A reliable probabilistic sleep stager based on a single EEG signal

Arthur Flexer [a,b,*,1], Georg Gruber [a], Georg Dorffner [a,c]

[a] *The Austrian Research Institute for Artificial Intelligence, Freyung 6/6, A-1010 Vienna, Austria* [2]

[b] *Swartz Center for Computational Neuroscience, Institute for Neural Computation, University of California San Diego 0961, La Jolla, CA 92093-0961, USA*

[c] *Department of Medical Cybernetics and Artificial Intelligence, Medical University of Vienna, Freyung 6/2, A-1010 Vienna, Austria*

Abstract

Objective: We developed a probabilistic continuous sleep stager based on Hidden Markov Models using only a single EEG signal. It offers the advantage of being objective by not relying on human scorers, having much finer temporal resolution (one second instead of 30 seconds), and being based on solid probabilistic principles rather than a predefined set of rules (Rechtschaffen & Kales)

Methods and Material: 68 whole night sleep recordings from two different sleep labs are analysed using Gaussian Observation Hidden Markov Models.

Results: Our unsupervised approach detects the cornerstones of human sleep (wakefulness, deep and rem sleep) with around 80% accuracy based on data from a single EEG channel. There are some difficulties in generalizing results across sleep labs.

Conclusion: Using data from a single electrode is sufficient for reliable continuous sleep staging. Sleep recordings from different sleep labs are not directly comparable. Training of separate models for the sleep labs is necessary.

Key words: Time series processing, Sleep Analysis, Hidden Markov Models, EEG

1 Introduction

Since every human needs to sleep, sleep is not a matter of choice. The need for sleep varies between humans with seven to nine hours being optimal for most of us. Disturbed sleep has a decisive influence on health, behavior, and mood (see e.g. [1]). Sleep loss can be caused by a number of factors of which not all are subject to an individual's choice (e.g. long hours of work, shift work, stress, family responsibilities, illness including sleep disturbances). A recent study [2], consistent with other national studies, reported about one-third of US Americans had some type of sleep problem. Lack of sleep reduces reaction time, vigilance, alterness and concentration as well as the quality of decision making and learning. Sleep recordings are of vital importance for the analysis,

* Corresponding author. Tel.: 001 - 858 - 458 1927, Fax.: 001 - 858 - 458 1847.
Email addresses: arthur@sccn.ucsd.edu, arthur@oefai.at (Arthur Flexer), georgg@oefai.at (Georg Gruber), georg@ai.univie.ac.at (Georg Dorffner).

[1] Arthur Flexer is supported by an Erwin Schrödinger Fellowship provided by the Austrian Science Fund (FWF), project J2221-N04.

[2] The recordings for this work were done within the BIOMED-2 BMH4-CT97-2040 project SIESTA, funded by the EC DG XII. The Austrian Research Institute for Artificial Intelligence is supported by the Austrian Federal Ministry of Education, Science and Culture and the Austrian Federal Ministry for Transport, Innovation and Technology.

diagnosis and treatment of various kinds of sleep disturbances.

Sleep staging is one of the most important steps in sleep analysis and is usually done using the traditional Rechtschaffen & Kales [3] (R&K) rules. It is a very time consuming task consisting of classifying all 30 second pieces of an approximately 8 hour recording into one of six sleep stages: wake, rapid eye movement (rem) sleep, S1 (light sleep), S2, S3, S4 (deep sleep). A sleep recording is made with a minimum setting of four channels: electro-encephalogram (EEG) from electrodes C3 and C4, electro-myogram (EMG) and electro-oculogram (EOG). In order to classify each 30 second segment of sleep, the human scorer looks for defined patterns of waveforms in the EEG, for rapid eye movements in the EOG and for EMG level. Quite some work has already been done in trying to replicate R&K sleep staging with automatic methods (see [4] for an overview) including neural networks (e.g. [5] or [6]).

There is however a considerable dissatisfaction within the sleep research community concerning the very basics of R&K sleep staging: R&K is based on a predefined set of rules leaving much room for subjective interpretation; it is a very time consuming and tedious task; it is designed for young normal subjects only; it has a low 30 second temporal resolution; it is defined in terms of six stages neglecting the micro-structure of sleep; it cannot be automized reliably due to the large inter-scorer variability and insufficient rules for staging.

Our aim is to built an automatic continuous sleep stager, based on probabilistic principles which overcomes the known drawbacks of traditional R&K sleep staging. In previous efforts [7] we tried to find a new description of human sleep which is based on the comparably unambiguous “extreme” cornerstones of traditional sleep staging rather than merely automating and replicating

R&K sleep staging. We used a Hidden Markov Model (HMM) to produce three continuous probability traces P(wake), P(deep) and P(rem) with a one second resolution. The newly obtained continuous sleep profiles were compared to traditional R&K scoring. The two "extreme" R&K stages "wake" and "deep" could be detected very satisfactory with an accuracy of above 80%. However, we had great problems discriminating rem sleep from wakefulness and stages 1, 2 and even 3. The mean accuracy for detection of rem sleep was as low as 26%.

