EEG-Based Lapse Detection With High Temporal Resolution

Paul R. Davidson*, *Member, IEEE*, Richard D. Jones, *Senior Member, IEEE*, and Malik T. R. Peiris

***Abstract*—A warning system capable of reliably detecting lapses in responsiveness (*lapses*) has the potential to prevent many fatal accidents. We have developed a system capable of detecting lapses in real-time with second-scale temporal resolution. Data was from 15 subjects performing a visuomotor tracking task for two 1-hour sessions with concurrent electroencephalogram (EEG) and facial video recordings. The detector uses a neural network with normalized EEG log-power spectrum inputs from two bipolar EEG derivations, though we also considered a multichannel detector. Lapses, identified using a combination of video rating and tracking behavior, were used to train our detector. We compared detectors employing tapped delay-line linear perceptron, tapped delay-line multilayer perceptron (TDL-MLP), and long short-term memory (LSTM) recurrent neural networks operating continuously at 1 Hz. Using estimates of EEG log-power spectra from up to 4 s prior to a lapse improved detection compared with only using the most recent estimate. We report the first application of a LSTM to an EEG analysis problem. LSTM performance was equivalent to the best TDL-MLP network but did not require an input buffer. Overall performance was satisfactory with area under the curve from receiver operating characteristic analysis of 0.84 ± 0.02 (mean ± SE) and area under the precision-recall curve of 0.41 ± 0.08.**

***Index Terms*—Alertness monitoring, artificial neural networks, EEG, lapses of responsiveness, microsleeps, visuomotor tracking.**

I. Introduction

A lapse in psychomotor performance at the wrong moment can have catastrophic consequences. The fatigue process is associated with gradual deterioration in perceptual, cognitive, and sensorimotor performance [1], [2] but it is also common to observe rapid, temporary lapses of responsiveness, particularly in deeper fatigue states. These are typically accompanied by other behavioral sleep signs, such as head nodding, slow eye movements (SEM), loss of facial tone, and partial or full eye closure [3], followed rapidly by resumption of acceptable performance [4]. These episodes are often termed *lapses* or *microsleeps* [5], and indicate temporary deactivation of the cortical networks responsible for task performance [6]. A device capable of detecting or predicting lapses has the potential to markedly improve public safety.

Research into EEG-based lapse detection has been encouraged by studies showing lapses are correlated with changes in EEG spectra [7]–[11]. However, the short-term temporal dynamics of these changes tend to be considered too variable to be useful. Consequently, most studies have aimed to estimate alertness level by averaging performance at discrete auditory or visual vigilance tasks over broad 1–2 min time windows [10], [12], [13]. Studies using continuous compensatory tracking tasks have also estimated alertness by smoothing the tracking error with a moving window of 1 [14] or 2 [15] min duration. This windowing approach provides approximately minute-scale temporal resolution, which is appropriate for detecting slow shifts in arousal but does not provide sufficient temporal specificity to detect lapse events lasting only a few seconds. In this paper we report work on detection of lapses with finer temporal resolution. Our novel approach is to utilize the EEG patterns occurring in the seconds leading up to a lapse that might be obscured by the averaging process.

While the terms "lapse" and "microsleep" are often used as synonyms, there is an important distinction between microsleeps defined by EEG and behavioral criteria. EEG-defined microsleeps, usually identified via bursts of theta activity, can occur without any noticeable changes in task performance [16]. Behavior-defined microsleeps are less easily detected. They occur when key attentional or sensorimotor pathways required for responding to a given task are temporarily deactivated [17]. While it may be associated with EEG-defined microsleep, this deactivation process has no consistent EEG markers identifiable by human experts [18]. Despite this, we aimed to identify subtle spatio-temporal patterns in the EEG power spectrum that may by overlooked by a human EEG observer.

We have developed a system to detect lapses in real-time with second-scale temporal resolution based on continuous EEG data. The system was tested using a data set collected in a previous study of lapsing during a visuomotor pursuit tracking task [19]. The task was selected for its similarity to driving a car, though we hope our detector will generalize beyond this to other related tasks. Lapse episodes during the task were identified using a simple combination of tracking and video measures.

Our detector uses a neural network to identify lapses given only the EEG log-power spectrum. Unlike comparable systems operating with lower temporal resolution (e.g., [3], [12], [20], and [21]), our system makes use of the temporal dynamics of the EEG log-power spectrum, which we show improves lapse

Manuscript received August, 2005; revised October 15, 2006. *Asterisk indicates corresponding author.*

*P. R. Davidson is with the Van der Veer Institute for Parkinson's and Brain Research, Christchurch, New Zealand, the Department of Medical Physics and Bioengineering, Christchurch Hospital, Christchurch, New Zealand, and the Department of Electrical and Computer Engineering, University of Canterbury, Christchurch, New Zealand (e-mail: p.davidson@ieee.org).

R. D. Jones is with the Van der Veer Institute for Parkinson's and Brain Research, Christchurch, New Zealand, the Department of Medical Physics and Bioengineering, Christchurch Hospital, Christchurch, New Zealand, the Department of Electrical and Computer Engineering, University of Canterbury, Christchurch, New Zealand, and the Department of Medicine, Christchurch School of Medical and Health Sciences, University of Otago, Christchurch, New Zealand.

M. T. R. Peiris is with the Van der Veer Institute for Parkinson's and Brain Research, Christchurch, New Zealand, the Department of Medical Physics and Bioengineering, Christchurch Hospital, Christchurch, New Zealand, and the Department of Electrical and Computer Engineering, University of Canterbury, Christchurch, New Zealand.

Digital Object Identifier 10.1109/TBME.2007.893452

detection. We report results using tapped delay-line linear perceptron (TDL-Linear), tapped delay-line multilayer perceptron (TDL-MLP), and long short-term memory (LSTM) [22]–[24] networks to classify our EEG data. LSTM is a promising recurrent neural network architecture which, as far as we are aware, has not previously been applied to EEG analysis. Unlike TDL-MLP, LSTM networks do not employ a fixed memory representation and can learn complex temporal relationships over arbitrary time scales.

