An Approach to the Determination of Differences between Good and Bad Sleepers by Means of an Automatic Sleep Stage Scoring

J. L. NAVARRO-MESA, A. G. RAVELO-GARCÍA, F. D. LORENZO-GARCÍA,
S. I. MARTÍN-GONZÁLEZ, E. HERNÁNDEZ-PÉREZ, P. QUINTANA-MORALES
Departamento de Señales y Comunicaciones. Universidad de Las Palmas de Gran Canaria
Campus de Tafira 35017 Las Palmas de Gran Canaria. SPAIN
http://www.dsc.ulpgc.es

Abstract: This paper presents a sleep stage scoring method based on a Hidden Markov Model (HMM) with the goal of obtaining differences between good and bad sleepers according to the Self Rating Questionnaire for Sleep and Awakening Quality (SSA). For the design of the model, we study several parameterization techniques, the model topology and the training strategy for optimum performance. The system uses only one electroencephalographic channel (EEG), which represents an improvement over manual and automatic classifiers that use several channels. We adopt in our study the sleep stages W, S1/REM, S2 and S3/S4 according to the R&K standard. The experiments show that our system performs well compared with the inter scorer agreement. The experiments are performed over 24 recordings from SIESTA database.

Key-Words: - Sleep Stage Scoring, Hidden Markov Models, feature extraction, Sleep Quality.

1 Introduction

One of the main problems which is presented in Sleep studies is the automatic human sleep stage classification. This representation is called hypnogram and the manual classification scored by an expert, is a hard task.

Rechtschaffen & Kales (R&K) [1] is the set of rules which defines the sleep process divided in six stages or epochs: W (wake), S1 (light sleep), S2, S3 and S4 (deep sleep) and REM activity (Rapid Eye Movement).

Construction of an automatic sleep stage scoring is a difficult task, because the R&K rules are subjective resulting in low inter scorer agreement. For example two independent manual hypnogams scored by two expert scorers have an agreement rate between 51% and 87% [2]. The reason of this result is that the rules are not useful in older people or in cases where sleep disorders exist. Moreover R&K has a poor time resolution (30 seconds). Despite of these drawbacks, R&K continues to be the gold standard in sleep clinics.

Many automatic sleep scoring devices use the electroencephalogram (EEG), electromyogram (EMG) and electrooculogram (EOG), as minimum system for the automatic sleep scoring, although the inclusion of other polysomnographic signals, can improve the classification rate. We have three goals in this work:

Firstly, to find a feature extraction technique with high discrimination capacity, taking only EEG recordings. Secondly, to develop the base of a classification system with HMM that could be useful in an automatic sleep scoring device. Thirdly, to discriminate between two groups of people that can be differentiated by the sleep quality. For this purpose, we obtain parameters from our automatic classification system according to R&K and try to observe differences between groups.

The performance of the proposed model is compared with R&K manual scoring (we only have epochs labelled according to this standard). Related to some restrictions of R&K, our system could take into account a better time resolution (3 seconds) and results shown as probabilities instead of fix epochs could be studied.

2 Database

The database consists of 24 recordings of subjects with ages between 20 and 69 years. Recordings belong to SIESTA database [3] with 16 EEG signals (C3-M2) with a sampling frequency of 100 Hz. According to a SSA criteria (Self-Rating questionnaire for Sleep and Awakening Quality), we have separated our database with 16 recordings with good sleep quality and 8 recordings with poor sleep quality. Eight recordings with good sleep quality were elected to train the model (group TR). On the other hand 8 recordings with good sleep quality (group TG) and others 8 recordings with poor quality (group TP) were chosen as test recordings to validate the model.

Database contains 130 manual hypnograms that will be used to propose hypothesis about possible transition among stages.

3 Feature Extraction

In the feature extraction process, it is necessary to segment the EEG signal. This segmentation takes into consideration the stationary criteria of the signal. Traditionally 1-30 seconds length EEG segments have been used. In our case we have proposed 3 seconds length, since it presents a good compromise between stationary and time resolution.

We apply different parameterization methods. The features that were considered are:

Filter Banks log-energies (Fbank). It exists a very clear correspondence between sleep stages and spectral power. This situation allows us to suggest an analysis with Fbank that covers the whole frequency band.

A FFT-based bank of equally spaced filters is applied to obtain the power at each band. To avoid dependencies with signal dynamics, we normalise each filter output with the total signal power. A logarithm operation is applied to the power values in order to reduce the dynamic range and therefore keeping the whole frequency information.

