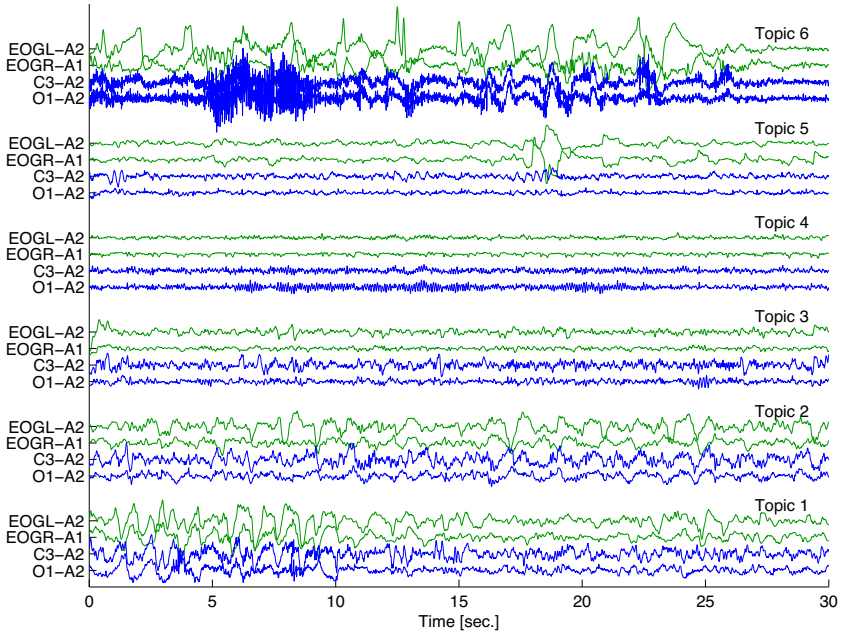

Fig. 4. Examples of epochs containing one topic with high certainty. Dominating topic certainty in the respective epochs are: topic 6 $p_6 = 46.8\%$, topic 5 $p_5 = 88.9\%$, topic 4 $p_4 = 68.9\%$, topic 3 $p_3 = 82.9\%$, topic 2 $p_2 = 86.0\%$ and topic 1 $p_1 = 34.7\%$.

Table 4
Model accuracy (% $\mu \pm \sigma$) and Cohen's kappa coefficient ($\mu \pm \sigma$) using mSVM.

	Control		PLM		iRBD		PD		Overall	
	Acc.	κ	Acc.	κ	Acc.	κ	Acc.	κ	Acc.	κ
Base term matrix										
Total	73.5 ± 4.62	0.72 ± 0.05	72.6 ± 3.06	0.71 ± 0.03	68.1 ± 8.68	0.67 ± 0.09	60.9 ± 12.1	0.59 ± 0.13	68.8 ± 9.02	0.67 ± 0.09
W	47.4 ± 34.7	0.45 ± 0.36	55.6 ± 29.8	0.54 ± 0.31	50.0 ± 26.5	0.48 ± 0.28	62.2 ± 19.9	0.61 ± 0.21	53.8 ± 27.4	0.52 ± 0.29
REM	79.9 ± 5.73	0.79 ± 0.06	76.4 ± 9.10	0.75 ± 0.09	55.4 ± 19.9	0.54 ± 0.21	45.4 ± 18.3	0.43 ± 0.19	64.3 ± 19.9	0.63 ± 0.21
NREM	71.4 ± 7.70	0.68 ± 0.09	71.7 ± 3.21	0.68 ± 0.04	70.7 ± 8.84	0.67 ± 0.10	62.1 ± 11.0	0.57 ± 0.12	69.0 ± 8.73	0.65 ± 0.10
N1	15.0 ± 9.44	0.12 ± 0.10	17.8 ± 11.7	0.14 ± 0.12	28.2 ± 24.3	0.25 ± 0.25	32.0 ± 20.1	0.29 ± 0.21	23.2 ± 18.0	0.20 ± 0.19
N2	75.1 ± 9.93	0.74 ± 0.10	74.3 ± 9.53	0.73 ± 0.10	72.4 ± 11.7	0.71 ± 0.12	65.3 ± 11.9	0.64 ± 0.12	71.8 ± 11.0	0.71 ± 0.11
N3	81.1 ± 8.26	0.80 ± 0.09	81.4 ± 14.2	0.81 ± 0.15	72.8 ± 24.0	0.72 ± 0.25	75.6 ± 17.8	0.75 ± 0.19	78.0 ± 16.1	0.77 ± 0.17
Validation										
Total	69.0 ± 4.62	0.68 ± 0.05	70.1 ± 5.10	0.69 ± 0.05	67.2 ± 8.30	0.66 ± 0.09	67.7 ± 9.07	0.66 ± 0.09	68.3 ± 7.44	0.67 ± 0.08
W	47.8 ± 26.2	0.46 ± 0.27	56.1 ± 29.7	0.54 ± 0.31	57.5 ± 25.4	0.56 ± 0.26	68.8 ± 23.2	0.67 ± 0.24	59.3 ± 26.1	0.58 ± 0.27
REM	64.7 ± 16.0	0.63 ± 0.17	64.4 ± 17.4	0.63 ± 0.18	60.3 ± 16.2	0.59 ± 0.17	38.9 ± 23.0	0.36 ± 0.24	54.8 ± 21.7	0.52 ± 0.23
NREM	69.8 ± 5.51	0.7 ± 0.06	70.2 ± 5.16	0.70 ± 0.05	67.4 ± 9.20	0.67 ± 0.09	67.2 ± 8.56	0.67 ± 0.09	68.3 ± 7.68	0.64 ± 0.09
N1	29.2 ± 11.6	0.26 ± 0.12	22.0 ± 18.7	0.19 ± 0.19	25.1 ± 14.0	0.22 ± 0.15	22.1 ± 16.4	0.19 ± 0.17	24.3 ± 15.3	0.21 ± 0.16
N2	73.1 ± 8.51	0.72 ± 0.09	72.2 ± 8.54	0.71 ± 0.09	68.0 ± 11.5	0.67 ± 0.12	67.6 ± 12.7	0.66 ± 0.13	69.6 ± 11.0	0.68 ± 0.11
N3	80.2 ± 14.3	0.79 ± 0.15	82.0 ± 14.1	0.81 ± 0.15	77.7 ± 20.3	0.77 ± 0.21	74.7 ± 21.7	0.74 ± 0.23	78.1 ± 18.4	0.77 ± 0.19

