Hearables: Automatic overnight sleep monitoring with standardised in-ear EEG sensor

Takashi Nakamura, Yousef D. Alqurashi, Mary J. Morrell and Danilo P. Mandic

Abstract—*Objective:* **Advances in sensor miniaturisation and computational power have served as enabling technologies for monitoring human physiological conditions in real-world scenarios. Sleep disruption may impact neural function, and can be a symptom of both physical and mental disorders. This study proposes wearable in-ear electroencephalography (ear-EEG) for overnight sleep monitoring as a 24/7 continuous and unobtrusive technology for sleep quality assessment in the community.** ***Methods:*** **Twenty-two healthy participants took part in overnight sleep monitoring with simultaneous ear-EEG and conventional full polysomnography (PSG) recordings. The ear-EEG data were analysed in the both structural complexity and spectral domains; the extracted features were used for automatic sleep stage prediction through supervised machine learning, whereby the PSG data were manually scored by a sleep clinician.** ***Results:*** **The agreement between automatic sleep stage prediction based on ear-EEG from a single in-ear sensor and the hypnogram based on the full PSG was 74.1 % in the accuracy over five sleep stage classification; this is supported by a Substantial Agreement in the kappa metric (0.61).** ***Conclusion:*** **The in-ear sensor is both feasible for monitoring overnight sleep outside the sleep laboratory and mitigates technical difficulties associated with PSG. It therefore represents a 24/7 continuously wearable alternative to conventional cumbersome and expensive sleep monitoring.** ***Significance:*** **The 'standardised' one-size-fits-all viscoelastic in-ear sensor is a next generation solution to monitor sleep – this technology promises to be a viable method for readily wearable sleep monitoring in the community, a key to affordable healthcare and future eHealth.**

I. Introduction

Sleep is an essential process for human well-being, and its quality reflects both a person's lifestyle as well as various medical conditions. In our modern 24/7 society, sleep quality has become a major issue which affects the state of body and mind, with implications on both general health and economy. Consequently, quality of sleep is considered one of the most important current topics in sleep medicine [1–3], and is typically investigated by recording a person's sleep patterns in a sleep clinic. However, current clinical sleep monitoring is expensive, cumbersome to administer, and prohibitive to performing recordings continuously over days and weeks; this also affects the subsequent diagnosis and treatment. For example, Beebe *et al.* designed and conducted sleep monitoring for three weeks in home using a wrist-worn actigraphy, and found that the daytime behavioural problems for adolescents may have been caused by inadequate sleep [4]. Trotti *et al.* monitored periodic leg movements (PLMs) in restless legs syndrome (RLS) patients using actigraphy for 2-3 weeks in home, and found significant variability of PLMs within a single RLS patient from night to night [5]. Vazir *et al.* conducted cardiorespiratory monitoring from patients with heart failure at patients' home, and reported shifting type of sleep-disordered breathing (SDB) over four nights [6]. These issues and previous findings have spurred the development of unobtrusive, wearable sensors capable of long-term monitoring of physiological variables related to sleep [7]. On the other hand, such miniaturisation of sensors inevitably affects the quality of recorded data; this calls for advanced signal processing and machine learning tools, throughout the process, from data conditioning to automatic sleep staging.

The polysomnography (PSG) is a standard clinical methodology to diagnose sleep disorders [8]. The so-acquired sleep profiles of individuals are rigorous and comprehensive, however, the expensive and cumbersome nature of PSG may even disturb patients' normal sleep, thus affecting diagnosis and treatment. During the analysis, the recorded PSG data are scored manually by a trained sleep clinician; given the scale of sleep disorders in our modern society, this imposes unrealistic demands on their time and incurs significant economic costs. Therefore, from a point of view of continuous widely affordable healthcare in the community and the future eHealth, conventional sleep monitoring based on the PSG is not realistic. To address this issue, multiple potential solutions have been proposed which fall under the two main categories:

1) Employ standard PSG but replace the clinician with an automatic scoring system [9],
2) Use wearable sensing, possibly with a reduced number of sensing modalities (e.g. only EEG or actigraphy), and perform the analysis either by the clinician or automatically based on machine learning [10].

To address the first issue of time-consuming sleep stage scoring process performed by clinicians, automatic sleep staging systems have been proposed based on full PSG recordings [19, 20] – these include the electroencephalogram (EEG), electrooculogram (EOG), and chin electromyogram (EMG) – or more recently based on single channel EEG recordings [21]. The corresponding automatic sleep stage classification approaches employ various machine learning algorithms and have been validated on both publicly available datasets [22] as well as on proprietary data recorded as part of various research projects [21]. Most studies have used sleep recordings

Takashi Nakamura and Danilo P. Mandic are with Department of Electrical and Electronic Engineering, Imperial College London, London, SW7 2AZ, United Kingdom. Yousef D. Alqurashi and Mary J. Morrell are with Sleep and Ventilation Unit, National Heart and Lung Institute, Imperial College London, and NIHR Respiratory Disease Biomedical Research Unit at the Royal Brompton and Harefield NHS Foundation Trust, and Imperial College London, London, SW3 6NP, United Kingdom , email: {takashi.nakamura14, y.alqurashi15, m.morrell, d.mandic}@imperial.ac.uk

TABLE I
COMPARISON OF EXISTING APPROACHES TO SLEEP RESEARCH USING IN-EAR/AROUND-EAR WEARABLE SENSING TECHNOLOGY

Author	# of participants	Earpiece	Recording	Details
[11]	4	In-ear viscoelastic	Daytime nap	Comparison of manually scored hypnograms based on ear-EEG and scalp-EEG
[12]				Comparison of automatic sleep scoring based on ear-EEG and scalp-EEG
[13]	23		Daytime naps	Comparison of sleep latency based on ear-EEG and scalp-EEG
[14]				Automatic detection of drowsiness based on ear-EEG
This study	**21**	**In-ear viscoelastic**	**Overnight**	**Automatic overnight sleep staging using ear-EEG**
[15]	8	In-ear viscoelastic	Overnight	Automatic sleep scoring based on ear-EEG (accuracy not given)
[16]	1	In-ear hardshell		Comparison of manually scored hypnograms based on ear-EEG and scalp-EEG
[17]	9			Automatic overnight sleep staging using ear-EEG
[18]	1	Around-ear		Demonstration of EEG patterns in different sleep stages (activities)

of healthy participants, and the state-of-the-art [23] indicates the possibility of correctly discriminating five sleep stages – wake (W), non-Rapid Eye Movements sleep (NREM1-3), and REM – based on a single EEG (channel).

Regarding the second issue of a wider deployment of wearable devices, commercial products already exist based on wrist activity (i.e. actigraphy) [24, 25]. Information from such wearable devices is proven to be sufficient to distinguish between wakefulness and sleep, and the agreement between the actigraphy based wake/sleep stages and those manually scored based on standard PSG recordings can be as high as 91 % [26]. More recently, data from commercial wrist-worn devices were found to allow for identifying multiple sleep parameters, such as 'sleep efficiency' and 'total sleep time' [27]. The smartphone accelerometers (i.e. off-body sensing) have also indicated the possibility of monitoring 'sleep duration', although such studies typically do not simultaneously record standard PSG as a 'ground-truth' [7].

Despite relative success, current proof-of-concept achievements based on 'wearables' are not yet capable of faithfully providing the much more complex information regarding clinically valid sleep analyses, that is, to discriminate between wakefulness, Non-REM sleep, and REM sleep.

With the development in sensor technology, one of the most convenient wearable solutions for physiological monitoring introduced in the research community is based on sensing from inside or around the ear, the so-called 'hearables' [28]. The original in-ear system in [29, 30] was shown to offer unobtrusive and robust brain monitoring (ear-EEG). The so-recorded data have been validated and compared to conventional on-scalp EEG (scalp-EEG) in different scenarios, including evoked potentials, brain-computer interface, and person authentication [31–33]. Multiple strategies have also been proposed for ear-EEG based sleep research, as summarised in Table I.