Linear Frequency Cepstral Coefficients (LFCC). Discrete Cosine Transform (DCT) has the property of decorrelating power values and reducing the dimensionality of the feature vectors while preserving the relevant information. This fact is determinant in achieving good classification scores with HMM classifiers. Thus, a matrix transformation is applied to the Fbank vectors where matrix coefficients are obtained with the DCT.

Autoregressive Coefficients (AR). This is a classical method of spectral estimation and modeling. It has been very useful in EEG spectral estimation so it will be used also in our feature extraction process.

4 HMM Sleep Modelling

A Hidden Markov Model (HMM) has been proposed as a classification method which allows to model time sequences.

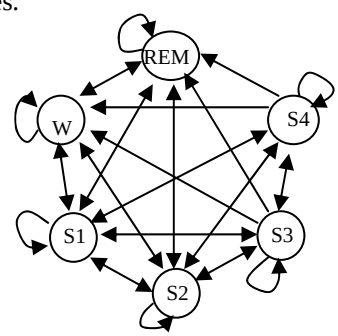

Fig. 1. Example of transition diagram among sleep stages

The model represents the state features and their dynamic evolutions in the sleep period. This model has two main components.

Firstly, the transition probability matrix with defines the probabilistic nature of the dynamic state transitions (fig. 1). Secondly a mixture of probability density functions (pdf) that characterise the probabilistic nature of the features in each state.

We propose an association between states and sleep stages. This strategy makes easier our work from two points of view. For one thing, we give sense to the concept of "state" and on the other hand, it makes easier the training process, since we have a manual classification scored by experts.

To design a HMM, we must estimate the parameters (A,B,π) which optimize the probability of the training observation vectors set.

$$\lambda = (A, B, \pi) \qquad (1)$$

where A, B and π are, respectively, the state transition probability matrix, the mixtures of probability density functions in each state and the probabilities to be in a initial state in the initial time.

For A matrix we consider a model with a topology as shown in figure 1 where all the transitions are possible except W→S3, W→S4, REM→S3 and REM→S4. We get to this conclusion after analyzing the transition matrix which is obtained by observing more than 130 hypnograms corresponding to the same number of different recordings (table 1).

TABLE 1. TRANSITION PROBABILITIES OBTAINED AFTER OBSERVING MORE THAN 130 HIPNOGRAMS

Actual/ Next	W	S1	S2	S3	S4	REM
W	84.54	13.09	1.71	0	0	0.31
S1	11.57	50.31	32.38	0.03	0.01	5.16
S2	2.39	5.34	87.52	3.01	0.03	1.12
S3	0.64	0.25	18.98	70.66	8.75	0.10
S4	0.51	0.11	1.11	9.12	88.41	0.03
REM	2.21	4.22	0.91	0	0	91.90

In the initial instant all the patients are in W state. Thus the probabilities vector π is

$$\pi = [1,0,\cdots 0] \qquad (2)$$

With respect to the mixtures of pdf in each state, we use gaussians. To train this mixture, we use a hypnogram per each one of the 8 training EEG recordings, classified in consensus by 2 human experts according to R&K. The recording are divided in 3 seconds segments, parameterized and grouped in a stage based on the hypnogram.

Now our goal is centred in estimating the parameters of the gaussians which define B in our HMM. In particular, for each state, we have a mixture with the following aspect:

$$b_j(o_t) = \sum_{l=1}^{M} c_{jl} b_{jl}(o_t) = \sum_{l=1}^{M} c_{jl} N\left(o_t / \mu_{jl}, \Sigma_{jl}\right), \; 1 \le j \le N \qquad (3)$$

where given a feature vector *ot* in the instant "t", the probability of being observed in the state p is given by the sum of M gaussians, each one defined by its particular measure vector μ_{jl}, its covariance matrix Σ_{jl} and the weighted coefficients c_{jl} of each gaussian of the mixture.

For the estimation of $N(\mu_{jl}, \Sigma_{jl})$, we proceed with the application of the Expectation Maximization (EM) algorithm. In our training phase, the gaussian mixtures showed good performance in B with EM isolated from A matrix. This matrix was designed based on the same recordings used for B. According to a Bayesian Information Criteria (BIC), we estimate empirically the optimum number of gaussians with a margin between 1 and 7, which results a good compromise between a good modelling, consistency in estimations and computational charge. The optimum number found has been 7.

In the test phase, using the Viterbi algorithm, the most probable state sequence is decodified. A detailed documentation can be studied in [4].

We extract features every 3 seconds. To be compatible with R&K, we analyze 10 segments of 3s included in a 30s epoch and the histogram of the classification outputs is obtained.