II. Methods

A. Tracking Study

In a previously reported study [19], 15 normal male volunteers aged 18–36 years performed a visuomotor tracking task [25] while EEG, video of facial features, and tracking behavior were recorded. Subjects were asked to keep a cursor as close as possible to a repeating pseudorandom target ($\text{bandwidth} = 0.164$ Hz, $\text{period} = 128$ s) scrolling down a screen. The cursor was located at the bottom of the screen and subjects had an 8-s preview of the scrolling target. Subjects moved the cursor horizontally by rotating a steering wheel. A 25 Hz analog video camera, time locked to the tracking, recorded head and facial features of subjects during the task. Sixteen channels of scalp EEG, and horizontal and vertical EOG, were recorded continuously during all sessions ($\text{sampling rate} = 256$ Hz, $\text{bandwidth} = 0.1$–100 Hz). EEG electrodes were placed according to the international 10–20 system. Each subject attended two sessions, held on separate days in which they performed the tracking task continuously for one hour. They were asked to stay alert and perform to the best of their ability.

As part of the same study, 30 hr. of video were rated by a human expert on a 7-point scale indicating probable lapses, sleep, deep drowsiness, light drowsiness, forced eye closure, distraction, and alertness. The rater marked transitions between levels with 1-s accuracy. The lapse and sleep categories were rated conservatively, with the subject's eyes having to be closed before these categories were assigned. This ensured that the subject was definitely unresponsive to new stimuli when a lapse or sleep episode was marked. The video analysis revealed that 8 of 15 subjects lapsed at some time during the two sessions, and these subjects were used in our subsequent analysis. Of those that lapsed, the median rate was 44 lapses per hour.

B. Lapse Identification

The video rating represents our most reliable conservative indication of when a lapse had occurred but we also observed clear lapses in the tracking response. This tracking information was used to improve identification of lapses. Subjects frequently stopped moving the response cursor just prior to a video lapse event, and these tracking *flat-spots* usually continued throughout, and frequently beyond the end of, a video lapse. Tracking flat-spots were a more reliable and specific indication of lapsing than simple tracking error, at least for our 1-D task, for several reasons. The low frequency of the target meant there was often a delay of several seconds between the start of a lapse and a clear increase in tracking error. Also, because the target was constrained within a limited range, tracking error periodically dropped to zero, even when the response cursor was not moving. Tracking error tolerance also varied both between and within individuals, with a higher error tolerance being characteristic of drowsiness—an observation consistent with vigilance studies showing shifts in both response criterion and stimulus sensitivity with vigilance level [6].

Identification of tracking flat-spots was made more difficult because of periodic stationary points in the target (where the velocity dropped to zero). At these times a tracking flat-spot may reflect appropriate behavior. Consequently, an algorithm was developed to distinguish "appropriate" from "inappropriate" tracking flat-spots. A lapse was deemed to be occurring when a subject was unresponsive according to the video rating and/or their tracking response exhibited an "inappropriate" flat-spot.

To identify flat-spots the target and response signals were first low-pass filtered with a cutoff at 5 Hz using an 8th-order bidirectional Butterworth filter. Target flat-spots were identified as intervals of at least 300 ms duration in which the target moved less than 1.5 mm. Similarly, response flat-spots were identified as intervals of at least 1500 ms duration in which the response cursor moved less than 0.8 mm. A "start-zone" and an "end-zone" were also marked for each response flat-spot. The start zone extended forward 2.0 s from the beginning of the response flat-spot. The end zone extended back 1.3 s from the end of the response flat-spot.

A tracking flat-spot was classified as "appropriate", and therefore excluded from our lapse measure, if 1) the "start-" and the "end-zone" overlapped with one or more target flat-spots; 2) the RMS error between the target and the response during the event was less than 15.0 mm; 3) the duration of the event was <6.0 s. The RMS error threshold in "2" was necessary for cases where a clearly "inappropriate" tracking flat-spot coincided with two or more target flat-spots. Careful visual inspection of the tracking data confirmed that an RMS error threshold of 15.0 mm was sufficiently sensitive to detect these flat-spots without introducing false positives. The duration of the longest contiguous target flat-spot was 5.0 s, so any tracking flat-spots longer than 6.0 s were considered inappropriate.

The measure performed well and missed only the early stages of a few clear lapses. These included cases where the video rating did not indicate a lapse was occurring yet the response cursor drifted incoherently. An example from a typical subject is shown in Fig. 1.

C. EEG-Based Lapse Detection

Epochs exhibiting clear electrode pop were marked as artifact using a simple algorithm which detected a change of greater than 0.4 mV in EEG amplitude within a single sample (3.9 ms). Standard longitudinal bipolar montage derivations were then calculated and used for further analysis. Signals from all derivations were divided into sequential, non-overlapping 1-s windows. Power spectral density across each window was calculated using the covariance method to form a fortieth-order autoregressive (AR) model. The covariance method was selected as it is resistant to noise and works well for short data sequences [19]. The model order was selected by iteratively increasing the order until δ, θ, α, and β band spectral peaks were clearly defined based on 10 min random samples of a single EEG derivation from all subjects. The high model order selected reflects our requirement for sufficient frequency resolution to discriminate between the standard EEG bands. Investigation with linear classifiers confirmed that a fortieth-order AR model provided

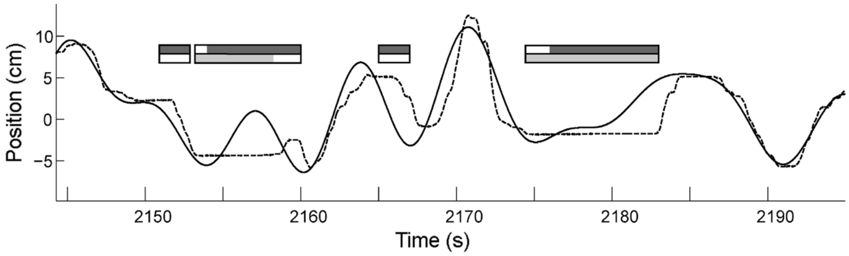

Fig. 1. Typical tracking behavior and identified lapses from the middle of a subject's first session. The tracking target (black line) and response (dashed line) are shown below rectangles indicating lapse events. The black outer border of the rectangles indicate a lapse interval as identified by our algorithm. The dark gray upper bars indicate video lapses and the light gray lower bars indicate inappropriate response flat-spots; a lapse was marked when either or both of these were identified. In the example, the first lapse is identified from video alone. This was because our algorithm did not classify the tracking flat-spot as inappropriate, since both the 2 s start- and 1.3 s end-zones of the flat-spot overlapped with a low velocity target.