Reviewing our old results lead us to the hypothesis that the problems we encountered so far might be due to problems in the data base. This paper reports about the training of separate models for individual sleep labs instead of generalizing to data from diverse sleep labs. Concentrating on data from separate sleep labs enabled us to detect all three cornerstones of human sleep (wakefulness, deep and rem sleep) with around 80% accuracy based on data from a single EEG channel. This result could be obtained for data from the sleep lab for which we already achieved the best results so far. Experiments with data from the worst sleep lab so far cannot be improved by training a separate model. We conclude that our previous problem of detecting rem sleep is not a general problem of our method but rather due to insufficient information in the data for some of the sleep labs. Earlier versions of this manuscript have been published at two conferences [8] [9].

2 Data

In our previous efforts [7] our data base consisted of nine whole night sleep recordings from a group of healthy adults (total sleep time = 70.5h, age ranges

from 20 to 60, 5 females and 4 males). We used reflection coefficients and stochastic complexity computed for EEG channels C3 and C4 and a measure of EMG level (altogether five features) for analysis with an HMM. The nine recordings had been recorded in five different European sleep laboratories during the SIESTA project [1]. The results were as described above: satisfactory detection of wakefulness and deep sleep, but accuracy as low as 26% for rem sleep.

The SIESTA project resulted in yet another sleep stager [10], which is based on a semi-supervised approach using Gaussian kernels plus sensor fusion to fuse information from different channels. Without giving any further detail it should suffice to say that its outputs are again three continuous probability traces. This sleep stager has been evaluated on data from eight different sleep labs. Plotting the mean entropy of the probability traces per sleep lab reveals clear lab effects (see Fig. 1). The mean entropy $\bar{E}_L$ for sleep lab L is defined as:

$$\bar{E}_L = \bar{P}_L(wake)\log_2\left(\frac{1}{\bar{P}_L(wake)}\right) + \bar{P}_L(rem)\log_2\left(\frac{1}{\bar{P}_L(rem)}\right) + \bar{P}_L(deep)\log_2\left(\frac{1}{\bar{P}_L(deep)}\right) \quad (1)$$

$$\bar{P}_L(wake) = \frac{1}{T}\sum_{t=1}^{T} P_t(wake) \quad (2)$$

where $\bar{P}_L(wake)$ is the average of the probability trace for wakefulness with $t = 1, \ldots, T$ being a time index across all data from sleep lab L. Computation of $\bar{P}_L(rem)$ and $\bar{P}_L(deep)$ follows the same principles. Since entropy is largest for equally probable variables, high entropy values indicate that all three prob-

ability traces stay around .33 most of the time and show only little variation. Obviously, the sleep stager seems to work quite well for some of the sleep labs and quite bad for others.

To gather further evidence for this hypothesis, we decided to concentrate our analysis on data from single sleep labs. Sleep lab A is the one with best results so far according to the entropy plot (left most in Fig. 1) and sleep lab B the worst so far (second from the right in Fig. 1).

Our new data base A consists of 40 whole night sleep recordings from a group of healthy adults (total sleep time = 326.4h, age ranges from 22 to 86, 22 females and 18 males). We use only EEG channel C3 for further analysis. Given the large amount of data we decided to use only one training and test set for further analysis rather than running a full cross validation. In order to avoid the effect of the particular selections of data we matched both the training and the test set for sex and age. This should ensure that both data sets are representative for the population under examination. Twenty recordings are used to train our automatic sleep stager (training set A), twenty are set aside to evaluate it (test set A).

Data base B consists of 28 whole night sleep recordings from a group of healthy adults (total sleep time = 224.2h, age ranges from 22 to 95, 14 females and 14 males). We use only EEG channel C3 for further analysis. Fourteen recordings are used to train our automatic sleep stager (training set B), fourteen are set aside to evaluate it (test set B). Both sets are again matched for sex and age.

The EEG channels were recorded versus the contralateral mastoid electrode (M1 or M2) with a sampling rate of 200 Hz (high-pass filter: 0.3 Hz, low-pass filter: 85 Hz) in lab A and with a sampling rate of 256 Hz (high-pass filter:

0.1 Hz, low-pass filter: 75 Hz) in lab B.

Data from both labs has been re-sampled to 100 Hz before any further processing to ensure comparable results.

3 Methods

HMMs [11] allow analysis of non-stationary multi-variate time series by modeling, both, the probability density functions of locally stationary multi-variate data and the transition probabilities between these stable states. In the context of sleep analysis, the locally stable states can be thought of as sleep stages. Our choice to use an HMM allows us to exploit the probabilistic dependence of successive sleep stages. After all, transitions between some of the sleep stages are much more likely than between others. For e.g. a transition from stage "wake" directly to stage "deep" is quite unlikely. A transition from stage S3 to stage "deep" (S4) is quite common.

