Multimodal Ambulatory Sleep Detection

Weixuan Chen[1,*], Akane Sano[1,*], Daniel Lopez Martinez[1,2], Sara Taylor[1], Andrew W. McHill[3], Andrew J. K. Phillips[3], Laura Barger[3], Elizabeth B. Klerman[3], and Rosalind W. Picard[1]

***Abstract*— Inadequate sleep affects health in multiple ways. Unobtrusive ambulatory methods to monitor long-term sleep patterns in large populations could be useful for health and policy decisions. This paper presents an algorithm that uses multimodal data from smartphones and wearable technologies to detect sleep/wake state and sleep episode on/offset. We collected 5580 days of multimodal data and applied recurrent neural networks for sleep/wake classification, followed by cross-correlation-based template matching for sleep episode on/offset detection. The method achieved a sleep/wake classification accuracy of 96.5%, and sleep episode on/offset detection F1 scores of 0.85 and 0.82, respectively, with mean errors of 5.3 and 5.5 min, respectively, when compared with sleep/wake state and sleep episode on/offset assessed using actigraphy and sleep diaries.**

I. INTRODUCTION

Inadequate sleep impairs quality of life and results in increased risk of morbidity and mortality. Chronic sleep disturbances have been associated with medical problems, including diabetes [16], obesity, and psychological conditions[17]. There is a need for better tools to enable accurate long-term evaluation of sleep timing and duration in daily life.

While sleep measurements based on polysomnography (PSG) are currently the gold standard, existing PSG technologies are impractical for long-term home use. Smartphones and wearables that measure acceleration, skin temperature, skin conductance, light exposure, and behavioral parameters offer possibilities for easy-to-use, long-term daily sleep monitoring.

We present a novel method to automatically detect sleep/wake and sleep episode on/offset times using multimodal data from a smartphone and a wrist-worn sensor. While previous methods have leveraged accelerometer [15], smartphone [1], [2], or biosensor [6] data alone, our method combines the multimodal ambulatory physiological and behavioral data offered by both smartphone and wrist-worn sensors. Our method consists of binary sequential classification of epochs by recurrent neural networks, commonly used in deep learning, and is followed by a sleep episode on/offset detection algorithm based on cross-correlation with trained templates.

The main contributions of this work are: (1) comparisons of sleep/wake and sleep episode on/offset detection performance using physiological and behavioral data from a mobile phone and a wearable sensor with data from actigraphy and sleep diaries, which are two methods frequently used in ambulatory sleep studies; (2) use of recurrent neural networks for ambulatory sleep state detection.

II. METHODS

A. Data acquisition

186 undergraduate students in 5 cohorts participated in an ∼30-day study (120 males, 66 females, age: 18-25) that produced 5580 days of data. Participants were recruited through email. During the ∼30-day experiment, participants (i) wore a wrist sensor on their dominant hand (Q-sensor, Affectiva, USA) to measure 3-axis acceleration (ACC), skin conductance (SC), and skin temperature (ST) at 8 Hz; (ii) installed an Android phone application using the *funf* open source framework [7] to measure timing of calls, timing of short message service (SMS), location, and timing of screen-on; (iii) wore a wrist actigraphy monitor on their non-dominant hand (Motion Logger, AMI, USA) to measure activity and light exposure levels every 1 minute; and (iv) completed a sleep diary every morning to record bed time, sleep latency, wake time, and the number and timing of awakenings. The sleep diary was inspected by an experimenter every day to check completion and to obtain corrections or clarifications from the student if there were any clear errors or missing data.

We used a previously established method to score sleep from diaries and actigraphy data [18]. An experienced investigator first reviewed the data and selected analysis windows for potential sleep episodes based on the combined diary and activity data. Software (Action-W) set the sleep episode on/offset times and classified each epoch as sleep or wake. Based on the sleep episode time and duration, the investigator labeled each sleep episode as either a main sleep or a nap. From this procedure, we obtained (1) a classification of sleep or wake for every 1-min epoch, (2) sleep episode on/offset times, (3) whether a sleep episode was a main sleep or a nap. These labels were used as "ground truth". Fig. 1 shows an exemplary day of raw data collected in our study in which the first stage labels are superimposed. These assessments were used to train and test results from the Q-sensor (i, above) and phone data (ii, above).

This work was supported by NIH (R01GM105018, K24HL105664, P01AG009975, R01HL114088, R00HL119618, F32DK107146, and R21HD086392), NSBRI (HFP02802, HFP00006 and HFP04201), the Harvard-Australia Fellowship, MIT Media Lab Consortium, and Samsung Electronics.

