Sleep Staging Automaton Based on the Theory of Evidence

JOSE C. PRINCIPE, MEMBER, IEEE, SUNIT K. GALA, AND TAE G. CHANG, MEMBER, IEEE

Abstract—This paper addresses sleep staging as a medical decision problem. It develops a model for automated sleep staging by combining signal information, human heuristics in the form of rules, and a mathematical framework. The EEG/EOG/EMG events relevant for sleep staging are detected in real time by an existing front-end system and are summarized per minute. These token data are translated, normalized, and constitute the input alphabet to a finite state machine (automaton).

The processed token events are used as partial belief in a set of anthropomorphic rules, which encode human knowledge about the occurrence of a particular sleep stage. The Dempster–Shafer theory of evidence weighs the partial beliefs and attributes the minute sleep stage to the machine state transition that displays the highest final belief. Results are briefly presented.

I. INTRODUCTION

Sleep staging is probably the most important piece of information in the analysis of sleep. A computer system for the automated scoring of sleep is much needed not only as a consistent, quantitative tool but also as a labor-saving device (a night’s sleep produces roughly 288 m of papers with four channels of data that need to be scored by human experts). During sleep the EEG undergoes cyclic modifications that can be catalogued in five sleep stages plus awake [1]. In order to classify each minute of sleep, the human scorer looks for patterns involving combination of waveforms, i.e., delta waves, K-complexes, theta, alpha, sigma, and beta spindles in the EEG channels, REM and muscle bursts in the EOG channel, and EMG level in the EMG channel when available. Sleep staging is a peculiar microworld because it is a relatively small problem domain (there are eight EEG/EOG/EMG activities with at most 61 possible values that must be combined to produce one out of six states). However, sleep staging also contains human heuristics, which makes it particularly attractive as a test ground for knowledge encoding strategies.

A microcomputer-based system was developed in our laboratory to detect the EEG/EOG/EMG waveforms in real time and to send to a host computer the information pertaining to their occurrence (for sigma, REM, and K-complexes) or to their duration (for all others) [2]. These minute summarized data are called the token data. Here it is assumed that they contain all valuable information to score sleep with automated methods. The purpose of this paper is to describe a model that combines the token data with the human knowledge about sleep and assigns a sleep stage to every minute of data with a high degree of accuracy.

Automated sleep staging systems have used strategies dependent upon the implementation of signal processing methodology like thresholds on spectral amplitude in several frequency bands [3]; fuzzy metrics [4] and more conventional linear discriminant analysis methods [5], [6]; or an algorithm based on waveform detection [7] derived from the Rechtschaffen and Kales rules.

II. THE AUTOMATED SLEEP STAGING MODEL

The information about the sleep process is condensed into sleep stages. The human scorer uses combination rules involving EEG/EOG waveforms and EMG level to decide which is the sleep stage for a given minute. These rules, which are called anthropomorphic rules, are invoked depending upon the particular waveforms present in the minute. They borrow heavily from the rules stated in the standard manual of Rechtschaffen and Kales (R + K) [1].

To automate this procedure a finite state machine is defined. The machine will start in stage 0, the awake state, and depending upon the particular combination of waveforms detected by the front-end system it will loop or jump to another state. Therefore, each sleep stage is mapped onto a machine state and the minute summarized token data are mapped onto the input alphabet.

To complete the formal definition of the automaton, the next state transition is computed using the mathematical theory of evidence [8] (also called the Dempster–Shafer (D–S) theory of evidence). The theory requires the selection of evidence and the computation of beliefs. Fig. 1 depicts the important aspects of the new sleep scoring model which we call the belief automaton. In this model, the EEG tokens are used as evidence. First, a set of heuristic rules, called fuzzification functions, translate EEG token values into partial beliefs that a particular sleep stage is occurring.

The anthropomorphic rules select whichever of these partial beliefs are relevant for a given sleep stage transition. The firing of such rules associate the partial belief to the confirmation or disconfirmation of sleep stage tran-
versions. Furthermore, the D–S theory combines these partial beliefs and computes the final belief in each sleep stage transition.

The next machine state, which becomes the sleep stage for the minute, is attributed to the transition that displays the highest final belief. The final belief also represents the model confidence in the scoring. This procedure repeats itself every minute until the end of the night.