Following the classical text by Rabiner and Juang [11], an HMM can be characterized as having a finite number of N states Q:

$$Q = \{q_1, q_2, \ldots, q_N\} \tag{3}$$

A new state q_j is entered based upon a transition probability distribution A which depends on the previous state (the Markovian property):

$$A = \{a_{ij}\}, a_{ij} = P(q_j(t+1) \mid q_i(t)) \tag{4}$$

where $t = 1, \ldots, T$ is a time index with T being the length of the observation sequence. After each transition an observation output symbol is produced according to a probability distribution B which depends on the current state. Although the classical HMM uses a set of discrete symbols as observation output, Rabiner and Juang [11] already discuss the extension to continuous observation symbols. Such a Gaussian Observation HMM (GOHMM) [12] has already been proposed as a model for EEG analysis and the model we now describe is the same we used for our previous work on sleep staging [7]. We use a GOHMM where the observation symbol probability distribution for state j is given by

$$B = \{b_j(x)\}, b_j(x) = \mathcal{N}[x, \mu_j, U_j] \tag{5}$$

where $\mathcal{N}$ is the normal density and μ_j and U_j are the mean vector and covariance matrix associated with state j. Please note that this a simple version of the Gaussian M-component mixture given in [11] with M equal one. The Expectation-Maximization (EM) algorithm is used to train the GOHMM thereby estimating the parameter sets A and B as well as the μ_j and U_j. Viterbi decoding is used to identify most likely state sequences corresponding to a particular time series and enables the computation of the probabilities of being in any of the N states at each point in time. Full details of the algorithms can be found in [11].

A GOHMM is defined over the first reflection coefficient of the EEG channel at C3. Reflection coefficients are the coefficients of the order recursive representation of autoregressive (AR) processes [13]. We used a lattice filter representation of an AR process. The inferred a-posteriori distribution over

model coefficients are the reflection coefficients (see [10] for full detail). The reflection coefficient is computed with a one second resolution.

Our aim is not to replicate R&K scoring but to find a new description of human sleep which is based on the comparably unambiguous "extreme" cornerstones of traditional sleep staging. Since R&K sleep staging is based on a predefined set of rules which leave much room for subjective interpretation there can be considerable disagreement between human scorers analyzing the same sleep recording. Some studies on inter-scorer reliability report overall good agreement of around 90% for all sleep stages [14]. Others [15] report high inter-scorer reliability of around 80% only for stages "wake", "deep" and "rem" and low reliability of 40% to 60% for all other R&K stages (S1, S2, S3). An extensive study [16] on inter-scorer reliability between scorers working on SIESTA data [1] from subjects with different sleep disorders points in the same direction. Overall agreement was 74.6% with agreement being highest for "rem", "wake" and "deep". The neccesity of a continuous sleep analyser working at least on a one second time scale which is able to measure the sleep/wake process on a scale from 0% to 100% as well as to determine the REM-sleep on/off process has been highlighted in the literature [17] before. Related research [18] on continuous sleep staging also confirmed that only the three "extreme" R&K sleep stages "wake", "deep" and "rem" are relatively unambiguous and can be detected most reliably by human scorers.

We therefore model the human sleep as a mixture of three different processes: wakefulness, deep sleep and rem sleep. The other three stages (S1, S2 and S3) can be seen as mixtures of the three basic processes. Consequently, we use a fully connected 3-state GOHMM to build our sleep stager. A separate GOHMM is trained on all available data from the training sets A and B. The

probabilities of being in any of the 3 states are computed at each point in time using the posterior state probabilities, i.e. the probability that an observation x_t came from state k given the observed sequence x:

$$P_t(wake) = P(q_t = k \mid x). \tag{6}$$

Computation of P_t(rem) and P_t(deep) follows the same principles. Thereby we obtain 3 continuous probability plots (P(wake),P(deep),P(rem)) which indicate the amount of wakefulness, rem and deep sleep with a one second resolution [3].

Although it is not the purpose of our approach to replicate R&K sleep staging, we nevertheless like to compare our results to the R&K standard. We construct a classifier as suggested in [18] by computing mean values of P(wake), P(rem) and P(deep) for each of the six human scored R&K stages of all the recordings in a training set:

$$\bar{P}_i(wake) = \frac{1}{\sum_{t=1}^{T} \delta_{it}} \sum_{t=1}^{T} \delta_{it}\ P_t(wake) \tag{7}$$

$$\delta_{it} = 1 \ \ if \ \ R\&K(t) = Stage_i \ \ else \ \ \delta_{it} = 0 \tag{8}$$

where $R\&K(t)$ is the $R\&K$ sleep stage at time point t and $i = 1, \ldots, 6$ indicates the respective R&K sleep stage. Please note that $t = 1, \ldots, T$ is a time index across all data in the training set. Computation of $\bar{P}_i$(rem) and

[3] Please note that these probability plots are being smoothed with a moving average window. The length of the window is 361 seconds.

$\bar{P}_i$(deep) follows the same principles. We therefore compute six sets of mean values $\bar{P}_i$(wake), $\bar{P}_i$(rem) and $\bar{P}_i$(deep). For classification of recordings from a test set we find the minimum Euclidean distance between these mean values and the current probabilities:

$$GOHMM(t) = \min_i \Big[\Big(P_t(wake) - \bar{P}_i(wake) \Big)^2 + \Big(P_t(rem) - \bar{P}_i(rem) \Big)^2 + \Big(P_t(deep) - \bar{P}_i(deep) \Big)^2 \Big]^{\frac{1}{2}} \quad (9)$$

where $GOHMM(t)$ is the sleep stage as classified by the GOHMM at time point t and the minimum is taken over the $i = 1, \ldots, 6$ sets of mean values $\bar{P}_i(wake \mid rem \mid deep)$. $GOHMM(t)$ is computed for all $t = 1, \ldots, T$ with t being a time index across all data in the test set. Please note that whereas setting up and learning the parameters of the GOHMM as well as obtaining the continuous probability plots is totally unsupervised (i.e. it does not use use any label information from the human scorers) the construction of the classifier is not. This last classification step is only needed to allow for comparison with R&K scoring thereby illustrating the merits and problems of our unsupervised approach.