an additional high delta activity sleep stage S4 (Rechtschaffen and Kales, 1968).

The probability of each sleep stage changes gradually within NREM stages whereas the transitions involving wakefulness and REM sleep seem more abrupt. The physiological understanding of sleep control has been described by the flip–flop model (Saper et al., 2010), which later was approached by Luppi et al. (2011) who included more integrated neural pathways. The topic diagrams are in accordance with both theories and an extension of this study's model may contribute to the research in sleep control.

Clear visual concordances between the AASM sleep stages and sleep topics as well as agreements between the frequency content and the eye movements were found. This indicates similar sleep scoring, however, the sleep topics are not equivalent to the manually scored sleep stages due to the model's data-driven approach. The simpler topics, topic 4 and 6, rarely reach high certainties and visual inspection showed that both topics often contain minor movement artefacts. Epochs with high certainty of topic 4 resemble N1 but also contain either topic 3, 5 or 6. The corresponding PSG signals express either EEG characteristic of N2 or alpha activity with anti-correlated EOG. This indicates that topic 4 is a brief transition state and the concurrent dominating topic in the respective epoch may contain information about the previous/following sleep state. Likewise, epochs with high certainty of topic 6 contain either topic 1 or topic 3. The PSG signals include large amplitude EOG or alpha activity and anti-correlated EOG but no N2 characteristics in the EEG.

It may be beneficial to determine the words describing different micro-sleep events such as sleep spindles and K-complexes. In future work a throughout analysis of different morphology characteristics may reveal information about micro-sleep events and their correlations with sleep states and subject groups.

The model does not correct for eventual changes in electrode impedance but updating the SAX boundaries may handle this issue. However, the distribution of the amplitude spectrum follows an ultradian pattern and by updating the SAX boundaries, the model may eliminate the temporal changes throughout the night. In addition, the boundary update would be determined by the sleep stage in the "update area" and thereby introduce uncertainties due to a changing word definition throughout the night.

Inclusion of electromyography (EMG) is likely to improve the model's performance when applied to control/PLM subjects, especially through improved REM sleep detection. This study's main focus was to develop a general model and the changed muscle activity in iRBD/PD patients during REM sleep would make the model not suited for this patient group. Therefore the model does not include EMG but relies on EOG for REM sleep detection, and the results showed a general good performance classifying REM sleep epochs.

The model performs in the lower range of automatic sleep classifiers but the data-driven approach, similar scoring and large degree of detail indicate that this study's sleep classifier has potential as a novel method to analyze sleep.