The original study by Looney *et al.* [11] recorded daytime naps with simultaneous ear-EEG and scalp-EEG systems, from four healthy participants. The corresponding manually scored hypnograms conclusively validated the feasibility of ear-EEG for sleep monitoring. The same data were also used for automatic sleep stage classification in [12], which further demonstrated the possibility of out-of-clinic sleep monitoring with ear-EEG. After this initial proof-of-concept stage, Alqurashi *et al.* [13] conducted comprehensive multiple daytime nap recordings to establish the degree of matching of the corresponding sleep latencies based on ear-EEG and scalp-EEG under two conditions: 1) after normal sleep and 2) after sleep restriction. The same nap data over twenty three participants were used by Nakamura *et al.* [14] to establish the potential of ear-EEG in automatic detection of drowsiness (i.e. to distinguish between wakefulness and light sleep). Nguyen *et al.* [15] conducted overnight sleep recordings over eight participants to evaluate their in-ear sensing system; their sensors were able to record the EEG, EOG, and EMG, key physiological variables for sleep monitoring. It is important to highlight that the sleep studies in [11–15], together with this study, were conducted using one-size-fits-all viscoelastic in-ear sensors, which are not optimised for a particular user but are convenient for wide deployment and promise an affordable out-of-clinic solution. Owing to their flexibility and favourable stress-strain properties (memory foam) [28], these viscoelastic earpieces can be squeezed and shaped up to fit comfortably any ear; such a 'generic' in-ear sensor is readily applicable to a large population, a pre-requisite for the future eHealth in the community. With custom-made hardshell earpieces, a technology derived from hearing aids earpieces, Looney *et al.* established the ear-EEG concept [29], while more recently Zibrandtsen *et al.* [16] monitored overnight sleep EEG activity from a single participant, and confirmed the similarities in temporal and spectral features between ear-EEG and conventional scalp-EEG. Recently, Mikkelsen *et al.* [17] validated automatic overnight sleep staging using hardshell binaural ear-EEG recordings over nine participants. The recent around-ear EEG device (cEEGrid) [18], which strictly speaking records scalp-EEG behind the ear, has also been utilised for overnight sleep recordings. It has been shown to be capable of monitoring specific sleep patterns, such as the K-complex, theta activity, and delta activity in NREM 3 stage sleep.

With our own one-size-fits-all in-ear sensing system [28], we here further establish and validate the feasibility of sleep monitoring in the community using ear-EEG in the following setups:

1) 'Standardised' viscoelastic in-ear sensors [34] for off-the-shelf sleep monitoring of a relatively large population of young healthy adults;
2) A 'real-world' out-of-clinic overnight sleep scenario, namely in participants' homes to reflect their normal sleep patterns (community based screening);
3) In conjunction with the conventional PSG for hypnogram generation, which then serves as the 'ground-truth' for further analyses.

For rigour and feasibility considerations, in this study, we employ the exact same shape of 'standardised' earpieces (size, materials) throughout the recordings on multiple participants. The recordings were undertaken in participants' homes; such familiar environments are a key to truly representative sleep monitoring, as this minimises the stress and inconvenience of the participants while maximising the likelihood of exhibiting usual sleep patterns. Our comprehensive setup involves simultaneous ear-EEG and PSG recordings; the ear-EEG data were used for automatic sleep stage classification, whilst the PSG data were scored manually by Author YDA. To benchmark the performance of ear-EEG against scalp-EEG, two channels of scalp-EEG were extracted from PSG and used for automatic sleep stage classification. Through a rigorous comparative examination of performance metrics of an automatic sleep staging method based on ear-EEG and the manually scored hypnogram (based on the standard PSG), we conclusively confirm the feasibility of automatic overnight sleep monitoring in out-of-clinic scenarios with readily deployable 'standardised' one-size-fits-all in-ear sensors, a prerequisite for affordable eHealth.

II. Methods

Figure 1 presents the flowchart for this study, whereby after simultaneous ear-EEG and PSG recording in the first step, two channels of scalp-EEG were extracted from the PSG data. The ear-EEG and scalp-EEG data were preprocessed through down-sampling, bandpass filtering, and removal of noisy epochs. Then, both structural complexity and frequency domain features for classification were extracted. The full PSG recordings were manually scored, and the so-obtained hypnogram was used as the 'ground-truth' of sleep stages for their automatic classification.

A. Data acquisition

The ear-EEG and PSG data were simultaneously recorded between October 2017 and June 2018 under the ethics approval, ICREC 17IC4150, Joint Research Office at Imperial College London. In total, twenty two healthy participants (aged 23.8 ± 4.8 years) were recorded, after an informed consent was obtained. Only two participants had worn our in-ear sensor prior to this study, whilst none of participants had ever participated in overnight PSG recordings.

Participants were visited in their own home at night (approximately two hours before their usual bedtime) to setup the ear-EEG and PSG sensors; after the sensor setup, the clinician started the recording and left the participant's home. Each participant went to bed as per usual and had their normal overnight sleep. The next morning, the clinician visited to detach the sensors; the participants were instructed to detach the sensors at any point during the night should they feel any discomfort.

The ear-EEG and PSG were recorded simultaneously from two data acquisition systems, and these two amplifiers were manually controlled to start and stop each recording. To ensure data alignment, the agreements between their time stamps were checked before every recording. For the ear-EEG, the g.tec g.USBamp amplifier with 24-bit resolution at a sampling frequency $fs = 1200\,\mathrm{Hz}$ was used for the recordings. The 'standardised' in-ear sensor was in the form of a one-size-fits-all viscoelastic earplug with two flexible electrodes, the details can be found in [28, 34]. The size of in-ear sensors was the same for all participants, approximately 25 mm in length and 12 mm in diameter. Before insertion, a participant's ear canal was cleaned with a cotton bud to remove ear wax; then conductive gel was applied to the electrode. The in-ear sensor was inserted into either participant's left or right ear, according to their preference, within a monaural setup. After the insertion, the sensor adapted snugly to the shape of the ear canal. Standard gold-cup electrodes were used as a reference (behind the ipsilateral helix) and ground (the

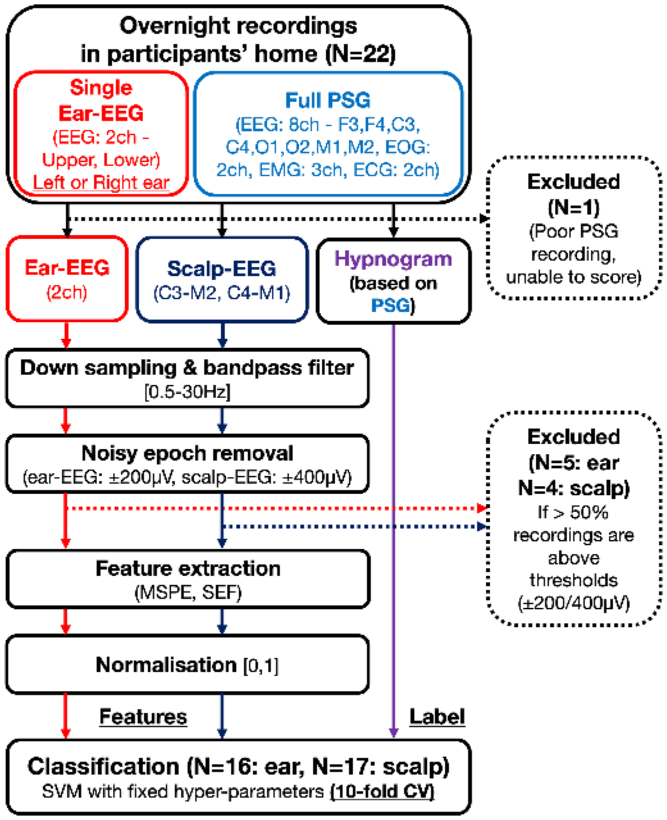
Overnight recordings in participants' home (N=22)
Single Ear-EEG (EEG: 2ch - Upper, Lower) Left or Right ear
Full PSG (EEG: 8ch - F3,F4,C3, C4,O1,O2,M1,M2, EOG: 2ch, EMG: 3ch, ECG: 2ch)
Excluded (N=1) (Poor PSG recording, unable to score)
Ear-EEG (2ch)
Scalp-EEG (C3-M2, C4-M1)
Hypnogram (based on PSG)
Down sampling & bandpass filter [0.5-30Hz]
Noisy epoch removal (ear-EEG: ±200μV, scalp-EEG: ±400μV)
Excluded (N=5: ear N=4: scalp) If > 50% recordings are above thresholds (±200/400μV)
Feature extraction (MSPE, SEF)
Normalisation [0,1]
Features
Label
Classification (N=16: ear, N=17: scalp)
SVM with fixed hyper-parameters (10-fold CV)

Fig. 1. Flowchart for the automatic sleep stage classification in this study.