*Both authors contributed equally to this work

[1]Affective Computing Group, Media Lab, Massachusetts Institute of Technology cvx@media.mit.edu

[2]Harvard-MIT Division of Health Sciences and Technology, Massachusetts Institute of Technology.

[3]Division of Sleep and Circadian Disorders, Brigham and Women's Hospital, Harvard Medical School.

978-1-5090-4179-4/17/$31.00 ©2017 IEEE

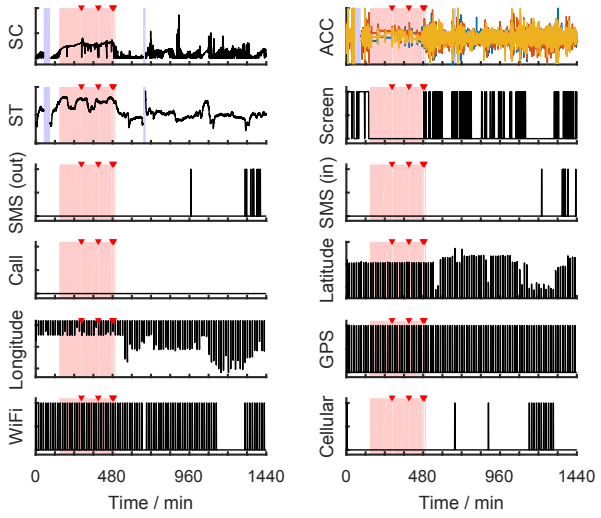

Fig. 1. Raw data streams from an exemplary day. The pink bars mark sleep epochs and the red triangles indicate waking up during the night as determined from actigraphy and sleep diary. The blue bars denote missing data. (SC = skin conductance, ACC = accelerations of three axes, ST = skin temperature)

TABLE I

FEATURE SETS FOR SLEEP DETECTION

Source	Modality	Feature variables
Wrist sensor	Skin conductance (SC)	Mean, SD, power within 0-0.1, 0.1-0.2, 0.2-0.3, 0.3-0.4, and 0.4-0.5Hz bands, the number of SC responses, storm flag, elapsed time since a storm started
	Acceleration (ACC)	Mean, SD
	Skin temperature (ST)	Mean, SD
Phone	Screen	Screen was on, the time the screen was turned on
	SMS	Sent a message
	Call	On a call, missed a call
	Location	Movement index, connected to WiFi, connected to cellular nets
Time	Time	Elapsed minutes since 12:00 AM

B. Feature preparation

Table I shows the features we computed for each time window. A window length of 20 or 30 seconds is the convention for PSG sleep scoring [8], while other studies using ambulatory data adopt 10-min [1] or 5-min [2] windows. In this study, we used a window length of 1 minute without overlap to match the scale of our ground truth labels. There were several reasons why we chose these feature variables. First, it has been shown that SC is more likely to have periods of high frequency activity called "storms" during deep sleep [3]; we used algorithms developed to automatically detect storms in SC data [4]. Therefore, for the SC modality we computed mean, standard deviation, and frequency-domain features, including powers for five frequency bands (0-0.5 Hz, in 0.1 Hz intervals), and three storm features according to [4], including the number of SC responses, storm flag (whether we observe a storm in that minute), and the elapsed time since a storm started. Also, ST rises during sleep in individuals in living environments similar to those in our experiment [11]. Second, our phone app recorded time stamps both when phone users sent a SMS and also when they received a SMS. Since receiving a SMS is a passive behavior that could happen during sleep, we only kept SMS-sending events as a feature variable. Third, the raw location data acquired in our experiments were the latitude and longitude of phones, whose absolute numbers are nearly meaningless for sleep estimation across subjects. Hence, we developed a movement index for each minute, formulated as the arithmetic mean of the variances of the latitude and the longitude, to indicate whether a user was actively moving in that minute.

We had missing data lasting from a few minutes to several hours because of phone and sensor charging, and activities such as removal for a shower. We used a two-step strategy to solve this problem. First, a 25% missing tolerance threshold was applied to the wrist-worn sensor data: if any modalities had a missing rate higher than 25% within a day, the whole day's data were dropped. This rule was not adopted on the phone data for sporadic events such as sending a SMS, because we cannot discriminate if such events did not happen or were missed. Second, for the remaining days, we filled minutes with missing data using the average of the same feature variable over the remaining part of the same day. After dealing with the missing data, we had 3439 days of data left. To train and evaluate a machine learning model, we needed to split our features and labels into a training set and a test set. The simplest way to do this is to randomly assign every minute to either the training or test set. However, sleep/wake detection can be improved by using both past and future information. If we assign two consecutive days to the training and test sets respectively, the first period of the second day will lose its past information. Therefore, we connected consecutive days of each subject into chunks, and then randomly cut the first or last 20% of each chunk as the test set. Finally, we had 2772 days in the training set, and 667 days in the test set. To equalize features and help with gradient descent optimization, every feature variable was also normalized to the $[0,1]$ range within each day.