4.2. W/REM/NREM detection

Clinically, a coarse picture of a patient's sleep pattern can be achieved by looking at a sleep recording in the W/REM/NREM domain. Even though the model presented does not include a frontal EEG channel known to be good at capturing slow wave activity, the model showed good performance predicting N3. The model is complex and multiple reasons may explain why the frontal EEG channel does not improve model performance: (1) the large amplitude slow waves may be captured by the EOG and/or C3-A2 channel, (2) the large slow wave amplitudes may make the EOG signals correlated and (3) the lack of impacting words characteristic for the remaining stages will decrease the probability of other sleep topics.

It is a common problem for automatic classifiers to separate W, REM and N1 due to similar EEG and/or EOG characteristics. Inclusion of EOG in this model ensured good prediction of REM sleep and distinguishability of N1 and REM due to slow versus rapid eye movements described by the cross-correlation measure. However, the N1 dominating topic was sporadically present during wakefulness and REM sleep, which may be caused by similar EEG/EOG characteristics during W, REM and N1. The model predicts wakefulness and N1 poorly and no certain combination of topics describes wakefulness accurately. Future work may include optimizing certainty thresholds for topics related to wakefulness in order to predict wake epochs visually (the topics of interest are 6, 4 and 1). Otherwise, a simple W/REM/NREM topic diagram may be produced by eliminating the N1 topic and summing the remaining NREM topic probabilities. Inclusion of EMG may improve prediction of wakefulness but EMG was excluded due to the motivation to build a model applicable on patients with iRBD/PD.

4.3. General application of sleep classifier

The AASM standard is criticized for being difficult to apply on patients. In specific, patients with motor abnormalities during sleep do not fit into the scoring standard because lowered muscle activity is required to score REM sleep. To accommodate this, the model only includes EEG and EOG. The challenges of sleep scoring in

iRBD/PD patients are reflected by the large accuracy variations in all sleep stages and the low mean accuracy in REM sleep (Table 4).

The iRBD/PD patient's individual disease progress may introduce different EEG and EOG characteristics. Between-subject differences in the level of the spectral and eye correlation values were eliminated by SAX and due to the hypothesis, that control subjects express similar and "true" sleep characteristics, training was carried out only on control subjects. Visual inspection of the topic diagrams indicated more state transitions in iRBD/PD which may be caused by notation of too few state switches in the manual AASM scorings (AASM requires certain events to notate a stage switch Iber et al., 2007) or this study's sleep model is sensitive to noise.

The examples in Fig. 2 indicate more unstructured transitions and less systematic sleep architecture in iRBD/PD subjects compared to the control/PLM subjects. The NREM and REM bouts seem longer in control/PLM subjects in contrast to iRBD/PD patients, who seem to fluctate in and out of sleep stages. This was a general pattern in patients and may be caused by the brainstem degeneration in iRBD/PD patients. The control/PLM and iRBD/PD groups are not well balanced for age or sex and it cannot be rejected that some of the qualitative differences may be due to factors other than disease status.

The neurodegeneration may introduce deviations between the EEG/EOG characteristics in control/PLM and iRBD/PD subjects. In general, iRBD/PD subjects show less certainty because the model is trained to fit control subjects and this reflects diverse EEG/EOG characteristics. Training a topic model for each disease group may improve the sleep classification in the individual subject groups, because each model will be adapted to the disease specific EEG/EOG phenotypes. However, this will make the model not general and hinder direct comparison between groups. The EEG and EOG characteristics between the subject groups may be changed to a degree, where training a new topic model will reveal topics that describe different sleep characteristics. This study aimed at developing a general sleep classification model but in future work it would be interesting to generate disease specific models at different disease stages and thereby reveal specific EEG/EOG phenotypes. This may indicate potential biomarkers for the different stages in the disease progress.

In Christensen et al. (submitted for publication) a similar data-driven approach was used on sleep EEG and EOG separately in order to investigate the latent diversities between control/PLM and iRBD/PD. Features aimed at reflecting the changes caused by neurodegeneration were designed and the study reported differences between control/PLM and iRBD/PD subjects. This indicates the large potential of this study's innovative approach, which may lead to new findings as well as supportive conclusions on neurological aspects.

