Sleep and Wake Classification With Actigraphy and Respiratory Effort Using Dynamic Warping

Xi Long, *Member, IEEE*, Pedro Fonseca, Jérôme Foussier, *Member, IEEE*, Reinder Haakma, and Ronald M. Aarts, *Fellow, IEEE*

***Abstract*—This paper proposes the use of dynamic warping (DW) methods for improving automatic sleep and wake classification using actigraphy and respiratory effort. DW is an algorithm that finds an optimal nonlinear alignment between two series allowing scaling and shifting. It is widely used to quantify (dis)similarity between two series. To compare the respiratory effort between sleep and wake states by means of (dis)similarity, we constructed two novel features based on DW. For a given epoch of a respiratory effort recording, the features search for the optimally aligned epoch within the same recording in time and frequency domain. This is expected to yield a high (or low) similarity score when this epoch is sleep (or wake). Since the comparison occurs throughout the entire-night recording of a subject, it may reduce the effects of within- and between-subject variations of the respiratory effort, and thus help discriminate between sleep and wake states. The DW-based features were evaluated using a linear discriminant classifier on a dataset of 15 healthy subjects. Results show that the DW-based features can provide a Cohen's Kappa coefficient of agreement $\kappa = 0.59$ which is significantly higher than the existing respiratory-based features and is comparable to actigraphy. After combining the actigraphy and the DW-based features, the classifier achieved a κ of 0.66 and an overall accuracy of 95.7%, outperforming an earlier actigraphy- and respiratory-based feature set ($\kappa = 0.62$). The results are also comparable with those obtained using an actigraphy- and cardiorespiratory-based feature set but have the important advantage that they do not require an ECG signal to be recorded.**

***Index Terms*—Dynamic warping (DW), feature extraction, respiratory effort, sleep and wake classification, unobtrusive monitoring.**

I. Introduction

SLEEP plays an important role in human's emotional well-being and physical health. Many people live with sleep-related problems that have a primary implication of one's health condition [1]–[3]. Objective assessment of sleep is often based on the monitoring of sleep and wake states throughout the entire night during bedtime [4], [5]. According to the guidelines of the American Academy of Sleep Medicine (AASM) [6], the sleep stages consist of rapid-eye-movement (REM) and non-REM (NREM, including N1, N2, and N3).

Over-night polysomnography (PSG) recordings with manually annotated hypnograms are considered a "gold standard" for objectively analyzing the sleep architecture and occurrence of specific sleep-related problems [1]. A PSG typically comprises physiological data such as the electroencephalogram (EEG), electrocardiogram (ECG), electromyogram, electrooculogram, oxygen saturation, and respiratory effort. When used for sleep staging, recorded signals are typically split into nonoverlapping epochs of 30 s each in accordance with the Rechtschaffen and Kales rules [1], and also the more recent guidelines of the AASM [6].

Although PSG is the gold standard for sleep assessment, it has several drawbacks such as the high costs of laboratory facilities, disruption of "normal" sleep, and impossibility to perform long-term monitoring. This has motivated the investigation of sensors/methods that allow for a reliable acquisition of physiological modalities in an unobtrusive or at least more comfortable and convenient way. In particular, actigraphy and cardiorespiratory signals have been often considered in the context of automatic sleep monitoring [2], [4].

Actigraphy is a less-unobtrusive way of measuring body movement based on an accelerometer, which is typically worn on wrist. It has been extensively studied [7]–[12] and is considered a standard method for sleep assessment when PSG is not available [10]. However, researchers argue that actigraphy accounts for error when compared with PSG [11]; and it cannot cope with the misclassification of "quiet-wake" with low body activity, resulting in low accuracy in detecting wake state [9], [12]. Since actigraphy only measures body movement, it reflects limited physiological information. It has been shown that cardiorespiratory signals contain relevant physiological information which can help improve actigraphy-based sleep and wake classification [4], [13], [14]. More importantly, these signal modalities can be acquired in an unobtrusive circumstance in different ways (e.g., ballistocardiogram [15], Doppler radar [16], near-infrared camera [17], under-pillow sensor [18], and bed sensor [19]). For example, acquiring cardiorespiratory information using a static-charge-sensitive bed [20], [21] has been investigated, and in recent years it has become more popular for unobtrusive monitoring of sleep [19], [22]. However, difficulty has been found in separating wake and REM sleep [23] when only using cardiorespiratory signals. So it is necessarily

Manuscript received June 18, 2013; revised August 29, 2013; accepted September 30, 2013. Date of publication October 4, 2013; date of current version June 30, 2014.

X. Long, P. Fonseca, and R. M. Aarts are with the Department of Electrical Engineering, Eindhoven University of Technology, 5612 AZ Eindhoven, The Netherlands, and also with the Philips Research, HTC, 5656 AE Eindhoven, The Netherlands (e-mail: xi.long@philips.com; pedro.fonseca@philips.com; ronald.m.aarts@philips.com).

J. Foussier is with the Philips Chair for Medical Information Technology, RWTH Aachen University, 52074 Aachen, Germany (e-mail: foussier@hia.rwth-aachen.de).

R. Haakma is with the Philips Research, HTC, 5656 AE Eindhoven, The Netherlands (e-mail: reinder.haakma@philips.com).

Digital Object Identifier 10.1109/JBHI.2013.2284610

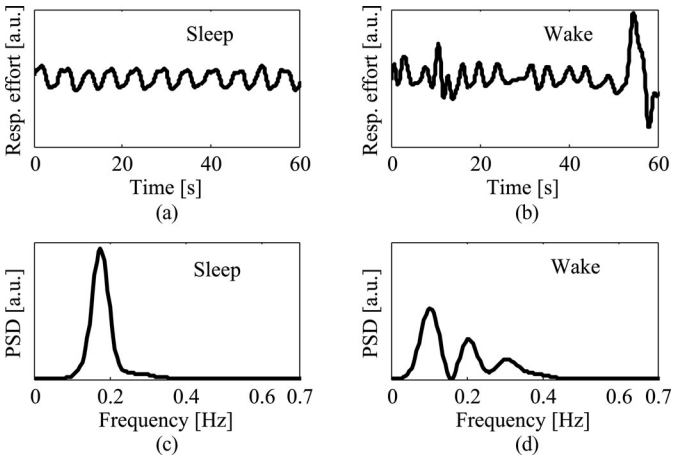


Fig. 1. Typical examples of respiratory time series (a) during sleep and (b) during wake in a period of 1 min, and respiratory PSD series (c) during sleep and (d) during wake.

important to improve the sleep and wake classification when actigraphy is absent. On the other hand, cardiac activity is relatively difficult to capture reliably in an unobtrusive manner, particularly when compared with body movement and respiratory activity [21]. For example, a novel radio frequency sensing system [24], which can only capture the respiratory effort, was developed for sleep/wake measurement. Thus, enhancing the sleep and wake classification performance without cardiac activity is also of importance. This paper therefore addresses the problem of obtaining a reliable sleep and wake classification based on the following physiological signal modalities: 1) only the respiratory effort and 2) the combination of actigraphy and respiratory effort.

As presented in previous studies, a large amount of features have been explored for sleep and wake classification [2], [4], [7]. As long as either ECG or actigraphy is excluded, the classification performance will degrade to a certain degree [4], [5], [13]. In this paper, we present new features based on the respiratory effort, which result in a classification performance not only better than the previous actigraphy and respiratory features, but also comparable to the cardiorespiratory feature set described in [4]. Compared to that work, this study does not require ECG signals, which is particularly well suited to the problem of unobtrusive sleep and wake classification.

It is known that the breathing rhythm is usually more stable and more regular during sleep than when awake [25], [26]. After observing different respiratory effort signals in the time and the frequency domains, we found that the morphology of the respiratory waveform and the properties of its power spectral density (PSD) differ between sleep and wake epochs. As illustrated in Fig. 1, the respiratory effort is more regular during sleep than during wake. Additionally, the PSD of the respiratory effort signal of a sleep epoch is typically distributed with a clear peak indicating the dominant respiration frequency, while that of a wake epoch often distributes with multiple peaks. Therefore, it is assumed that, a sleep epoch is more similar to another sleep epoch and less similar to a wake epoch from the perspective of "series shape," regardless of being in the time or in the frequency domain. We thereby concentrate on two questions: 1) how to quantify the "(dis)similarity" between two series in terms of their morphological properties and 2) which template best reflects the shape of a specific state (sleep/wake)?

Dynamic warping (DW) algorithms have been used to assess (dis)similarity of two data series with respect to their values. In particular, dynamic time warping (DTW) [27] is a signal matching algorithm that represents the time-alignment between two time series via dynamic programming by means of a total cumulative distance function. It can therefore be used to establish the degree to which two patterns match. Dynamic frequency warping (DFW) [28] is an exact analog of DTW but applied in the frequency domain, where it aims at aligning two PSD curves (often known as spectrogram frames). When used with respiratory effort signals, DTW is expected to find a good match between the waveforms of the respiratory effort in two separate sleep periods. In contrast, it should not find any good match of the respiratory waveform between a sleep and a wake period, or even between two distinct wake periods. This is simply because the breathing pattern during wake is usually not as regular as it is during sleep, and sometimes it is more related to body motion artifacts. Analogously to DTW, DFW can help distinguish respiratory PSD curve between a sleep and a wake state. Using DTW and DFW, we can express the (dis)similarity of signals in the time and in the frequency domains, and accordingly capture properties of the respiratory effort signals which are characteristic of sleep and wake.

In this paper, we propose two respiratory-based features based on DW algorithms to discriminate the respiration pattern between a sleep and a wake state. More concretely, one feature uses DTW to calculate dissimilarity scores in the time domain and is applied on the respiratory (effort) time series; the other uses DFW to calculate dissimilarity scores in the frequency domain and is applied on the respiratory PSD series. Both algorithms find an optimal alignment between two discrete data series allowing variations in two dimensions, e.g., scaling or shifting, and amplitude or offset [27]–[29]. For a given epoch from a subject's recording, the features search for the most similar epoch (i.e., optimally aligned epoch) as a template over some other epochs of the same recording based on DW, instead of using a globally predefined template for all subjects. This may possibly reduce the impact of the physiological differences between subjects. Besides, because these epochs are all taken from the same subject and the properties of the respiratory activity will not change dramatically throughout the night, the impact of within-subject variation might be small. Consequently, these would potentially increase the classification performance across the entire dataset.