4 Results

The GOHMM trained with data from sleep lab A is evaluated using twenty whole night recordings from test set A. The newly obtained continuous sleep profiles (using Equ. 9) are compared to traditional R&K scoring. R&K scores are taken as true scores and for each sleep stage separately the percentages of GOHMM classification into each of the 6 stages are given in Tab. 1. We

expected that the GOHMM would be able to correctly classify data from the unambiguous "extreme" R&K stages "wake", "rem" and "deep". As can be seen in Tab. 1, this is indeed the case. The accuracies are 79% for wake, 82% for deep sleep and 68% for rem sleep. This is a clear improvement over our previous results which read as 86% (wake), 81% (deep) and 26% (rem). Probability plots plus R&K and HMM scoring for one whole night recording (subject from the test data group) are given in Fig. 2. The overall structure of sleep plus short periods of wakefulness are clearly visible in the probability plots. Note the respective high values of P(wake) and P(deep) aligned with R&K stages wake and S3 or S4 respectively. There still is some mix up between rem sleep and S1 and S2 at the end of the night, which can also be read from Tab. 1.

Three possible explanations come to mind: (i) It is known that detection of rem is difficult from EEG alone. The very definition of rem sleep includes rapid eye movements visible in the EOG and changes in EMG level. In Fig. 3 a boxplot of all training and test data from sleep lab A (first reflection coefficient at C3) is shown for the different sleep stages. Stages wake, S1, S2, S3 and deep seem to be distinguishable quite well. Data from rem sleep on the other hand seems to be very similar to data from stages S1 and S2. (ii) Classification of S2 according to R&K is done based on very short events (spindles and K-complexes) and although the rest of a 30 second segment might look like S1, S3 or rem, it is nevertheless judged as S2 in its entirety. (iii) S2 has already been described as a "compound" state not easily discriminable from other states [18]. Supervised approaches which use label information from the human scorer for training of their models seem to be able to overcome this problem of rem sleep detection. Quite satisfactory results using neural networks have been reported even when

only a single EEG signal is being used [6]. However, supervised approaches can per definition not go beyond replication of R & K sleep staging. As for the other sleep stages, S3 is mainly mixed up with deep sleep, which is as expected.

Additional experiments using more than a single feature (i.e. first reflection coefficient at C3) did either not change the results (including the second reflection coefficient at C3) or even worsen them considerably (including a feature of EMG level or EOG activity). Including data from C4 does not seem to make much sense since it is highly correlated with data from C3 anyway. The same holds for stochastic temporal complexity compared to the first reflection coefficient.

The GOHMM trained with data from sleep lab B is evaluated using fourteen whole night recordings from test set B. Again the continuous sleep profiles obtained using Equ. 9 are compared to traditional R&K scoring. R&K scores are taken as true scores and for each sleep stage separately the percentages of GOHMM classification into each of the 6 stages are given in Tab. 2. The accuracies for the unambiguous "extreme" R&K stages now read as only 25% for wake, 87% for deep sleep and 61% for rem sleep. The bad result for detection of wakefulness is due to a severe mix-up with stages of light sleep (S1) and rem sleep.

In Fig. 4 a boxplot of all training and test data from sleep lab B (first reflection coefficient at C3) is shown for the different sleep stages. The biggest overlap can be found between stages wake, S1 and rem. In comparison to the same boxplot for data from sleep lab A in Fig. 3 it is noticeable that the general overlap between all sleep stages is bigger for sleep lab B. This might explain

why for data from sleep lab B the separate training of a GOHMM did not improve the performance.

Figs. 3 and 4 make it clear that based on our single channel of data (first reflection coefficient at C3), discrimination between sleep stages is not easy especially for rem sleep. The probabilistic dependence between successive sleep stages is therefore a fundamental aspect to consider and taking advantage of the Markovian property in the data seems to make the difference. The transition probabilities $A = \{a_{ij}\}$ (see Equ. 4) are given in Tabs. 3 and 4 for sleep labs A and B. The numbers for sleep lab A in Tab. 3 show that the GOHMM did learn a quite distinct temporal structure:

- transition probabilities are of course highest for transitions of states on to themselves (main diagonal);
- there are some transitions that are almost never taken (from wake to deep, from deep to wake);
- there is a preferred overall temporal structure (starting with a transition from wake to rem, going back and forth between rem and deep, finally going from rem to wake again).