C. Sleep/wake detection

Our goal is to automatically classify every minute of data as sleep or wake, which is a binary sequential classification problem. We wish to use a model that exploits how current feature variables can depend on both past and future ones. For example, if a participant turned on her phone screen at 11:01pm, it would be highly likely that she was still awake at 11:00pm. One powerful tool to solve this kind of problem is a recurrent neural network (RNN). Long Short-Term Memory networks (LSTMs), first proposed in [10], are a special kind of RNN, capable of learning long-term dependencies. LSTMs have recently shown great success in sequence learning tasks such as speech recognition [12] and machine translation [13], [14].

Fig. 2 shows the structure of the bidirectional LSTM neural network we used for sleep detection. The vector x_t contains all the features at time t, and y_t is a binary label indicating sleep or wake for each minute. The activation

function used in the fully-connected layer is rectified linear units. The bidirectional neural network was trained using RMSprop [9] with binary cross-entropy loss. We set the past- and future-looking window lengths to 30 min each based on the work of Min et al. [1]. The whole algorithm was implemented using deep learning frameworks Theano 0.8.2 and Keras 1.0.5.

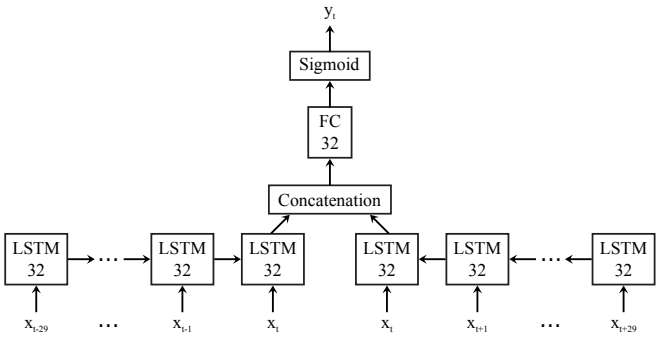

Fig. 2. Our bidirectional LSTM model for sleep detection. Output dimensions are denoted in each box. (FC = fully-connected layer)

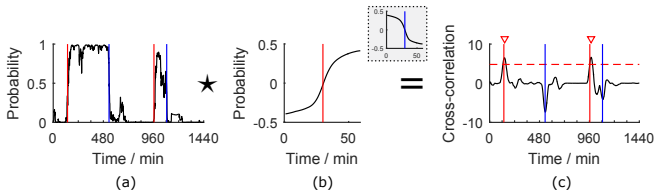

Fig. 3. An exemplary day showing our sleep episode on/offset points estimation method. The red and blue lines indicate the sleep episode on/offset points from "ground truth" method. (a) Sleep detection results in probabilities. (b) 59-min templates around a sleep episode onset or offset (in the gray box) point. (c) Cross-correlation results with the optimized peak detection threshold and the detected peaks denoted.

D. Sleep on/offset points estimation

After sleep/wake detection, we estimated sleep episode on/offset points. Fig. 3 shows an example of our method using cross-correlation and peak detection algorithms. First, to summarize the pattern of a sleep episode onset point (sleep-to-wake transition point) or sleep episode offset point (wake-to-sleep transition point) in the sleep detection results, we sampled an L-min window of detection probabilities (Fig. 3 (a)) around every sleep episode onset or offset point in the training set, and computed the average of them respectively to form two templates T_s and T_w (Fig. 3 (b)). After removing mean values from the templates, the templates were compared to a 1-min-stride sliding window of the detection probabilities using cross-correlation to find time points with the highest similarities to them. Let $P[t]$ be the sleep detection probability of time t. The cross-correlation is defined as

$$(P \star T_{s/w})[t] = \sum_{\tau=1}^{L} P[\tau + t - \frac{L+1}{2}] \cdot T_{s/w}[\tau] \qquad (1)$$

in which the translation factor $\frac{L+1}{2}$ is for centering the template around time t.