DW has been widely applied to recognize patterns in various topics such as speech processing [30], fingerprint verification [31], and gene expression [32]. However, to our knowledge, studies exploring the application of DW in association with sleep staging do not seem to exist. Preliminary results of this paper were previously published in [33].

TABLE I
SUBJECT DEMOGRAPHICS

Parameter	Mean ± Std	Range
Sex	5 males and 10 females	
Age [yrs]	31.0 ± 10.4	23 – 58
Body Mass Index (BMI) [kg/m^2]	24.4 ± 3.3	20.2 – 31.2
Total recording time [hrs]	7.2 ± 1.1	4.2 – 9.1
Number of total epochs	866.2 ± 135.6	507 – 1092
Sleep efficiency* [%]	92.3 ± 3.8	86.0 – 97.9

Note: For some subjects, only a portion of recording was used because EEG electrodes fell off during the night.
*Ratio between total sleep time and total time in bed (here equal to the recording length) based on the manual scores.

II. METHOD: DATASET

The dataset was comprised of single-night PSG recordings and actigraphy (Actiwatch, Philips Respironics) of 15 *healthy* adults. Inclusion in the data collection trial was defined by a score lower than six on the Pittsburgh sleep quality index. For each subject, full PSG was recorded according to the AASM guidelines [6]. Nine subjects were monitored in the Sleep Health Center, Boston, USA, during 2009 (Alice 5 PSG, Philips Respironics) and of six in the Philips Experience Lab of the High Tech Campus in Eindhoven, The Netherlands, during 2010 (Vitaport 3 PSG, TEMEC). The subject demographics are presented in Table I. The Ethics Committee of the two sleep laboratories (or labs) approved the study protocol and all subjects signed an informed consent form.

Actigraphy was obtained with the wrist-worn Actiwatch where acceleration data, caused by body movements, were recorded and converted into activity counts per second (influenced by the intensity and frequency of acceleration) [34], [35]. The thoracic respiratory effort signal was recorded using respiratory inductance plethysmography with a sampling rate of 10 Hz. Note that the recordings from the Actiwatch were synchronized with those from the PSG, using markers in both the Actiwatch and the PSG clocks.

Sleep stages were scored on 30-s epochs by sleep experts based on the guidelines of the AASM [6] as *wake*, *REM*, and each of the *NREM* stages (*N1–N3*). *NREM* and *REM* stages were further combined in a single *sleep* class. Each PSG recording was manually clipped to the time interval comprised between the instant when the subject turned the lights OFF with the intention of sleeping until the moment the lights were turned ON before the subject got out of bed in the morning.

III. METHOD: DYNAMIC WARPING ALGORITHM

A. Dynamic Warping Distance

DW computes a distance between two series by nonlinearly aligning them in a given dimension. Consider two series:

$$A = \{a_1, a_2, \ldots, a_i, \ldots, a_n\} \quad (\text{length } n) \tag{1}$$

$$B = \{b_1, b_2, \ldots, b_j, \ldots, b_m\} \quad (\text{length } m). \tag{2}$$

These two series can be arranged such that they form a warping path in an n-by-m "warping matrix," where each element of the matrix (i, j) is given by a distance function D, expressing the squared distance between a_i and b_j

$$D(i, j) = (a_i - b_j)^2. \tag{3}$$

The warping path maps the elements of A and B through the matrix so that the total cumulative distance between them is minimized. The warping path W belongs to a set Ω including all possible warping paths, and is denoted as

$$W = \{w_1, w_2, \ldots, w_k, \ldots, w_K\} \quad (\text{length } K) \tag{4}$$

where $w_k = (i, j)_k$ is the kth element of the warping path W and $\max(n, m) \leq K \leq m + n - 1$. The DW distance between the two series is the minimum measure based on W such that

$$\text{DW}(A, B) = \min \left[\frac{1}{K} \sqrt{\sum_{k=1}^{K} w_k} \right], \quad W \in \Omega \tag{5}$$

where the distance is normalized by a factor K (path length). Fig. 2 illustrates an example of the dynamic warping procedure between two series A and B in a 2-D space.

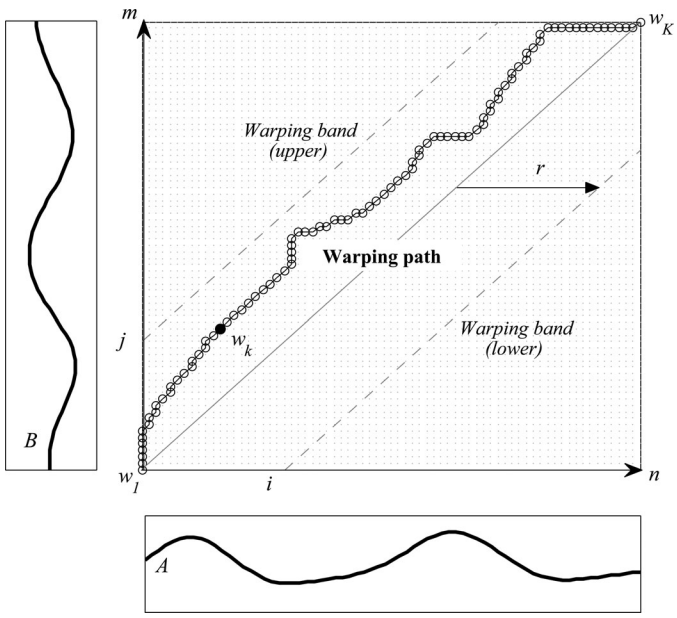


Fig. 2. Example of the DW process between two series A and B, where the warping path (circle markers) and the Sakoe–Chiba warping bands with the size of r (dash lines) are indicated.

B. Warping Conditions

Since the DW algorithm searches for an optimal warping path through all possible paths, the number of possible combinations quickly explodes with the length of the series. The search space can be reduced by means of "conditions," which help to effectively mitigate the quadratic complexity of the algorithm [27]. Several conditions are used to decrease the number of possible paths including *continuity*, *monotonicity*, *slope constraint*, and *boundary constraint* [27], [36]. They can be used to construct a warping path specified by a recurrence

$$\Delta(i, j) = D(i, j) + \min[\Delta(i-1, j-1), \Delta(i-1, j), \Delta(i, j-1)] \tag{6}$$

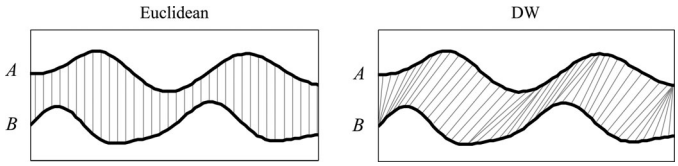


Fig. 3. Example of the alignment between two series (A and B) when computing the Euclidean (Left) and DW (Right) distances.

where the cumulative distance $\Delta(i, j)$ is defined as the sum of the distance $D(i, j)$ found in a warping step with the minimum of the cumulative distances of the adjacent elements on the warping matrix.

Additionally, the warping path can be restricted by a band of size r (i.e., $|i_k - j_k| \leq r$) on both sides of the diagonal points of the warping matrix to reduce computational complexity of a DW procedure (i.e., to reduce search space of the warping matrix). It is called *warping band* condition, and the corresponding band is commonly known as the Sakoe–Chiba band [37] (see Fig. 2). In regard to the *warping band* condition, using a band size r that is too large often results in "over warping" the periodic series with multiple cycles and thus introducing artificial features [38]. These artificial features usually occur when the warping path takes excessive numbers of nondiagonal (i.e., vertical or horizontal) moves. While a very small band size may account for "under-warping" between two series (the extreme cases is the Euclidean alignment that corresponds to the diagonal line of the warping matrix) [29]. Over warping and underwarping are both undesirable. To determine a suitable band size r, we search for the parameter value that would result in the highest feature discriminative power. This will be presented in Section V-B.

C. DW Versus Euclidean

The Euclidean distance (computed as a sequential mapping of two series) is a special case of the DW distance, where the warping path coincides with the diagonal line of the warping matrix. It is known to be sensitive to distortion in the horizontal dimension of a series [36]. Fig. 3 depicts an example of the Euclidean and the DW alignments between series A and B. It illustrates that the DW allows them to scale or shift along the horizontal dimension. Thus, in this example, the DW distance is smaller than the Euclidean distance.

D. DTW and DFW Distance

When the DW algorithm is used to compute the distance between two time series A^T and B^T, it is called "DTW algorithm" with corresponding DTW distance. Similarly, it is called "DFW algorithm" with corresponding DFW distance when used to compute the distance between two frequency (or PSD) series A^F and B^F. The superscripts indicate the time series (T) and PSD series (F). These two distance measures can be obtained based on (5) and (6) described previously.

IV. Method: Sleep and Wake Classification

A. Signal Preprocessing and PSD Estimation

Before feature extraction, the respiratory effort signal of each recording is first low-pass filtered (using a tenth-order Butterworth filter with a cutoff frequency of 0.7 Hz) to eliminate high-frequency noise, after which the baseline is removed by subtracting the median peak-to-trough amplitude estimated over the entire recording. On the other hand, for each epoch, a short-time Fourier transform (STFT) can be used to estimate a PSD based on the resulting preprocessed respiratory effort signal according to the following procedure: the resulting signal is first divided in 60-s frames centered on the epoch of interest, with an frame-to-frame overlap of 50%; after that, a Hanning window with a length of 60 s is used to reduce spectral leakage; the spectrum is then computed using the fast Fourier transform; finally, the absolute spectral values along the positive frequency axis are squared, yielding the PSD estimate for this epoch.

