Comparative assessment of sleep quality estimates using home monitoring technology

Jose M. Perez-Macias – *IEEE Student Member*, Holly Jimison – *EMBS Member*, Ilkka Korhonen – *EMBS Senior Member*, and Misha Pavel, *Senior IEEE Fellow*

***Abstract*— Poor sleep quality is associated with chronic diseases, weight increase and cognitive dysfunction. Home monitoring solutions offer the possibility of offering tailored sleep coaching interventions. There are several new commercially available devices for tracking sleep, and although they have been tested in sleep laboratories, little is known about the errors associated with the use in the home. To address this issue we performed a study in which we compared the sleep monitoring data from two commercially available systems: Fitbit One and Beddit Pro. We studied 23 subjects using both systems over a week each and analyzed the degree of agreement for different aspects of sleep. The results suggest the need for individual-tailoring of the estimation process. Not only do these models address improved accuracy of sleep quality estimates, but they also provide a framework for the representation and harmonization for monitoring data across studies.**

I. Introduction

Sleep is an important factor in our health and wellbeing. Lack of sleep has been related to obesity, diabetes, heart diseases, dementia, and depression [1]. In 2008 a report from the Centers for Disease Control and Prevention warned that 11.3 % of Americans reported not having sufficient sleep or rest during the last 30 days [2]. An important step in improving sleep is a reliable assessment. Polysomnography (PSG) is considered the *de facto* gold standard and the primary tool for sleep monitoring [3]. Among its disadvantages are the high cost, its intrusiveness, time consumed and the disturbance of the usual bed environment [4]. To overcome some of these disadvantages, new methods have been developed for monitoring sleep in the home with approaches including passive infrared sensors [5], pressure sensitive mats, EEG electrode headbands, and wrist actigraphs. The cost of these systems have been much reduced in recent years, making them available for routine consumer use. However, the accuracies of these systems vary according to the context. Some are more intrusive than others and require user adherence to work properly.

This research was supported by the National Technology Agency of Finland (TEKES), distinguished Professor Programme, and by the Tampere University of Technology Foundation.

Jose M. Perez-Macias is with the Department of Signal Processing, Tampere University of Technology, Tampere, 33720 Finland (phone: +358 414913404; e-mail: jose.perez-macias.eng@ieee.org).

Holly Jimison and M. Pavel, are with the College of Computer & Information, Northeastern University, MA CO 02115 USA, and the Department of Signal Processing, Tampere University of Technology, Tampere 33720 Finland (e-mails: h.jimison@neu.edu, and m.pavel@neu.edu).

Ilkka Korhonen, is with the Department of Signal Processing, Tampere University of Technology and the VTT Technical Research Centre, Tampere, 33720 Finland (e-mail: ilkka.korhonen@tut.fi).

Many of the new sleep devices designed for consumer sleep tracking use accelerometer technology (e.g., Fitbit One, Traxmeet, Lark, WakeMate, Jawbone UP, SleepTracker). There are applications (iSleep, SleepCycle, and SleepRate among others) to estimate sleep time and quality using cell phones' accelerometers. MyZeo (out of business in 2013) used both actigraphy and electro-encephalography (EEG) to estimate sleep parameters. Additionally, Beddit Pro [6], uses a fully unobtrusive piezoelectric sensor strip (placed under the bed sheets) in conjunction with ballistocardiography (BCG) analysis to estimate sleep parameters by combining heart rate (HR), breathing and movement. Our lab has tested Beddit Pro, Fitbit One, Traxmeet, Firstbeat devices and self-reporting for comparative analysis to characterize the accuracy of the different types of sleep monitoring approaches. Previous studies have found that Fitbit overestimated the population average of the total sleep time (TST) and sleep efficiency (SE) when compared to the laboratory standard of PSG [7]. Beddit has proven to provide accurate estimates of heart rate (HR) and breathing [8] that together can be used to estimate sleep parameters [9]. The present study was conducted to examine sleep assessment by Fitbit in comparison to Beddit while exploring the possibility of tailoring the estimation technique for individual participants.