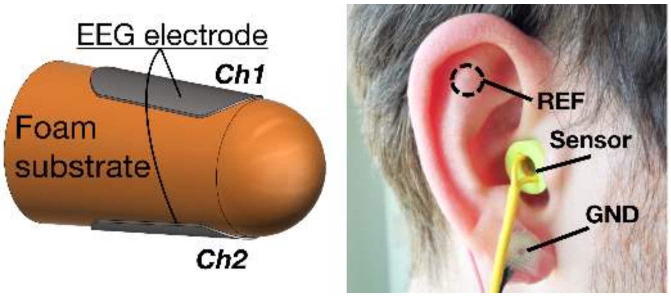

Fig. 2. The in-ear sensor used in our study. *Left*: Wearable in-ear sensor with two flexible electrodes. *Right*: Placement of the in-ear sensor.

ipsilateral earlobe). Figure 2 shows the in-ear sensors with two electrodes (left) and the placement of earpiece (right). Before the overnight recordings, the quality of ear-EEG responses was inspected during the participants' resting state with their eyes closed, and the electrode impedance was also checked.

The PSGs were recorded by the SOMNO Screen device, by SOMNO medics. The electro-physiological sensors were placed onto participants' scalp (including eight channels of EEG: F3, F4, C3, C4, O1, O2, M1, and M2, according to the international 10-20 system), face (including two channels of EOG), chin (three channels of EMG), chest, abdomen, and legs (including two channels of electrocardiography (ECG)). The ground and reference electrodes were attached on the forehead. Multiple other signals were monitored, such as participants' movement, body position, pulse rate and the pulse waveform (pulse oximeter), Naso/Oral flow (thermistor), and snoring sound. The data were recorded at a sampling frequency of 256 Hz, and transmitted to a laptop wirelessly.

B. Manual scoring

The recorded PSG data, including EEG, EOG, EMG, and respiration, were analysed by the Domino Plus system, by SOMNO medics. The obtained PSGs were bandpass filtered with the passband frequencies from $0.2 - 35$ Hz, and such processed PSG data were then manually scored by a clinician based on the American Academy of Sleep Medicine (AASM) criteria [35]. The so-labeled sleep stages were Wakefulness (W), NREM1 (N1), NREM2 (N2), NREM3 (N3), REM, and Movement. A PSG recording by one participant was not scored due to high frequency artefacts; therefore, overall, 21 out of 22 participants' overnight recordings were used for further analyses.

C. Pre-processing

The recorded signals were first aligned in accordance with the time stamps from the separate ear-EEG and PSG amplifiers. In total, approximately 165 hours (i.e. approximately eight hours per participant) of ear-EEG and PSG data were used for further analyses. For the classification, we used: 1) two ear-EEG channels (upper and lower channel of the earpiece), and 2) two scalp-EEG channels (C3-M2 and C4-M1) from the PSG recordings. The ear-EEG and scalp-EEG signals were downsampled to 120 Hz and 128 Hz, respectively; the downsampled frequencies of two systems were different due to the recording sampling rates (1200 Hz for ear-EEG, and 256 Hz for scalp-EEG). After downsampling, the EEG signals were bandpass filtered using a fourth-order Butterworth filter with the passband from $0.5-30$ Hz. Figure 3 illustrates different EEG sleep features, including alpha, theta, K-complex, and delta activities, from an on-scalp and the in-ear EEG channel.

Epochs scored as 'Movement' were removed. In this analysis, we considered the standard epochs (i.e. 30 s segment of recordings) while epochs which contained amplitudes of more than $\pm 200\,\mu$V for ear-EEG, and $\pm 400\,\mu$V for scalp-EEG, were deemed to be contaminated by noise. This is because, for the scalp-EEG, the amplitude of the K-complex, a signature response of NREM 2 sleep, is normally less than $\pm 400\,\mu$V [36]; we also assumed $\pm 200\,\mu$V was applicable for such amplitude thresholding in ear-EEG, since the amplitude of ear-EEG is smaller than that of scalp-EEG [29], as also seen in Figure 8. In order to remove noisy epochs from further analyses, the epochs for classification were selected using the criteria below:

TABLE II
PROPORTION OF SCORED SLEEP STAGES IN DIFFERENT SYSTEMS

	Ear-EEG (N=16)					
	Wake	N1	N2	N3	REM	All
# of epochs	2568	183	5332	2048	1479	11610
Ratio (%)	*22.1*	*1.6*	*45.9*	*17.6*	*12.7*	*100*
	Scalp-EEG (N=17)					
	Wake	N1	N2	N3	REM	All
# of epochs	2385	163	6510	1902	2080	13040
Ratio (%)	*18.3*	*1.2*	*49.9*	*14.6*	*16.0*	*100*

1) Find all epochs which contain amplitudes larger than the threshold (i.e. $\pm 200\,\mu$V for ear-EEG, and $\pm 400\,\mu$V for scalp-EEG) in each channel;
2) Remove an epoch if *at least* one channel of ear-EEG (or scalp-EEG) has amplitudes above the threshold;
3) Count the number of removed epochs (noisy epochs);
4) Do not consider a trial if the number of noisy epochs is more than 50 % of the entire recording from a participant.

We applied this epoch/participant rejection strategy to ear-EEG and scalp-EEG separately, therefore, the total numbers of epochs in the ear-EEG and scalp-EEG systems for further analyses were different, with details of their proportion given in Table II. The number of participants and epochs for further analysis of ear-EEG were respectively $N = 16$ and 11610 (out of 15120 – 76.8 % remaining epochs), while for scalp-EEG we had $N = 17$ and 13040 (out of 15970 – 81.7 % remaining epochs). This, in turn, means that the remaining participants for ear-EEG and scalp-EEG were not necessarily the same, since the removed participants were chosen based on amplitude thresholding, as explained earlier.

D. Feature extraction

After the pre-processing stage, feature extraction was performed using: 1) a complexity science feature, multi-scale entropy (MSE) [37], and 2) a frequency feature, the spectral edge frequency (SEF) [38]. These metrics were calculated for each epoch of both ear-EEG and scalp-EEG data. This combination of multi-scale permutation entropy (MSPE) and SEF was proven to be particularly successful in our previous automatic sleep staging work [23] which considered a publicly available overnight Sleep-EDF [expanded] dataset [39]. Based on two channels of scalp-EEGs from 61 participants, the achieved accuracy was 88.6 % with the corresponding kappa coefficient [40] of $\kappa = 0.84$ (Almost Perfect Agreement) in the 5-class sleep stage classification. For continuity, the same feature extraction methodology was applied in this study.