B. Feature Extraction

Actigraphy and respiratory effort are considered, from which features are extracted for sleep and wake classification. First, an actigraphy feature can be extracted from the output (activity counts per second) of the Actiwatch. Second, we introduce two new features: a DTW feature "minimum DTW distance" and a DFW feature "minimum DFW distance," extracted from the preprocessed respiratory effort signal.

1) Actigraphy Feature: The actigraphy feature (`ac`) is first calculated as the sum of activity counts over one epoch with 30 s; then it is smoothed across nine epochs via a weighted moving average method to eliminate noise introduced during measurement and/or when converting acceleration data into activity counts [4], [34], [35]. This feature gives an indication of gross body movements during sleep.

2) DTW Feature: The DTW feature (`dtw`) is computed based on the respiratory time series of a subject with the DTW algorithm described previously. For each epoch, it measures the maximum similarity in the time domain between that epoch and a "time-series template" having the same length. Assume that the respiratory effort data recorded for a given subject is split into L nonoverlapping epochs. Each of them consists of a collection of N data points in the time domain with a length of 30 s, such that

$$E^T(L) = \{E_1^T, E_2^T, \ldots, E_p^T, \ldots, E_L^T\} \tag{7}$$

where $E_p^T = \{x_{p,1}, x_{p,2}, \ldots, x_{p,N}\}$ is the time series of the pth epoch ($p \in \mathbf{Z}_+$ and $1 \leq p \leq L$) and N is the number of data points per epoch ($N = 300$ at a signal sample rate of 10 Hz). In order to compute the feature value for a given epoch of a recording, the template needs to be determined. We search for the template based on a window Λ^T with a size of $2\lambda^T$ ($<2\lambda^T$ when $p < \lambda^T$ or $p > L - \lambda^T$) centered on the given epoch ($\pm\lambda^T$), where this epoch itself should be excluded to avoid "self-alignment." Thus, for the pth epoch E_p^T, the time

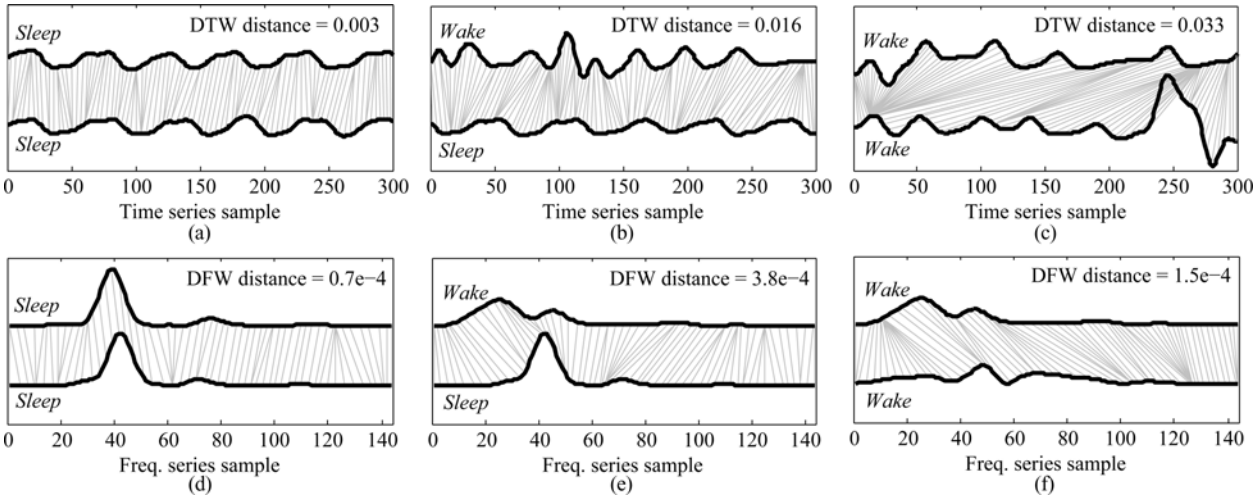


Fig. 4. Examples of DTW alignments of the respiratory time series (a)–(c) and of DFW alignments of respiratory PSD series (d)–(f), respectively, between two *sleep* epochs (S–S), between a *wake* and a *sleep* epoch (W–S), and between two *wake* epochs (W–W). Each time series lasts 30 s sampled at 10 Hz and each PSD series contains 144 samples falling within a frequency range of 0 to ∼0.7 Hz. The values of corresponding DTW and DFW distances are indicated. The alignments were done without applying the *warping band* condition for visual comparison.

series template Γ_p^T is selected using

$$\Gamma_p^T = \arg\min_{E_q^T} \mathrm{DW}(E_p^T, E_q^T)$$
$$\text{for all } q \in \mathbf{Z}_+, |q-p| \le \lambda^T, \text{ and } q \ne p \tag{8}$$

where λ^T is a positive integer with $1 \le \lambda^T \le L-1$. Then, the feature value of the pth epoch is computed by

$$\mathtt{dtw}(p) = \mathrm{DW}(E_p^T, \Gamma_p^T). \tag{9}$$

It means that we choose, as the feature value, the minimum of all DTW distances between the given epoch E_p^T and all the other epochs within a searching window Λ^T.

3) DFW Feature: The DFW feature (`dfw`) is computed based on the DFW algorithm. The procedure of computing `dfw` is the same as that of computing `dtw`, but for a respiratory PSD series rather than its time series. This feature compares the shape of the PSD curve between a given epoch and a "frequency-series template" with an indication of maximum similarity in the frequency domain. Therefore, the feature value of `dfw` for the pth epoch is obtained as

$$\mathtt{dfw}(p) = \mathrm{DW}(E_p^F, \Gamma_p^F) \tag{10}$$

where $E_p^F = \{\varphi_{p,1}, \varphi_{p,2}, \ldots, \varphi_{p,M}\}$ is the PSD series of the pth epoch ($p \in \mathbf{Z}_+$ and $1 \le p \le L$), containing M frequency bins and Γ_p^F is the selected frequency-series template. Here, the template searching window of the DFW feature is Λ^F with a size of $2\lambda^F$ epochs. As explained previously, the PSD series are obtained after STFT, for each of which the number of frequency bins is $M = 144$ in a frequency range between 0 and ∼0.7 Hz (a subset of the original spectrum with 1024 frequency bins in the range of 0–5 Hz). We limit the comparison of the PSD of each epoch to this frequency range since it can be observed that the frequency components of a healthy subject's respiration during sleep are usually below 0.7 Hz. We experimentally found that including higher frequency components would result in a lower discriminative power of the feature since they carry very small but unexpected nonzero noise that would contaminate the DFW alignment.

The use of template searching window is to reduce the computational complexity when extracting the DW features, restricting the search for minimum DW value to that window. An assumption here is that, for a given epoch, it will always offer a suitable template by searching from all the other epochs within the window except the given epoch. The procedure of determining λ^T and λ^F will be presented in Section V-B.

C. Understanding of DW-Based Features

Intuitively, there should be higher similarities of respiratory waveform and PSD shape between any two sleep epochs than between a wake and a sleep epoch or between two wake epochs. This will be expressed by the minimum DTW and minimum DFW distances found for each epoch. To further understand this, we consider two simple cases: the current epoch is sleep or is wake. Then, the feature `dtw` (or `dfw`) of this epoch may have three possible situations, where the minimal DTW (or DFW) distance may occur: between two sleep epochs (S–S), between a wake and a sleep epoch (W–S), or between two wake epochs (W–W). Regarding the DTW feature, we can state the following:

1) if the current epoch is sleep, it is likely to find a small value of DTW distance after searching for similarities of signal waveform between this epoch and the remaining epochs in a certain window, since S–S may happen;
2) if the current epoch is wake, it is not likely to obtain a small feature value, because W–S or W–W may happen.

For this reason, this feature will in turn have discriminative power for distinguishing sleep and wake states. Regarding the DFW feature, the same reasoning applies; but instead of (dis)similarities of respiratory waveform, this feature expresses (dis)similarities in the shape of PSD series. Fig. 4 depicts two examples of the alignment found by DTW and DFW between epochs, in the three situations (S–S, W–S, and W–W). Note that

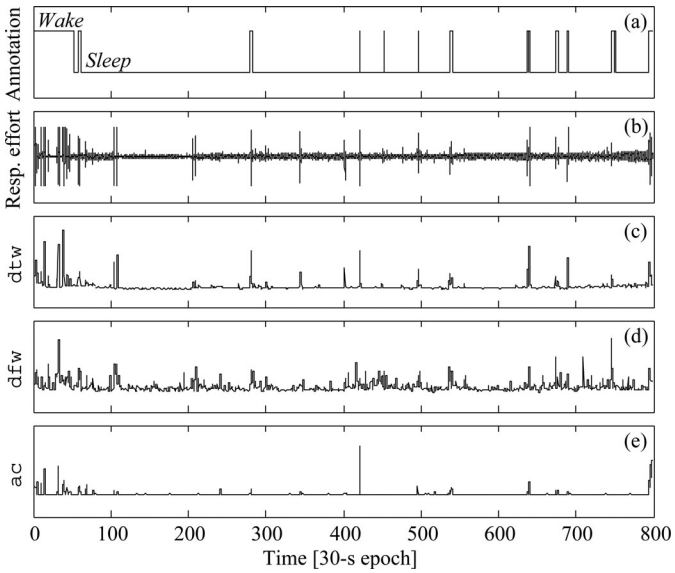


Fig. 5. Example of (a) manually scored *sleep*/*wake* annotation, (b) respiratory effort recording at 10 Hz, (c) feature values of `dtw`, (d) `dfw`, and (e) `ac` for each 30-s epoch of a healthy subject.

for the visual comparison, the alignments in the figure were done without applying the *warping band* condition, therefore, over warping seems occurring between two wake epochs [e.g., Fig. 4(c)]. Even though, this DTW distance value is still larger than the others [i.e., Fig. 4(a) and (b)].