II. Material and Methods

A. Subjects

The study sample consisted of 23 healthy subjects (18 males and 5 females). Subjects' age ranged from 21 to 31 years. They were recruited from the Tampere University of Technology in Finland by word of mouth and by advertisement in internal mailing lists. Inclusion criteria included age 18 - 65 years and good health. Participants who shared the bed with a partner, had a BMI > 28, or were pregnant were excluded from the trial. Additionally, those who were away from home during the study were removed from analysis. All volunteers agreed to participate in this research and signed a written informed consent.

B. Protocol

The trial had duration of 7 to 10 consecutive days. Participants were instructed to carry a Fitbit One (Fitbit Inc., CA, USA) in a pocket or on the wrist during the day and to wear it in a band on the wrist and activate the *night mode* when "they decided to go to sleep" and after watching television or reading. The Beddit Pro (Beddit, Espoo, Finland) sensor strip was installed directly under the bed sheets, and the set-top box next to the bed. This device does

978-1-4244-7929-0/14/$26.00 ©2014 IEEE

not require any type of intervention from the user and it is programmed to record data during a specific timeframe. Participants were also fitted with a reference actigraph (Traxmeet, Helsinki, Finland) on the dominant hand, a Firstbeat Bodyguard 2 HR monitor (Firstbeat, Jyväskylä, Finland), and were instructed to fill a sleep and activity diary daily and a subjective sleep quality assessment questionnaire before and after the trial. In this paper we analyzed Fitbit and Beddit data to explore accuracy and inference issues. Remaining recordings were not taken into account in the current study. However, questionnaires were however used to better interpret the results.

C. Data recording

Fitbit One uses a 3-D Micro-Electro Mechanical System (MEMS) accelerometer. It has two modes: day mode and sleep mode. Sleep mode is enabled by pushing a button at the beginning and in the end of sleep. Using proprietary algorithms, it estimates steps, distance, climbed stairs, burned calories throughout the day, and sleep parameters when night mode is activated. The device also displays a growing flower graphic showing progress toward a daily target activity goal.

Beddit Pro has four types of sensors: a temperature sensor, a microphone, a light sensor, and a piezoelectric sensor. The first three sensors are located in a set-top box, which is connected to the piezoelectric sensor by a cable. This force sensor is a flexible 4cm x 70cm x 0.4mm foil placed under the bed sheets. Beddit measures both body signal variables and environmental variables. Body signal variables include heart rate, sleep structure (light, REM, and deep sleep), stress, total sleep time (TST), and presence by using post-processing algorithms from the piezoelectric sensor (sampling frequency of 140 Hz, 16 bit). Environmental variables include temperature (every 5 min), noise (every 5 min) and light (every 5 min).

Both systems, Beddit and Fitbit, compute a similar set of sleep variables detailed in their *API* documentation. Data were recorded and synchronized with the manufacturers' servers using their software. Recorded data were downloaded from Beddit and Fitbit servers to a mash-up server W2E (Wellness Warehouse Engine) [11-13]. Finally, data were downloaded from the W2E for each of the participants for analysis. Data quality assessments and validations were performed visually: we selected nights where both Fitbit and Beddit data were within the expected limit and available simultaneously.

TABLE I. STANDARD VARIABLES AND THEIR CORRESPONDENCE IN FITBIT AND BEDDIT APIS.

Variable	API's variable names
Beddit- TST, minutes	time_sleeping
Beddit- SE, %	sleep_efficiency
Fitbit-TST, minutes	minutesAsleep
Fitbit-SE, %	summary.totalMinutesAsleep /fitbitsleep{p}.summary.totalTimeInBed*100
Beddit- sleep stages, array	sleep_stages
Fitbit-NWAK, no.	awakeningsCount
Fitbit-SOL, minutes	minutesToFallAsleep

D. Sleep parameters

Four basic sleep parameters were compared: Total Sleep Time (TST), Sleep Efficiency (SE), Sleep-Onset Latency (SOL), and number of awakenings (NWAK). TST is the total amount of time that the subject spent sleeping, SE is the number of minutes of sleep divided by the number of minutes in bed, SOL is the amount of time in minutes that is required to transit from full wakefulness to sleep, and NWAK is the number of times the person goes out of bed after the sleep onset and before the final awakening.

Whereas TST and SE are available directly from both APIs, Beddit does not directly provide SOL and NWAK. Instead, it provides the sleep stages estimates in the form of an array of tags with their corresponding timestamp.