1) Structural complexity feature: The MSE method was shown to be able to quantify the degree of correlation in a time series, therefore it can be used to estimate structural complexity in data. The original MSE was designed to estimate

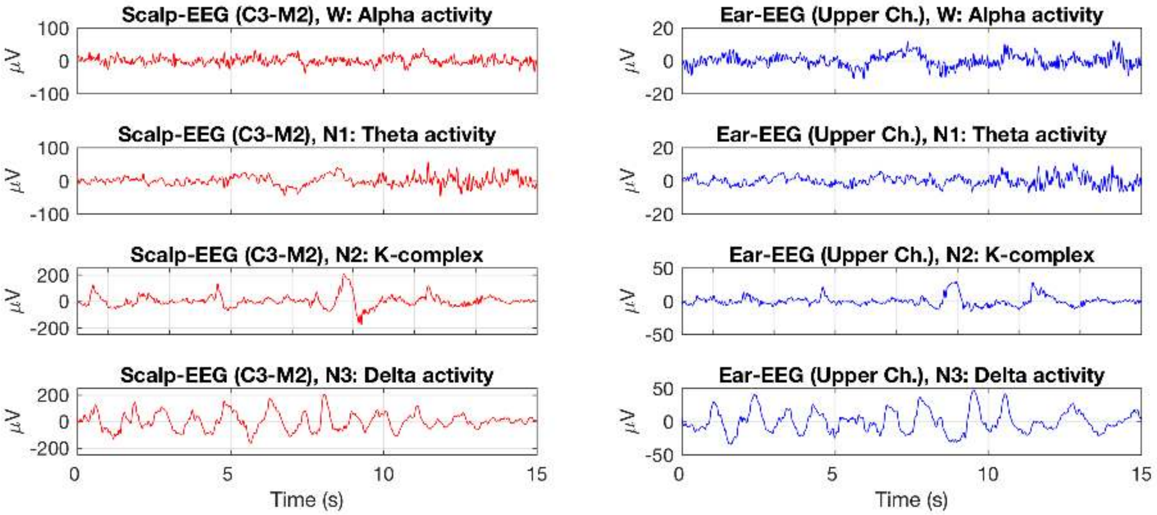

Fig. 3. EEG recordings from a single participant in different sleep stages (Red: the on-scalp C3-M2 channel, Blue: in-ear upper channel).

sample entropy of coarse grained time series, while its multivariated generalisation (MMSE) has been proposed to assess structural complexity of noisy multi-channel physiological data [41, 42]. As a structural complexity entropy based method for this study, we employed the permutation entropy (PE) [43]. The PE is a metric for detecting dynamical changes and for estimating the information contained in a time series based on comparing consecutive values of a time series. Compared to other entropy metrics, the PE requires less computational time and is robust to noise in the measurements; hence, the method is suited to time series with poor stationarity characteristics, such as physiological signals [44].

The details of MSPE can be found in our earlier related work [23]; the same parameters (i.e. the scale: $\tau = 20$, the embedding dimension: $d = 5$, and the time delay: $L = 1$) were used for this study. Figure 4 illustrates the MSPE analysis for ear-EEG and scalp-EEG channels of overnight data for one participant. The trends of two MSPEs for ear-EEG and scalp-EEG were similar, as evidenced by lower complexity in the N3 sleep stage which is due to the 'deterministic' dominant delta activity ($0.5-4$ Hz), and especially for slow wave activity ($0.5 - 2$ Hz).

2) Spectral feature: As a frequency domain feature, the r % spectral edge frequency (SEF), denoted by SEFr was used. The SEFr is defined as the frequency value which contains r % of the power in a given frequency range, that is

$$\sum_{f=f_{low}}^{f_{high}} \|magnitude(f)\|^2 \times \frac{r}{100} = \sum_{f=f_{low}}^{SEFr} \|magnitude(f)\|^2.$$

Owing to its robustness and ease of calculation, the SEF metric is now commonly used in physiological data analyses, especially in studies of EEG [38]. Relevant to this work, recently Imtiaz *et al.* [45] utilised the SEF methods in a sleep study, and proposed using the difference between SEF95 and SEF50, called SEFd, to detect the REM stage effectively, whereby $SEFd = SEF95 - SEF50$.

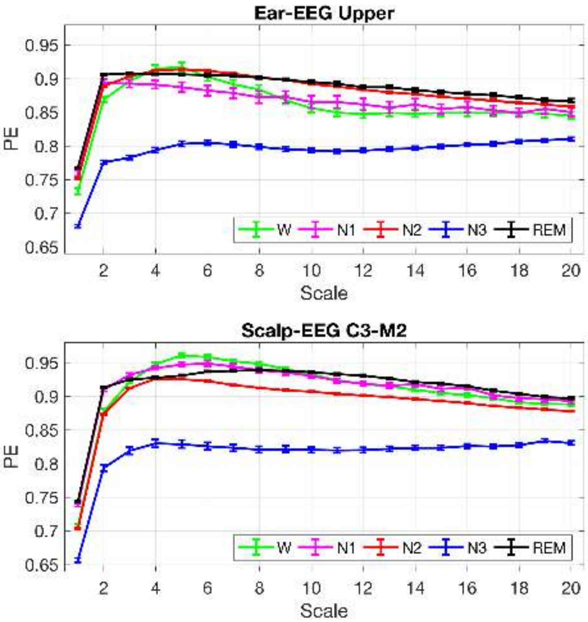

Fig. 4. Averaged multi-scale permutation entropy (MSPE) for the overnight sleep of one participant for an in-ear upper channel (ear-EEG, Top panel) and an on-scalp C3-M2 channel (scalp-EEG, Bottom panel) over different sleep stages. The error bars represent the standard error.

For our study, the power spectral density (PSD) of each epoch (30 s) was obtained using Welch's averaged periodogram method with the window length of 4 s and 50 % window overlap. Following the analysis in [23], we chose the same frequency ranges for SEF50, SEF95, and SEFd in the following bands; $\delta - \beta = 0.5 - 30$ Hz, $\delta - \alpha = 0.5 - 16$ Hz, $\theta = 2-8$ Hz, $\alpha = 8-15$ Hz, $\alpha_l = 8-11$ Hz, $\alpha_h = 11-15$ Hz, and $\beta = 16 - 30$ Hz.

E. Classification

The extracted features were normalised to between $[0, 1]$ participant-wise before performing classification. The multi-class support vector machine (SVM) with the radial basis function (RBF) kernel was employed as a classifier. The regularisation parameter was set to $C = 3$, and the hyper-parameters of the RBF kernel were set to $\gamma = 1$. The same hyper-parameters were used throughout the analysis.

F. Evaluation

The pre-processing and feature extraction analyses were conducted in Matlab 2016b, and the classification was implemented in Python 2.7.12, Anaconda 4.2.0 (x86_64) operated on an iMac with 2.8GHz Intel Core i5, and 16GB of RAM. In order to evaluate the classification performance of the proposed study, we utilised two metrics: 1) class-specific performance and 2) overall performance.

The *class-specific* performance metrics used were the sensitivity, $SE = TP/(TP + FN)$, and precision, $PR = TP/(TP + FP)$, where TP (true positive) represents the number of positive (target) epochs correctly predicted, FN (false negative) designates the number of positive epochs incorrectly predicted as negative class, and FP (false positive) is the number of negative epochs incorrectly predicted as positive class.

The *overall* performance was evaluated by the accuracy (AC) and Kappa coefficient (κ) metrics [40], defined as:

$$AC = \frac{\sum_{i=1}^{M} TP_i}{N_{epoch}}, \;\; \kappa = \frac{AC - \pi_e}{1 - \pi_e},$$

$$where \;\; \pi_e = \frac{\sum_{i=1}^{M} \{(TP_i + FP_i)(TP_i + FN_i)\}}{{N_{epoch}}^2}.$$

The parameter M denotes the number of classes (e.g. $M = 5$ class: Wake, N1, N2, N3, REM), and N_{epoch} is the total number of epochs.