The need for fuzzification functions stems from the fact that humans use thresholds like “above/below X sec of Y activity” to define sleep stages. When implementing these definitions in a computer another aspect surfaces: the precise behavior of the computer versus the flexible behavior of the human given a sleep stage definition involving a threshold. An example illustrates this point well. Alpha has a threshold of 30 s to identify awake (also called stage zero). If alpha is greater than or equal to 30 s, sleep stage is zero; if alpha is less than 30 s, principle, sleep stage is not zero. From the point of view of the implementation using the D–S model, the following anthropomimetic rule could be developed:

\[
\text{IF } (\text{ALPHA} > 30 \text{ SEC}) \text{ THEN } (\text{CONFIRM STG 0} = 1).
\]

However, humans use thresholds as guidelines. In fact, whether alpha time is 28 or 32 s is irrelevant (no scorer measures precisely the alpha duration in a minute). Moreover, the “belief” that the subject is in the awake state is not dichotomic: it does not jump from 0 to 1 for a small variation around 30 s of alpha time.

Three very simple types of fuzzification functions were implemented: the staircase function, the piecewise linear, and the adaptive piecewise linear. They aim at representing three different kinds of human interpretation of sleep staging rules. The added advantage is the intrinsic normalization obtained (the outputs are values between [0, 1]) which makes the result directly usable by the D–S theory.

III. DESCRIPTION OF THE MODEL

A. The Automaton for Sleep Staging

Fig. 2 shows the automaton implemented for this study. Since the tests were performed in a normative population some of the transitions were disallowed (i.e., stage 0 to stage 5) as a simplifying factor. The differentiation between sleep stages 1 and 5, and between sleep stages 3 and 4 was accomplished in a latter part of the program, so conceptually each pair can be grouped into a single machine state.

B. Dempster–Shafer Theory Applied to Sleep Staging

Bayesian statistics require the \textit{a priori} knowledge about the problem domain to calculate the \textit{a posteriori} probabilities. It also automatically assigns one minus the probability of an event to the complement of the event. Therefore, there is no mechanism to differentiate between uncertainty about an event and the probability of its complement, i.e., lack of evidence for or against the event. These two aspects limit the application of conventional
statistical decision theory to some medical problems [9].
An alternate approach to the quantification of nondeter-
ministic phenomena was proposed by Dempster [8].
Evidence, which in the sleep staging problem is the occu-
rence of a particular waveform, is combined in a
mathematical framework to arrive at the belief of a par-
ticular outcome, i.e., one of the five sleep stages plus
awake that globally constitute the sleep EEG frame of
discernment $H$:

$$H = \{1, 2, 3, 4, 5, \text{awake}\}.$$

Each sleep stage is called a singleton hypothesis in the D-
S theory.

One of the strong points of the theory is the efficient
use of the available evidence without forcing unknown
assumptions about the problem domain. It does that by
narrowing the hypotheses, a process that characterizes
classification (diagnostic) reasoning in medicine. The D-
S theory uses a number in the range $[0, 1]$ to indicate
belief in a hypothesis given a piece of evidence. This
number is the degree of belief to which the evidence
supports the hypothesis. Also, evidence against a hypothesis
is regarded as evidence for the negation of the hypothesis.
For example, it is known that alpha is related to the awake
state. Let us say that when alpha is 20 s, one can attribute
a partial belief of 0.6 to the classification of the minute
as sleep stage zero, i.e., $\text{Bel}(SS0) = 0.6$. The rest of the
belief in D–S is uncommitted and waits for further evi-
dence gathering. One simplification in the sleep staging
is that the evidence contained in the tokens supports only
one singleton hypothesis and the remainder belief can then
be assigned to $H$. So, in this case, $\text{Bel}(H) = 1 - 0.6 =
0.4$. These assignments of belief constitute a simple ex-
ample of a basic probability assignment function or bpa,
which has values between $[0, 1]$. The Bayesian approach
would have forced the assignment of the remainder belief
to the complement of sleep stage zero, i.e., $H - \{550\}$.
Moreover precise probabilities should have been assigned
to each sleep stage.

