Towards a Robust Estimation of Respiratory Rate from Pulse Oximeters

Marco A.F. Pimentel*, Alistair E.W. Johnson, Peter H. Charlton, Drew Birrenkott, Peter J. Watkinson, Lionel Tarassenko, David A. Clifton

***Abstract*—Objective: Current methods for estimating respiratory rate (RR) from the photoplethysmogram (PPG) typically fail to distinguish between periods of high- and low-quality input data, and fail to perform well on independent "validation" datasets. The lack of robustness of existing methods directly results in a lack of penetration of such systems into clinical practice. The present work proposes an alternative method to improve the robustness of the estimation of RR from the PPG. Methods: The proposed algorithm is based on the use of multiple autoregressive models of different orders for determining the dominant respiratory frequency in the three respiratory-induced variations (frequency, amplitude and intensity) derived from the PPG. The algorithm was tested on two different datasets comprising 95 8-minute PPG recordings (in total) acquired from both children and adults in different clinical settings, and its performance using two window sizes (32 and 64 seconds) was compared with that of existing methods in the literature. Results: The proposed method achieved comparable accuracy to existing methods in the literature, with mean absolute errors (median, 25^{th}-75^{th} percentiles for a window size of 32 seconds) of 1.5 (0.3-3.3) and 4.0 (1.8-5.5) breaths per minute (for each dataset respectively), whilst providing RR estimates for a greater proportion of windows (over 90% of the input data are kept). Conclusion: Increased robustness of RR estimation by the proposed method was demonstrated. Significance: This work demonstrates that the use of large publicly-available datasets is essential for improving the robustness of wearable-monitoring algorithms for use in clinical practice.**

***Index Terms*—Respiratory rate, photoplethysmography, pulse oximetry, patient monitoring, mobile health**

I. Introduction

Respiratory rate (RR) is a known antecedent of many episodes of physiological deterioration in patients [1], [2], and its accurate estimation in a non-invasive manner is therefore of substantial importance in many settings including mobile health and home monitoring applications. These are introduced below, in sections A-C.

A. Monitoring Hospital In-Patients

Hospitals and clinics often use *early warning scores* that involve the observation of patients' vital signs, including RR, throughout a patient's stay, with scores assigned to the observed values [3]. In the UK, for example, the use of such systems has been recommended in national clinical guidelines [4], [5].

These scores form an integral part of patient care. Review of the patient by senior members of the clinical staff is prompted if the overall score exceeds some pre-determined threshold. While attempts have been made to make these scoring systems less heuristic [3], [6], [7], little research has been performed on making the inputs to these scoring systems (i.e., the values of the vital signs) more robust. Manual observations of the vital signs are typically performed every 4 - 6 hours within UK hospitals, and so continuously-measured data from patient monitoring systems could be used to provide early warning scores on a continuous basis. This will provide an opportunity to track patient condition second-by-second, between manual observations, thereby increasing the potential for early detection of patient deterioration. Improved early recognition of physiological deterioration in patients leads to improved patient outcomes [3]. However, while heart rate (HR) and peripheral blood oxygen saturation (SpO_2) can be measured continuously using pulse oximetry, continuous estimation of RR relies on the use of extra equipment (via capnometry or measurement of gas flow). There is, therefore, a need to improve the robustness of RR estimation from the electrocardiogram (ECG), the photoplethysmogram (PPG) acquired from pulse oximeters, or other biosignals that are known to be modulated by the respiratory cycle, and hence which may be used to estimate RR.

B. m-Health and Monitoring at Home

Mobile healthcare (or m-health) is an area of patient monitoring that has received much attention in recent years [8], [9]. Patients at home will typically not tolerate wearing adhesive sensors such as ECG electrodes for extended periods. Instead, m-health applications often include the use of pulse oximetry [10], [11], whereby patients can easily insert their finger into a pulse oximeter probe.