G. Validation setup

A 10-fold cross-validation (CV) approach was utilised; EEG recordings from all participants were concatenated into one large matrix, which was then randomly split into the training data (90 %) and the test data (10 %). We repeated the validation 10 times with changing the selection of training and test data.

III. Results

The feature matrices based on two ear-EEG channels and two scalp-EEG channels were classified by a multi-class SVM with fixed hyper-parameters as explained in Section II-E.

Figure 5 shows the classification results for $M = 5$ classes (W, N1, N2, N3, and REM) using the ear-EEG and scalp-EEG. In the ear-EEG setup, the overall accuracy was 74.1 % with the corresponding kappa value of $\kappa = 0.61$, which indicates Substantial Agreement, whereas the accuracy and κ of scalp-EEG were respectively 85.9 % and 0.79 (Substantial Agreement). The sensitivities to each sleep stage of ear-EEG were approximately 10 % lower than those of scalp-EEG, except for the REM condition. The sensitivities to REM of ear-EEG and scalp-EEG were 45.8 % and 81.8 %, respectively. Notice that more epochs labeled REM were misclassified as N2 (679 epochs) than correctly classified as REM (677 epochs) in the ear-EEG setup.

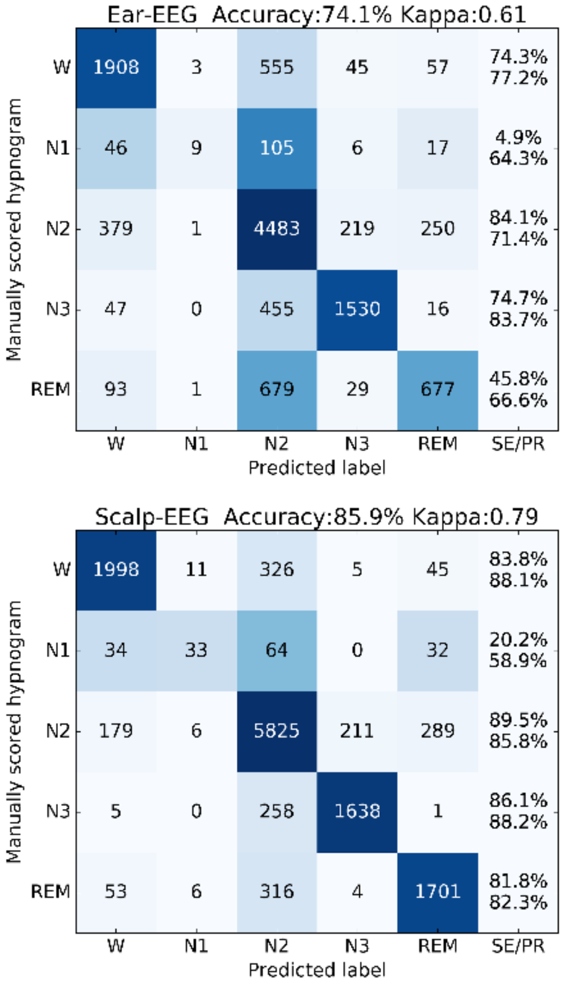

Fig. 5. Confusion matrix for the classification results. Upper: Ear-EEG, Lower: Scalp-EEG. The symbols SE/PR on the bottom right denote respectively sensitivity (above) and precision (below).

Figure 6 depicts the participant-wise classification accuracy (blue bar plot) and the corresponding κ (red dot plot) for both the ear-EEG and scalp-EEG scenarios. Although not without variations, overall, data from all participants were amenable to being automatically classified using the proposed methods.

Figure 7 illustrates the overnight hypnograms of two participants for both the ear-EEG and scalp-EEG system; the graphs show the manually scored hypnograms based on the full PSG recordings (blue) and the automatically predicted label based on the proposed algorithm (red). The black crosses denote the epochs removed from the analyses due to amplitude thresholding (see Section II-C). In Figure 7A, notice the REM condition between time stamps 03:30 and 04:30; regarding the predicted label based on ear-EEG, a large portion of epochs was misclassified as N2 sleep, whereas the majority of epochs

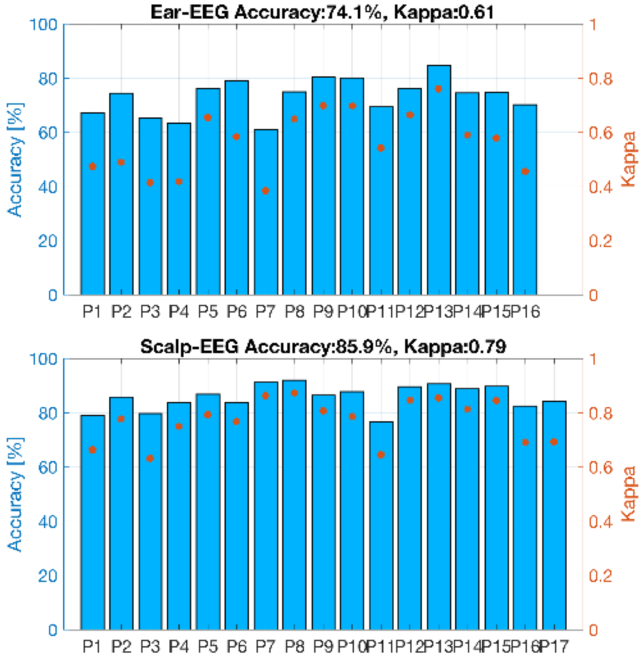

Fig. 6. Classification accuracy (blue bars) and kappa values (red dots) for individual participants, P1-17. Upper: Ear-EEG based results, Lower: Scalp-EEG based results.

was correctly classified as REM in the prediction based on scalp-EEG. For further illustration, in a hypnogram of another participant in Figure 7B, multiple sleep cycles, e.g. three REM conditions, can be observed.

Table III displays the classification results over different number of classes ($M = 2, 3, 5$). The class-wise sensitivity and precision (*in italic*), and overall accuracy and κ are also provided. The notation $M = 2$ means the 2-stage classification (Wake vs Sleep), whereas $M = 3$ denotes 3-stage classification (Wake vs NREM sleep vs REM sleep). The accuracy of ear-EEG ranged from 74.1 % to 89.9 % with the corresponding, κ, from 0.60 to 0.68. The accuracies and kappa coefficients of scalp-EEG were higher than those of ear-EEG.

TABLE III
SENSITIVITY AND *PRECISION* (*in italic*) FOR SLEEP CLASSIFICATION IN DIFFERENT CLASS SCENARIOS, M=2, 3, 5.

Ear-EEG	$M = 2$	$M = 3$	$M = 5$
Wake	66.7 *84.1*	70.4 *80.6*	74.3 *77.2*
N1	96.4 *91.1*	92.0 *82.3*	4.9 *64.3*
N2			84.1 *71.4*
N3			74.7 *83.7*
REM		42.1 *68.1*	45.8 *66.6*
Accuracy	89.9	80.9	74.1
κ	0.68	0.60	0.61

Scalp-EEG	$M = 2$	$M = 3$	$M = 5$
Wake	81.6 *91.3*	82.9 *89.3*	83.8 *88.1*
N1	98.3 *96.0*	94.3 *91.7*	20.2 *58.9*
N2			89.5 *85.8*
N3			86.1 *88.2*
REM		80.3 *83.1*	81.8 *82.3*
Accuracy	95.2	90.0	85.9
κ	0.83	0.80	0.79

IV. Discussion

This study has proposed an overnight sleep monitoring system using a ‘standardised’ in-ear sensor, and has validated the feasibility of automatic sleep staging based on ear-EEG. Compared to the gold standard – manually scored hypnogram based on a standard PSG recording – the obtained classification accuracy using ear-EEG features was 74.1 % with the corresponding κ value of 0.61, which indicates Substantial Agreement.