Beddit distinguishes between A-away, W-wake, L-light, D-deep sleep, R-REM sleep and M-missing. SOL was computed as the time difference between the first W and the first L, D, or R tags. NWAK was computed as the number of times the letter A appears after the first L, D or R stages, and before the last L, D or R. Table 1 lists the sleep variables used in this paper and their corresponding variables in each of the manufacture's APIs.

E. Analysis

Ideally, the two systems would yield the same data for a given subject and date. To examine this hypothesis we analyzed, for each subject, the sleep measures from Fitbit as a function of the measures from Beddit for TST, SE, SOL and NWAK. This choice was made since Beddit is more likely to provide more direct information regarding the participants in and out of bed transition.

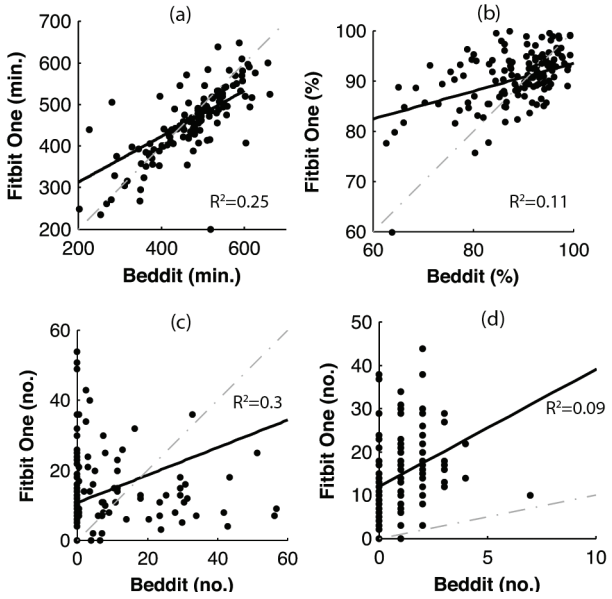

Figure 1. Scatter plots representing (a) TST, (b) SE, (c) SOL and (d) NWAK as measured by Fitbit and Beddit for all subjects and all nights. Each night is indicated by a filled circle; the regression model is represented by solid lines and the ideal linear transformation (identity) is shown as dashed lines.

To assess the degree of agreement between the two systems we first computed a linear regression for all participants that recorded at least three nights. The coefficient of correlation and the coefficient of determination R^2 were used to represent the degree of agreement between the two systems. We used two models for this comparison, one with the intercept set to zero (strong model) and the other estimated intercepts for each individual *i* (strong model). The strong model, $f = b$, in which a Fitbit feature *f*, is predicted by the corresponding Beddit feature *b*, is used to assess the ability to provide absolute estimates assuming that the two systems measure exactly the same features. The second model, with subject-dependent parameters, $f = c_i + d_i\, b$, was used to estimate our ability to harmonize data from different devices with minimal measurement errors. Both models were evaluated by computing the root mean squared error (RMS) and the coefficient of determination (R^2). The statistical analyses were performed using software packages MATLAB (version 7.0; Mathworks, Natick, MA) and SPSS (Release 20.0.0; IBM Corp, Armonk, NY), respectively.

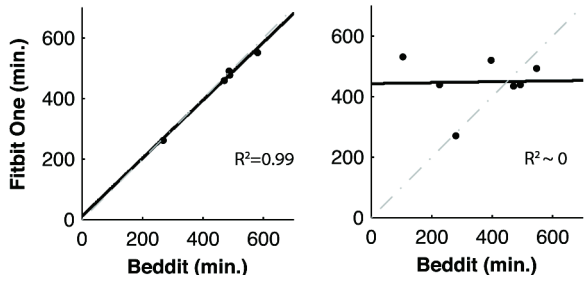

Figure 2. Scatterplots of Fibit and Beddit for TST for two subjects. Each point represents the Beddit and the Fitbit estimate for 1 night of data. The dashed line represents the diagonal line (perfect agreement) between the two measurements. The solid line represents the best-fit regression line.