The work of M.A.F. Pimentel was supported by a Health Innovation Challenge Fund from the Wellcome Trust and the Department of Health. The work of P.H. Charlton was supported by the EPSRC [Grant EP/F058845/1] and the National Institute for Health Research (NIHR) Biomedical Research Centre based at Guy's and St Thomas' NHS Foundation Trust and Kings College London. D. Birrenkott is funded by the Rhodes Trust. The work of D.A. Clifton was supported by the Centre of Excellence in Personalised Healthcare funded by the Wellcome Trust and EPSRC under Grant WT 088877/Z/09/Z, the Royal Academy of Engineering, and Balliol College, Oxford.

*M.A.F. Pimentel, D. Birrenkott, L. Tarassenko, and D. A. Clifton are with the Institute of Biomedical Engineering, Department of Engineering Science, University of Oxford, Oxford OX3 7DQ, UK (e-mail: {marco.pimentel; drew.birrenkott; lionel.tarassenko; david clifton}@eng.ox.ac.uk).

A. E. W. Johnson is with the Institute for Medical Engineering & Science, Massachusetts Institute of Technology, Boston, MA 02139 USA (e-mail: aewj@mit.edu).

P. H. Charlton is with the Guy's & St. Thomas' Hospital, London, UK, and King's College, London, UK (e-mail: peter.charlton@gstt.nhs.uk).

P. J. Watkinson is with the Oxford University Hospitals NHS Trust, Oxford OX3 9DU, UK (e-mail: peter.watkinson@ndcn.ox.ac.uk).

While the pulse oximeter provides robust estimates of HR and SpO_2, methods for the robust estimation of RR from the PPG waveform are lacking. The challenges for accurate RR-estimation methods in m-health are significant, primarily because of movement artefact.

C. Upcoming Technologies

Finally, we note that methods for estimating RR from the PPG may also be used in a number of recent technologies which acquire "PPG-like waveforms". The consumer electronics market is now populated by a large number of fitness trackers and other similar devices, where the aim of using such systems is not to detect physiological deterioration, but to maintain (and perhaps optimise) the "wellness" of healthy subjects. This is of particular interest to the producers of consumer electronics, because wellness applications do not require the costly, time-consuming clinical validation needed to certify devices for clinical use. While the majority of existing devices consist of simple accelerometers for tracking "activity", the latest generation of these devices includes other sensors such as pulse oximeters (e.g., in the new smart watch by Apple, Inc., USA) and bioimpedance sensors (e.g., in the *UP3* by Jawbone, USA). The availability of cardiosynchronous signals from such devices is currently being marketed as being useful only for determining the HR of the subject; however, if sufficiently robust methods are available, they could also be used to estimate RR, thereby increasing the number of physiological variables measured by such devices.

Moreover, a number of studies in the literature have recently described approaches to measuring HR and RR without any sensors or electrodes being attached to the patient, using standard video data [12]–[14]. The resulting "vPPG" (video-derived PPG) waveform provides the opportunity to estimate RR. Existing methods typically rely on tracking small regions of exposed skin in the video of the subject. However, the vPPG signal is substantially noisier than the equivalent PPG signal from pulse oximeters.

D. Overview of This Paper

Respiration is known to modulate the PPG in different ways [15], [16]. Many methods have been proposed in the literature to address the need for robust estimation of RR from analysis of the PPG signal. The state-of-the-art is reviewed in section II. Section III describes the proposed algorithm designed to provide improved robustness of RR-estimation. We then evaluate our method using two "independent" datasets, which are described in section IV. Section V then describes our evaluation of existing methods, along with our new proposed method. Finally, we discuss the implications of our study and its results in section VI.

II. Related work

Existing methods for estimating RR from the PPG are typically applied to moving windows of the time-series data, and an estimate of RR is produced for each window. They typically comprise up to four components, as illustrated in Figure 1.

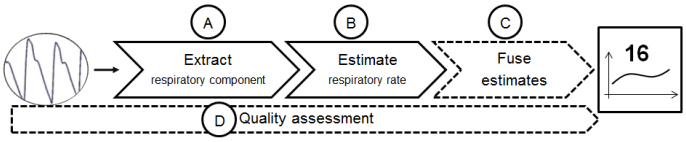


Fig. 1. Typical components of an algorithm for estimating RR (optional components are indicated by dashed lines): (A) a time-series (or multiple time-series) exhibiting respiratory variation is extracted from the PPG signal; (B) RR is estimated from the respiratory signal; (C) multiple RR estimates may be estimated from a single sensor, perhaps by considering multiple respiratory signals; these are then fused to obtain a single estimate; (D) quality assessment may be used to reject or mitigate against noisy estimates.