Assume now that for the same minute ART (artifact)
has a value of 10 s. Again one can attribute to this new
piece of evidence a partial belief of 0.5 that the minute is
stage zero. How can these two pieces of evidence be com-
bined? The Dempster rules [8], which are extensions of
the Bayes rule, combine pieces of evidence. The belief
automaton uses the same rules that have been proposed to
substitute the certainty factor (CF) model in MYCIN [9].
However, they were implemented using the Barnett

The D–S reasoning lends itself very well to imple-
mentation because the evidence gathering scheme can be
preserved, i.e., the final belief in a singleton hypothesis can
be directly composed from the pieces of evidence. The
belief automaton explores this characteristic. In fact, in
sleep staging not all the EEG/EOG/EMG tokens contain
information to define a given sleep stage so, for the sake of
simplicity, only the relevant information should be in-
cluded. This selection of evidence was accomplished by
incorporating anthropomimetic rules which represent ex-
actly which tokens are relevant for a given sleep stage
definition. Only the anthropomimetic rules that are fired
will contribute bpas to the Barnett scheme. Moreover, in
this scheme the order of combination of evidence is im-
material, the only requirement being that the set of hy-
potheses be mutually exclusive and exhaustive. There-
fore, an automaton structure was chosen instead of more
elaborate inference mechanisms.

C. Fuzzification of EEG Parameters

Let us state the initial assumption again: each EEG to-
ken conveys evidence about a particular sleep stage and
this evidence can be translated into a partial belief that
the subject is in one of the five sleep stages plus awake. Table
1 summarizes the ranges of broad categories ($NE$none-
existent; $L$ – low; $M$ – medium; $H$ – high) of waveform
tokens that can occur in each of the five sleep stages plus
awake.

The problem now is to translate and quantify this
information into our model. Towards this goal, piecewise
linear functions were created that associate each token data
($x$ axis) to the partial belief in a particular sleep stage ($y$
axis), as shown in Fig. 3. Therefore, the partial belief
$F(\text{act})$, (act = delta, alpha, beta, sigma, artifact (ART),
REM, or $K$-complex) translates the evidential belief con-
tained in the tokens.

A comparison of Fig. 3 with Table I (read column wise)
shows that the fuzzification of alpha and muscle artifact
was chosen to score awake (i.e., high partial beliefs were
associated with large values of the tokens that occur in
sleep stage 0); the fuzzification of sigma and $K$-complexes
to score stage 2; the fuzzification of beta and REM to
score stage 1/5; and the fuzzification of delta to score
3/4. The choices for the maximum values of (act) are
pretty much set from the $R + K$ manual. However, the
number of steps, the final and intermediate values of
$F(\text{act})$, were assigned in an empiric way, taking into con-
sideration our knowledge about the sleep process and the
reliability of the front-end system with respect to the de-
tection accuracy of EEG waveforms.

An example helps to clarify the fuzzification. From
$R + K$ it is established that two or more sigma spindles in
a minute define sleep stage two. Therefore, when the
sigma token is two the partial belief associated with this
evidence is set at 0.90; for three or more spindles, 0.95;
when sigma count is zero the partial belief that sleep stage
is two is almost zero, so we set it to 0.1 (assigning zero
can cause algebraic problems in the Dempster combina-
Two other fuzzification rules were implemented in order to test the sensitivity of the model performance to the fuzzification, which entailed mainly heuristic knowledge. The piecewise linear fuzzifiers were defined only for alpha and beta [12]. They have the same breakpoints as the
staircase fuzzifiers, but for the intermediate values of the tokens, \( F(\text{act}) \) is obtained by linear interpolation.

For the adaptive piecewise linear fuzzifiers, again defined only for alpha and beta, the token values where the breakpoints occur are computed adaptively according to the following heuristic: experimentally it was observed that steady-state counts of these variables ran approximately 2/3 of the maximum observed during awake (alpha) and sleep stage 1 (beta). Therefore, the minute data for alpha and beta are screened until the first minute containing two spindles (first stage 2) is encountered. The largest 5 min’s values are averaged and multiplied by 0.67 to obtain the value that will correspond to the belief 0.95. The two other breakpoints are scaled such that the proportion shown in Fig. 3 is kept (i.e., 1/3 and 2/3).

\[ 0.25 \text{ minutes} \times 0.67 = 0.1675 \text{ belief} \]

D. Anthropometric Rules for Sleep Staging

The knowledge about sleep staging of normal subjects was first standardized in the R + K manual. Sleep stages are defined in terms of patterns in the EEG/EOG/EMG signals that enable the human scorer to attribute a sleep stage to every minute of sleep data. It is not by chance that the book is extensively illustrated; the patterns are quite often defined ostensibly, i.e., by example. Following a lot of other cases dealing with biological signals, the rules constitute a consistent but incomplete set, empirically established.

In the belief automaton, the representative EEG/EOG waveforms according to R + K are used as evidence to compute the partial belief associated with each hypothesis (sleep stage). The anthropometric rules select and quantify the importance of the occurrence of EEG tokens into sleep stages. The firing of such rules associates the value of \( F(\text{act}) \), to the confirmation or disconfirmation of sleep stage transitions.