better discrimination then lower order AR models. The logarithm of the mean power in 7 standard frequency ranges was then calculated for each derivation: delta $(0.1 < \mathrm{f} \leq 4\ \mathrm{Hz})$, theta $(4 < \mathrm{f} \leq 8\ \mathrm{Hz})$, alpha $(8 < \mathrm{f} \leq 13\ \mathrm{Hz})$, low beta $(13 < \mathrm{f} \leq 18\ \mathrm{Hz})$, high beta $(18 < \mathrm{f} \leq 36\ \mathrm{Hz})$, gamma $(36 < \mathrm{f} \leq 44\ \mathrm{Hz})$, and higher $(\mathrm{f} > 44\ \mathrm{Hz})$. This selection was based on preliminary work indicating standard wide frequency-bands provided increased robustness and inter-subject generalization compared with narrower frequency bands. These values were then converted to z-scores, normalized by the first minute of EEG data from each subject, session, and derivation. Where principal components analysis (PCA) was applied, the resulting 112 element feature vector (7 bands × 16 derivations) was used as input. The target was set to $+1$ when a lapse was present and -1 otherwise.

All classifiers were implemented using PDP++ [26]. The TDL-Linear networks had two layers and a linear activation function in the output unit. The TDL-MLP and LSTM networks had three layers with a linear bypass from input to output and a sigmoidal activation function in their output units. Each LSTM hidden unit consisted of a single memory cell. Sequential on-line training was used except when a sample was identified as containing EEG artifact. In this case the LSTM network weights were not updated and the internal states of all memory cells were reset. MLP networks were trained using back-propagation with momentum, and LSTM networks were trained using a mixture of real-time recurrent learning (RTRL) and back-propagation through time (BPTT), as described in [24]. Higher-order training algorithms are unavailable in PDP++ but the training of our best TDL-MLP network was repeated using the Matlab implementation of the Levenberg-Marquadt algorithm. The resulting classifier was very similar, though convergence occurred in fewer iterations.

Data from all 8 subjects who had clear video lapses were used to train and test the networks. Performance was evaluated using several metrics. We calculated the area under the receiver operating curve (AUC-ROC) and the area under the precision recall curve (AUC-PR) using the ROCR package [27]. These measures are independent of operating point and both were considered when selecting the best classifier [28]. We also report the phi correlation coefficient (φ) [29] for the mean optimum threshold based on the training set when applied to the test set, sensitivity $(\mathrm{s_n} = \mathrm{TP}/[\mathrm{TP} + \mathrm{FN}]$, where TP and FP are the proportions of true and false positive samples respectively, and TN and FN are the true and false negative sample proportions), specificity $(\mathrm{s_p} = \mathrm{TN}/[\mathrm{TN} + \mathrm{FN}])$, and precision $(\mathrm{TP}/[\mathrm{TP} + \mathrm{FP}])$.

Classifier performance was assessed with leave-one-out cross-validation, in which the data from one subject was set aside and used to test a network trained using the remaining data. This was done once for each of our 8 subjects. The entire 8-fold cross-validation was then repeated three times with different initial random weights. Results reported here are means across those cross-validation repetitions. Paired t-tests were used to compare the performance of different detectors. All networks employed the same learning rate of $\mu = 0.0001$. Where over-fitting was detected and could not be eliminated by pruning the model structure, we report results using weight decay regularization [30].

To facilitate comparison of our results with those of systems with lower temporal resolution, we also smoothed our binary detector output using the same exponential filter applied by Jung *et al.* to generate a "local-error rate" estimate [12]. The filter was applied to both the target and the output of the neural network. The filter comprised an exponential moving window in which the gain decreased from 1.0 to 0.1 over 93.4 s, giving a half-life of 27.4 s.

III. Results

A. Lapse Identification

A lapse was marked whenever a video lapse event and/or an inappropriate tracking flat-spot was identified. By considering inappropriate tracking flat-spots we improved identification of the start and end of some lapses by several seconds compared with using the video alone (see Fig. 1.). Video lapse events occurred surprisingly frequently at 65.1 ± 16.8 (mean ± SE) events per hour, while the combined lapse measure gave 72.5 ± 16.9 lapses per hour. The difference in these rates was caused by inappropriate flat spots unaccompanied by video events, and probably reflects the conservative criteria used to identify video lapses. The duration of video-only lapse events was 4.0 ± 0.7 s, while the duration of combined lapse events was 4.4 ± 0.7 s reflecting the fact that video and tracking events typically overlapped.

B. Multichannel Analysis

Assuming over-fitting can be avoided, best performance is likely to be achieved using information from all channels. To limit the number of features with which the classifier models must work, we applied PCA to the log-power spectral data from all 16 bipolar derivations. 80% of the input variance was accounted for by the top 11 of 112 components and 90% by the

TABLE I
LAPSE DETECTION PERFORMANCE FOR TDL-LINEAR NETWORK WITH 30 PRINCIPAL COMPONENTS INPUT

Input Window Length (s)	AUC-PR mean ± SE	AUC-ROC mean ± SE
1.0	0.335 ± 0.088	0.759 ± 0.054
2.0	0.376 ± 0.099	0.777 ± 0.055
3.0	0.391 ± 0.102	0.785 ± 0.054
4.0	0.397 ± 0.104	0.788 ± 0.053
5.0	0.393 ± 0.104	0.791 ± 0.052
6.0	0.390 ± 0.104	0.794 ± 0.051

top 30. To confirm these additional 19 components contained information useful for lapse identification separate linear classification models were fitted with 11 and 30 input components. Classification performance was inferior using only 11 features ($\mathrm{AUC}-\mathrm{PR} = 0.29 \pm 0.07$, $\mathrm{AUC}-\mathrm{ROC} = 0.72 \pm 0.04$) compared to 30 features ($\mathrm{AUC}-\mathrm{PR} = 0.34 \pm 0.09$, $\mathrm{AUC}-\mathrm{ROC} = 0.76 \pm 0.06$). Consequently, we decided to continue our initial analysis using 30 components.