The first component of most RR-estimation algorithms is the extraction of a respiratory-induced variation signal (or signals) from the PPG. There are a number of methods for extracting the respiratory-induced variation, most of which rely on the identification of the peaks and troughs of the PPG waveform. The segmentation of the PPG into a series of peaks and troughs is a well-described procedure in the literature; for example, peak-trough detection in the time-domain [17] or time-domain segmentation methods [18], [19]. We define the time-series of peaks in the PPG to be a set of pairs $\{t_{\text{pk},i}, y_{\text{pk},i}\}_{i=1\ldots N_{\text{pk}}}$ and the time-series of troughs in the PPG to be a set of pairs $\{t_{\text{tr},i}, y_{\text{tr},i}\}_{i=1\ldots N_{\text{tr}}}$. We note that N_{pk} and N_{tr}, the number of peaks and troughs, respectively, need not be equal. Indeed, it is often the case that peak-trough detection algorithms fail to identify peaks or troughs in noisy signals, or identify spurious peaks or troughs, both of which cause $N_{\text{pk}} \neq N_{\text{tr}}$.

A. Extraction of respiratory components

The time-series of peaks and troughs may be used to derive three new time-series which represent different sources of information related to respiration:

RIIV: The *respiratory-induced intensity variation* is, straightforwardly, the time-series of amplitudes of the PPG peaks. It is believed that this effect is observed due to variations in intrathoracic pressure, leading to a change in the baseline of perfusion, which is shown as a change in the absolute amplitude of the PPG peaks [15], [16]. Therefore, $\mathbf{y}_{\text{RIIV}} = \{t_i, y_{\text{tr},i}\}_{i=1\ldots N_{\text{tr}}}$, where we reserve the use of bold variables for non-scalars.

RIAV: The *respiratory-induced amplitude variation* is the "height" of the PPG pulse, defined to be the difference in amplitude between the corresponding peak and trough, such that $\mathbf{y}_{\text{RIAV}} = \{t_i, y_{\text{RIAV},i}\}_{i=1\ldots N_{\text{RIAV}}}$, with $y_{\text{RIAV},i} = y_{\text{pk},j} - y_{\text{tr},k}$ for j, k defined to be the indices of consecutive peaks and troughs. The time t_i of the i^{th} pair in $\mathbf{y}_{\text{RIAV}}$ is typically set to be the timestamp of the j^{th} PPG peak, s.t. $t_i = t_j$. The RIAV effect is believed to be caused by changes in cardiac output, which have a direct consequence in the quantity of refill in the vessels at the periphery [15], [16].

RIFV: The *respiratory-induced frequency variation* is the change in the value of the instantaneous HR during the respiratory cycle. This phenomenon is known as

respiratory-sinus arrhythmia (RSA), which is regulated by the vagal nerve. The (scaled) instantaneous HR, and therefore the RIFV, may straightforwardly be found by determining the time between successive PPG pulses; i.e., $\mathbf{y}_{\text{RIFV}} = \{t_i, y_{\text{RIFV},i}\}_{i=1\ldots(N_{\text{pk}}-1)}$, and where $y_{\text{RIFV},i} = t_{j+1} - t_j$ for the times t_j in the time-series of peaks. The time of the RIFV signal is typically set to be equal to that of the corresponding PPG peak, s.t. $t_i = t_j$.

Other respiratory-induced variations in the PPG signal have been considered [20]. These rely on the extraction of additional features from the signal, such as the pulse width variability, which has been explored in [21], in order to estimate RR. Many techniques variously explore one or more of the respiratory-induced variations [22]. Digital filters [23], Fourier transforms [16], joint time-frequency analysis [24], auto-regressive modelling [25], [26], wavelet decomposition [27], and Gaussian process [28], [29] methods have all been used.

B. Estimation of Respiratory Rate

In the method proposed by Nilsson et al. [23], a 3rd-order bandpass Butterworth filter with a passband from f = 0.1 to 0.3 Hz (corresponding to 6 to 18 breaths per minute) is used. In the original study, individual breaths were identified manually in the resulting filtered PPG signal.