The rule base incorporates rules pertaining to the existence of waveforms to define a sleep stage (which we call static information) and also dynamic information (decrease or increase from previous minute) about alpha and beta tokens [12]. A total of 130 rules were implemented in the model. The rule base is structured in the following way.

First, the rules are grouped by present state, i.e., all rules pertaining to the transition of stage \( i \) to stage \( j \) are attached to the machine state \( i \). Since the rules are organized by subdomains (the machine states) and the D-S final belief is independent of the order with which evidence is combined, the triggering of rules was made sequential and there is no need for an inference engine. Within a particular transition, the rules are grouped according to the confirmation of the transition or its disconfirmation.

Finally, these rules are of two basic types:

1) Assignment rules that always attribute a single belief. Their form is

\[ \text{(DIS)CONFIRM } [I] [J] = F(\text{act}) \]

which reads (dis)confirm transition from state \( i \) to state \( j \) with a belief equal to a value given by \( F(\text{act}) \).

2) Rules with an if clause

\[ \text{IF } \text{(CONDITION) THEN} \]
\[ \text{(DIS)CONFIRM } [I] [J] = F(\text{ACT}) \]
\[ \text{ELSE } \text{(DIS)CONFIRM } [I] [J] = F1(\text{ACT}). \]

IV. Implementation Characteristics

The algorithm is conceptually divided in three passes. The first pass implements the fuzzification procedure with one of the three rules presented earlier. It uses as input the EEG tokens detected by the front-end system and creates an output vector containing the evidence associated with the minute token data.

The second pass merges the fuzzification process with the firing of the rules and the D-S model. It is the main body of the algorithm where the minute by minute scoring is performed. The first minute is scored as sleep stage 0 (the subject is assumed awake) which starts the automaton. Since the rules use evidence contained in the minute tokens, we refer to this pass as scoring by context free rules. Each minute is assigned to a sleep stage and the belief value is part of the computer printout scoring. It is also in this pass that the distinction between sleep stage 3/4 is done according to the amount of the delta token; and the distinction between sleep stages 1/5 is accomplished based on the existence of REM waves.

The third pass compensates for sleep staging details related to information that transcends the minute summarized data and also for peculiarities of the D-S model when applied to sleep scoring. We like to refer to this pass as scoring by contextual information because the rules take into consideration the context.

The first objective is the recognition of artifacts in the beginning of the record. In clean records (i.e., with no muscle artifact) of low alpha subjects, the belief automaton can score stage 5 when rapid eye movements are present in the eye channels while the subject is awake. The human scorer, when in the beginning of the record, does not associate REM waves with REM sleep but with awake (subject is simply looking around). So the model incorporates a routine that tests for the first occurrence of stage 2 and goes back from this point to the beginning of the record and does exactly the same association, i.e., scores minutes with REM as stage 0.

The second aspect improved in this pass is the stage 5 scoring. A human scored run of REM sleep does not possess necessarily REM waves in every minute. REM sleep is associated with a flat EEG with low amplitude mixed frequency components (scored by the machine as stage 1) and sporadic rapid eye movements in the eye channel. So, when a minute scored as stage 5 is encountered, a routine changes adjacent minutes scored as stage 1 into stage 5.

The final problem corrected arises in conjunction with the application of D-S theory to sleep staging. Sometimes there are minutes of data scored with a low belief. Experimentally, it was noticed that minutes scored with a
belief below 0.4 tended to be wrongly scored. A low belief (defined here as below 0.4) is associated with two situations; not enough evidence in the tokens or contradictory evidence expressed in the tokens, typical of minutes where the sleep stage changes. So, when within a run of a particular stage 1 or 2 min of another sleep stage are scored with a low belief, the machine should do what the human does, i.e., smoothing. The rationale is to keep the score of the run in the absence of other strong evidence indicating otherwise. It should be stressed that this feature can be incorporated only on account of the existence of the belief attached to the sleep score. However, we restricted this smoothing to 2 min. If there are more than two consecutive minutes scored with low belief, it was decided to keep the scoring attributed by the machine and flag the output.

V. Results and Conclusions

The algorithm was tested with five sleep records from our normative data set, which has been scored by human experts in their routine practice. The subject age ranged from 13 to 70 years to cover the changes in sleep patterns found with advancing age. Table II presents the total man–machine agreement with the three fuzzification rules: the adaptive, the piecewise linear, and the staircase. The largest agreement is 90.6 percent and the smallest 78.44 percent with an average of 84.74 percent which is substantially larger than the previous system agreement [10].

