33rd Annual International Conference of the IEEE EMBS
Boston, Massachusetts USA, August 30 - September 3, 2011

Automatic Sleep Spindles Detection – Overview and Development of a Standard Proposal Assessment Method

S. Devuyst, T. Dutoit, *Member, IEEE*, P. Stenuit, and M. Kerkhofs

Abstract— **Since the 1970s, various automatic sleep spindles procedures have been implemented and presented in the literature. Unfortunately, their results are not easily comparable because the databases, the assessment methods and the terminologies employed are often radically different. In this study, we propose a systematic assessment method for any automatic sleep spindles detection algorithm. We apply this assessment method to our own automatic detection process in order to illustrate and legitimate its use. We obtain a global sensitivity of 70.20%, for a false positive proportion (relative to the total number of visually scored sleep spindles) of only 26.44% (False positive rate= 1.38% and specificity = 98.62%).**

I. INTRODUCTION

SLEEP spindles (SS) consist in sinus-like bursts that increase and decrease progressively in amplitude, with minimum duration of 0.5 s (Fig. 1). Their frequencies have been defined between 12 and 14 Hz in the Rechtschafen and Kales (R&K) criteria [1]. However, this interval has been proved to be too narrow and it was extended in several studies (11.75-16Hz in [2], 11.5-15Hz in [3], 10-16Hz in [4]-[5], 11-16Hz in [6] and 11-15Hz in [7]). The American Academy of Sleep Medicine (AASM) proposed a wider frequency range in their new guidelines for visual sleep scoring: 11Hz-16Hz [8]. Neither AASM nor R&K standards specify an amplitude criterion as a necessary requirement for the definition of SS. However, many authors have suggested using a minimal peak to peak amplitude criterion from 5uV to 25uV [2]-[3], [7].

The Presence or absence of SS is crucial for the scoring of sleep, since, sleep spindles represents one of the most important hallmarks of stage 2. However their visual analysis is time-consuming and tedious since there are typically hundreds of spindles in a full night recording [5], [9]. Therefore, several automatic methods have been developed for their detection.

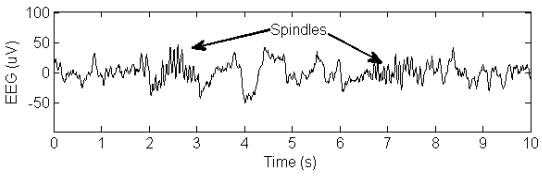

Fig. 1. Ten seconds of EEG recording (CZ-A1) with two spindles.

Manuscript received October 9, 2001. This work was supported in part by the Belgian Network DYSCO, funded by the Interuniversity Attraction Poles Programme, initiated by the Belgian State, Science Policy Office.

S. Devuyst and T. Dutoit are with the TCTS Lab, Université de Mons - UMONS, B-7000 Mons, Belgium (stephanie.devuyst@umons.ac.be);

P. Stenuit and M. Kerkhofs are with the SleepLaboratory CHU Vésale, Montigny-le-Tilleul, Belgium.

II. PAST WORKS

The earliest SS detectors were based on hardware. Several of them have combined pass-band filters together with a system performing the frequency detection (by searching the zero crossings [2] or by using a phase-locked loop [3]). By adding to this process other heuristic criteria in order to distinguish amoung spindles, EMG artifacts and alpha rhythm, Fish *et al.* obtained a sensitivity of 96% [3].

After these hardware detection systems, several software methods have been implemented. Two main approaches became prevalent: those using band-pass filtering and level detection, and procedures employing a first stage of features extraction, followed by decision-making for classification.

Concerning amplitude-based algorithms, selection of the threshold value is critical for the sensitivity of spindle detection. Traditionally, the amplitude threshold is fixed on the base of training recordings [6], [9]. However, we presently know that there is a considerable inter-subject variability in spindles amplitude and mean frequency, while sleep spindles characteristics are almost invariant from a night to another within an individual [10].