It should be kept in mind that the respiratory waveform and PSD shape might carry some information of body motion artifacts, which often appear during wake state. This would possibly lead to irregularity of a recorded respiratory effort. As illustrated in Fig. 5, some peaks (e.g., around the 420th epoch) of the DW-based features seem correlated with the actigraphy feature (`ac`), expressing the activity counts. It means that these two features might help detect body motion artifacts. On the other hand, some peaks (e.g., around the 750th epoch) of the DW-based features seem related to the wake epochs, but where no activity counts are observed. These peaks might possibly be in correspondence with irregular breathing rhythm.

D. Classifier

A linear discriminant (LD) classifier is adopted in this study. It has been previously proved to be appropriate for the task of sleep and wake classification using actigraphy, respiratory, and cardiac data [4], [23], [39], [40]. The details of an LD classifier can be found in [23] and [41]. Note that the classifier used here is based on epoch-by-epoch classification.

Regarding the prior probability in the LD classifier, it can be observed that the probabilities of different classes vary throughout the night. For example, the probability of being awake just right after sleep onset or at the end of the night is much higher than in the middle of the night. To exploit these variations, we compute a time-varying prior probability for each epoch by counting the relative frequency that specific epoch was annotated as each class [23], [39].

V. Method: Experiment and Evaluation

A. Experimental Validation

Due to the relatively small size of our dataset, it is not appropriate to split it into separate training and test sets. To alleviate this issue, a leave-one-subject-out cross-validation (LOSOCV) procedure [41] can be used to evaluate the performance of our sleep and wake classifier. Given a set of feature vectors, we first divide it into l subsets (corresponding to $l = 15$ subjects in this study). On each iteration, one subset is used as test set and the remaining subsets are used to train the classifier. The classifier is evaluated on each test set, obtaining its performance for each iteration of the cross validation. Finally, the results are averaged and pooled to obtain an indication of the overall performance.

B. Evaluation

1) Classifier: To evaluate the performance of our classifier, overall accuracy (i.e., ratio of correctly identified samples to the total number of samples) used in a binary classification problem is not the most adequate. The reason is that during a recording of a whole night the number of epochs of the *wake* class (accounting for 7.6% of all epochs) is much smaller than that of the *sleep* class (accounting for 92.4% of all epochs), in what is usually called an "imbalanced class distribution" [42]. Thus, we also consider the metrics specificity (proportion of correctly identified actual negatives), sensitivity or recall (proportion of correctly identified positives), and precision (ratio of true positives to true positives plus false positives). Besides to these metrics, the Cohen's Kappa coefficient of agreement κ [43] provides a more insightful measure of the general performance of the classifier (0–0.20: slight, 0.21–0.40: fair, 0.41–0.60: moderate, 0.61–0.80: substantial, and 0.81–1: almost perfect agreement); but it only represents a single point in the entire solution space [44]. In order to have an overview of the performance across the entire solution space, we use a Precision–Recall (PR) curve [45], which plots precision versus recall by varying the classifier's decision-making threshold. Compared with the well-known receiver operating characteristic curve that has been shown to be overoptimistic when the dataset is heavily imbalanced between classes [46], a PR curve gives a more conservative view of the classifier's performance. The corresponding "Area under the PR curve" ($\mathrm{AUC}_{\mathrm{PR}}$) can then be estimated [46]. In the remainder of this paper, we will consider *wake* and *sleep* as the positive and negative classes, respectively.

2) Features: An absolute standardized mean difference (ASMD) metric is utilized to evaluate the discriminative power (i.e., separability) of a single feature. It computes as the absolute mean difference of the feature values between sleep and wake epochs divided by the standard deviation among that of all epochs. A Mann–Whitney unpaired (1-sided) test is applied to check whether the feature values of the two classes significantly differ. Moreover, the Spearman's rank correlation coefficient (denoted as ρ) measures the correlation between features. The significance of correlation can be examined with a Student's t-test.

In addition to evaluating the feature discriminative power between sleep and wake epochs, more specifically, we will also evaluate that between wake and REM epochs and between sleep and quiet-wake epochs. This is because the sleep and wake misclassification often occurs between wake and REM epochs by means of the traditional (cardio)respiratory features [23], [47], and between quiet-wake and sleep epochs when using actigraphy only [9]. Here, the quiet-wake is defined as the wake with computed activity counts lower than 4.5, approximate to the mean value of all the sleep epochs.

3) Parameters: The ASMD metric can also be used to determine the parameters (i.e., the Sakoe–Chiba band size r^T and r^F and the template searching window size λ^T and λ^F) for computing the DW-based features. For each feature, a grid search method is applied for the two parameters that optimize the feature's ASMD value. To obtain an unbiased determination, the grid search is therefore run on each training set during the LOSOCV procedure. Then, for each parameter, the determined value is in the majority of the optimal values occurred on different training sets.

4) Sleep Statistics: For the purpose of objectively assessing different aspects of sleep, it makes sense to evaluate the performance of the classifier in respect to its ability to deliver good estimates of so-called "sleep statistics." The sleep statistics include: total sleep time (TST), total wake time (TWT), sleep efficiency (SE) computed as the ratio of TST to total time in bed, sleep onset latency (SOL) computed as the time it took before the subject fell asleep, wake after sleep onset (WASO), and snooze time (ST). Since we are considering exclusively sleep and wake states in this study, SOL is defined as the period between the beginning of a recording and the first epoch that is annotated (or classified) as *sleep* according to the AASM guidelines [6]. For the computation of ST, we follow a similar criterion, measuring the period between the last epoch that is annotated (or classified) as *sleep* and the end of the recording. Keep in mind that the recordings are restricted to the intervals from lights OFF until lights ON. For each statistic, we compute the error as the difference value (estimation bias) and as the absolute difference value (absolute error) between the reference (computed based on the PSG-based manual annotation) and the estimate (computed based on the classification result). Furthermore, we apply Bland–Altman scatter plots to assess the degree of agreement between the PSG-based and estimated statistics.

C. Classification Performance Comparison

The actigraphy and the DW-based features used in this study are first compared. They are denoted as F_{AC}, F_{DTW}, and F_{DFW} for comparison with other feature sets (see Table II).

Our earlier studies [4], [39] have considered a large amount of features for sleep and wake classification. In those studies, a subset of features was selected from them based on the feature selection method described in [39]. It consists of five features—an actigraphy feature activity counts (`ac`); three respiratory features including standard deviation of respiratory frequency over nine epochs (`sdf`), high frequency components (`hfc`) [2], and nonlinear measure by means of sample entropy (`se`) [48]; and a cardiac feature mean heart rate (`mhr`). However, these selected respiratory features do not or less reflect characteristic morphological properties of the respiratory effort waveform or their variation over time by means of PSD shape. Those properties will be exploited with the introduction of the new DW-based features.

TABLE II
SUMMARY OF FEATURE SETS

Feature set	Features (#)	Modality*
F_{AC}	`ac` (1)	A
F_{DTW}	`dtw` (1)	R
F_{DFW}	`dfw` (1)	R
F_{R1}	`sdf`, `hfc`, `se` (3)	R
F_{R2}	`sdf`, `hfc`, `se`, `dtw`, `dfw` (5)	R
F_{DW}	`dtw`, `dfw` (2)	R
F_{AR1}	`ac`, `sdf`, `hfc`, `se` (4)	A, R
F_{AR2}	`ac`, `sdf`, `hfc`, `se`, `dtw`, `dfw` (6)	A, R
$F_{\mathrm{AC\text{-}DW}}$	`ac`, `dtw`, `dfw` (3)	A, R
F_{ARC1}	`ac`, `sdf`, `hfc`, `se`, `mhr` (5)	A, R, C
F_{ARC2}	`ac`, `sdf`, `hfc`, `se`, `dtw`, `dfw`, `mhr` (7)	A, R, C

*A: actigraphy data; R: respiratory effort data; C: cardiac data.

To understand whether the new DW-based features add discriminative power to a sleep and wake classifier that uses the selected features extracted from different signal modalities, we consider three respiratory-based feature sets and three actigraphy- and respiratory-based feature sets, in which the features included are presented in Table II . For the comparison of classification performance with and without cardiac information, two feature sets F_{ARC1} and F_{ARC2} including all the previously selected features (or together with the DW-based features) are also considered.

Since our data were collected from two distinct sleep labs (Boston or Eindhoven), the lab-effect (possibly caused by the difference of PSG setup during measurement between labs) on sleep and wake classification is then analyzed by using one dataset for training and the other for testing.

D. Computational Complexity of DW-Based Features

The original dynamic programming (without any conditions) is extraordinarily computationally intensive because it searches through all possible warping paths [27]. The use of the warping conditions can, to a great extent, speed up the DW computation [27], [29]. To compare the computational complexity when extracting DW-based features, three approaches are considered as follows.

1) The most commonly used DW approach is the one with the warping conditions (see Section III-B) but without the *warping band* condition [49]. It requires a computational complexity of $O(N^2)$, where the two series have the same length N. When we use exhaustive template searching, the complexity of computing a DW-based feature value becomes $O(LN^2)$, in which L is the epoch number of a recording. This approach is denoted as A1.
2) The Sakoe–Chiba *warping band* condition brings down the computational complexity to $O(LrN)$ instead of

TABLE III
PARAMETER DETERMINATION PROCEDURE

Parameter	Symbol	Grid Search			Determined
		Min	Max	Step	Value
Sakoe-Chiba	r^T	0	300	5	60 samples†
warping band	r^F	0	144	1	5 freq. bins‡
Template searching	λ^T	25	500*	25	200 epochs§
window size (1-side)	λ^F	25	500*	25	250 epochs♮

*The maximal window size could be limited to the total number of epochs when computing features.
†6 (6.4 ± 0.9) seconds; ‡∼0.024 (0.026 ± 0.004) Hz; §100 (94.4 ± 10.3) mins; ♮125 (134.4 ± 23.3) mins.

TABLE IV
FEATURE DISCRIMINATIVE POWER (ASMD)

Feature	*Sleep* vs. *Wake*	*Quiet-wake* vs. *Sleep*	*Wake* vs. *REM*
ac	1.77*	0.16**	0.92*
dtw	1.75*	0.48*	1.03*
dfw	1.39*	0.74*	0.70*

Note: Significance of difference between classes was examined with a Mann-Whitney test (*$p < 0.0001$ and **$p < 0.005$).