The belief automaton presented here has several distinct features. It creates a mathematical framework where signal features are combined with heuristic domain knowledge. The forte of computer based diagnostic is objectivity. On the other hand, computer models generally are analytic and have difficulties to cope with the variability of biological data, the absence of precise definitions and incomplete knowledge. The approach presented here provides a way of including flexibility in the data processing without losing objectivity (i.e., the fuzzification rules can be customized and the knowledge base expanded to accommodate specific laboratory requirements). It also provides more information than previous sleep analyzing systems, because the sleep stage is qualified by a number that reflects the belief the computer attributes to the scoring.

In terms of performance, the goal is to have the belief automaton perform at the 90 percent agreement, which is the level required in most laboratories for interhuman scoring agreement. Some of the man–machine disagreement observed occurs in sleep transitions and in the distinction between sleep stage 3 and 4. Disagreement in the minute of a stage transition can be traced to time skew between the paper and the FM tape recorder. The distinction between sleep stage 3 and 4 has been identified as one of the greatest sources of inconsistencies between human scorers. If stages 3 and 4 are combined in the same stage the average agreement between the belief automaton and the human jumps to 87.14 percent. An improvement would be to extend the fuzzification to the definition of sleep stages 3 and 4 instead of using a simple threshold. The belief automaton also produced disagreement with the human scoring which could not be easily accounted for [12].

An improvement would be to incorporate in the model contextual (across minute) information, which the human subconsciously uses in the visual analysis. The existence of beliefs as qualifiers to the minute scoring, which is a feature not yet explored in this implementation, opens up the way for a hierarchical processing approach. The belief in the sleep stage using the present within one minute information can be considered as evidence when across minute rules are established.

The algorithm was coded in C and runs in a VAX 780 in less than 20 s for an all night sleep record (the input tokens were stored in files). The program length is approximately 1 000 lines, occupying 50 K words of memory. Such characteristics make the algorithm portable to much smaller computers like the IBM compatibles.

References


Jose C. Principe (M’83) was born in Porto, Portugal, in 1950. He received the B.S.E.E. degree from the University of Porto, Portugal, in 1972 and the M.Sc. and Ph.D. degrees from the University of Florida, Gainesville, in 1975 and 1979, respectively. He was a postdoctoral fellow at the University of Florida from 1979 to 1980. Since 1980 he has been a faculty member of the Department of Electrical Engineering, University of Aveiro, Portugal, where he is a Professor of Electrical Engineering. He joined the Faculty of the Electrical Engineering Department of the University of Florida in 1985. His research interests are biological signal processing and modeling, in particular the electroencephalogram. He is also interested in machine intelligence and neural networks.

Sunit K. Gala was born in India in 1964, Bombay, India. He received the M.Sc. (Tech.) Instrumentation degree from Birla Institute of Technology and Science, Pilani, India, in July 1985. He received the M.S. degree in electrical engineering from the University of Florida, Gainesville, in August 1986, where he is currently pursuing the Ph.D. degree.

His current research interests are in the areas of database systems, data modeling, knowledge representation, knowledge based systems, logic and artificial intelligence. He has worked at Tata Research, Design and Development Center (a subsidiary of Tata Consultancy Services), Pune, India and at the IBM T. J. Watson Research Center, Yorktown Heights, NY.

Tae G. Chang (S’79–M’88) received the B.S. degree in 1979 from the Seoul National University, Seoul, Korea, the M.S. degree in 1981 from the Korea Advanced Institute of Science and Technology, Seoul, Korea, and the Ph.D. degree in 1987 from the University of Florida, Gainesville, all in electrical engineering.

From 1981 to 1983, he was with the Hyundai Engineering Company, Seoul, Korea, as a Control Systems Design Engineer. From 1983 to 1984, he was with the Hyundai Electronics Inc., Seoul, Korea, as a Computer Systems Design Engineer. In July 1987, he joined the faculty of Tennessee State University, Nashville, TN, where he is currently an Assistant Professor at the Center of Excellence in Information Systems Engineering. His interests include artificial intelligence, pattern recognition, digital signal processing, and intelligent control for robotics.

Dr. Chang is a member ofEta Kappa Nu, the Computer Society of IEEE, and the Associations of Computing Machinery.