The use of joint time-frequency analysis methods has been extensively demonstrated [24], [27], [30], [31]. In the method described by Shelley et al. [24], for example, the PPG signal is analysed with a short-time Fourier transform using a moving Hann window of 82s duration. The maximum frequency in the range of plausible respiratory frequencies is identified as that which corresponds to RR. More recently, Garde et al. [30] proposed an algorithm based on the time-varying correntropy spectral density function (CSD) applied to the PPG. The CSD is a generalisation of the power spectral density using correntropy, which is a similarity measure that models time-varying structure and the statistical characteristics of a signal. From applying this method to each window of data, the heart rate is estimated by detecting the maximum frequency peak f_{HR} within the cardiac frequency band, and then removed from the signal (using a zero-phase 5th-order lowpass filter with a cutoff frequency of 0.1 Hz below f_{HR}). The RR is finally estimated by detecting the maximum frequency peak within the respiratory frequency band.

Auto-regressive (AR) modelling has also been used to identify the frequency contained within a respiratory signal [25], [26], [32]. In these methods, the respiratory signal is typically extracted by applying a lowpass (or bandpass) filter that attenuates the component at the cardiac frequency in the PPG signal (as in the method by Nilsson et al. [23]). The poles generated by the ("all-pole" filter) AR model correspond to resonant frequencies, where the frequency is determined by the pole's angle. The respiratory pole (and accompanying frequency) can be identified as the pole with the greatest magnitude within the plausible range of respiratory frequencies [25], [26]. Alternatively, AR modelling can be used to calculate the power spectral density (PSD) and evaluate the frequency content of the respiratory signal [32].

Among the most recently-proposed approaches, Karlen et al. [16] describe a method for estimating the respiratory rate from PPG recordings obtained from pulse oximetry. The three respiratory-induced variations (RIIV, RIAV, RIFV) described above are extracted from the PPG signal using an incremental merge-segmentation algorithm [33], which is also used to identify abnormal pulse periods caused by noise and motion artefacts. The proposed *smart fusion* method analyses the frequency content of each respiratory-induced variation using the Fast Fourier Transform (FFT) and combines the results of the three estimations by taking their mean. Estimations containing artefacts or which are deemed to be of low quality (in which the standard deviation of the three estimations exceeds 4 breaths per minute) are discarded. While the fusion method improved the robustness of the estimation (as demonstrated with a publicly-available benchmark dataset), it also substantially reduced the number of windows for which good-quality estimations were possible.

In this study, we propose an algorithm that combines the results of the three respiratory-induced variations described above, with the goal of providing robust estimates of RR while retaining a larger number of estimations than with the current state-of-the-art.

III. Proposed Algorithm

This section describes the novel algorithm presented by this paper. We first introduce the pre-processing procedure that is applied to the PPG signal prior to data analysis, including a short description of a signal quality metric that is used to identify sections of the PPG waveform that are artefactual (section III-A). Then, we describe the proposed approach to combine the RR estimations from the three respiratory-induced variations based on autoregressive modelling techniques (section III-B).

A. Pre-processing Procedure

The pre-processing procedure involves the use of the three waveforms derived from the PPG: the RIIV, RIAV, and RIFV. PPG beat detection was performed using the segmentation algorithm proposed in [18]. This algorithm was originally developed for identifying peaks and troughs in the arterial blood-pressure waveform, and is directly applicable to the PPG, given the similar morphology of the two waveforms. From the resulting time-series of peaks and troughs, the three derived waveforms were calculated as described in section II-A. The resulting time-series $\mathbf{y}_{\text{RIIV}}$, $\mathbf{y}_{\text{RIAV}}$, and $\mathbf{y}_{\text{RIFV}}$ were then resampled at $f_s = 4$ Hz, using linear interpolation, noting that the originals are unevenly-sampled time-series (see Figure 2). This resampling is performed so that autoregressive modelling may be used straightforwardly. Finally, to have all three signals with the same dynamic range, each resampled time-series is normalised using a zero-mean unit-variance transformation, $\mathbf{y}^* = (\mathbf{y} - \bar{\mathbf{y}})/\sigma\mathbf{y}$, where $\mathbf{y}^*$ corresponds to the normalised time-series $\mathbf{y}$, and $\bar{\mathbf{y}}$ and $\sigma\mathbf{y}$ correspond to the mean and standard deviation of the time-series $\mathbf{y}$, respectively.