$O(LN^2)$, where r is the warping band size and typically $r \ll N$ [49]. This approach is denoted as A2.

3) Setting a template searching window Λ with a size of 2λ can reduce the complexity to $O(\lambda r N)$, where $\lambda < L$. This approach is denoted as A3.

These three DW approaches will be compared in terms of average computation time of extracting a DW-based feature for one epoch, implemented in a MEX-compiled C routine used in MATLAB (Mathworks, Natick, MA, USA). All computations were carried out in a laptop computer with a single Intel(R) Core(TM) i5 processor (2.53 GHz) and 4 GB-RAM memory.

VI. RESULTS

Table III indicates the determined parameter values obtained by the grid search method. Since the determination was based on the training set of each iteration during the LOSOCV procedure, the optimal values for each iteration might differ. Their means and variances (over grid search iterations) are also indicated in the table.

Table IV shows the pooled discriminative power (as measured by ASMD) of the selected features for all subjects in separating the *sleep* and *wake* classes. As confirmed with the Mann–Whitney unpaired (1-sided) test, the differences of the features between these two classes are significant. The table also indicates that the DW-based features perform much better than actigraphy when discriminating between quiet-wake and sleep; and the feature dtw offers a higher discriminative power compared with the other features for *wake* and *REM* separation. Fig. 6 illustrates the box plots of the three features (ac, dtw, and dfw) for sleep and wake epochs for every subject and the pool of all these subjects. It clearly shows how the features can help discriminate (albeit not perfectly) between the two classes. Classification errors will occur for feature values where the box plots overlap. Besides, the in-between feature correlations for all subjects are presented in Table V, indicating a higher correlation between ac and dtw.

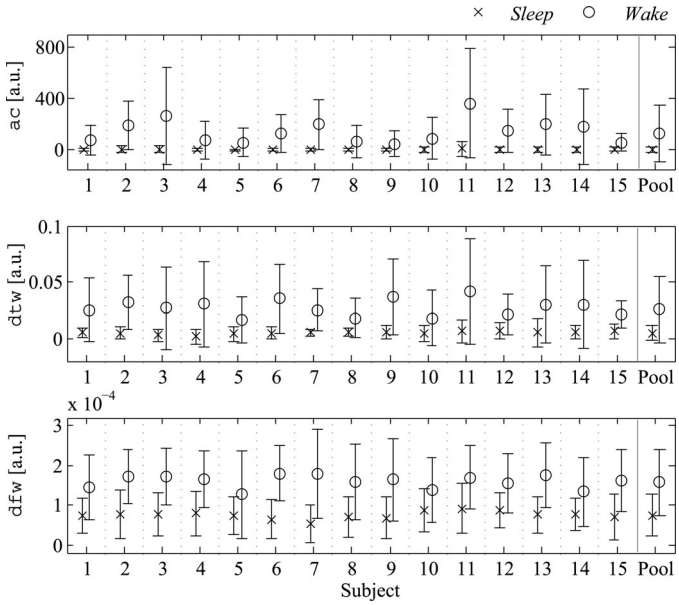


Fig. 6. Box plots (means and stds) of the feature values of ac, dtw, and dfw for *sleep* and *wake* epochs for each of the 15 subjects and for the pool of all these subjects.

TABLE V
FEATURE CORRELATION MATRIX

Correlation (ρ)*	ac	dtw	dfw
ac	1	0.32†	0.26†
dtw	–	1	0.26†
dfw	–	–	1

*Spearman's rank correlation coefficient.
†Significance of correlation was examined with a *t*-test, $p < 0.0001$.

The classification results obtained with each of the feature sets after LOSOCV are summarized in Table VI, where both "averaged" and "pooled" results are presented. Note that, for each feature set, the decision threshold (i.e., operating point) of the classifier was chosen to optimize the Kappa coefficient (based on training sets) rather than the overall accuracy due to the between-class imbalance of our data. As it can be seen in the table, for instance, the two DW-based features (i.e., F_{DW}) provide a pooled κ of 0.59, which seems to be comparable with the actigraphy feature (corresponding to a κ of 0.58). Combining them with the actigraphy feature in $F_{\mathrm{AC-DW}}$, we achieved a pooled κ of 0.66 and a pooled accuracy of 95.7%. The table also presents the classification results obtained with F_{AR1} and F_{AR2}, indicating that the addition of DW-based features significantly improves the classification performance. It also shows that the feature set $F_{\mathrm{AC-DW}}$ performs significantly better than F_{AR1} and comparably with F_{AR2}. For comparison, the results based on the feature sets comprising actigraphy, respiratory, and cardiac features are also provided in Table VI. No significant difference was found between $F_{\mathrm{AC-DW}}$, F_{ARC1}, and F_{ARC2}. Fig. 7 compares the pooled PR curves using different feature sets.

The classifier's learning curves (based on $F_{\mathrm{AC-DW}}$) using LOSOCV are displayed in Fig. 8. It is plotted as pooled κ versus the number of subjects (varying from 2 to 15). The results on

TABLE VI
SUMMARY OF SLEEP AND WAKE CLASSIFICATION PERFORMANCES USING LOSOCV

Feature set	Precision [%]	Sensitivity [%]	Specificity [%]	Accuracy [%]	AUC_{PR}*	Kappa (κ)*
F_{AC}	64.5 (66.2 ± 20.9)	57.5 (61.8 ± 16.1)	97.4 (97.3 ± 2.0)	**94.4 (94.2 ± 1.7)**	**0.66 (0.73 ± 0.09)**	**0.58 (0.57 ± 0.07)**
F_{DTW}	50.7 (51.0 ± 17.9)	62.9 (67.7 ± 17.0)	95.0 (94.9 ± 2.5)	92.5 (92.4 ± 2.1)	0.55 (0.60 ± 0.13)	0.52 (0.51 ± 0.11)
F_{DFW}	43.3 (42.6 ± 11.8)	50.8 (53.7 ± 11.9)	94.5 (94.3 ± 2.3)	91.2 (91.0 ± 2.9)	0.43 (0.44 ± 0.10)	0.41 (0.41 ± 0.08)
F_{R1}	45.2 (51.6 ± 20.2)	54.3 (52.9 ± 16.8)	94.6 (93.8 ± 6.9)	91.5 (90.9 ± 6.0)	0.52 (0.55 ± 0.14)	0.45 (0.44 ± 0.12)
F_{R2}	64.2 (66.8 ± 20.5)	56.0 (55.0 ± 19.3)	97.3 (96.6 ± 3.0)	94.2 (94.1 ± 2.6)	0.64 (0.67 ± 0.16)	0.57 (0.55 ± 0.17)
F_{DW}	63.5 (64.0 ± 18.9)	59.9 (63.4 ± 16.2)	97.3 (97.2 ± 1.9)	**94.3 (94.2 ± 2.2)**	**0.64 (0.68 ± 0.12)**	**0.59 (0.58 ± 0.11)**
F_{AR1}	70.6 (75.3 ± 29.2)	60.5 (62.4 ± 18.7)	97.8 (97.6 ± 2.7)	95.0 (94.8 ± 2.3)	0.68 (0.75 ± 0.12)	0.62 (0.61 ± 0.12)
F_{AR2}	75.0 (80.2 ± 20.1)	62.9 (64.2 ± 20.6)	98.1 (97.9 ± 2.8)	95.6 (95.3 ± 2.4)	0.73 (0.78 ± 0.13)	0.66 (0.64 ± 0.15)
$F_{AC\text{-}DW}$	77.3 (79.1 ± 16.5)	61.2 (64.5 ± 20.6)	98.5 (98.3 ± 1.9)	**95.7 (95.5 ± 2.0)**	**0.74 (0.78 ± 0.12)**	**0.66 (0.65 ± 0.13)**
F_{ARC1}†	76.9 (81.4 ± 17.0)	60.3 (59.8 ± 19.1)	98.5 (98.4 ± 2.3)	95.6 (95.5 ± 2.9)	0.72 (0.77 ± 0.12)	0.65 (0.64 ± 0.14)
F_{ARC2}†	75.4 (79.8 ± 16.6)	63.2 (63.0 ± 18.8)	98.3 (98.1 ± 2.3)	**95.7 (95.5 ± 2.2)**	**0.74 (0.77 ± 0.13)**	**0.67 (0.65 ± 0.12)**

Note: For each metric, the pooled and the averaged (between brackets) results over subjects are provided. Results were chosen to optimize κ.
*Significance of difference between feature sets was examined with a *t*-test (with 14 degrees of freedom and $p < 0.05$). Normality of the results was confirmed with a *Q-Q* plot method.
†Compared to the previous work [4], [33], a larger data set with 15 subjects was used; the features (except the DW-based features) were selected from a larger feature set based on the selection method described in [39].

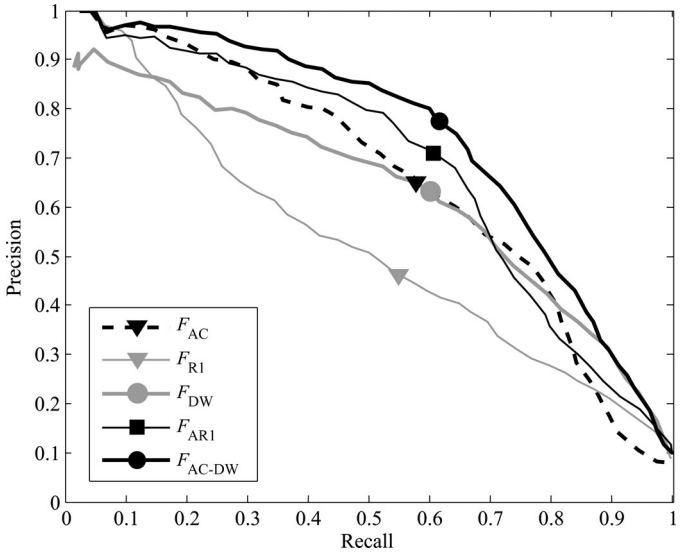


Fig. 7. PR curves with features or feature sets with their corresponding operating points of classifier (representing κ) are marked.