To provide a baseline for assessing neural network classifier performance, we investigated two-layer linear perceptron networks with tapped delay-line inputs. Table I shows leave-one-out cross-validation results for TDL-Linear networks with input windows between 1.0 and 6.0 s. Since the EEG log-power spectrum is updated at 1 Hz, this corresponds to between 1 and 5 delay-line taps. Hence, an input window of 1.0 s provides only the most recent spectrum estimate, with no history. The mean AUC-PR of the network output was larger with a 2-s than a 1-s input window (paired t-test; $\mathrm{p} = 0.0093$) and with a 4-s than a 2-s window ($\mathrm{p} = 0.033$), but did not differ when the window was extended from 4.0 s to 6.0 s ($\mathrm{p} = 0.31$). These results show that temporal information is able to improve detector performance but the slight trend to a lower AUC-PR for windows longer than 4 s indicates over-fitting may be an issue even for linear networks with tapped delay line inputs. Best performance was achieved with a 4-s input window ($\mathrm{AUC}-\mathrm{PR} = 0.40 \pm 0.10$, $\mathrm{AUC}-\mathrm{ROC} = 0.79 \pm 0.05$).

An LSTM recurrent neural network with 1 unit in the hidden layer and a linear bypass connection was subsequently trained until full convergence, but gave poorer performance compared to that of the TDL-Linear networks ($\mathrm{AUC}-\mathrm{PR} = 0.29 \pm 0.08$, $\mathrm{AUC}-\mathrm{ROC} = 0.74 \pm 0.04$) due to as over-fitting. To address this, we employed weight decay regularization [31], iteratively increasing the weight decay constant by factors of 10 until performance stopped improving (which occurred at $\lambda = 0.01$). This led to an $\mathrm{AUC}-\mathrm{PR} = 0.43 \pm 0.09$ and an $\mathrm{AUC}-\mathrm{ROC} = 0.81 \pm 0.04$. Adding further units to the hidden layer of the LSTM network did not improve classifier performance, and the LSTM classifier had lower mean RMS error than the best linear classifier (0.25 vs 0.26, $\mathrm{p} = 0.018$). This indicates the lapse classification problem exhibits mildly nonlinear EEG log-power spectrum dynamics.

We also investigated the performance of TDL-MLP networks. With a single unit in the hidden layer the window length was increased until performance differed from linear. Performance was worse with a window length of 2 s compared with 1 s so weight decay regularization was added and the procedure repeated. AUC-PR and AUC-ROC were lower than for our best LSTM result with one to three windows, but did not differ with a 4 s or longer input windows ($\mathrm{AUC}-\mathrm{PR} = 0.41 \pm 0.11$, $\mathrm{AUC}-\mathrm{ROC} = 0.79 \pm 0.05$). Adding additional units to the hidden layer did not improve performance of the TDL-MLP networks, further emphasizing that the problem is only mildly nonlinear.

Because over-fitting had a strong influence on our results, we repeated the analysis using only the top 11 components, explaining 80% of the variance in the input data. Fitting an LSTM network with a single unit in the hidden layer resulted in a network that performed better than the equivalent network with 30 components as input ($\mathrm{AUC}-\mathrm{PR} = 0.32 \pm 0.08$, $\mathrm{AUC}-\mathrm{ROC} = 0.75 \pm 0.04$), indicating that less over-fitting had occurred, but did not perform better than TDL-Linear with 4 s input window. Adding weight decay improved the results slightly but performance remained worse than the equivalent LSTM network with 30 components as input ($\mathrm{AUC}-\mathrm{PR} = 0.40 \pm 0.09$, $\mathrm{AUC}-\mathrm{ROC} = 0.80 \pm 0.03$).

Overall, the PCA results showed that adding temporal information improves classification performance and that adding nonlinear elements only provides a slight improvement. It should be noted that the training procedure for the LSTM network was simpler, as we did not need to iterate over a range of input window lengths.

C. *Limited Channel Subset Analysis*

The 16-channel analysis was intended to give an indication of the best performance we could obtain from the available data. With lapse data from only 8 subjects, and PCA unable to yield fewer than 30 features while retaining greater than 90% variance, the resulting classifiers either over-fit the data or discard an unacceptable proportion of the input variance. Consequently, some doubt remained as to whether optimum performance had been achieved.

Since one aim was to build a portable lapse detector, we also wanted to minimize the size and complexity of the detector unit. To achieve this we aimed to reduce the number of EEG derivations and keep the electrodes clustered as close together as possible. Consequently, we tried reducing the input features by simply limiting the number of input derivations.

To select the best EEG derivations we fitted linear classification models to data from each derivation in isolation. These results are shown in Table II.

These show a trend to better classification performance from more posterior derivations. Best performance was achieved with P4-O2, so we started by fitting an LSTM model to data from this derivation alone. Since the model only has 7 inputs there is substantially less risk of over-fitting

TABLE II
LEAVE-ONE-OUT CROSS-VALIDATION RESULTS. LINEAR CLASSIFIERS TRAINED WITH LOG POWER SPECTRAL DATA FROM EACH DERIVATION INDIVIDUALLY

Channel	AUC-PR	AUC-ROC
	mean ± SE	mean ± SE
Fp1 - F7	0.091 ± 0.031	0.478 ± 0.038
F7 - T3	0.114 ± 0.031	0.556 ± 0.048
T3 - T5	0.185 ± 0.039	0.660 ± 0.032
T5 - O1	0.232 ± 0.067	0.722 ± 0.031
Fp2 - F8	0.091 ± 0.030	0.482 ± 0.037
F8 - T4	0.107 ± 0.043	0.558 ± 0.037
T4 - T6	0.162 ± 0.049	0.649 ± 0.037
T6 - O2	0.274 ± 0.071	0.741 ± 0.034
Fp1 - F3	0.101 ± 0.034	0.466 ± 0.046
F3 - C3	0.142 ± 0.038	0.586 ± 0.035
C3 - P3	0.214 ± 0.054	0.678 ± 0.036
P3 - O1	0.254 ± 0.058	0.745 ± 0.026
Fp2 - F4	0.0864 ± 0.03	0.431 ± 0.038
F4 - C4	0.141 ± 0.040	0.572 ± 0.047
C4 - P4	0.225 ± 0.060	0.677 ± 0.039
P4 - O2	0.285 ± 0.065	0.754 ± 0.030