Compared to the classification performance based on ear-EEG, the results based on scalp-EEG were better, especially regarding the sensitivity to REM stage. As seen in Figure 5, the majority of manually labeled REM epochs were misclassified as N2 in the ear-EEG setup. Figure 8 illustrates the averaged power spectral density for the ear-EEG (left) and the scalp-EEG (right) of two participants. For this analysis, the recorded signals were manually selected in order to compare N2 vs N3 vs REM; for the top panel (Participant 1), a single consecutive 90 minutes of sleep data were selected, whereas two consecutive 60 minutes of sleep data were selected (i.e. 120 minutes in total) for the bottom panel (Participant 2). The PSDs were obtained using Welch’s averaged periodogram method, the window length was 4 s with 50 % of window overlap. The trends in ear-EEG and scalp-EEG analyses were similar and included: 1) high-alpha ($12 - 15$ Hz) activities in N2 sleep, 2) prominent delta activities ($0.5 - 4$ Hz), and especially slow wave activity ($0.5 - 2$ Hz) in N3 sleep, and 3) relatively lower EEG amplitude in REM. However, the slow wave activities of N2 and REM sleep in the ear-EEG were similar, as evidenced by an overlap in their spectrum, whereas clear visual separation was present in the scalp-EEG setup. This overlap might have caused the lower discrimination performance for N2 and REM in the ear-EEG scenario. We would like to highlight that the scalp-EEG montage (C3-M2 and C4-M1) is the gold standard for sleep medicine, and has been studied and validated over decades. Also, the algorithm applied in this study was originally tested and developed on a publicly available dataset of scalp-EEG [23]. In [23], the classification performance based on two scalp-EEG channels over 61 participants from a publicly available dataset was 88.6 % in accuracy with the corresponding κ of 0.84 in a 5-class sleep stage classification, which was similar to the results in this study – the accuracy and κ were respectively 85.9 % and 0.79 in a 5-class sleep staging using two channels of scalp-EEG over 17 participants. Future work will consider introducing a fine-tuned classifier, specifically designed for ear-EEG. Collecting a very large cohort of in-ear sleep EEG data will allow us to examine more practical validation setups, such as leave-one-participant-out CV in addition to K-fold CV.

Our study leaves room for improvement; for example, some noisy epochs were removed by amplitude thresholding, and some participants were removed from the analyses as mentioned in Section II-C. According to visual inspection of the shape of the recorded signal, the noise was categorised into: 1) abrupt electrode noise and 2) physiological noise from respiration. The first type of noise might have been caused

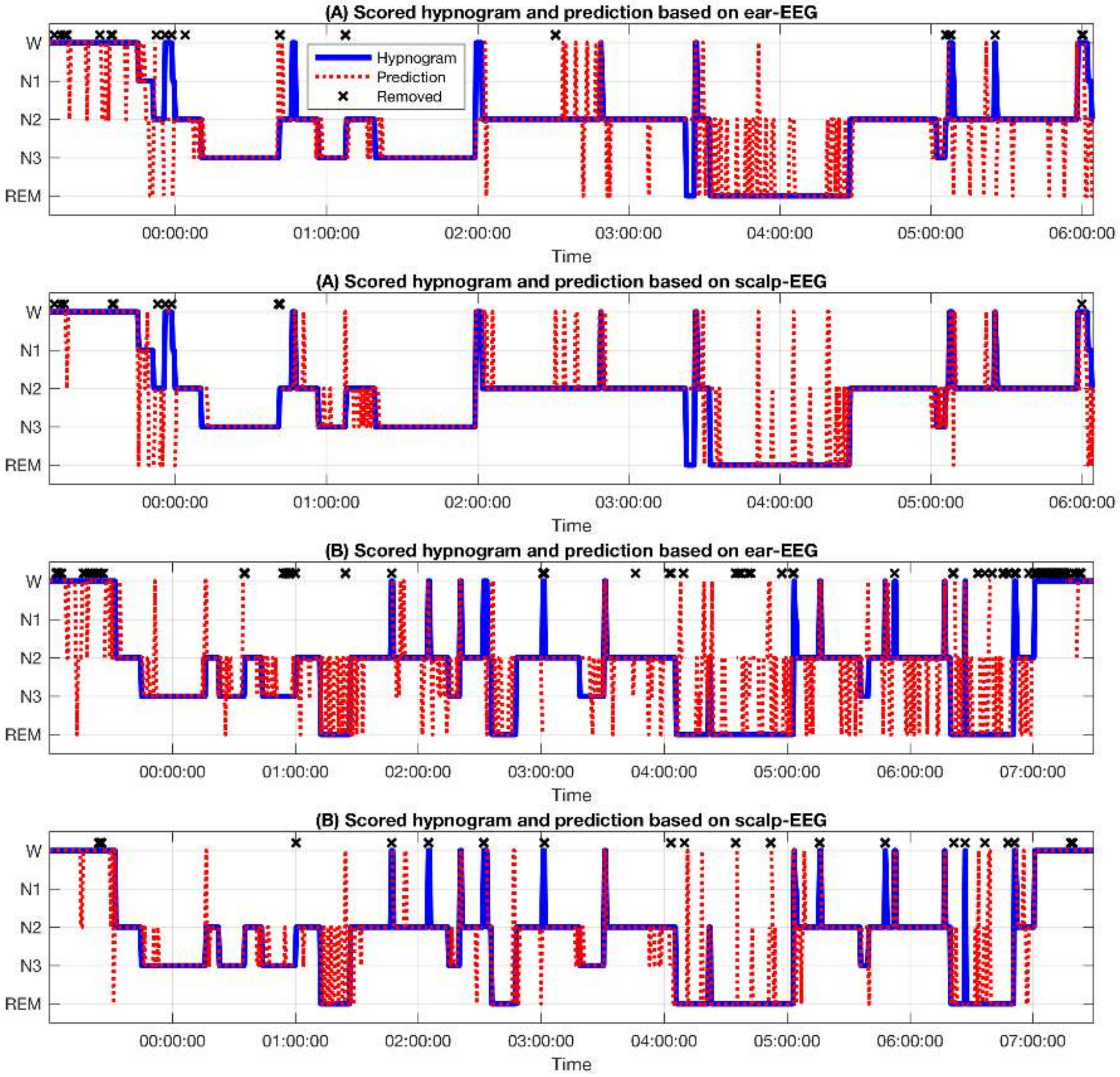

Fig. 7. Manually scored hypnograms of overnight sleep based on the full PSG (blue) and the automatically predicted label by the proposed algorithm with the ear-EEG and scalp-EEG features (red) from two participants (A and B). The black crosses above the W stage indicate the epochs removed from the analyses due to the amplitude thresholding.

by participants' movement. Note that our in-ear sensor has been extensively validated in a research lab when the participants were in the resting state, but only a few studies were conducted under participants' movements (e.g. jaw movement [28]). These issues can be resolved with a future advanced sensor design. The second type of noise was due to the placement of the in-ear sensor on the participant's head. In certain recordings, the recorded ear-EEG signal was overlaid by a slow oscillation of large amplitude, which represents an artefact from respiration. Our study utilised a monaural setup in order to minimise both the time for technical setup and participants' inconvenience, however, this might have interfered with the quality of recordings. As shown in our recent work [46], in addition to one more degrees of freedom in ear-EEG recording, a binaural setup would also allow for the monitoring of other physiological parameters such as ECG and respiration [28].