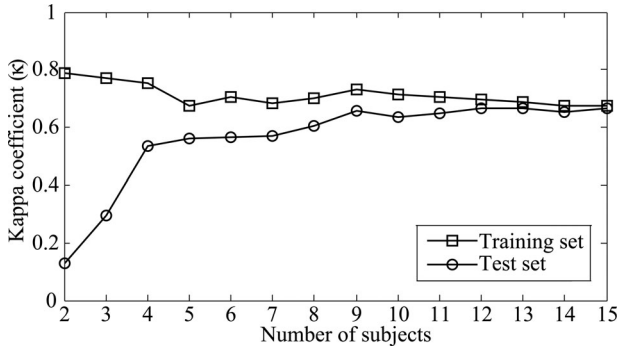


Fig. 8. Learning curves with LOSOCV by varying the number of subjects.

TABLE VII
CLASSIFICATION RESULTS WITH SPLIT TRAINING AND TEST SETS

Training	Test	Accuracy [%]	AUC_{PR}	Kappa (κ)
Boston	Boston	96.0	0.80	0.71
Boston	Eindhoven	95.4	0.66	0.61
Eindhoven	Eindhoven	95.2	0.65	0.59
Eindhoven	Boston	95.5	0.78	0.67

TABLE VIII
COMPARISON OF SLEEP STATISTICS—ABSOLUTE ERROR AND ESTIMATION BIAS (MEAN ± STD OVER SUBJECTS)

Sleep Statistics	Absolute Error		Estimation Bias*	
	F_{AR1}	$F_{AC\text{-}DW}$	F_{AR1}	$F_{AC\text{-}DW}$
SE [%]	3.3 ± 2.4	2.8 ± 2.1	-0.1 ± 4.1	-1.3 ± 3.3
TST [min]	13.7 ± 10.4	11.2 ± 7.9	-0.43 ± 17.6	-6.5 ± 12.4
TWT [min]	13.7 ± 10.4	11.4 ± 8.0	0.43 ± 17.6	-6.9 ± 12.4
SOL [min]	6.4 ± 7.1	5.0 ± 6.8	3.9 ± 8.8	3.4 ±7.8
WASO [min]	12.2 ± 10.4	7.1 ± 5.3	-3.6 ± 15.9	3.6 ± 8.2
ST [min]	1.0 ± 1.3	1.1 ± 1.7	0.17 ± 1.7	-0.13 ± 2.0

*For each subject, estimation bias was computed as the reference value minus estimated value.

training and test sets start converging rapidly from four or five subjects and become stable at 13 subjects, ultimately achieving a κ of ∼0.66. This confirms the unsuitability of splitting separate training and test sets in our experiment.

Table VII shows the classification results (pooled overall accuracy, AUC_{PR}, and Kappa) of using our actigraphy- and respiratory-based feature set F_{AC-DW} by splitting training and test sets with regard to lab (i.e., using the Boston set to train the classifier and testing it on the Eindhoven set, and the other way around).

The results (absolute error and estimation bias) of the sleep statistics over subjects using different actigraphy and respiratory feature sets (F_{AR1} and F_{AC-DW}) are summarized and compared in Table VIII. Using F_{AC-DW}, we achieved significantly lower absolute errors (after *t*-test, $p < 0.05$) in estimating the sleep statistics compared with that using F_{AR1}, with an exception of ST. To compare the degree of agreement, the Bland–Altman scatter plots were produced in Fig. 9. It can be seen that the

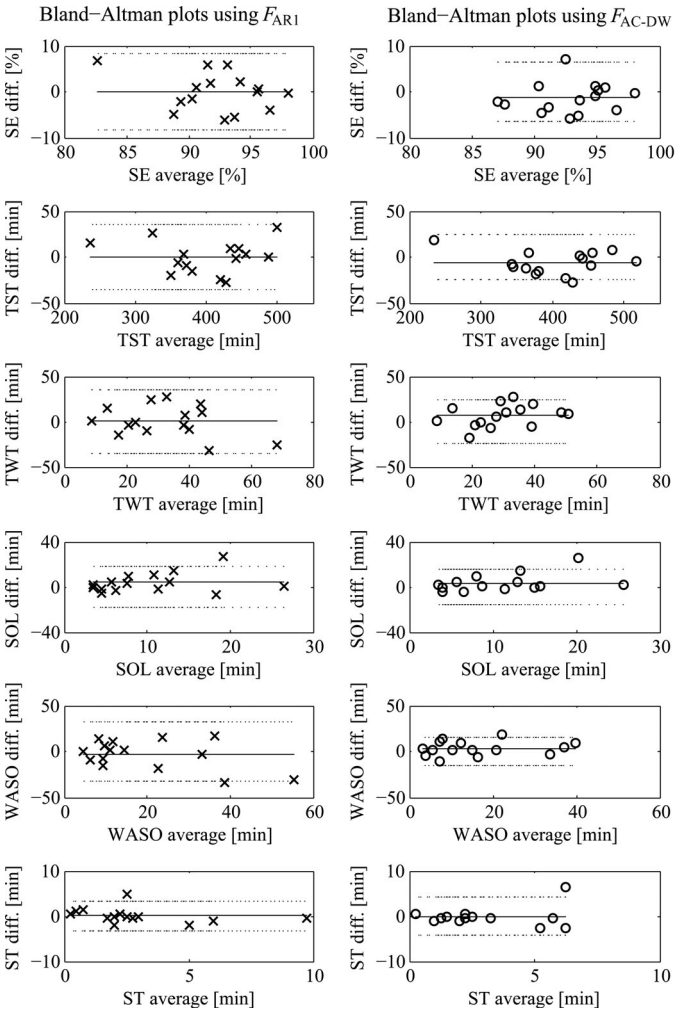


Fig. 9. Bland Altman plots of for sleep statistics estimated using F_{AR1} (Left) with data points marked by "×" and $F_{\mathrm{AC-DW}}$ (Right) with data points marked by "○." Data points in a plot represent different subjects. Mean bias and 95% limits (±1.96 std) are shown as solid and dash lines, respectively.

TABLE IX
COMPARISON OF COMPUTATIONAL COMPLEXITY FOR DIFFERENT DW-BASED FEATURE EXTRACTION APPROACHES

DW Approaches	Computational Complexity	Average Computation Time [s]	
		`dtw` (one epoch)	`dfw` (one epoch)
A1	$O(LN^2)$	2.29 ± 0.08	0.44 ± 0.02
A2	$O(LrN)$*	1.43 ± 0.05	0.21 ± 0.01
A3	$O(\lambda rN)$*	0.53 ± 0.04	0.10 ± 0.03

*Here the parameters are $r^T = 60$, $r^F = 5$, $\lambda^T = 200$, and $\lambda^F = 250$.

difference values of SE, TST, TWT, SOL, and WASO are more converging when using $F_{\mathrm{AC-DW}}$ than using F_{AR1}, indicating less variances (or higher degree of agreement) when estimating the sleep statistics with $F_{\mathrm{AC-DW}}$.

Table IX compares the computational complexity of different DW approaches (A1, A2, and A3). It means that, when extracting DW-based features, using a warping band for DW and constraining the template searching range can significantly reduce the computation time (after a *t*-test, $p < 0.001$). On average, it takes approximately 7.5 min for the DTW feature and 1.5 min for the DFW feature to compute all their feature values of one night per subject.

VII. DISCUSSION

During the training step of each LOSOCV iteration, some parameters, evaluated by the pooled $\mathrm{AUC_{PR}}$, were determined. The determined Sakoe–Chiba warping band for DTW ($r^T = 60$) is much larger than that for DFW ($r^F = 5$). This is because, when computing the DTW distance between two respiratory time series, they usually start and end with different phases of a breathing cycle. A larger DTW warping band allows a larger signal variation (caused by breathing phase, length, amplitude differences, etc.) between two epochs. It helps compensating for the signal variation and thus enables to find a better alignment between them. On the other hand, when computing the DFW distance, the respiratory PSDs were normalized between 0 and 1 so that the amplitude variation between epochs would be no more existing (no improvement on classification performance was observed without normalizing them). Also, they usually have less peaks and no troughs compared with time series (see Fig. 1). These would yield a higher similarity between two respiratory PSD series than between two respiratory time series. Besides, using a smaller warping band for DFW is able to avoid overalignment between two PSD series, which still enables to discriminate between *sleep* and *wake* with respect to their minimum distance.

The searching window sizes for extracting DW-based features were also determined with the use of the grid search method. Since we relied on the observation that the minimum DW distance for a sleep epoch is small, this potential disadvantage of restricting the search space are alleviated by the fact that sleep epochs are usually not isolated in time, i.e., there are, very likely, other sleep epochs close to any given (sleep) epoch during the night. Furthermore, a larger searching window might not provide a better separation between *sleep* and *wake* classes. For instance, when analyzing a wake epoch, the inclusion of more distant (in time) candidate templates might increase the likelihood of selecting a more similar wake template. This would result in a smaller DW distance and thus decrease the feature's discriminative power. Here, we found that the discriminative power of these two features did not dramatically change when $\lambda > 25$ epochs.

The DW-based features performed well for sleep and wake classification. These features can effectively encode differences in the waveform and PSD shape of the respiratory effort between sleep and wake states. As shown in Table VI, when considering the use of only the respiratory effort, our DW-based feature set $F_{\mathrm{AC-DW}}$ offers around relative 31% increase of κ compared to the existing respiratory feature set F_{R1} (i.e., κ of 0.59 versus 0.45); and it is comparable with the well-known actigraphy ($\kappa = 0.58$). After combining actigraphy with the respiratory effort signal, our DW-based features improved the classification performance from $\kappa = 0.62$ to $\kappa = 0.66$, yielding a higher relative increase ($\sim$14%) when compared with actigraphy. The reason might be that the DW-based features (particularly the

DFW feature) better help distinguish between quiet-wake and sleep (see Table IV).