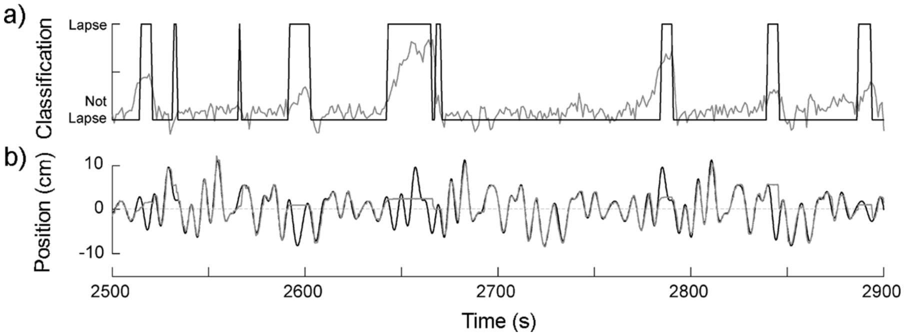

Fig. 2. Example of LSTM lapse detector performance. (a) Detector output (gray line) and target (black line). (b) Corresponding tracking behavior with target (black line) and response (gray line).

compared with the 30 component models assessed previously. This was confirmed as the network with a single LSTM unit in the hidden layer performed substantially better than the simple linear model for P4-O2 shown in Table II ($\mathrm{AUC-PR} = 0.35 \pm 0.07, \mathrm{AUC-ROC} = 0.82 \pm 0.02$), and the RMS error over the test set did not increase during training. With 2 units in the hidden layer we again observed over-fitting ($\mathrm{AUC-PR} = 0.33 \pm 0.07, \mathrm{AUC-ROC} = 0.82 \pm 0.02$).

The next best derivation based on the linear classifier analysis (Table II) were T6-02, according to AUC-PR, and P3-01, according to AUC-ROC, though these derivations did not differ from each other in either statistic ($\mathrm{p} > 0.05$). Consequently, we decided to train a detector with P3-O1 and P4-O2 as input channels because they are in different hemispheres and did not share a reference electrode, so seemed less likely to contain redundant information. Training a linear classifier gave $\mathrm{AUC-PR} = 0.30 \pm 0.06$ and $\mathrm{AUC-ROC} = 0.78 \pm 0.02$, while and a single hidden unit LSTM unit network gave $\mathrm{AUC-PR} = 0.36 \pm 0.07, \mathrm{AUC-ROC} = 0.83 \pm 0.02$, which was a slight improvement over the single derivation case. There was evidence of over-fitting, so we repeated the analysis and added weight decay, giving $\mathrm{AUC-PR} = 0.41 \pm 0.08, \mathrm{AUC-ROC} = 0.84 \pm 0.02$. This was our best overall classification result.

The analysis was repeated with the best four channels, P4-02, P3-01, T6-O2 and T5-O1, which gave a very similar result ($\mathrm{AUC-PR} = 0.41 \pm 0.07, \mathrm{AUC-ROC} = 0.83 \pm 0.03$). The 2-derivation classifier is preferred as it is more economical in terms of electrode usage (four electrodes versus eight).

To confirm the advantage for the 2-derivation LSTM classifier over a simple linear system, the linear delay analysis was repeated with two channels. The same pattern emerged as in the PCA analysis (Table I), with best performance being achieved with a 4-s input window ($\mathrm{AUC-PR} = 0.39 \pm 0.08, \mathrm{AUC-ROC} = 0.82 \pm 0.03$). Compared with this classifier, the 2-derivation LSTM classifier had a larger AUC-ROC ($\mathrm{p} = 0.021$), higher phi coefficient ($\mathrm{p} = 0.0063$), and lower RMS error ($\mathrm{p} = 0.036$), but no difference on AUC-PR. This classifier also bettered our best 30 component, multichannel linear and LSTM classifiers in AUC-ROC and RMS error ($\mathrm{p} < 0.05$ in both cases) but not in AUC-PR.

D. *Classifier Performance*

Having selected our best classifier model, we characterized its overall performance using several methods. Fig. 2. shows a typical output from the LSTM network over a 6.7-min period

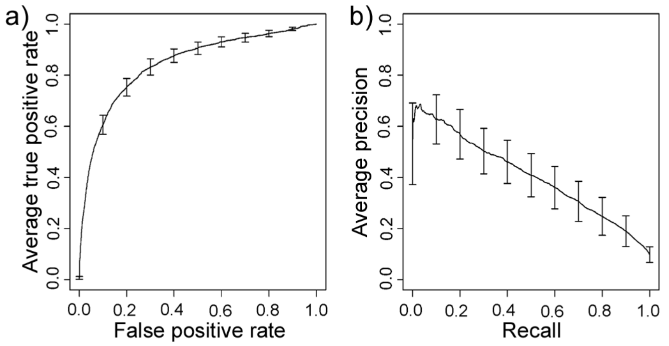

Fig 3. (a) Mean ROC curve. (b) Mean Precision—recall curve. On both graphs vertical bars indicate standard error.

from a subject 42 min into the second session. Fig. 3 shows a full ROC and precision-recall curves for this detector.

To assess classification performance without tuning to individual subjects we calculated an optimal threshold based only on the training data and applied this to the test data. To avoid bias, an optimum threshold based on phi correlation was calculated for each subject in the training set and the mean optimum threshold was then applied to the test data. This showed a moderate overall phi correlation ($\text{mean } \varphi = 0.38 \pm 0.05$, range 0.152–0.621). The system was moderately sensitive ($\text{mean } s_n = 0.63 \pm 0.05$, range 0.409–0.875) and highly specific ($\text{mean } s_p = 0.89 \pm 0.02$, range 0.732–0.946) but exhibited relatively poor precision ($\text{mean precision} = 0.33 \pm 0.06$, range 0.05–0.68), particularly for those subjects who lapsed only a few times during their two hours. Low precision is tolerable in a lapse detection system, as false alarms have low cost and are preferable to missed lapses.

Our system operates on a much shorter time scale than other similar systems in the literature. For comparison, we applied a 93.6-s exponential moving window to the binary network output and target to give an indication of performance under less stringent temporal resolution requirements. Smoothing the network output resulted in substantially higher correlation with the smoothed target than for the unsmoothed results ($\text{mean r} \pm \text{SE} = 0.61 \pm 0.09$, range 0.23 to 0.91). Smoothing the output yielded a very strong correlation for three of the eight subjects ($r > 0.8$).

IV. Discussion

We have reported results from the first system capable of detecting lapses in responsiveness in real-time and with second-scale temporal resolution. The system operates continuously and requires only 2 bipolar channels of EEG, which we have shown performs similarly to a system using 16 bipolar channels. We showed that using temporal information prior to a lapse improves detector performance. LSTM has the ability to detect patterns at arbitrary time-scales although comparison with TDL-MLP and TDL-Linear networks suggests essentially all the information for detection is contained within a 4.0-s window prior to a lapse. While current lapse detection performance is encouraging, we consider that the system is not yet sufficiently reliable for general use.