V. Conclusion

We have proposed and validated an automatic overnight sleep monitoring system with readily deployable 'standardised' one-size-fits-all viscoelastic in-ear sensors. Full standard PSG and in-ear EEG have been simultaneously recorded for twenty-two healthy participants, who participated in overnight sleep recordings at their own home in order to both minimise participants' inconvenience and provide a 'real-world' out-of-clinic scenario. The scalp-EEG and ear-EEG have been shown to exhibit a high degree of similarity in both the structural complexity and spectral domains. The agreement between manually scored hypnograms based on full PSG and automatic 5-class sleep stage prediction based on ear-EEG was 74.1 % in the accuracy with the kappa coefficient of 0.61 (Substantial Agreement), whereas the obtained accuracy and κ based on scalp-EEG were 85.9 % and 0.79 (Substantial Agreement), respectively. This study has demonstrated that a

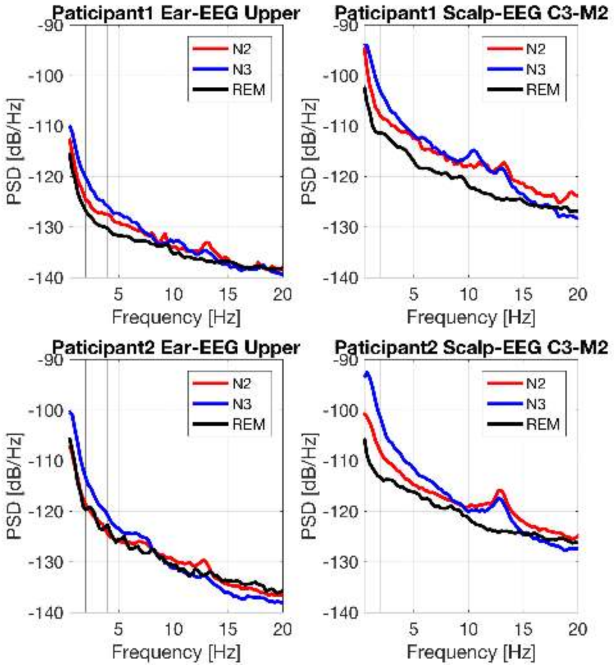

Fig. 8. Averaged periodograms of ear-EEG and scalp-EEG during selected consecutive epochs of two different participants. The vertical lines denote the 2 and 4 Hz frequencies in order to designate the delta activity $(0.5 - 4\,\mathrm{Hz})$ and slow wave activity $(0.5 - 2\,\mathrm{Hz})$.

single in-ear sensor is capable of monitoring overnight sleep in an unobtrusive and inexpensive way, and that one-size-fits-all viscoelastic sensor promises to become a viable eHealth community-based alternative to conventional sleep monitoring in a clinic.

References

[1] F. P. Cappuccio, L. D'Elia, P. Strazzullo, and M. A. Miller, "Quantity and quality of sleep and incidence of type 2 diabetes," *Diabetes Care*, vol. 33, no. 2, pp. 414–420, 2010.

[2] I. Rosenzweig, M. Glasser, D. Polsek, G. D. Leschziner, S. C. Williams, and M. J. Morrell, "Sleep apnoea and the brain: A complex relationship," *The Lancet Respiratory Medicine*, vol. 3, no. 5, pp. 404–414, 2015.

[3] N. F. Watson, M. S. Badr, G. Belenky, D. L. Bliwise, O. M. Buxton, D. Buysse, D. F. Dinges, J. Gangwisch, M. A. Grandner, C. Kushida, R. K. Malhotra, J. L. Martin, S. R. Patel, S. F. Quan, and E. Tasali, "Joint consensus statement of the American Academy of Sleep Medicine and Sleep Research Society on the recommended amount of sleep for a healthy adult: Methodology and discussion," *Sleep*, vol. 38, no. 8, pp. 1161–1183, 2015.

[4] D. W. Beebe, G. Fallone, N. Godiwala, M. Flanigan, D. Martin, L. Schaffner, and R. Amin, "Feasibility and behavioral effects of an at-home multi-night sleep restriction protocol for adolescents," *Journal of Child Psychology and Psychiatry and Allied Disciplines*, vol. 49, no. 9, pp. 915–923, 2008.

[5] L. M. Trotti, D. L. Bliwise, S. A. Greer, A. P. Sigurdsson, G. B. Gudmundsdóttir, T. Wessel, L. M. Organisak, T. Sigthorsson, K. Kristjansson, T. Sigmundsson, and D. B. Rye, "Correlates of PLMs variability over multiple nights and impact upon RLS diagnosis," *Sleep Medicine*, vol. 10, no. 6, pp. 668–671, 2008.

[6] A. Vazir, P. C. Hastings, I. Papaioannou, P. A. Poole-Wilson, M. R. Cowie, M. J. Morrell, and A. K. Simonds, "Variation in severity and type of sleep-disordered breathing throughout 4 nights in patients with heart failure," *Respiratory Medicine*, vol. 102, no. 6, pp. 831–839, 2008.

[7] Z. Chen, M. Lin, F. Chen, N. D. Lane, G. Cardone, R. Wang, T. Li, Y. Chen, T. Choudhury, and A. T. Campbell, "Unobtrusive sleep monitoring using smartphones," in *Proceedings of the 7th International Conference on Pervasive Computing Technologies for Healthcare and Workshops*, pp. 145–152, 2013.

[8] M. H. Silber, S. Ancoli-Israel, M. H. Bonnet, S. Chokroverty, M. M. Grigg-Damberger, M. Hirshkowitz, S. Kapen, S. A. Keenan, M. H. Kryger, T. Penzel, M. R. Pressman, and C. Iber, "The visual scoring of sleep in adults," *Journal of Clinical Sleep Medicine*, vol. 3, no. 2, pp. 121–131, 2007.

[9] R. Agarwal and J. Gotman, "Computer-assisted sleep staging," *IEEE Transactions on Biomedical Engineering*, vol. 48, no. 12, pp. 1412–1423, 2001.

[10] R. Dick, T. Penzel, I. Fietze, M. Partinen, H. Hein, and J. Schulz, "AASM standards of practice compliant validation of actigraphic sleep analysis from SOMNOwatch versus polysomnographic sleep diagnostics shows high conformity also among subjects with sleep disordered breathing," *Physiological Measurement*, vol. 31, no. 12, pp. 1623–1633, 2010.

[11] D. Looney, V. Goverdovsky, I. Rosenzweig, M. J. Morrell, and D. P. Mandic, "A wearable in-ear encephalography sensor for monitoring sleep: Preliminary observations from nap studies," *Annals of the American Thoracic Society*, vol. 13, no. 12, pp. 2229–2233, 2016.

[12] T. Nakamura, V. Goverdovsky, M. J. Morrell, and D. P. Mandic, "Automatic sleep monitoring using ear-EEG," *IEEE Journal of Translational Engineering in Health and Medicine*, vol. 5, no. 1, pp. 1–8, Article no. 2800108, 2017.

[13] Y. D. Alqurashi, T. Nakamura, V. Goverdovsky, J. Moss, M. I. Polkey, D. P. Mandic, and M. J. Morrell, "A novel in-ear sensor to determine sleep latency during the Multiple Sleep Latency Test (MSLT) in healthy adults with and without sleep restriction," *Nature and Science of Sleep*, vol. 10, pp. 385–396, 2018.

[14] T. Nakamura, Y. D. Alqurashi, M. J. Morrell, and D. P. Mandic, "Automatic detection of drowsiness using in-ear EEG," in *Proceedings of IEEE International Joint Conference on Neural Networks (IJCNN)*, pp. 5569–5574, 2018.

[15] A. Nguyen, R. Alqurashi, Z. Raghebi, F. Banaei-kashani, A. C. Halbower, and T. Vu, "A lightweight and inexpensive in-ear sensing system for automatic whole-night sleep stage monitoring," in *Proceedings of the 14th ACM Conference on Embedded Network Sensor Systems*, pp. 230–244, 2016.

[16] I. Zibrandtsen, P. Kidmose, M. Otto, J. Ibsen, and T. W. Kjaer, "Case comparison of sleep features from ear-EEG and scalp-EEG," *Sleep Science*, vol. 9, no. 2, pp. 69–72, 2016.