From the histogram, it is estimated as the state in the epoch, the most frequent state. To present the results, we have grouped similar stages from the EEG point of view. So we have the stages W, S1/REM, S2 and S3/S4. These are the epochs we can separate if we only have one EEG channel.

5 Experiments and Results

The scoring results are given in % of agreement with *E1-E2* (agreement rate between human experts who score the recordings separately) and are presented in table 2 considering the recordings which are present in the training and test phase. Mean, maximum and minimum values are given for each parameterization.

The optimum values in each parameterization are the following: AR (order = 6), Fbank (number of filters = 43), DCT (43 filters, order = 40).

As can be observed, the feature with the best results is LFCC. This parameter has a global success rate of 86.61% for the training patients set. The coincidence rate between experts is 85.7% (E1-E2). For the test recordings with good sleep quality, the rate is 80.1%. In this case, E1-E2 is 86.51%. In case of poor sleep quality the rate is 73.4% and E1-E2 is 85.02%.

In figure 2a we can observe a representation of the sleep stage evolution for one recording of our test group from the experts' point of view, and in figure 2b, the result of our automatic system. LFCC has been the parameterization used in that experiment.

Table 3 represents the confusion matrix for test recordings with good sleep quality.

TABLE 2. CLASSIFICATION RESULTS IN % FOR EACH PARAMETERIZATION TECHNIQUE IN TRAINNING AND TEST PHASE

Training (TR)	AR	**Fbank**	**LFCC**	**E1-E2**
Maximum rate %	82.02	85.85	88.98	87.79
Minimum rate %	69.43	77.1	80.67	77.49
Mean rate %	77.59	83.1	86.61	85.7
Test (Good Sleep Quality) (TG)	AR	Fbank	LFCC	E1-E2
Maximum rate %	80	89.54	92.76	89.88
Minimum rate %	48.69	48.48	67.61	62.1
Mean rate %	61.65	75	80.1	86.51
Test (Poor Sleep Quality) (TP)	AR	Fbank	LFCC	E1-E2
Maximum rate %	75.12	78.55	81.39	87.23
Minimum rate %	48.21	66.11	63.89	64.26
Mean rate %	69.9	72	73.4	85.02

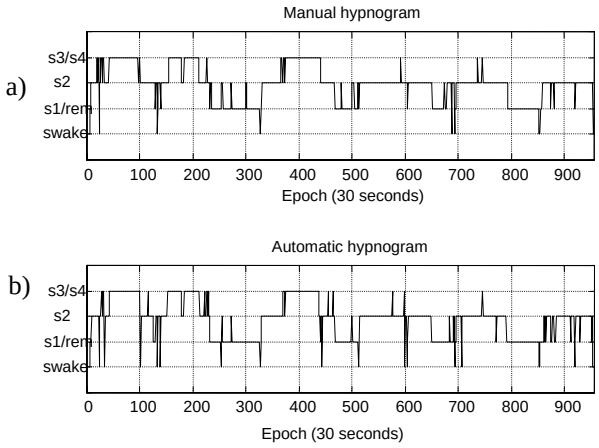

Fig. 2. a) Manual hypnogram b) Automatic hypnogram

TABLE 3. CONFUSION MATRIX FOR TEST RECORDINGS WITH GOOD SLEEP QUALITY

	W	S1/REM	S2	S3 /S 4
W	664	121	8	1
S1/REM	114	1736	76	0
S2	38	579	2943	82
S3 /S 4	3	89	381	669

6 Differences between Sleepers

Table 4 shows the parameters which can be obtained directly from the inspection of the hypnogram on the three recording groups (TR, TG and TP). These parameters are:

- Time percentage in stage W, S1/REM, S2 and S3/S4.
- TIB (Time in bed).
- Sleep period time (SPT, Time in bed (from lights off) minus wake before sleep onset and minus wake after last epoch of sleep).
- Total Sleep Time (TST, Sleep period time minus wake time after sleep onset).

- Sleep latency (SL, Time between lights off and the first epoch of S2).
- Sleep efficiency index ([Total sleep time/time in bed (from lights off to lights on)] * 100).