Therefore, Ray *et al.* and Huuponen *et al.* proposed to compute a recording-specific amplitude threshold, before performing the level detection. This threshold is derived from amplitudes distribution of some sleep spindles previously detected on the recording (either by an expert [11], or automatically thanks to spectral characteristics [9]). By this way, Huuponen *et al.* reported a sensitivity of 73.5% and a specificity of 98.5%, while Ray *et al.* obtained a sensitivity of 98.96% for a specificity of 88.49%.

Band-pass filtering is not the unique method used to capture the spindle activity before carrying out a level detection. Discrete wavelet transform was also used [12], as well as Matching pursuit (MT) [7].

Concerning algorithms based on features extraction followed by classification, it is broadly recognized that Short Time Fourier Transform (STFT) is an adequate tool to study the changes of the frequency content which characterized sleep spindles. In some cases, STFT coefficients are directly used as inputs of the classifier. By this way, Görür *et al*. obtained an agreement rate of 88.7% with a multilayer perceptron (MLP) classifier and a mean agreement rate of 95.4% with support vector machine (SVM) [13]. In other cases, features are extracted from the STFT (e.g. mean amplitude in different frequency bands). Anderer *et al*. 2004 reported up a specificity of 80% and a sensitivity of 86% by

using subsequently a linear discriminant analysis [6].

Adaptive autoregressive (AR) modeling is another commonly used method for features extraction. Average performance was hereby reached between 88.8 and 93.6% with MLP and between 93.3 and 96.0% with SVM [13].

Finally, let us remark that independent component analysis (ICA) was also investigated to separate spindle activity from multichannel EEG recording [14].

III. PERFORMANCE COMPARISON

All these algorithms performance are not easily comparable because criteria used for sleep spindles identification are inconsistent across studies. Moreover, performance measurements are different, and classical terms usually employed to report the results do not always have the same significance. For example, algorithms for classification problem ([6], [13]) use EEG segment of fixed length containing or not the micro-event to be detected. The corresponding false positive rate is calculated as follow:

$$FPrate = 1 - specificity = \frac{\#False\ Positive}{(\#False\ Positive\ + \#True\ Negative)} \quad (1)$$

Algorithms for detection problem use a whole night recording. In this case, the false positive rate can either be calculated on a fixed resolution like above [4], or approximated by looking at the proportion of false positives relative to the number of real sleep spindle events [5]:

$$"FPrate" \approx FPproportion = \frac{\#False\ Positive}{(\#True\ positive + \#False\ Negative)} \quad (2)$$

Lastly, some others have also used the term "false positive rate" to indicate the amount of false alarms among all automatic detections [9]:

$$"FPrate" \approx FPamount = \frac{\#False\ Positive}{(\#True\ Positive + \#False\ Positive)} \quad (3)$$

This multiple definition of the false positive rates must be taken carefully, especially when we compare algorithms performance. Indeed, a method presenting a false positive rate (FPproportion) of 30% according to (2) is definitely preferable to a method presenting a false positive rate of 30% computed with (1) (since there is generally less than 4 spindles per 30s during sleep stage 2). To avoid any confusion in the future, we propose in this paper an assessment method which allows computing these three parameters, and we will refer to them respectively as FPrate, FPpropotion and FPamount.

IV. STANDARD PROPOSAL ASSESSMENT METHOD

To allow evaluation and comparison between studies concerned with sleep spindles detection, we have published our database (as well as the associated visual scorings) on Internet. In addition, we propose below a general assessment method from which all parameters often reported in the literature can be extracted.

http://www.tcts.fpms.ac.be/~devuyst/DataBaseSpindles/

A. Recordings

Data were recorded at the Sleep Laboratory of the André Vésale hospital (Belgium). They consist of six whole-nights recordings coming from patients (3 men and 3 women aged between 31 and 54) with different pathologies (dysomnia, PLMS, insomnia, apnea syndrome, etc.). Two EOG channels, three EEG channels and one submental EMG channel were recorded. The sampling frequency was 200Hz. A segment of 30 minutes was extracted from each night from the central EEG channel for spindles scoring. No effort was made to select good spindle epochs or noise free epochs, in order to reflect reality as well as possible. These excerpts were given independently to two experts for sleep spindles scoring. The total number of identified spindles was 289 for scorer 1 and 409 for scorer 2.