Our method for identifying lapses strikes a compromise between conflicting requirements for temporal resolution and simplicity. Other researchers have used simpler behavioral measures based on resultant tracking error [12] to judge alertness. The nature of our task prevented us using tracking error alone but, with full synchronous video of the face available, we were able to achieve acceptable temporal resolution. Inter- and intra-subject variation in tracking ability makes setting a reliable threshold on tracking error difficult and necessitates quite severe temporal smoothing of the error to achieve a meaningful metric. The nature of our 1-D driving-like tracking task made simple tracking error particularly inappropriate, as low tracking error can occur by chance when the target happens to move close to the response cursor. Nevertheless, we were able to achieve approximately second scale resolution by conservatively identifying lapses in the video and augmenting these with a simple measure based on tracking behavior—inappropriate tracking “flat-spots.”

This is the first reported application of the promising LSTM recurrent neural network [24] to EEG analysis. Unlike TDL-MLP, LSTM networks do not employ a fixed memory representation and can learn complex temporal relationships over arbitrary time scales. The LSTM architecture employs continuous internal states, which should allow then to represent more complex systems than discrete-state Hidden Markov Models as applied to the related sleep-staging problem [32], [33]. LSTM networks also overcome the “vanishing gradient” problem affecting most other recurrent neural network architectures when required to learn patterns over long time-lags. Given their ability to detect temporal patterns we were surprised to find that LSTM networks did not detect lapses from EEG any better than a relatively simple TDL-MLP network with a 4-s input window. This suggests EEG-log power spectrum patterns on longer time-scales are not useful for improving detector performance. We emphasize, however, that this study employed a limited parametrization of the EEG signal (relative log power in fixed bands at 1-s sampling interval). We intend to continue to explore the application of LSTM to EEG analysis with alternative parametrizations of the EEG. In particular, we believe that by increasing the sampling rate, the system may be able to resolve and use subtler temporal patterns occurring within the 4-s window prior to a lapse.

Several previous studies have looked at using EEG to detect lapses. Sommer *et al.* [3] used learning vector quantization to discriminate clear behavioral microsleeps from clear non-microsleeps in a night driving simulator. They achieved excellent classification rates (90.4%) by averaging the power spectrum over a long time window (8-s duration, starting 4 s before an event). In their design, data from all subjects were lumped together so that the performance figure disproportionately reflected those subjects who lapsed most frequently. In particular, by selecting only clear examples of lapses and attentive responsiveness, and ignoring the intermediary states, the discrimination task is made substantially easier. The clearest lapses, where the eyes close and the head drops forward, are more likely to be accompanied by EEG microsleep which, being clearly visible in the EEG, is easier to detect. These limitations need to be considered in interpreting their performance result. Other recent systems have focused on distinguishing low and high arousal levels as distinct from lapse episodes [20], [21].

Jung *et al.* [12] showed it is possible to use EEG log-power spectra applied to an MLP neural network to estimate alertness for an auditory vigilance task. They smoothed the missed stimulus time series using a 93.4-s long exponential moving window to derive a local error rate metric. Their system was able to estimate the local error rate with acceptable accuracy based on data from 2 EEG electrodes. While their results were promising, the detector was individualized (requiring training before it could be applied to a different individual) and had relatively limited temporal resolution. Their detector employed a static neural network, leaving open the possibility that temporal patterns in the power spectrum might be used to improve performance. Our results suggest their results could be improved by modeling log-power spectrum dynamics.

One of our aims was to design a detector able to generalize well to new subjects. For this reason, we formed a single between-subjects model with wide frequency bands to encourage generalization. The accuracy of within-subjects models, tuned to the idiosyncratic EEG rhythms of each subject, is likely to be superior but could not be used in a device without an extensive "training mode". A useful hybrid approach might be to automatically tune the algorithm to individuals, perhaps using unsupervised algorithms to identify subject specific spectral peaks. Alternatively, a better between-subjects model could be formed using a "stacked" approach, building many well tuned, narrow spectral band within-subject models, then forming a second-level between-subjects models to identify commonalities.

Eye-blink artifacts are not filtered out in our system, as we found the extra complexity was not warranted. Independent components analysis (ICA) was briefly investigated for eye-blink artifact removal [34], but despite the extra computational effort involved, removing eye blinks did not improve classifier performance. Consequently, we believe eye blink-artifacts, while present, do not strongly influence the reported results. Muscle artifacts were not removed from the EEG data and, consequently, the system may be using correlated changes in EMG activity to enhance lapse identification. While visual inspection showed little EMG activity in the parietal-occipital derivations, further investigation is required to properly address the influence of EMG activity on our results.

Efforts to interpret EEG concomitant with lapses tend to highlight confusion over the relationship between clinical EEG based estimates of cortical arousal and corresponding behavior [18], [35]. Clinical sleep staging [36] provides a global measure of the brain's level of arousal but sleep stage is only weakly correlated with behavior, particularly in the transitional stages between alertness and sleep [18]. This may be related to the apparent anatomical and functional independence of the arousal and attentional systems [16], [17], [35]. While responsiveness is generally better during high cortical arousal, there is evidence that attentional networks can operate at very low levels of arousal, perhaps even during apparent EEG sleep [37]. Maintenance of attention, and hence performance, during low arousal seems to depend on compensatory activation of anatomy common to the arousal and attention systems in the thalamus [38]. In an fMRI study, Portas *et al.* [38] showed increased activation of the thalamus when attention was maintained despite a state of low arousal. They suggested this may be related to the subjective experience of greater mental effort. We speculate that some lapses may be interpreted as a rapid disengagement of sustained attentional networks due to sudden relaxation in the compensatory activity of the thalamus [17]. Conversely, some lapses may also be caused by fatigue specific to attentional networks, regardless of the state of arousal. In support of this, we observed occasional tracking lapses that were not accompanied by signs of low arousal in the video. These may represent a distinct class of "attention-only" lapses.

While we do not include analysis of the characteristics of EEG-power fluctuations associated with lapses here, analysis of data from the same study is included in another recent paper [39]. The paper showed that lapses in this task are associated with increased power and positive correlations in the delta, theta, and alpha bands and decreased power in the beta, gamma, and higher bands. The finding of stronger correlations in the lower frequency bands is consistent with findings from similar studies [10], [12], and the well established association between slowing of the EEG rhythms and sleep-like states [36].