Fig. 3 (c) displays the cross-correlation results of the example, in which the peaks are sleep episode on/offset point candidates. To localize them, we applied the *findpeaks* function in MATLAB R2015b to the signal to find local maxima satisfying certain conditions. On one hand, to eliminate potential false positives, the height of a detected peak needs to be higher than a threshold, which was optimized on the training set towards a higher F_1 score (introduced below) and applied to the test data. On the other hand, the distance between two peaks needs to be longer than 30 min. This rule was set to avoid false positives, since the shortest time interval between two neighboring sleep episode on/offset points was 45 min in our data.

The detected peaks in the cross-correlation data of the test set were then compared to the ground truth quantitatively. If the distance between a peak and its closest sleep episode on/offset point in the ground truth were less than 30 min, we defined the peak as a true positive. Based on this, we computed the precision, recall, and F_1 score (the harmonic mean of precision and recall) of our estimation. For all the true positive points, we also reported the average of their distances to the ground truth as the estimation errors.

III. Results

We summarize sleep detection performance in Table II, comparing two main cases: whether the algorithm could not use the clock time ("without time") or could use the clock time as an input feature ("with time"). This quantifies how much the algorithm is biased by sleep being more likely to occur at night. ACC + ST showed the best performance, 96.2% and 96.5% accuracy without and with time, respectively, and phone features showed the worst performance.

Table II also shows a summary of sleep episode on/offset detection with a 59-min template. We obtained the best F_1 score with ACC + ST features both for sleep episode on/offset without and with time features. Without time features, the error was smallest for the wrist sensor + phone for sleep episode onset and for ACC + ST for sleep episode offset. With time features, the ACC showed the smallest error for sleep episode onset and ACC + ST for sleep episode offset.

For detecting sleep episode on/offset, we also compared different lengths of the template, from 9 to 239 min. Here, we describe the best template lengths for sleep episode on/offset with ACC + ST + time. For sleep episode onset, F_1 score was highest (0.87) at the template length of 169 min (the mean error 5.4 min), and for sleep episode offset, F_1 score was highest (0.85) when the template length was 199 min (error 6.0 min). On the other hand, we observed that the mean errors of sleep episode onset or offset increased as we increased the length of the template beyond 79 min.

In order to test whether our model's performance differed between main sleep and naps, we computed the percent of main sleep and naps that were successfully detected. Our results showed that with ACC + ST and time features, 93% of epochs within main sleep were correctly classified (3% sleep was misrecognized as wake, 4% wake was misrecognized

TABLE II

SLEEP DETECTION, SLEEP EPISODE ON/OFFSET DETECTION PERFORMANCE WITH BEST RESULTS HIGHLIGHTED IN BOLD (ME: MEAN ERROR (MIN))

Feature combinations	Sleep detection accuracy		Sleep episode on/offset detection							
			Without time				With time			
	Without time	With time	Sleep onset		Sleep offset		Sleep onset		Sleep offset	
			F_1	ME	F_1	ME	F_1	ME	F_1	ME
Wrist sensor	95.9	96.5	0.83	5.0	0.80	5.8	0.83	5.3	0.81	5.7
Phone	76.7	89.1	0.24	11.5	0.16	16.2	0.41	9.6	0.35	11.7
Wrist + Phone	96.0	96.3	0.83	**4.7**	0.80	5.8	0.83	5.7	0.80	6.0
ACC + EDA	95.6	96.3	0.83	5.1	0.80	6.3	0.84	5.4	0.81	5.9
ACC + ST	**96.2**	**96.5**	**0.84**	4.9	**0.81**	**5.1**	**0.85**	5.3	**0.82**	**5.5**
EDA + ST	92.5	94.5	0.70	7.3	0.69	7.3	0.73	6.9	0.71	6.3
ACC	95.5	96.3	0.82	5.2	0.79	7.2	0.82	**5.1**	0.79	6.3
EDA	90.8	93.7	0.67	8.1	0.64	8.1	0.70	6.8	0.68	7.3
ST	86.7	90.7	0.49	10.7	0.48	11.6	0.51	9.6	0.56	9.1

as sleep) and 65% of epochs within naps (33% sleep was misrecognized as wake, 2% wake was misrecognized as sleep) were successfully detected. As expected, given the irregular timing of naps, the performance in detecting naps was lower when using time features.

IV. DISCUSSION

Our classifier with LSTM showed the best accuracy (96.5%) with features from ACC, ST, and time. The combination of movement features (ACC) with ST helps distinguish sleeping and being still from being awake and still. Features from phones showed the lowest accuracy. In prior work [1], only features from phones were used to detect sleep. Since in our study we used different information from the phone, we could not simply compare our accuracy from phone data with those in the previous studies.