This picture is much less clear for sleep lab B (Tab. 4). The transition probability from state wake on to itself is much smaller. The transitions from wake to deep and vice versa are much higher than for sleep lab A. The same holds for the transition from wake to rem. We have no reason to believe that the overall sleep profiles for subjects from sleep lab A are any different from those of sleep lab B. The differences in transition probabilities rather reflect the bigger overlap between sleep stages in feature space for sleep lab B and explain the poorer performance of the GOHMM for sleep lab B.

5 Discussion

We presented an approach towards automatic sleep staging that goes beyond mere replication of the traditional R&K standard and offers a new continuous description of human sleep which is based on probabilistic principles. It is therefore in line with previous recommendations [17] and work [18] on continuous sleep staging. The approach is based on Hidden Markov Models, is totally unsupervised and uses only a single channel of EEG.

We like to draw two main conclusions from the empirical results presented in this paper:

(i) Sleep recordings from different sleep labs are not directly comparable. Training of separate models for the sleep labs is necessary. Even then there are great differences in performance when comparing different sleep labs. These conclusions are of course based on a specific data base (recorded during the SIESTA project [1]) and a specific method (GOHMM defined over reflection coefficients). But one should bear in mind that the empirical results in this paper are based on 68 whole night recordings from two labs and the entropy results depicted in Fig. 1 even on 590 whole night recordings from eight different labs. Our GOHMM results and the entropy results have been obtained with different methods based on different parameterizations of sleep data and still point in the same direction of huge variation across sleep labs. Although a lot of effort had been put into harmonization of sleep labs and recording protocol, there seem to be differences in hardware and also in filter settings which are still visible in EEG and other signals. See e.g. the differences in high- and low-pass filter settings and even sampling frequency for sleep labs

A and B described above.

(ii) Using data from a single electrode is sufficient for reliable continuous sleep staging. For data from sleep lab A, the output in the form of three continuous probability traces clearly captures the three main processes in human sleep: wakefulness, deep sleep and rem sleep with an accuracy of around 80%. This performance is superior compared to our own previous results [7] and has been made possible by realizing that there exist clear lab effects in our data base of sleep recordings.

References

[1] Kloesch G., Kemp B., Penzel T., Schloegl A., Rappelsberger P., Trenker E., Gruber G., Zeitlhofer J., Saletu B., Herrmann W.M., Himanen S.-L., Kunz D., Barbanoj M., Roeschke J., Vaerri A., Dorffner G.: The SIESTA Project Polygraphic and Clinical Database, IEEE Eng. in Medicine & Biology Magazine, 20(3)51-57, 2001.

[2] Ancoli-Israel S., Roth T.: Characteristics of insomnia in the United States: Results of the 1991 National Sleep Foundation Survey, Sleep, Volume 22, Supplement 2, Pages S347-S353, 1999.

[3] Rechtschaffen A., Kales A.: A Manual of Standardized Terminology, Techniques and Scoring System for Sleep Stages of Human Subjects, U.S. Dept. Health, Education and Welfare, National Institute of Health Publ. No.204, Washington, 1968.

[4] Penzel T., Stephan K., Kubicki S., Herrmann W.M.: Integrated Sleep Analysis, with emphasis on automatic methods, in Degen R., Rodin E.A. (eds): Epilepsy, Sleep and Sleep Deprivation, Elsevier, pp. 177-203, 1991.

[5] Koprinska I., Pfurtscheller G., Flotzinger D.: Sleep classification in infants by decision tree-based neural networks, Artificial Intelligence in Medicine, Volume 8, Issue 4, pp. 387-401, 1996.

[6] Groezinger M., Roeschke J.: The automatic recognition of REM sleep - a challenge and some answers, Methods Find Exp Clin Pharmacol, 24, Suppl D:33-5, 2004.

[7] Flexer A., Sykacek P., Rezek I., Dorffner G.: An automatic, continuous and probabilistic sleep stager based on a Hidden Markov Model, Applied Artificial Intelligence, 16(3), pp.199-207, 2001.

[8] Flexer A., Gruber G., Dorffner G.: Continuous Unsupervised Sleep Staging Based on a Single EEG Signal, in Dorronsoro J.R.(ed.), Artificial Neural Networks - ICANN 2002, Lecture Notes in Computer Science, Springer, LNCS 2415, pp. 1013-1018, 2002.

[9] Flexer A., Gruber G., Dorffner G.: Improvements on continuous unsupervised sleep staging, in Bourlard H., Adali T., Bengio S., Larsen J., Douglas S.(eds.), Neural Networks for Signal Processing XII, Institute of Electrical and Electronics Engineers, Inc., New York, NY, pp. 687-695, 2002.

[10] Sykacek P., Roberts S., Rezek I., Flexer A., Dorffner G.: A Probabilistic Approach to High-Resolution Sleep Analysis, in Dorffner G., Bischof H., Hornik K. (eds.), Artificial Neural Networks - ICANN 2001, International Conference, Vienna, Austria, Lecture Notes In Computer Science 2130, Springer, pp. 617-624, 2001.

[11] Rabiner L.R., Juang B.H.: An Introduction To Hidden Markov Models, IEEE ASSP Magazine, 3(1):4-16, 1986.

[12] Penny W.D., Roberts S.J.: Gaussian Observation Hidden Markov Models for EEG analysis, Technical Report, Imperial College, London, TR-98-12, 1998.