Our results suggest a way forward in the development of an EEG-based lapse detection system. Since temporal information on the scale of 4 s is useful in detecting lapses, our future work will focus on this time-scale and attempt to identify EEG dynamics that reliably herald an imminent lapse for all subjects.

References

[1] D. de Waard and K. A. Brookhuis, "Assessing driver status: A demonstration experiment on the road," *Accid. Anal. Prev.*, vol. 23, pp. 297–307, 1991.

[2] S. Porcu, A. Bellatreccia, M. Ferrara, and M. Casagrande, "Sleepiness, alertness and performance during a laboratory simulation of an acute shift of the wake-sleep cycle," *Ergonomics*, vol. 41, pp. 1192–1202, 1998.

[3] D. Sommer, T. Hink, and M. Golz, "Application of learning vector quantization to detect drivers dozing-off," in *Proc. Eunite. Symp.*, Albufeira, Portugal, 2002, pp. 99–103.

[4] S. K. Lal and A. Craig, "Driver fatigue: Electroencephalography and psychological assessment," *Psychophysiology*, vol. 39, pp. 313–321, 2002.

[5] Y. Harrison and J. A. Horne, "Occurrence of microsleeps during daytime sleep onset in normal subjects," *Electroencephalogr. Clin. Neurophysiol.*, vol. 98, pp. 411–416, 1996.

[6] R. Parasuraman*, The Attentive Brain*. Cambridge, MA: MIT Press, 1998.

[7] S. Makeig and M. Inlow, "Lapses in alertness: Coherence of fluctuations in performance and EEG spectrum," *Electroencephalogr. Clin. Neurophysiol.*, vol. 86, pp. 23–35, 1993.

[8] L. Torsvall and T. Akerstedt, "Sleepiness on the job: Continuously measured EEG changes in train drivers," *Electroencephalogr. Clin. Neurophysiol.*, vol. 66, pp. 502–511, 1987.

[9] G. Kecklund and T. Akerstedt, "Sleepiness in long distance truck driving: An ambulatory EEG study of night driving," *Ergonomics*, vol. 36, pp. 1007–1017, 1993.

[10] R. S. Huang, L. L. Tsai, and C. J. Kuo, "Selection of valid and reliable EEG features for predicting auditory and visual alertness levels," *Proc. Nat. Sci. Council Republic of China B*, vol. 25, pp. 17–25, 2001.

[11] S. Makeig and T. P. Jung, "Tonic, phasic, and transient eeg correlates of auditory awareness in drowsiness," *Brain Res. Cogn. Brain Res.*, vol. 4, pp. 15–25, 1996.

[12] T. P. Jung, S. Makeig, M. Stensmo, and T. J. Sejnowski, "Estimating alertness from the EEG power spectrum," *IEEE Trans. Biomed. Eng.*, vol. 44, no. 1, pp. 60–69, Jan 1997.

[13] S. Makeig and T. P. Jung, "Changes in alertness are a principal component of variance in the EEG spectrum," *Neuroreport*, vol. 7, pp. 213–216, 1995.

[14] K. F. Van Orden, T. P. Jung, and S. Makeig, "Combined eye activity measures accurately estimate changes in sustained visual task performance," *Biol. Psychol.*, vol. 52, pp. 221–240, 2000.

[15] S. Makeig, T. P. Jung, and T. J. Sejnowski, "Awareness during drowsiness: Dynamics and electrophysiological correlates," *Can. J. Exp. Psychol.*, vol. 54, pp. 266–273, 2000.

[16] P. Tassi, A. Bonnneford, A. Hoeft, R. Eschenlauer, and A. Muzetand, "Arousal and vigilance: Do they differ? Study in a sleep inertia paradigm," *Sleep Res. Online*, vol. 5, pp. 83–87, 2003.

[17] J. R. Foucher, H. Otzenberger, and D. Gounot, "Where arousal meets attention: A simultaneous fmri and EEG recording study," *Neuroimage*, vol. 22, pp. 688–697, 2004.

[18] R. D. Ogilvie, "The process of falling asleep," *Sleep Med. Rev.*, vol. 5, pp. 247–270, 2001.

[19] M. T. R. Peiris, R. D. Jones, G. J. Carroll, and P. J. Bones, "Investigation of lapses of consciousness using a tracking task: Preliminary results," in *Proc. 26th Annu. Int. Conf. IEEE Engineering in Medicine and Biology Society (EMBC 2004)*, San Francisco, CA, 2004, pp. 4721–4724.

[20] A. Vuckovic, V. Radivojevic, A. C. Chen, and D. Popovic, "Automatic recognition of alertness and drowsiness from EEG by an artificial neural network," *Med. Eng. Phys.*, vol. 24, pp. 349–360, 2002.

[21] S. K. Lal, A. Craig, P. Boord, L. Kirkup, and H. Nguyen, "Development of an algorithm for an EEG-based driver fatigue countermeasure," *J. Safety Res.*, vol. 34, pp. 321–328, 2003.

[22] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, pp. 1735–1780, 1997.

[23] J. Schmidhuber, F. Gers, and D. Eck, "Learning nonregular languages: A comparison of simple recurrent networks and LSTM," *Neural Comput.*, vol. 14, pp. 2039–2041, 2002.

[24] F. A. Gers, J. Schmidhuber, and F. Cummins, "Learning to forget: Continual prediction with LSTM," *Neural Comput.*, vol. 12, pp. 2451–2471, 2000.

[25] R. D. Jones, "Measurement of sensory-motor control performance capacities: Tracking tasks," in *The Biomedical Engineering Handbook*, J. Bronzino, Ed., 3rd ed. Boca Raton, FL: CRC Press, 2006, pp. 77:1–77:25.

[26] R. C. O'Reilly, C. K. Dawson, and J. L. McClelland, Pdp++ Neural Network Simulator 2003.

[27] T. Sing, O. Sander, N. Beerenwinkel, and T. Lengauer, "Rocr: Visualizing classifier performance in R," *Bioinformatics*, vol. 21, pp. 3940–3941, 2005.

[28] M. Goadrich, L. Oliphant, and J. Shavlik, "Learning ensembles of first-order clauses for recall-precision curves: A case study in biomedical information extraction," presented at the 14th Int. Conf. Inductive Logic Programming (ILP), Porto, Portugal, 2004.