A previous study [8] presented a novel actigraphy-based algorithm for sleep and wake classification, in which the authors reported an overall accuracy of ∼86%, a sleep accuracy of ∼91%, and a wake accuracy of ∼69% for a group of 38 normal subjects. In [12], the overall accuracy was ∼87% (measured in 14 healthy subjects). In this study, to perform an even comparison, we varied the operating point of our classifier and obtained comparable results based on only actigraphy. After combining it with the DW-based respiratory features, as shown in Table VI, we achieved much better results.

It is known that the wrist actigraphy ultimately measures the body (or more precisely, wrist) movements during sleep, which proved to be an indication of wake state [9], [12]. To a certain extent, they would often be reflected in the respiratory effort signal as body motion artifacts during measurement. This can be observed in Fig. 5, which suggests a relatively high correlation between peaks in the actigraphy feature and respiratory effort series. As mentioned, the respiratory waveform and PSD shape not only reflect the respiration information but also contain some information about body motion artifacts. It means that the DW-based features might encode the artifact information in both of the time and the frequency domains. Table V confirms this due to the significant correlation between `ac` and `dtw` ($\rho = 0.32$) and between `ac` and `dfw` ($\rho = 0.26$). These two features (particularly the DTW feature) might help separate wake and REM sleep, resulting in an improved classification when actigraphy is not provided (see Table VI).

The inclusion of the cardiac feature (i.e., using F_{ARC2}) did not significantly improve the performance of sleep and wake classification (see Table VI). It means that a good performance is still possible to be obtained when using fewer physiological signal modalities. However, it is still encouraged to explore new cardiac features containing additional information that can better discriminate between sleep and wake states, for which these information is not contained by actigraphy and respiratory activity. Moreover, the κ of 0.59 with only DW-based respiratory features is comparable with that of 0.60 reported in [23], where they used not only respiratory but also cardiac information.

The results of using F_{AR2} (with six features) are comparable with that using $F_{\mathrm{AC-DW}}$ (with three features). Since we aimed at evaluating the proposed new DW-based features, they were simply combined with the other preselected features. Often, using more features does not necessarily guarantee a better performance, and in some cases it may even decrease. This is because features may be mutually correlated with some extent, and thus some features are likely redundant. As a consequence, they may hardly contribute to (or even be against) the classification when the additionally useful information they carried is limited compared to the increase of noise level. Therefore, selecting features from a larger feature set aiming at removing the feature set redundancy (e.g., correlation-based feature selection [50]) merits further investigation.

As shown in Table VII, the sleep and wake classification results obtained on the Eindhoven set remain worse compared with those on the Boston set, regardless of either set used for training. This might be associated with the between-subject variability instead of lab-effect. Thus, it is not sufficiently confident to conclude about the existence of lab-effect based on our dataset with a small number of subjects. Although results have been shown to be consistent between labs [8], it is encouraged to be further studied on a larger-sized dataset .

By choosing a different classifier operating point, we can obtain results that prefer a higher specificity or sensitivity. In practice, this often depends on the requirement of accuracy in estimating sleep statistics, which can be delivered to subjects. For example, it should be chosen to optimize the estimate of SOL for subjects who might have insomnia; while for overall assessment of sleep, one can choose to optimize the estimate of SE.

In addition, we focused on the healthy subjects with high sleep efficiencies (>86%) rather than, e.g., the insomniacs with low sleep efficiencies. However, it has been indicated that distinguishing between sleep and wake states is more difficult in insomniacs than in healthy subjects when using cardiorespiratory activity [24], [47], or actigraphy [51]. Although the DW-based features perform well in sleep and wake classification for the healthy subjects, it is necessarily required to further evaluate how robust they are against low SE.

Finally, although the DW-based features seem computationally intensive based on our settings, it is still practically feasible to achieve an offline classification of sleep and wake. In fact, recent research has developed a set of techniques that can make the DW computation much faster and comparable with the Euclidean alignment, so that DW is applicable on large-sized datasets in real time [52]. Nevertheless, speeding up our algorithms using these techniques will be carried on in our future work.

VIII. Conclusion

In this study, actigraphy and respiratory effort were used to classify sleep and wake states during bedtime at night due to that they can be acquired easily and unobtrusively. To enhance the performance of sleep and wake classification, we proposed two new features extracted from the respiratory effort based on DW algorithms. The features compare the shape (dis)similarity between two series (in time and frequency domain) for a given 30-s epoch with the other epochs within a predetermined window from an entire-night respiratory effort recording. The minimal dissimilarity (measured by a DW distance) was computed as the feature value for this epoch. To evaluate the sleep and wake classification performance, an LD classifier was tested with an LOSOCV. By combining the two DW-based features with a well-known actigraphy feature, we obtained a significantly increased Cohen's Kappa coefficient ($\kappa = 0.66$) compared with the use of the actigraphy feature and the traditional respiratory features ($\kappa = 0.62$), and it significantly outperforms that only with actigraphy ($\kappa = 0.58$). It is comparable with that of 0.67, obtained with a feature set comprising the DW-features and the previously used actigraphy and cardiorespiratory features. Furthermore, when using the respiratory signal only, the

DW-based features provided a large improvement compared with the existing respiratory features (κ of 0.59 versus 0.45).

ACKNOWLEDGMENT

The authors would like to thank three anonymous reviewers and T. Leufkens from Philips Research Laboratories for their insightful comments.

REFERENCES

[1] E. A. Rechtschaffen and A. Kales, *A Manual of Standardized Terminology, Techniques and Scoring System for Sleep Stages of Human Subjects*. Bethesda, MD, USA: National Institutes of Health, 1968.

[2] S. J. Redmond and C. Heneghan, "Cardiorespiratory-based sleep staging in subjects with obstructive sleep apnea," *IEEE Trans. Biomed. Eng.*, vol. 53, no. 3, pp. 485–496, May 2006.

[3] S. Banks and D. F. Dinges, "Behavioral and physiological consequences of sleep restriction," *J. Clin. Sleep Med.*, vol. 3, pp. 519–528, Aug. 2007.

[4] S. Devot, R. Dratwa, and E. Naujokat, "Sleep/wake detection based on cardiorespiratory signals and actigraphy," in *Proc. IEEE 32nd Ann. Int. Conf. Eng. Med. Biol. Soc.*, Buenos Aires, Argentina, Aug. 2010, pp. 5089–5092.

[5] W. Karlen, C. Mattiussi, and D. Floreano, "Sleep and wake classification with ECG and respiratory effort signals," *IEEE Trans. Biomed. Circuits Syst.*, vol. 3, no. 2, pp. 71–78, Apr. 2009.

[6] C. Iber, S. Ancoli-Israel, A. L. Chesson, and S. F. Quan, *The AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology and Technical Specifications*. Darien, IL, USA: American Academy of Sleep Medicine, 2007, [Online]. Available: www.aasmnet.org

[7] R. J. Cole, D. F. Kripke, W. Gruen, D. J. Mullaney, and J. C. Gillin, "Automatic sleep/wake identification from wrist activity," *Sleep*, vol. 15, no. 5, pp. 461–469, Oct. 1992.

[8] J. Hedner, G. Pillar, S. D. Pittman, D. Zou, L. Grote, and D. P. White, "A novel adaptive wrist actigraphy algorithm for sleep-wake assessment in sleep apnea patients," *Sleep*, vol. 27, no. 8, pp. 1560–1566, Aug. 2004.

[9] S. Ancoli-Israel, R. Cole, C. Alessi, M. Chambers, W. Moorcroft, and C. P. Pollak, "The role of actigraphy in the study of sleep and circadian rhythms," *Sleep*, vol. 26, no. 3, pp. 342–392, May 2003.

[10] 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, Apr. 2007.

[11] W. W. Tryon, "Issues of validity in actigraphic sleep assessment," *Sleep*, vol. 27, no. 1, pp. 158–165, Feb. 2004.

[12] C. P. Pollak, W. W. Tryon, H. Nagaraja, and R. Dzwonczyk, "How accurately does wrist actigraphy identify the states of sleep and wakefulness?" *Sleep*, vol. 24, no. 8, pp. 957–965, Dec. 2001.

[13] W. Karlen, C. Mattiussi, and D. Floreano, "Improving actigraph sleep/wake classification with cardio-respiratory signals," in *Proc. IEEE 30th Ann. Int. Conf. Eng. Med. Biol. Soc.*, Vancouver, Canada, Aug. 2008, pp. 5262–5265.

[14] T. Penzel, N. Wessel, M. Riedl, J. W. Kantelhardt, S. Rostig, M. Glos, A. Suhrbier, H. Malberg, and I. Fietze, "Cardiovascular and respiratory dynamics during normal and pathological sleep," *Chaos*, vol. 17, no. 5, pp. 015116-1–015116-10, Mar. 2009.

[15] D. C. Mack, J. T. Patrie, P. M. Suratt, R. A. Felder, and M. A. Alwan, "Development and preliminary validation of heart rate and breathing rate detection using a passive, ballistocardiography-based sleep monitoring system," *IEEE Trans. Inf. Technol. Biomed.*, vol. 13, no. 1, pp. 111–120, Jan. 2009.

[16] G. Matthews, B. Sudduth, and M. Burrow, "A non-contact vital signs monitor," *Crit. Rev. Biomed. Eng.*, vol. 28, no. 1–2, pp. 173–178, 2000.

[17] Y. M. Kuo, J. S. Lee, and P. C. Chung, "A visual context-awareness-based sleeping-respiration measurement system," *IEEE Trans. Inf. Technol. Biomed.*, vol. 14, no. 2, pp. 255–265, Mar. 2010.

[18] W. Chen, X. Zhu, T. Nemoto, Y. Kanemitsu, K. Kitamura, and K. Yamakoshi, "Unconstrained detection of respiration rhythm and pulse rate with one under-pillow sensor during sleep," *Med. Biol. Eng. Comput.*, vol. 43, pp. 306–312, Mar. 2005.