B. Assessment algorithm

The assessment algorithm uses, as inputs, the beginnings and durations of the micro-events scored by expert 1, by expert 2 and automatically detected. Then it identifies the quantity of each possible covering illustrated on Fig. 2.

These various possible configurations are gathered in 4 categories: type T1 corresponds to a correct automatic detection since at least one of the two experts has scored the event like such; type T2 corresponds to a false detection; type T3 corresponds to a missing detection with respect to one or both experts; and type T5 corresponds to multiple coverings implying automatic detection.

Once the number of these various types is known, it is easy to deduce the number of true positives (#TP), the number of false positives (#FP) and the number of false negatives (#FN) of the different confusion matrices as illustrated in Table I. Furthermore, if we consider that the mean duration of sleep spindles is about 1 second, we can approach the number of true negatives (#TN) by:

$$\#TN \approx total\ duration\ of\ the\ database\ in\ second - \#FP - \#TP - \#FN \quad (4)$$

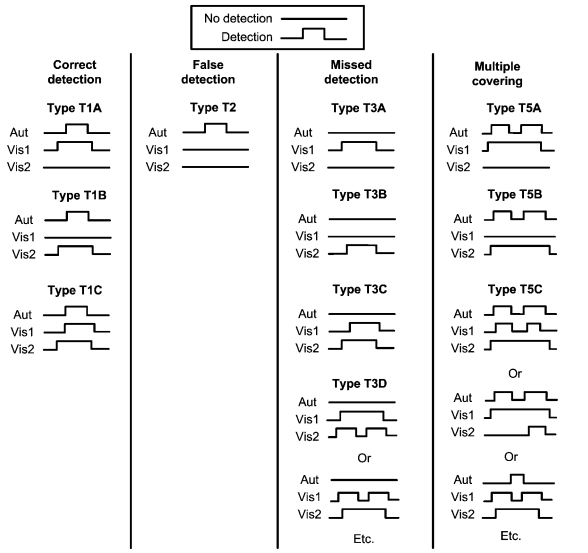

Fig. 2. Various possible coverings between the diverse micro-events scored by expert 1 (Vis1), scored by expert 2 (Vis2) and automatically detected (Aut).

TABLE I
CONFUSION MATRICES

CONFUSION MATRIX AUTOMATIC/VISUAL1

	Yes Vis1	NoVis1
Yes aut	#TP = #T1A + #T1C+#T5A+#T5C	#FP = #T2 + #T1B+#T5B
No aut	#FN = #T3A + #T3C + #T3D	~ #TN = 6*30*60 - #FP -#TP-#FN

CONFUSION MATRIX AUTOMATIC/VISUAL2

	Yes Vis2	NoVis2
Yes aut	#TP = #T1B + #T1C+#T5B+#T5C	#FP = #T2 + #T1A+#T5A
No aut	#FN = #T3B + #T3C + #T3D	~ #TN = 6*30*60 - #FP -#TP-#FN

CONFUSION MATRIX VISUAL1/VISUAL2

	Yes Vis2	NoVis2
Yes Vis1	#TP = #T1C + #T3C + #T3D+#T5C	#FP = #T1A + #T3A+#T5A+#T5A
No Vis1	#FN = #T1B + #T3B+#T5B	~ #TN = 6*30*60 - #FP -#TP-#FN