[29] D. Sheskin*, Handbook of Parametric and Nonparametric Statistical Procedures*. Boca Raton, FL: CRC Press, 1997.

[30] T. S. Rögnvaldsson, "A simple trick for estimating the weight decay parameter," in *Neural Networks: Tricks of the Trade*, K. Muller and G. Orr, Eds., 1st ed. Berlin, Germany: Springer, 1998, pp. 71–92.

[31] D. C. Plaut, S. J. Nowlen, and G. E. Hinton, Experiments on Learning by Backpropagation Carnegie Mellon Univ., Pittsburgh, PA, Tech. Rep. CMU-CS-86-126, 1986.

[32] G. Gruber, A. Flexer, and G. Dorffner, "Unsupervised continuous sleep analysis," *Meth. Find Exp. Clin. Pharmacol.*, vol. 24, no. Suppl D, pp. 51–56, 2002.

[33] A. Flexer, G. Gruber, and G. Dorffner, "A reliable probabilistic sleep stager based on a single EEG signal," *Artif. Intell. Med.*, vol. 33, pp. 199–207, 2005.

[34] T. P. Jung, S. Makeig, C. Humphries, T. W. Lee, M. J. McKeown, V. Iragui, and T. J. Sejnowski, "Removing electroencephalographic artifacts by blind source separation," *Psychophysiology*, vol. 37, pp. 163–178, 2000.

[35] M. Sarter, B. Givens, and J. P. Bruno, "The cognitive neuroscience of sustained attention: Where top-down meets bottom-up," *Brain Res. Rev.*, vol. 35, pp. 146–160, 2001.

[36] A. Rechtschaffen and A. Kales*, A Manual of Standardized Terminology, Techniques, and Scoring System for Sleep Stages of Human Subjects*. Los Angeles: Univ. California, Brain Inf. Service/Brain Res. Inst., 1968.

[37] O. Winter, A. Kok, J. L. Kenemans, and M. Elton, "Auditory event-related potentials to deviant stimuli during drowsiness and stage 2 sleep," *Electroencephalogr. Clin. Neurophysiol.*, vol. 96, pp. 398–412, 1995.

[38] C. M. Portas, G. Rees, A. M. Howseman, O. Josephs, R. Turner, and C. D. Frith, "A specific role for the thalamus in mediating the interaction of attention and arousal in humans," *J. Neurosci.*, vol. 18, pp. 8979–8989, 1998.

[39] M. T. R. Peiris, R. D. Jones, P. R. Davidson, G. J. Carroll, and P. J. Bones, "Frequent behavioural microsleeps during an extended visuomotor tracking task in non-sleep-deprived subjects," *J. Sleep Res.*, vol. 15, no. 3, pp. 291–300, 2006.

Paul R. Davidson (S'95–M'01) was born in New Zealand in 1977. He received the B.E. (Hons.) and Ph.D. degrees in electrical and electronic engineering from the University of Canterbury, Christchurch, New Zealand, in 1998 and 2001, respectively.

He is Deputy Director of the Christchurch Neurotechnology Research Programme, based at the Van der Veer Institute for Parkinson's and Brain Research, Christchurch. He is also a Biomedical Engineer and Neuroscientist with the Department of Medical Physics & Bioengineering of the Canterbury District Health Board and an Adjunct Fellow in the Department of Electrical & Computer Engineering at the University of Canterbury. His research interests include machine learning for biomedical applications, human motor control and learning, and biomedical signal processing.

Dr. Davidson is a Member of the Australasian College of Physical Scientists and Engineers in Medicine. He was a Brain Physiology and Modeling Track Co-Chair for EMBC 2005 in Shanghai.

Richard D. Jones (M'87–SM'90) received the B.E. (Hons.) and M.E. degrees in electrical and electronic engineering from the University of Canterbury, Christchurch, New Zealand, in 1974 and 1975, respectively, and the Ph.D. degree in medicine from the Christchurch School of Medicine, University of Otago, Christchurch, in 1987.

He is Director of the Christchurch Neurotechnology Research Programme, a Biomedical Engineer and Neuroscientist with the Department of Medical Physics & Bioengineering of Canterbury District Health Board, a Research Associate Professor in the Department of Medicine at the Christchurch School of Medicine & Health Sciences of the University of Otago, and an Adjunct Associate Professor in the Department of Electrical & Computer Engineering at the University of Canterbury. He is Research Director of the Brain Research Division of the Van der Veer Institute for Parkinson's and Brain Research (http://www.vanderveer.org.nz), in which he is based.

Dr. Jones's research interests and contributions fall largely within neural engineering and the neurosciences, and particularly within human performance engineering—development and application of computerized tests for quantification of upper-limb sensory-motor and cognitive function, particularly in brain disorders (stroke, Parkinson's disease, traumatic brain injury) and driver assessment; eye movements in brain disorders; computational modelling of the human brain in relation to purposive movements; signal processing in clinical neurophysiology—EEG analysis for detection of epileptic activity and lapses of responsiveness; virtual reality approaches to neurorehabilitation.

Dr. Jones is a Fellow of the Institution of Professional Engineers New Zealand, a Fellow and a Past President of the Australasian College of Physical Scientists and Engineers in Medicine, a Fellow of American Institution for Medical and Biological Engineering, and a Fellow of the Institute of Physics (U.K.). He has been a member of most of the IEEE Engineering in Medicine & Biology Society's International Conference Committees since 1988, and was Co-Chair of Neural Engineering Theme at EMBC 2005 in Shanghai. He is on Editorial Board of the *Journal of Neural Engineering*, an Associate Editor of IEEE TRANSACTIONS ON NEURAL SYSTEMS AND REHABILITATION ENGINEERING. and a past Associate Editor of IEEE TRANSACTIONS ON BIOMEDICAL ENGINEERING.

Malik T. R. Peiris was born in Sri Lanka in 1978. He received the B.E. (Hons.) degree in electrical and electronic engineering from the University of Canterbury, Christchurch, New Zealand, in 2000. He is currently working towards the Ph.D. degree in electrical and electronic engineering at the University of Canterbury.

He is a member of the Brain Research Division of the Van der Veer Institute for Parkinson's and Brain Research (www.vanderveer.org.nz) in Christchurch, New Zealand and currently works as a Product Development Engineer for Fisher & Paykel Healthcare, Auckland, New Zealand. His research interests include biomedical signal processing applied to detecting lapses in responsiveness from the EEG and detecting respiratory-related events in patients with obstructive sleep apnea.