[17] K. B. Mikkelsen, D. B. Villadsen, M. Otto, and P. Kidmose, "Automatic sleep staging using ear-EEG," *BioMedical Engineering OnLine*, vol. 16, no. 1, pp. 1–15, Article no. 111, 2017.

[18] M. G. Bleichner and S. Debener, "Concealed, unobtrusive ear-centered EEG acquisition: cEEGrids for transparent EEG," *Frontiers in Human Neuroscience*, vol. 11, no. 4, pp. 1–14, Article no. 163, 2017.

[19] A. Patanaik, J. L. Ong, J. J. Gooley, S. Ancoli-Israel, and M. W. Chee, "An end-to-end framework for real-time automatic sleep stage classification," *Sleep*, vol. 41, no. 5, pp. 1–11, 2018.

[20] H. Sun, J. Jia, B. Goparaju, G.-B. Huang, O. Sourina, M. T. Bianchi, and M. B. Westover, "Large-scale automated sleep staging," *Sleep*, vol. 40, no. 10, 2017.

[21] B. Koley and D. Dey, "An ensemble system for automatic sleep stage classification using single channel EEG signal," *Computers in Biology and Medicine*, vol. 42, no. 12, pp. 1186–1195, 2012.

[22] G. Zhu, Y. Li, and P. P. Wen, "Analysis and classification of sleep stages based on difference visibility graphs from a single-channel EEG signal," *IEEE Journal of Biomedical and Health Informatics*, vol. 18, no. 6, pp. 1813–1821, 2014.

[23] T. Nakamura, T. Adjei, Y. Alqurashi, D. Looney, M. J. Morrell, and D. P. Mandic, "Complexity science for sleep stage classification from EEG," in *Proceedings of IEEE International Joint Conference on Neural Networks (IJCNN)*, pp. 4387–4394, 2017.

[24] T. Morgenthaler, C. Alessi, L. Friedman, J. Owens, V. Kapur, B. Boehlecke, T. Brown, A. Chesson, J. Coleman, T. Lee-Chiong, J. Pancer, and T. J. Swick, "Practice parameters for the use of actigraphy in the assessment of sleep and sleep disorders: An update for 2007," *Sleep*, vol. 30, no. 4, pp. 519–529, 2007.

[25] M. Willetts, S. Hollowell, L. Aslett, C. Holmes, and A. Doherty, "Statistical machine learning of sleep and physical activity phenotypes from sensor data in 96,220 UK Biobank participants," *Scientific Reports*, vol. 8, no. 1, pp. 1–10, Article no. 7961, 2018.

[26] L. D. Souza, A. A. Benedito-Silva, M. L. N. Pires, D. Poyares, S. Tufik, and H. M. Calil, "Further validation of actigraphy for sleep studies," *Sleep*, vol. 26, no. 1, pp. 81–85, 2003.

[27] M. T. Smith, C. S. McCrae, J. Cheung, J. L. Martin, C. G. Harrod, J. L. Heald, and K. A. Carden, "Use of actigraphy for the evaluation of sleep disorders and circadian rhythm sleep-wake disorders: An American Academy of Sleep Medicine systematic review, meta-analysis, and GRADE assessment," *Journal of Clinical Sleep Medicine*, vol. 14, no. 7, pp. 1231–1237, 2018.

[28] V. Goverdovsky, W. von Rosenberg, T. Nakamura, D. Looney, D. J. Sharp, C. Papavassiliou, M. J. Morrell, and D. P. Mandic, "Hearables: Multimodal physiological in-ear sensing," *Scientific Reports*, vol. 7, no. 1, pp. 1–10, Article no. 6948, 2017.

[29] D. Looney, P. Kidmose, C. Park, M. Ungstrup, M. Rank, K. Rosenkranz, and D. P. Mandic, "The in-the-ear recording concept: User-centered and wearable brain monitoring," *IEEE Pulse*, vol. 3, no. 6, pp. 32–42, 2012.

[30] P. Kidmose, D. P. Mandic, M. Ungstrup, D. Looney, C. Park, and M. L. Rank, "Hearing aid adapted for detecting brain waves and a method for adapting such a hearing aid," 2015. US Patent 9,025,800.

[31] P. Kidmose, D. Looney, M. Ungstrup, M. L. Rank, and D. P. Mandic, "A study of evoked potentials from ear-EEG," *IEEE Transactions on Biomedical Engineering*, vol. 60, no. 10, pp. 2824–2830, 2013.

[32] Y.-T. Wang, M. Nakanishi, S. L. Kappel, P. Kidmose, D. P. Mandic, Y. Wang, C.-K. Cheng, and T.-P. Jung, "Developing an online steady-state visual evoked potential-based brain-computer interface system using earEEG," in *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 2271–2274, 2015.

[33] T. Nakamura, V. Goverdovsky, and D. P. Mandic, "In-ear EEG biometrics for feasible and readily collectable real-world person authentication," *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 3, pp. 648–661, 2018.

[34] V. Goverdovsky, D. Looney, P. Kidmose, and D. P. Mandic, "In-ear EEG from viscoelastic generic earpieces: Robust and unobtrusive 24/7 monitoring," *IEEE Sensors Journal*, vol. 16, no. 1, pp. 271–277, 2016.

[35] C. Iber, *The AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology and Technical Specifications*. American Academy of Sleep Medicine, 2007.

[36] G. Bremer, J. R. Smith, and I. Karacan, "Automatic detection of the K-complex in sleep electroencephalograms," *IEEE Transactions on Biomedical Engineering*, vol. BME-17, no. 4, pp. 314–323, 1970.

[37] M. Costa, A. L. Goldberger, and C.-K. Peng, "Multiscale entropy analysis of complex physiologic time series," *Physical Review Letters*, vol. 89, no. 6, pp. 1–4, Article no. 68102, 2002.

[38] J. W. Sleigh and J. Donovan, "Comparison of bispectral index, 95% spectral edge frequency and approximate entropy of the EEG, with changes in heart rate variability during induction of general anaesthesia," *British Journal of Anaesthesia*, vol. 82, no. 5, pp. 666–671, 1999.

[39] PhysioNet, "The Sleep-EDF Database." https://physionet.org/physiobank/database/sleep-edf/. [Online. Last Accessed: 11-Apr-2019].

[40] J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data," *Biometrics*, vol. 33, no. 1, pp. 159–174, 1977.

[41] M. U. Ahmed and D. P. Mandic, "Multivariate multiscale entropy: A tool for complexity analysis of multichannel data," *Physical Review E*, vol. 84, no. 6, pp. 1–10, Article no. 61918, 2011.

[42] T. Chanwimalueang and D. P. Mandic, "Cosine similarity entropy: Self-correlation-based complexity analysis of dynamical systems," *Entropy*, vol. 19, no. 652, 2017.

[43] C. Bandt and B. Pompe, "Permutation entropy: A natural complexity measure for time series," *Physical Review Letters*, vol. 88, no. 17, pp. 1–4, Article no. 174102, 2002.

[44] T. Adjei, W. von Rosenberg, V. Goverdovsky, K. Powezka, U. Jaffer, and D. P. Mandic, "Pain prediction from ECG in vascular surgery," *IEEE Journal of Translational Engineering in Health and Medicine*, vol. 5, no. 1, pp. 1–10, Article no. 2734647, 2017.

[45] S. A. Imtiaz and E. Rodriguez-Villegas, "A low computational cost algorithm for REM sleep detection using single channel EEG," *Annals of Biomedical Engineering*, vol. 42, no. 11, pp. 2344–2359, 2014.

[46] W. von Rosenberg, T. Chanwimalueang, V. Goverdovsky, N. S. Peters, C. Papavassiliou, and D. P. Mandic, "Hearables: Feasibility of recording cardiac rhythms from head and in-ear locations," *Royal Society Open Science*, vol. 4, pp. 1–13, Article no. 171214, 2017.