5. Conclusion

This study proposes a generalized automatic sleep classifier developed using a data-driven approach. The model only used frequency characteristics from EEG and EOG as well as eye correlation and this produced a detailed description of sleep expressed as mixture of sleep topics. The model used six topics, corresponding to six latent sleep states, and performed similarly across the groups with an overall subject-specific accuracy of (%) 68.3 ± 7.44 and Cohen's kappa 0.67 ± 0.08. Physiological correlations to the AASM standard was obtained as well as smooth state transitions which is in accordance with the physiological understanding of sleep control (Saper et al., 2010; Luppi et al., 2011) but not reflected in the AASM standard. Visually, iRBD/PD subjects showed more unstable sleep states and less sleep topic certainty indicating different EEG/EOG characteristics in these patients. Analyzing sleep as mixture of sleep topics may contribute to the investigation of iRBD/PD and sleep disorders.

Conflict of interest

No conflicts of interest.

References

Ahmed B, Tafreshi R. Advances in automatic sleep analysis. In: IFMBE proc.; 2009. p. 422–6, http://dx.doi.org/10.1007/978-3-540-92841-6_104.

Blei DM, Ng AY, Jordan MI. Latent Dirichlet allocation. J Mach Learn Res 2003;3:993–1022.

Christensen JAE, Koch H, Frandsen R, Kempfner J, Arvastson L, Christensen SR, et al. Classification of iRBD and Parkinson's disease patients based on eye movements during sleep. In: Conf proc IEEE eng med biol soc; 2013. p. 441–4, http://dx.doi.org/10.1109/EMBC.2013.6609531.

Christensen JAE, Zoetmulder M, Koch H, Frandsen R, Arvastson L, Christensen SR, Jennum P, Sorensen HBD, et al. Data-driven modeling of sleep EEG and EOG reveals characteristics indicative of pre-Parkinson's and Parkinson's disease. J Neurosci Methods 2014, submitted for publication.

Crowley K. Sleep and sleep disorders in older adults. Neuropsychol Rev 2011;21(1):41–53, http://dx.doi.org/10.1007/s11065-010-9154-6.

Dyken ME, Afifi AK, Lin-Dyken DC. Sleep-related problems in neurologic diseases. Chest 2012;141(2):528–44, http://dx.doi.org/10.1378/chest. 11-0773.

Iber C, Ancoli-Israel S, Chesson A, Quan SF. The AASM manual for the scoring of sleep and associated events: rules, terminology, and technical specifications. Westchester: American Academy of Sleep Medicine; 2007.

Koch H, Christensen JAE, Frandsen R, Arvastson L, Christensen SR, Sorensen HBD, et al. Classification of iRBD and Parkinson's patients using a general data-driven sleep staging model built on EEG. In: Conf proc IEEE eng med biol soc; 2013. p. 4275–8, http://dx.doi.org/10.1109/EMBC.2013.6610490.

Lin J, Keogh E, Lonardi S, Chiu B. A symbolic representation of time series, with implications for streaming algorithms. In: Conf proc DMKD; 2003. p. 2–11, http://dx.doi.org/10.1145/882082.882086.

Luppi P-H, Clément O, Sapin E, Gervasoni D, Peyron C, Léger L, et al. The neuronal network responsible for paradoxical sleep and its dysfunctions causing narcolepsy and rapid eye movement (REM) behavior disorder. Sleep Med Rev 2011;15(3):153–63, http://dx.doi.org/10.1016/j.smrv.2010.08.002.

Micallef J, Rey M, Eusebio A, Audebert C, Rouby F, Jouve E, et al. Antiparkinsonian drug-induced sleepiness: a double-blind placebo-controlled study of L-dopa, bromocriptine and pramipexole in healthy subjects. Brit J Clin Pharmacol 2009;67(3):333–40, http://dx.doi.org/10.1111/j. 1365-2125.2008.03310.x.

Rechtschaffen A, Kales A. A manual of standardized terminology, techniques and scoring system for sleep stages of human subjects. Los Angeles: University of California; 1968.

Santamaria J, Högl B, Trenkwalder C, Bliwise D. Scoring sleep in neurological patients: the need for specific considerations. Sleep 2011;34(10):1283–4, http://dx.doi.org/10.5665/sleep.1256.

Saper CB, Fuller PM, Pedersen NP, Lu J, Scammell TE. Sleep state switching. Neuron 2010;68(6):1023–42, http://dx.doi.org/10.1016/j.neuron.2010.11.032.

Verbeek J, Matlab toolbox: Latent Dirichlet Allocation toolbox, 2006. http://lear.inrialpes.fr_verbeek/software.php [accessed 10.23.12].