Our results showed higher accuracy (96.5%) using devices which are less burden to users than that in previous work (93.1% with phone data [1], or 93.2% with electrocardiogram, respiration + ACC data [6]). We also had smaller mean errors (onset error: 5.1 min, offset error: 5.5 min) than reported in sleep episode on/offset (onset error 35 min, offset error 31 min) [1], and smaller total sleep time errors than the 40.2 min [6] and 42 min [2]).

One limitation is that the ground truth used was not PSG, the gold standard. We used diaries, plus actigraphy, with an experienced investigator reviewing the data; this process is time consuming and does not always produce the same results as PSG when PSG and actigraphy are collected simultaneously. Ensemble labels could also be obtained from multiple investigators to lower potential bias. In addition, our phone data did not distinguish if we had missing data (e.g., phone battery ran out and phone was off), or if participants did not use their phone or were not carrying their phone. Although we asked our participants to charge their phone every day, our data might include some time when a phone was not operating.

For future work, we need to further test how our choices of parameters (e.g., the missing tolerance threshold and the minimum peak distance) affected our results. In order to assess the generalizability of our method, the algorithm also needs to be trained and tested with data from different participants, and evaluated in other populations, including people who do not intensively use their mobile phones, people with medical conditions and/or on medications that may affect the Q-sensor data, and people whose primary wake episode is not during the day (e.g., night or shift workers). In addition, while here we describe a general model for all users, we could also build personalized models with individual data or keep updating a model while capturing daily sleep data. In this way, the model performance could improve especially for irregular sleepers, shift workers or frequent travelers. Sleep/wake detection performance when using only the phone data might further improve if our phone application was modified to collect acceleration, audio and ambient light.

REFERENCES

[1] J. Min *et al.*, "Toss 'N' Turn: Smartphone as Sleep and Sleep Quality Detector," Proc. SIGCHI Conf. Hum. Factors Comput. Syst., pp. 477–486, 2014.

[2] T. Hao *et al.*, "Unobtrusive Sleep Monitoring using Smartphones," in SenSys, 2013, p. 4:1–4:14.

[3] K. Asahina, "Paradoxical phase and reverse paradoxical phase in human sleep," J. Phys. Soc. Jpn., 24, pp.443–450, 1962.

[4] A, Sano *et al.*, "Quantitative analysis of wrist electrodermal activity during sleep," Int. J. Psychophysiol., vol. 94, no. 3, pp. 382–389, 2014.

[5] S. H. Hwang *et al.*, "Sleep Period Time Estimation Based on Electrodermal Activity," IEEE J. Biomed. Heal. Informatics, pp. 1–1, 2015.

[6] W. Karlen, "Adaptive wake and sleep detection for wearable systems," PhD Thesis, EPFL, 2009.

[7] N. Aharony *et al.*, "The social fMRI," in UbiComp, 2011, p. 445.

[8] American Academy of Sleep Medicine, "The AASM Manual for the scoring of sleep and associated events," Am. Acad. Sleep Med., 2007.

[9] T. Tieleman, and G. Hinton, "Lecture 6.5 - RMSProp," COURSERA: Neural Networks for Machine Learning. Technical report, 2012. fewer

[10] S. Hochreiter and J. Schmidhuber, "Long short-term memory," in Neural Computation, 1997.

[11] J. A. Sarabia *et al.*, "Circadian rhythm of wrist temperature in normal-living subjects," Physiol. Behav., vol. 95, no. 4, pp. 570–580, 2008.

[12] A. Graves and N. Jaitly, "Towards end-to-end speech recognition with recurrent neural networks," in ICML, 2014.

[13] I. Sutskever *et al.*, " "Sequence to sequence learning with neural networks,"" in NIPS, 2014.

[14] K. Cho *et al.*, "On the properties of neural machine translation: Encoder-decoder approaches," in SSST Workshop, 2014.

[15] R. J. Cole *et al.*, "Automatic sleep/wake identification from wrist activity," Sleep, vol. 15, no. 5, pp. 461–9, 1992.

[16] F. Zizi *et al.*, "Sleep duration and the risk of diabetes mellitus: Epidemiologic evidence and pathophysiologic insights," Curr. Diab. Rep., vol. 10, no. 1. pp. 43–47, 2010.

[17] T. Shochat, "Impact of lifestyle and technology developments on sleep," Nat. Sci. Sleep, vol. 4, pp. 19–31, 2012.

[18] L. K. Barger *et al.*, "Prevalence of sleep deficiency and use of hypnotic drugs in astronauts before, during, and after spaceflight: an observational study.," Lancet. Neurol., vol. 13, no. 9, pp. 904–12, 2014.