III. Results

A. Population Summary

A total of twenty-three subjects, 5 women and 18 men (mean age 25.8 ± 2.9 years, height 172.2 ± 24.7 cm, mass 74.3 ± 12.5 kg), participated in the trial. One subject was excluded from the study due to a device malfunction. A total of 138 nights were recorded. Average Beddit-TST time was 469.9 (SD ±96.0) min, SE was 87.8% (SD ± 9.2%), SOL was 17.6 (SD ± 36.5) min, and NWAK was 0.97 (SD ± 1.18) times. Average Fitbit-measured TST was 461.6 (SD ±106.4) min, and SE was 90.2% (SD ± 7.7%), SOL was 17.7 (SD ± 25.7) min, and NWAK was 14.7 (SD ± 10.7) times.

Scatterplots and the associated linear regression for the estimates, of the four sleep parameters from all subjects and for all nights are shown in Figure 1. The linear model (*Fitbit* = A_0 + A_1**Beddit*) was used to assess the agreement between the two systems. The scatter plots as well as the analysis suggest very poor agreement between the systems due to significant variability. In particular, the coefficient of determination suggests that the Fitbit estimates accounted only for less that 30% of the variance. Fitbit appeared to overestimate SE, NWAK, and SOL.

B. Individual Subjects

Although the population averages suggest a poor agreement between the systems, it is possible that the correspondence for some individuals is acceptable. To examine this hypothesis we performed a separate linear regression analysis for each participant. Examples of two scatterplots for two sample participants are shown in Fig. 2. The regression estimates for the participant on the left are excellent, while they are poor for the participant on the right, as demonstrated by the corresponding R^2 values. The overall quality of the fit and the accuracy of the estimates for all participants and all four features was represented by R^2 computed for each participant. The distributions of R^2 for each sleep parameter are shown in Fig. 3. These distributions are consistent with our hypothesis that the correspondence between the Fitbit and Beddit estimates are poor for some, but excellent for other participants. Agreement for TST was high ($R^2 > 0.70$) for twelve participants and not as good for the other sleep variables.

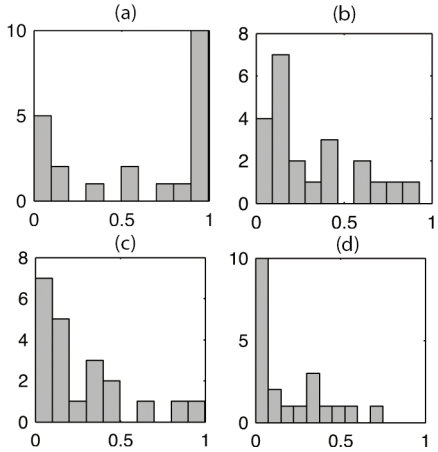

Figure 3. Distribution of R^2 for (a) TST, (b) SE, (c) SOL, and (d) NWAK for the 2 parameter weak model.

The conclusion from this analysis is that Fitbit may be useful for some subjects but not for others. This raises a question of how to determine which people would benefit from the sleep estimates derived from the Fitbit system. One approach would be to calibrate each person intending to use the Fitbit device using a Beddit or similar system. This approach would, however be costly and cumbersome. It would be significantly more convenient if it would be possible to estimate the validity of the Fitbit sleep estimates from the Fitbit data. To answer this question, we used a linear discriminant analysis of the Fitbit results. In particular, we examined whether a linear discriminant classifier can separate participants with low $R^2 < 0.75$ from those with high $R^2 \geq 0.75$. The linear discriminant classifier was computed using three features: Average total sleep time, average sleep efficiency and a correlation between adjacent TST values in the sequence of adjacent nights. The

preliminary results of the computation are promising in that the classifier could correctly classify more than 81% of the participants with respect to the validity of the Fitbit estimates. We note that these results are not conclusive because of the small number of participants in this study that prevented us to examine the generalizability of these results using cross-validation approaches.

IV. Discussion

The primary aim of this research was to compare two consumer sleep monitoring devices, Fitbit and Beddit. The motivation for the study was related to the fact that the accelerometer-based systems are very affordable, and are already used by many for activity measurement and do not require any additional modifications to the environment. In contrast, Beddit is more expensive and requires an installation in the participant's bed. When installed, the Beddit system is completely unobtrusive, whereas the Fitbit system requires the monitored individuals to signal their intention to sleep. Each of these devices is sensitive to certain type of movements and sleep disturbances that may guide their deployment.