[13] Ljung L.: System Identification, Theory for the User, Prentice-Hall, Englewood Cliffs, New Jersey, 1999.

[14] Kubicki S., Hoeller L., Berg I., Pastelak-Price C., Dorow R.: Sleep EEG Evaluation: A Comparison of Results Obtained by Visual Scoring and Automatic Analysis with the Oxford Sleep Stager, Sleep, 12(2):140-149, 1989.

[15] Kelley J.T., Reilly E.L., Overall J.E., Reed K.: Reliability of Rapid Clinical Staging of All Night Sleep EEG, Clinical Electroencephalography, Vol. 16, No. 1, 16-20, 1985.

[16] Danker-Hopfe H., Kunz D., Gruber G., Kloesch G., Lorenzo J. L., Himanen S. L., Kemp B., Penzel T., Roeschke J., Dorn H., Schloegl A., Trenker E., Dorffner G.: Interrater reliability between scorers from eight European sleep laboratories in subjects with different sleep disorders, Journal of Sleep Research, Volume 13, Issue 1, pp. 63-69, 2004.

[17] Kemp B.: A proposal for computer-based sleep/wake analysis, Journal of Sleep Research, 2, 179-185, 1993.

[18] Roberts S., Tarassenko L.: New Method of Automated Sleep Quantification, Medical and Biological Engineering and Computing, (5), 509-517, 1992.

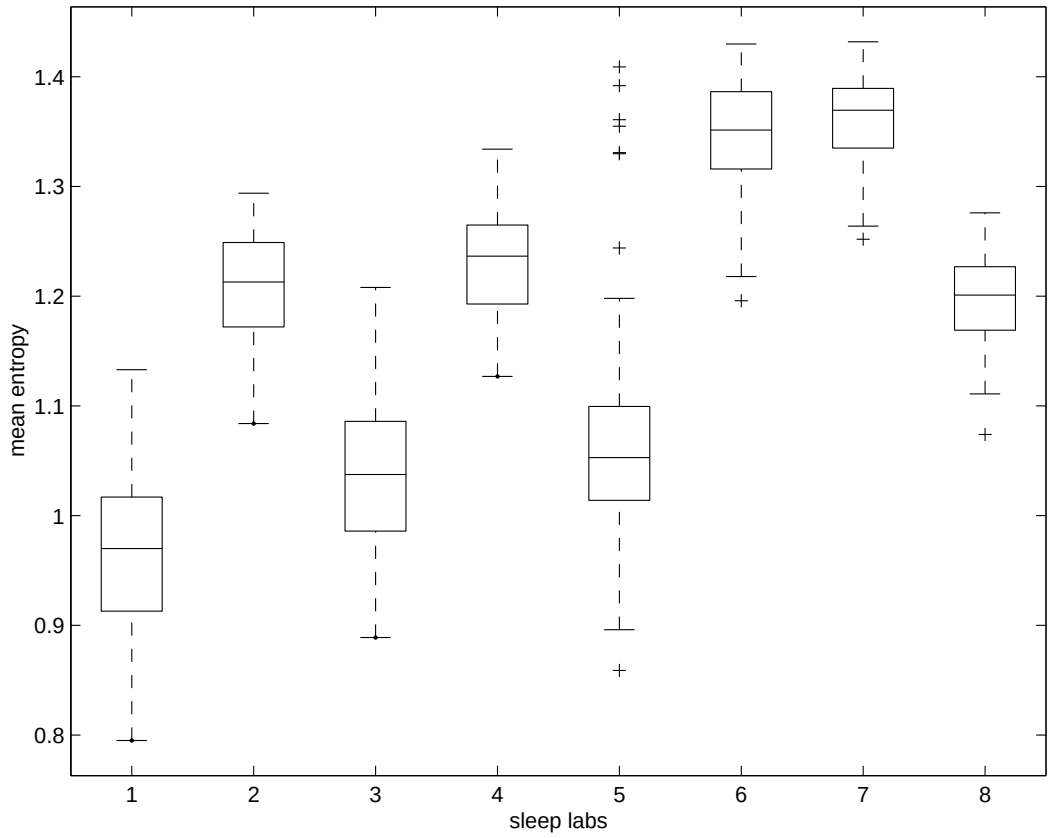

Fig. 1. Boxplot of the mean of the entropy of the probability traces given per sleep lab. Depicted are the medians and the 25% and 75% percentile per sleep lab.