Finally, we can deduce the various parameters generally employed in the literature from these confusion matrices:

$$sensitivity = TPrate = \frac{\#TP}{(\#TP + \#FN)} \quad (5)$$

$$specificity = \frac{\#TN}{(\#FP + \#TN)} \quad (6)$$

$$FPrate = 1 - specificity = \frac{\#FP}{(\#FP + \#TN)} \quad (7)$$

$$FPproportion = \frac{\#FP}{(\#TP + \#FN)} \quad (8)$$

$$FPamount = \frac{\#FP}{(\#TP + \#FP)} \quad (9)$$

V. SPINDLE DETECTION ALGORITHM

In order to illustrate and legitimate our assessment method, we have applied it to our own automatic sleep spindles detection procedure. This procedure is based on band-pass filtering and level detection, due to its simplicity.

In order to take the spindles amplitude variability into account, we used a recording-specific threshold. To do so, we operated a first distinction between spindles and non spindles on the basis of spectral features, as suggested by Huuponen *et al.* [9]. Then, we fixed the value of the recording-specific threshold by applying the Bayes' theory as suggested in [4]. By multiplying this "Bayes" value by a factor K we obtained the final recording-specific threshold. The corresponding ROC curve obtained by varying K from 0.1 to 2 by 0.1 intervals is illustrated in black on Fig. 3

As we can see, a sufficient sensitivity is only obtained by fixing a lower value of K, which unfortunately also supplies an important number of false detections. This is partially caused by high frequency alpha intrusions and EMG artifacts. Indeed, during muscles contraction, the power of EMG artifacts also contaminates the sigma band, increasing the amplitude of the filtered signal. To make the distinction between sleep spindles and these false detections, it is necessary to take the spindles power as related to the total power into account (and not only power in the sigma band).

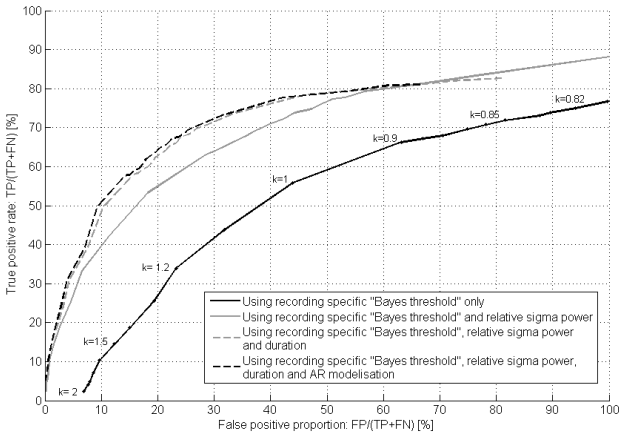

Fig. 3. Illustration of the ROC curves obtained with different methods.

To study this frequency content, the localized Fourier transform (STFT) is a particularly adapted tool.

By gathering the squared magnitude of the STFT (computed with an Hanning window of 0.5s long, shifted every 0.1s) we obtain the spectrogram S of the signal from which we can compute, at each instant, the spindle power related to the total power as follow:

$$relative\ spindle\ power\ (t) = \frac{\int_{11}^{15} S(f,t)\,df}{\int_{0.5}^{40} S(f,t)\,df} \quad (10)$$

Detection whose relative spindle power was inferior to 0.22 were removed, allowing thus to eliminate a big number of false detections as illustrated in Fig. 3 (solid grey line).

Then, we also removed detections of inadequate duration (<0.5s). The corresponding ROC curve is illustrated in dashed grey line in Fig. 3.

We computed lastly, for each possible sleep spindle, the corresponding autoregressive model of order 8 and we examined the frequency corresponding to the pole that has the maximum modulus, as suggested by Olbrich *et al*. [15]. We rejected all detections for which no pole had a corresponding frequency between 7 and 30Hz or those for which the frequency of the pole of maximum modulus was not in the interval 11-16Hz. The corresponding results were slightly improved as we can see in Fig. 3 (in dashed black).