TABLE 4 SLEEP PARAMETRS FOR DIFFERENT GROUPS

Sleep parameter Group TR	Manual Hypnogram	Automatic Hypnogram
Percentage stage W	7.62[4-11.9]	8.39[2.6-16.51]
Percentage stage S1/REM	33.4[28-45.2]	33.04[24.76-41.44]
Percentage stage S2	46.85[39.3-55.9]	39.78[26.1-51.41]
Percentage stage S3/S4	16[6-25.1]	18.99[4.7-36.32]
Total sleep time (TST)	437.31[410.5-460]	435.94[402-466.5]
Sleep latency (SL)	15.31[8-27]	14.87[6.5-27.5]
Sleep efficiency index (SEI)	92[88.4-96]	91.44[84.01-97.39]

Sleep parameter Group TG	Manual Hypnogram	Automatic Hypnogram
Percentage stage W	11.11[1.77-19.36]	11.92[2.82-24.95]
Percentage stage S1/REM	25.46[19.93-32.67]	33.41[19.26-46.39]
Percentage stage S2	48.17[37.91-64.71]	45.1[37.34-61.48]
Percentage stage S3/S4	15.1662[7.05-22.54]	10.05[0.31-20.67]
Total sleep time (TST)	411.875[304.5-469]	410.38[297.5-464]
Sleep latency (SL)	25.36[5.5-43]	31.41[6-57]
Sleep efficiency index (SEI)	87.76[69.75-97.91]	86.84[68.15-96.86]

Sleep parameter Group TP	Manual Hypnogram	Automatic Hypnogram
Percentage stage W	29.42 [17-48.4]	32.95[12.84-53.62]
Percentage stage S1/REM	20.4 [16.6-25.6]	29.71[16.83-48.33]
Percentage stage S2	28.11 [7.4-51.8]	29.3[19.83-38.41]
Percentage stage S3/S4	9.96 [0.7-18.3]	8.03[0.42-16.07]
Total sleep time (TST)	330.2 [212.5-397.5]	313.69[190.5-417.5]
Sleep latency (SL)	34.43 [4.5-73]	44[190.5-417.5]
Sleep efficiency index (SEI)	68.55 [41.6-83]	58.9[25.83-87.16]

We can establish a comparison between the results obtained by the automatic and manual hypnogram. The results express the mean value of the parameters for the total set of people which take part in each group. Minimum and maximum parameter values are also showed in table 4.

The first difference between groups is the sleep efficiency index (SEI). This is higher in sleepers with good sleep according to SSA criteria (TR and TG groups) than in sleepers with poor sleep quality (TP group). For group TR, SEI is 92% with manual classification and 91.44% with automatic classification. For group TG, SEI is 87.76% following the manual scoring and 86.84% with our automatic system. These four values are quite different from the ones obtained for TP (68.55% with manual scoring and 58.9% with automatic one). As consequence of SEI results, TST is higher in TR and TG groups than in TP group.

Similar conclusion can be obtained when direct manual and automatic hypnogram are inspected. Thus the percentage in W stage is greater in group TP than in groups TR and TG. This parameter is directly related to SEI since the higher the percentage in W, the lower the SEI. Percentage in S3/S4 is another measure which corresponds to a longer deep sleep period. In this sense, percentage in S3/S4 is grater in TR and TG groups than in TP group. Finally sleep latency is grater in TP group than in TR and TG what indicates longer time until S2 stage.

7 Conclusions

We have presented a study that describes a practical implementation of an automatic sleep stage scoring following R&K in order to obtain differences between good and bad sleepers. This model could be applied to another kind of representation as suggested in [5] since it could contain some improvements as a higher time resolution with 3 seconds segmentation and a possible probability stage representation instead of fix 30 s epochs.

The proposed model is validated with the only methodology which nowadays is used in most of sleep units and research centres, R&K. Thus we can measure the agreement rate of our system, since we have the manual hypnograms scored by experts. We propose an analysis based on an only EEG signal if we overlap S3 and S4 stages and on the other hand, S1 and REM stages.

Automatic scoring has been compared to the results obtained by the experts. Finally we have could detect and measure some differences between groups with good and bad sleep according to our automatic system.

8 Acknowledgement

We would like to thank Dr. Alpo Varri at Tampere University of Technology for providing us the database used in this research.

References:

[1] Rechtschaffen, A. and Kales, A. (1968) A manual of standardized terminology techniques and scoring system for sleep stages of human subjects. Brain Research Institute, UCLA, Los Angeles, USA

[2] K.D. Nielsen, *Computer assisted sleep analysis*, Doctoral Thesis, Aalborg University,Denmark, 1993

[3] http://www.ai.univie.ac.at/oefai/nn/siesta/

[4] L. R. Rabiner, B. H. Juang, *Fundamentals of Speech Recognition*. Prentice-Hall, 1993.

[5] A. Flexer, G. Dorffner, P. Sykacek, I.Re zek, An automatic, continuous and probabilistic sleep stager based on a hidden markov model, Applied Artificial Intelligence, Vol. 16, Num. 3, 2002, pp.199-207