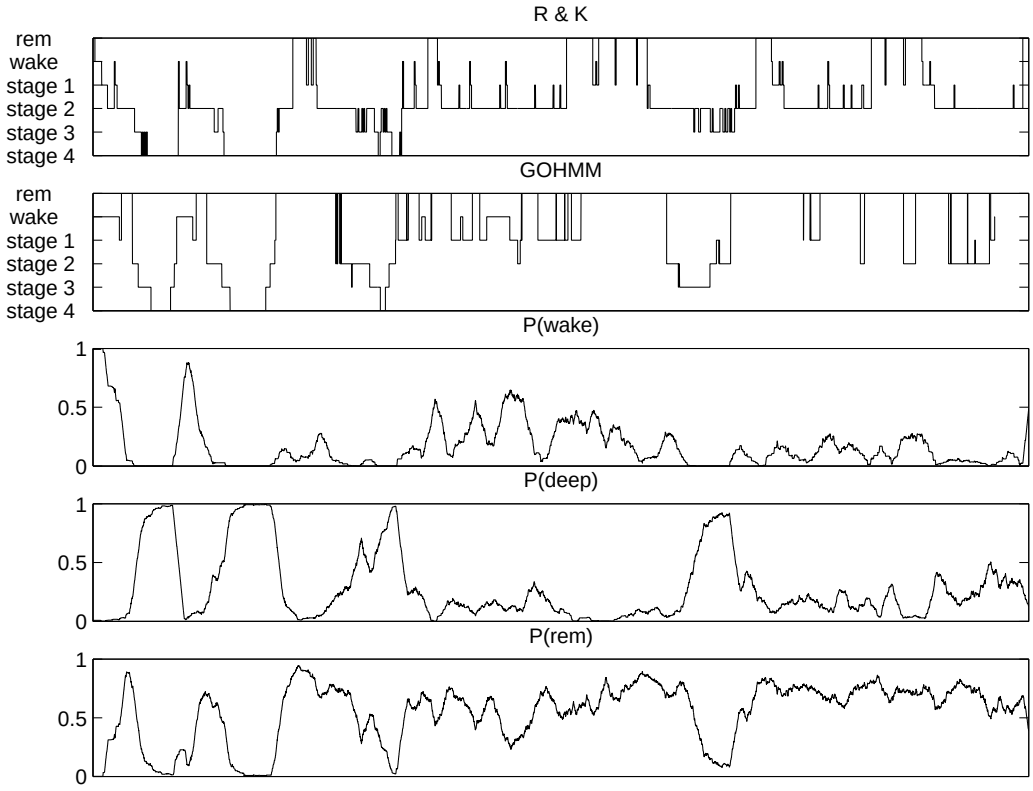

Fig. 2. Whole night results for one test subject from sleep lab A; from top to bottom: R&K scoring, GOHMM scoring, P(wake), P(deep), P(rem).

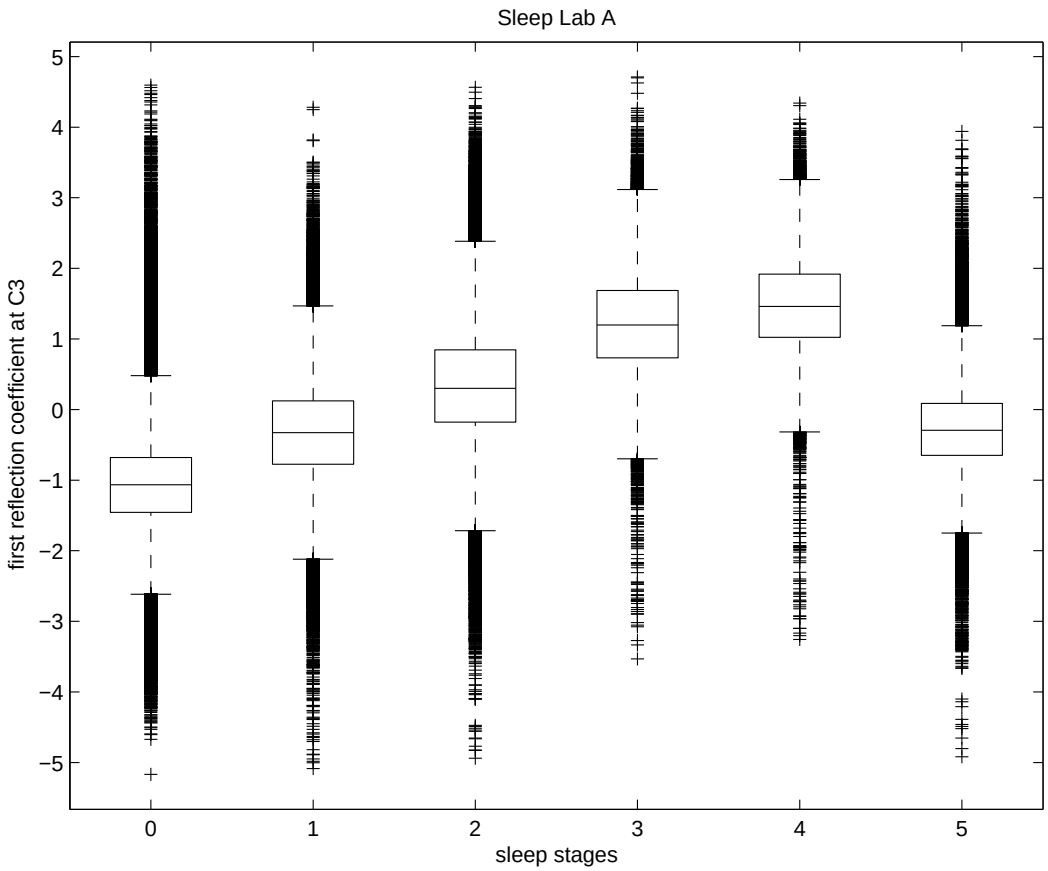

Fig. 3. Boxplot of data from sleep lab A (first reflection coefficient at C3) for different sleep stages: wake (0), S1 (1), S2 (2), S3 (3), deep (4), rem (5). Depicted are the medians and the 25% and 75% percentile per sleep stage.