VI. RESULTS AND DISCUSSION

To yield the results of our assessment method, we fixed the value of parameter K to 0.75, which corresponds to a sensitivity of 70%, while choosing the union of the visual scoring as reference. The corresponding quantities of possible coverings are reported in Table II and the corresponding confusion matrices are reported in Table III.

It can be noticed that for a total of 289 sleep spindles (SS) scored by scorer 1 and 409 SS scored by scorer 2, there is a mutual agreement on only 159. This corresponds to sensitivities of only 55.02 % and 38.88% respectively. This is much less than the 81% of inter-human agreement rate for

SS scoring reported Huupponen *et al.* [19]. Fortunately, the detection system agreed with 138 of these sleep spindles, which corresponds to an agreement rate of 86.79% (when a SS is considered as real when both scorers marked it as such).

If we consider this intersection of the visual scoring as reference, we can observe (on table III) that the number of false positives is elevated, leading to a FPproportion of 239.62%. However, by considering the union of the visual scoring, we clearly decrease this number of false detections, obtaining a FPproportion of only 26.44 % for a sensitivity of 70.20% (FPrate=1.38% and specificity= 98.62%). It seems therefore that extra automatic detections (relative to the intersection of the reference scorings) correspond to borderline cases that can be discussed, since most of them were classified as spindles by one of the 2 scorers. This also explains why the FPproportion obtained by considering only one visual scoring (111.81% for visual scoring 1 and 49.26% for visual scoring 2) exceeds the FPproportion obtained by considering the union of the reference scorings.

VII. CONCLUSION

We proposed in this paper a unique assessment method using a well defined terminology and from which it is possible to establish all the desired confusion matrices. In addition, we make our database and our visual scorings freely available on the web, to allow comparisons between other future works. Lastly, we applied, as example, our assessment method to our own automatic detection algorithm. The algorithm provides a sensitivity of 70.20% for a FPproportion of only 26.44 % (FPrate = 1.38% and specificity = 98.62%), that is quite suitable considering the inter-human agreement rate for sleep spindles scoring. In addition it shows an excellent repeatability. We do not exclude however the existence of more powerful detection processes and we encourage their authors to use our method of assessment to compare their results.

TABLE II
RESULTS ON THE 6 EXCERPTS OF THE DATABASE. K=0.75

	Sleep stages						Total
	Wake	REM	S1	S2	S3	S4	
Duration (s)	1260	0	1100	6080	1900	460	10800
Nbr. total scored by system	32	0	4	383	99	10	528
Nbr. total scored by scorer #1	31	0	5	236	13	4	289
Nbr. total scored by scorer #2	5	0	2	315	77	10	409
Nbr. scored by only system (T2)	18	0	3	84	35	2	142
Nbr. scored by only scorer #1 (T 3A)	20	0	4	44	1	1	70
Nbr. scored by only scorer #2 (T3B)	1	0	1	46	20	1	69
Nbr. scored by only system & scorer #1(T1A)	10	0	0	41	6	0	57
Nbr. scored by only system & scorer #2 (T1B)	3	0	0	119	52	6	180
Nbr. scored by only scorer #1 & scorer #2 (T3C)	0	0	0	19	1	1	21
Nbr. scored by system & scorer #1 & #2 (T1C)	1	0	1	127	5	2	136
Nbr. of type T3D	0	0	0	0	0	0	0
Nbr. of type T5A	0	0	0	2	0	0	2
Nbr. of type T5B	0	0	0	0	0	0	0
Nbr. of type T5C	0	0	0	2	0	0	2
Nbr.of automatic quotation implied in a multiple covering (3D, 5A,5B or 5C)	0	0	0	7	0	0	7
Nbr.of quotation of scorer 1 implied in a multiple covering (3D, 5A,5B or 5C)	0	0	0	5	0	0	5
Nbr.of quotation of scorer 2 implied in a multiple covering (3D, 5A,5B or 5C)	0	0	0	3	0	0	3