[19] T. Watanabe and K. Watanabe, "Noncontact method for sleep stage estimation," *IEEE Trans Biomed. Eng.*, vol. 51, no. 10, pp. 1735–1748, Oct. 2004.

[20] B. H. Jansen and K. Shankar, "Sleep staging with movement-related signals," *Int. J. Biomed. Comput.*, vol. 32, pp. 289–297, 1993.

[21] T. Kirjavainen, D. Cooper, O. Polo, and C. E. Sullivan, "Respiratory and body movements as indicators of sleep stage and wakefulness in infants and young children," *J. Sleep Res.*, vol. 5, pp. 186–194, 1996.

[22] J. M. Kortelainen, M. O. Mendez, A. M. Bianchi, M. Matteucci, and S. Cerutti, "Sleep staging based on signals acquired through bed sensor," *IEEE Trans Inf. Technol. Biomed.*, vol. 14, no. 3, pp. 776–785, May 2010.

[23] S. J. Redmond, P. de Chazal, C. OBrien, S. Ryan, W. T. McNicholas, and C. Heneghan, "Sleep staging using cardiorespiratory signals," *Somnologie*, vol. 11, pp. 245–256, Dec. 2007.

[24] P. de Chazal, N. Fox, E. O'Hare, C. Heneghan, A. Zaffaroni, P. Boyle, S. Smith, C. O'Connell, and W. T. McNicholas, "Sleep/wake measurement using a non-contact biomotion sensor," *J. Sleep Res.*, vol. 20, pp. 356–366, Jun. 2010.

[25] J. Krieger, "Breathing during sleep in normal subjects," *Clin. Chest Med.*, vol. 6, no. 4, pp. 577–594, Dec. 1985.

[26] SRS, *Basic of Sleep Guide*, 2nd ed. Darien, IL, USA: Sleep Research Society, 2011.

[27] D. J. Berndt and J. Clifford, "Using dynamic time warping to find patterns in time series," in *Proc. Assoc. Advancement Artif. Intell. Workshop Knowl. Discovery Databases*, 1994, pp. 229–248.

[28] E. P. Neuburg, "Frequency warping by dynamic programming," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 1988, pp. 573–575.

[29] E. Keogh and M. Pazzani, "Scaling up dynamic time warping for data-mining applications," in *Proc. 6th Assoc. Comput. Mach. Special Interest Group Knowl. Discovery Data Mining*, 2000, pp. 285–289.

[30] L. Rabiner, A. Rosenberg, and S. Levinson, "Considerations in dynamic time warping algorithms for discrete word recognition," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 26, no. 6, pp. 575–528, Dec. 1978.

[31] Z. M. Kovacs-Vajna, "A fingerprint verification system based on triangular matching and dynamic time warping," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 11, pp. 1266–1276, Nov. 2000.

[32] J. Aach and G. M. Church, "Aligning gene expression time series with time warping algorithms," *Bioinformatics*, vol. 17, pp. 495–508, 2001.

[33] X. Long, P. Fonseca, J. Foussier, R. Haakma, and R. M. Aarts, "Using dynamic time warping for sleep and wake discrimination," in *Proc. IEEE-EMBS Int. Conf. Biomed. Health. Inf.*, Hong Kong and Shenzhen, China, Jan. 2012, pp. 886–889.

[34] Philips Respironics Actiwatch, Philips Healthcare. (Nov. 2012). [Online]. Available: http://www.actiwatch.respironics.com

[35] R. Robillard, T. J. R. Lambert, and N. L. Rogers, "Measuring sleep-wake patterns with physical activity and energy expenditure monitors," *Biol. Rhythm Res.*, vol. 43, no. 5, pp. 555–562, Sep. 2012.

[36] C. A. Ratanamahatana and E. Keogh, "Making time-series classification more accurate using learned constraints," in *Proc. SIAM Int. Conf. Data Mining*, Apr. 2004, pp. 11–22.

[37] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. AASP-26, no. 1, pp. 43–49, Feb. 1978.

[38] D. Clifford, G. Stone, I. Montoliu, S. Rezzi, F. P. Martin, P. Guy, S. Bruce, and S. Kochhar, "Alignment using variable penalty dynamic time warping," *Anal. Chem.*, vol. 81, no. 3, pp. 1000–1007, Feb. 2009.

[39] J. Foussier, P. Fonseca, X. Long, and S. Leonhardt, "Automatic feature selection for sleep/wake classification with small data sets," presented at the Int. Joint Conf. Biomed. Eng. Syst. Technol., Barcelona, Spain, Feb. 2013.

[40] X. Long, P. Fonseca, R. Haakma, R. M. Aarts, and J. Foussier, "Time-frequency analysis of heart rate variability for sleep and wake classification," in *Proc. IEEE 12th Int. Conf. BioInf. BioEng.*, Larnaca, Cyprus, Nov. 2012, pp. 85–90.

[41] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed. Reading, MA, USA: Wiley, 2001.

[42] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, pp. 1263–1284, Sep. 2009.

[43] J. A. Cohen, "Coefficient of agreement for nominal scales," *Educ. Psychol. Meas.*, vol. 20, pp. 37–46, 1960.

[44] F. Provost, T. Fawcett, and R. Kohavi, "The case against accuracy estimation for comparing induction algorithms," in *Proc. 15th Int. Conf. Mach. Learn.*, 1998, pp. 445–453.

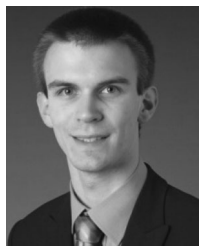
[45] T. Fawcett, "ROC graphs: Notes and practical considerations for data mining researchers," HP Labs, Palo Alto, CA, USA, Tech. Rep. MPL-2003-4, 2003.
[46] J. Davis and M. Goadrich, "The relationship between precision-recall and ROC curves," in *Proc. 23rd Int. Conf. Mach. Learn.*, 2006, pp. 233–240.
[47] K. Spiegelhalder, L. Fuchs, J. Ladwig, S. D. Kyle, C. Nissen, U. Voderholzer, B. Feige, and D. Riemann, "Heart rate and heart rate variability in subjectively reported insomnia," *J. Sleep Res.*, vol. 20, no. 1pt2, pp. 137–145, Mar. 2011.
[48] M. Costa, A. L. Goldberger, and C. K. Peng, "Multiscale entropy analysis of biological signals," *Phys. Rev. E*, vol. 71, no. 2, pp. 021 906:1–021 906:18, Feb. 2005.
[49] M. Muller, "Part 1: Analysis and retrieval techniques for music data—Dynamic time warping," in *Information Retrieval for Music and Motion*. Berlin, Germany: Springer-Verlag, 2007, ch. 4, pp. 69–84.
[50] M. A. Hall, "Correlation-based feature selection for machine learning" Ph.D. dissertation, Dept. Computer Science, The Univ. of Waikato, Hamilton, New Zealand, Apr. 1999.
[51] K. L. Lichstein, K. C. Stone, J. Donaldson, S. D. Nau, J. P. Soeffing, D. Murray, K. W. Lester, and R. N. Aquillard, "Actigraphy validation with insomnia," *Sleep*, vol. 29, no. 2, pp. 232–239, Feb. 2006.
[52] T. Rakthanmanon, B. Campana, A. Mueen, G. Batista, B. Westover, Q. Zhu, J. Zakaria, and E. Keogh, "Searching and mining trillions of time series subsequences under dynamic time warping," in *Proc. Assoc. Comput. Mach. Special Interest Group Knowl. Discovery Data Mining*, Aug. 2012, pp. 262–270.

Xi Long (M'09) was born in China, in 1983. He received the B.Eng. degree in electronic information engineering from Zhejiang University, Hangzhou, China, in 2006 and the M.Sc. degree in electrical engineering from the Eindhoven University of Technology, Eindhoven, The Netherlands, in 2009, where he is currently working toward the Ph.D. degree in electrical engineering, in association with Philips Research.

His research interests include biomedical signal analysis and pattern classification in healthcare.

Pedro Fonseca was born in Lisbon, Portugal, in 1979. He received the Dipl.-Ing. degree in electrical engineering and the M.Sc. degree in electrical and computer engineering from Instituto Superior Técnico, Technical University of Lisbon, Lisbon, Portugal.

He has eight years of R&D experience at Philips Research Eindhoven in the fields of image and video analysis and, more recently, personal healthcare. His current research interests include biosignal processing, machine learning, and unobtrusive physiological sensing.



Jérôme Foussier (M'11) was born in Cologne, Germany, in 1984. He received the Dipl.-Ing. degree in electrical engineering from RWTH Aachen University, Aachen, Germany, where he is currently working toward the Dr.-Ing. (Ph.D.) degree at the Chair of Medical Information Technology.

He is currently a Research Assistant at RWTH Aachen University. His research interests include signal processing and classification as well as physiological measurement techniques.

Reinder Haakma received the M.Sc. degree in electrical engineering from the University of Twente, Enschede, The Netherlands, in 1985, and the Ph.D. degree from the Eindhoven University of Technology, Eindhoven, The Netherlands, in 1998.

He is currently a Principal Scientist with Philips Research, Eindhoven, The Netherlands. His research interests include unobtrusive monitoring of health and sleep.

Ronald M. Aarts (M'95–SM'95–F'07) was born in Amsterdam, The Netherlands, in 1956. He received the B.Sc. degree in electrical engineering and the Ph.D. degree in physics from Delft University, Delft, The Netherlands, in 1977 and 1995, respectively.

He joined Philips Research Laboratories in 1977, working on CD, acoustics and audio, and various DSP-algorithms and applications. In 2003, he became a Philips Fellow at Philips Research, and extended his interests in engineering to medicine and biology in particular sensors, signal processing, and systems for ambulatory monitoring, sleep, lighting, and epilepsy detection. He is currently a part-time Full Professor at the Eindhoven University of Technology, Eindhoven, The Netherlands. He has published more than 200 papers and reports and holds more than 160 first patent application filings including over 50 granted U.S. patents in these fields.

Dr. Aarts received the Silver Medal and is the Fellow of the Audio Engineering Society.