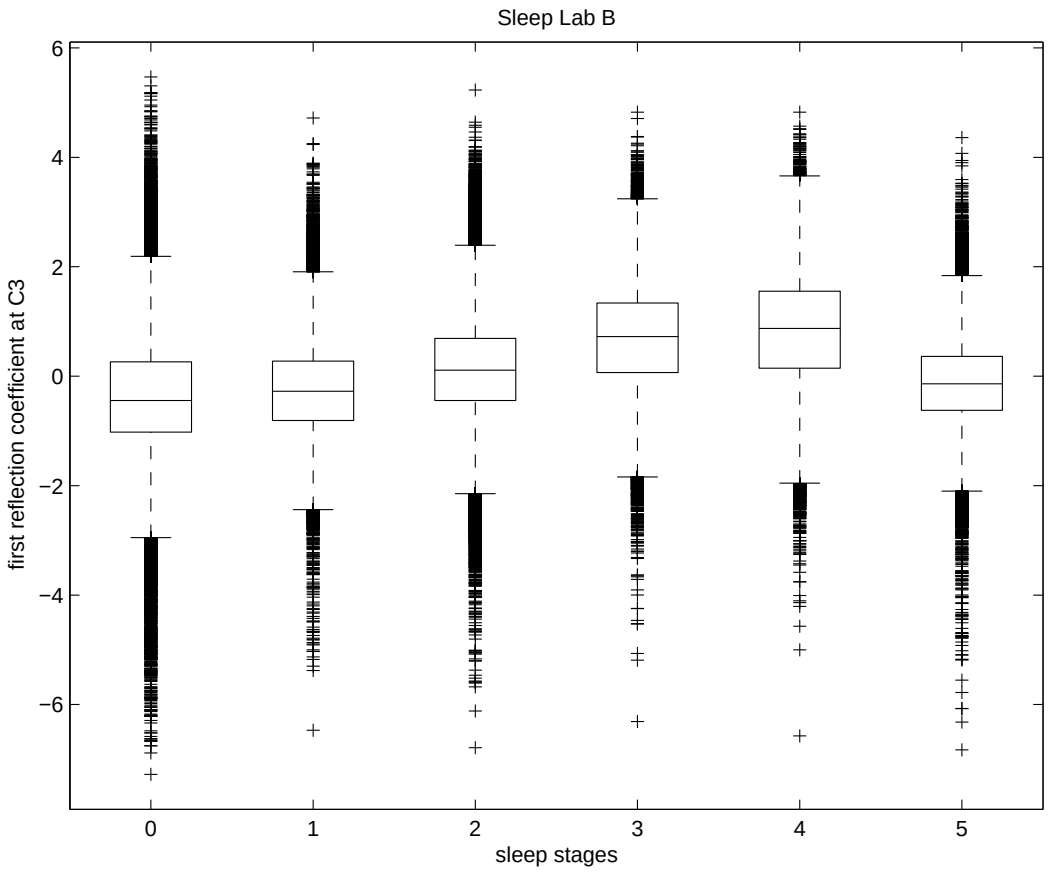

Fig. 4. Boxplot of data from sleep lab B (first reflection coefficient at C3) for different sleep stages: wake (0), S1 (1), S2 (2), S3 (3), deep (4), rem (5). Depicted are the medians and the 25% and 75% percentile per sleep stage.

Table 1

R&K scores vs. GOHMM classification for sleep lab A; GOHMM classification is given in percentages, separately for each sleep stage.

		GOHMM					
		wake	S 1	S 2	S 3	deep	REM
R & K	wake	**79**	10	4	0	0	7
	S 1	21	**24**	19	4	1	31
	S 2	3	8	**36**	16	8	29
	S 3	0	0	11	**35**	54	0
	deep	0	0	2	16	**82**	0
	REM	14	13	4	0	1	**68**

Table 2

R&K scores vs. GOHMM classification for sleep lab B; GOHMM classification is given in percentages, separately for each sleep stage.

		GOHMM					
		wake	S 1	S 2	S 3	deep	REM
R & K	wake	**25**	10	21	4	0	40
	S 1	12	**16**	17	3	0	51
	S 2	3	13	**38**	23	8	15
	S 3	0	0	4	**27**	69	0
	deep	0	0	1	12	**87**	0
	REM	6	16	16	2	0	**61**

Table 3

Transition probabilities a_{ij} of GOHMM for sleep lab A (a_{ij} is the probability to go from state i to state j).

		j		
		wake	deep	rem
i	wake	.954	.003	.042
	deep	.005	.929	.066
	rem	.019	.050	.931

Table 4

Transition probabilities a_{ij} of GOHMM for sleep lab B (a_{ij} is the probability to go from state i to state j).

		j		
		wake	deep	rem
i	wake	.749	.093	.158
	deep	.014	.938	.048
	rem	.021	.042	.937