The main outcome of this study is the finding that the accelerometer-based system (Fitbit) can be a useful device for assessment of relevant sleep parameters for a subset of the participants. In particular, for approximately 50% of the participants the data from the Fitbit system can be used to provide a good approximation of the total sleep time obtained from the Beddit system. For the remaining individuals, however, the Fitbit estimates were not very useful since they were almost independent of the Beddit estimates. The fact that individual differences determine the validity of the data implies that the deployment of such accelerometer-based systems will require individual-tailoring of the estimates. Finally, the results of the application of a linear classifier would suggest that it may be possible to use the Fitbit data to estimate the validity of the estimates. It is necessary to note, however, that the number of participants and the length of the present study are the limiting factors in interpretation of the results.

V. Conclusion

The key conclusion, based on the present study is that accelerometer-based devices may provide useful estimates of sleep parameters for a substantial subset of population. In order to utilize these data, it is necessary to model individuals and their behaviors. These individually-tailored models can be then used to harmonize data across devices and individuals.

These conclusions are subject to a number of limitations that include (a) the absence of a Beddit-PSG comparison, (b) a participant set comprising only young healthy adults, and (c) a small sample size.

In the future we need to enhance this model-based approach to incorporate estimates of user behavior, such as adherence. To address this issue, it might be possible to develop accelerometer-based devices that would be more context-aware and provide information useful in estimating the sleep starting time.

Acknowledgment

This research was supported by the National Technology Agency of Finland (TEKES) and by Tampere University of Technology Foundation. Technical support for gathering data was provided by the Heikki Peltola, Mika Saaranen, Arto Salminen, and Timo Aaltonen.

References

[1] Mahowald, M. W. (2007). Sleep Disorders and Sleep Deprivation: An Unmet Public Health Problem. *New England Journal of Medicine*, *356*(2), 199–200.

[2] McKnight-Eily LR, Liu Y, Perry GS, Presley-Cantrell LR, Strine TW, Lu H, Croft JB (2009) Perceived insufficient rest or sleep among adults—United States, 2008. MMWR Morbidity and Mortality Weekly Report, 58:1175–1179.

[3] A. Chesson et al. Practice parameters for the indications for polysomnography and related procedures. In Sleep, volume 20, pages 406–422, 1997.

[4] Agnew HW, Webb WB, Williams RL (1966) The first night effect: an EEG study of sleep. Psychophysiology 2:263–266.

[5] T.L. Hayes, T. Riley, M. Pavel and J.A. Kaye, A method for estimating rest-activity patterns using simple pyroelectric motion sensors, presented at 32nd Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Buenos Aires, Argentina, 2010.

[6] Paalasmaa, J., Waris, M., Toivonen, H., Leppakorpi, L., & Partinen, M. (2012). Unobtrusive online monitoring of sleep at home (pp. 3784–3788). Presented at the Engineering in Medicine and Biology Society (EMBC), 2012 Annual International Conference of the IEEE.

[7] Montgomery-Downs, H. E. H., Insana, S. P. S., & Bond, J. A. J. (2012). Movement toward a novel activity monitoring device. *Sleep & breathing = Schlaf & Atmung*, *16*(3), 913–917.

[8] Paalasmaa, J., Toivonen, H., & Partinen, M. (2014). Adaptive Heartbeat Modelling for Beat-to-beat Heart Rate Measurement in Ballistocardiograms. *Biomedical and Health Informatics, IEEE Journal of*, (99), 1.

[9] Karlen, W., Mattiussi, C., & Floreano, D. (2008). Improving actigraph sleep/wake classification with cardio-respiratory signals (pp. 5262–5265). Presented at the 2008 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, IEEE.

[10] Pollak, C. P., Tryon, W. W., Nagaraja, H., & Dzwonczyk, R. (2001). How accurately does wrist actigraphy identify the states of sleep and wakefulness? *Sleep*, *24*(8), 957–965.

[11] Wellness Warehouse Engine (W2E). [online] http://w2e.fi (accessed 27/03/14).

[12] Peltola, H., Salminen A. (2013). Towards a Reference Architecture for Server-Side Mashup Ecosystem. Presented at the 13th Symposium on Programming Languages and Software Tools, Sezged, Hungary, 2013.

[13] Saaranen M, Parak J, Honko H, Aaltonen T, Korhonen I (2014). W2E - Wellness Warehouse Engine for semantic interoperability of consumer health data. In: International Conference on Biomedical and Health Informatics, (BHI2014), Valencia, Spain, June 1-4, 2014.