TABLE III
CONFUSION MATRICES FOR K=0.75

	Yes Vis1	No Vis1		Yes Vis 2	No Vis 2		Yes Vis 2	No Vis 2
Yes Aut	197	322	Yes Aut	318	201	Yes Vis 1	159	129
No Aut	91	~10190	No Aut	90	~10191	No Vis 1	249	~10263

	Yes Vis1∪Vis2	No Vis1∪Vis2		Yes Vis1∩ Vis2	No Vis1∩ Vis2
Yes Aut	377	142	Yes Aut	138	381
No Aut	160	~10121	No Aut	21	~10260

REFERENCES

[1] A. Rechtschaffen ,and A. Kales, "A manual of standardised terminology and scoring system for sleep stages in human subjects," U.S. Government Printing Office, Washinghton, DC; 1968.

[2] JR. Smith, WF. Funke, WC. Yeo and RA. Ambuehl, "Detection of human sleep EEG waveforms," *Electroencephalogr. Clin. Neurophysiol.*, 38, pp 435-437, 1975.

[3] D. R. Fish, P.J. Allen and J.D. Blackie, "A new method for the quantitative analysis of sleep spindles during continuous overnight EEG recordings," *J. Sleep Res*, 70, pp 273-277, 1988.

[4] E. Huupponen, A. Värri, SL. Himanen, J. Hasan, M. Lehtokangas and J. Saarinen, "Optimization of sigma amplitude threshold in sleep spindle detection," *J Sleep Res*, vol. 9, pp 327-334, 2000.

[5] A. Nurettin and G. Cüneyt, "Automatic recognition of sleep spindles in EEG by using artificial neural networks," *Expert Systems with Applications*, vol. 27, pp 451-458, 2004.

[6] P. Anderer, T. Miazhynskaia, G. Gruber, S. Parapatics, M. Woertz, B. Saletu and G. Dorffner, "Automatic sleep spindle detection validated in 167 h of sleep recordings from 278 healthy controls and patients," *17th Congress of the European Sleep Research Society*, Prague, October 5-9, 2004, pp 313.

[7] Durka P.J., Blinowska K.J. "Analysis of EEG Transients by Means of Matching Pursuit," *Annals of Biomedical. Engineering*, Vol. 23, pp 608-611, 1995.

[8] C. Iber, S. Ancoli-Israel, A. Chesson and SF. Quan, "The AASM manual for the scoring of sleep and associated events : rules, terminology and technical specifications, " American Academy of Sleep Medicine, Westchester, Illinois (IL), 2007.

[9] E. Huupponen, G. Gomez-Herrero, A. Saastamoinen A. Varri, J. Hasan, S-L. Himanen, "Development and comparison of four sleep spindle detection methods," *Artificial Intelligence in Medicine,* vol. 40, pp 157-170, 2007.

[10] L. De Gennaro and M. Ferrara, "Sleep spindles: an overview," *Sleep Medicine Reviews*, vol. 7, no. 5, pp. 423–440, 2003.

[11] LB. Ray, SM. Fogel, CT. Smith and KR. Peters, "Validating an automated sleep spindle detection algorithm using an individualized approach," *J Sleep Res*., vol. 19, no. 2, pp 374-378, 2010.

[12] A. Akın and T.Akgül , "Detection of sleep spindles by discrete wavelet transform," *Proceedings of the IEEE 24th Annual Northeast Bioengineering Conference*, Hershey, 1998, pp. 15 -17.

[13] D. Görür, "Automated detection of sleep spindles," MSc thesis, Middle East Technical University, 2003.

[14] R. Rosipal, G. Dorffiner and E.Trenker, "Can ICA improve sleep spindles detection?," *Neural Networks World*, vol. 5, pp 539-547, 1998.

[15] E. Olbrich and P. Achermann, " Oscillatory events in the human sleep EEG – detection and properties," *Neurocomputing,* 58–60, pp 129–135, 2004.