In order to identify artefactual and, potentially, low-quality periods of the PPG waveform, a signal quality metric is subsequently used. The metric combines (A) a measure based on "flat-line" detection with (B) a measure determined using the approach described in [34], which evaluates the coincidence of the beats detected by two different PPG peak detectors. For (A), a hysteresis threshold was used to determine the smallest fluctuation that should be ignored; samples with fluctuations ranging below this threshold were set to be flat-lines. For (B), the beats detected by two peak detectors [18], [19] were said to be coincident if they fell within a 150 ms window. The final SQI value for a given window is then determined by SQI $= F1 \times K$, where $F1$ is the F1-score determined for that window (as a measurement of the agreement between the two peak detectors), and K is the proportion of samples in the same window that are *not* flat-lines. We note that this generates a number between 0 and 1, with 0 corresponding to a poor-quality window. A threshold value of 0.9 was used for the analysis described in this paper; i.e., windows of data with SQI < 0.9 were deemed to be of low-quality. This threshold was selected as it guarantees that if either 10% of a certain window corresponds to a flat-line, or there is a small (10%) disagreement between the two beat detectors, the window is discarded, and no RR is estimated for that window.

B. Respiratory Rate Estimation

The proposed method estimates the RR by combining spectral estimates of the three pre-processed outputs (RIIV, RIAV, RIFV) using multiple AR models. The method assumes that each input time-series is governed by an AR process, which in turn assumes that the current value of a time-series y_i may be defined as a linearly-weighted sum of the preceding p values,

$$y_i = \sum_{k=1}^{p} a_k y_{i-k} + e_i \tag{1}$$

where a_k are the weights and where e_i are errors assumed to be distributed $e \sim N(0, \sigma)$. The parameters a_k of the AR process are analogously the coefficients of the IIR filter that transforms the white-noise input e_i into the observed time-series y_i. Hence, an estimate of the spectrum of the time-series y_i can be obtained by factorising the denominator of the IIR filter's transfer function $H(\omega) = [1 - \sum_{k=1}^{p} a_k z^{-k}]^{-1}$, given by the polynomial in a_k [35]. The AR spectral estimate is parsimonious in the number of peaks in the spectrum, which therefore avoids the problems of peak detection that occur with noisier spectral estimates, such as the FFT.

The selection of the value of p is a problem of model-order selection, because the fit of the AR process to data y_i increases as p increases. For odd p, there are $(p-1)/2$ poles γ_k in the spectral estimate in the range of frequencies $[0\ f_s/2]$. If f_s and p are selected appropriately, the dominant pole corresponds to a peak in spectral energy, which yields (in the application described in this paper) the respiratory frequency. We note that this assumes that the dominant frequency component in the input time-series is due to respiration. However, determining the "appropriate" value of p is difficult, and selecting a single model *a priori* order may result in a poor ability to generalise to previously-unseen data. Model-selection techniques based on the asymptotic properties of time-series are sometimes used in such situations; these include, for example, regularisation methods such as the Akaike information criterion (AIC) or the Bayesian information criterion (BIC). Instead of selecting a single model order, we can fuse the results from many models, as suggested in [32]. In this work, we fit a range of AR processes (based on Burg's algorithm) with model orders $p = 2 \ldots 19$ to each of the three pre-processed signals. For each model, we obtain the corresponding estimate of the amplitude spectrum $|H(\omega)|_p$, which we evaluate at a set of N_ω equally-spaced points along the ω-axis between $[0\ \ 2\pi f_s]$, giving a set of pairs $\mathbf{H}_p = \{\omega_i, |H(\omega)|_i\}_{i=1\ldots N_\omega}$ (see Figure 2). We then define the median $\mathbf{H}_m$ of these amplitude spectra $\mathbf{H}_p$ for all three respiratory-induced variation signals to be the median of the $3 \times p$ values at each of the N_ω points on the ω-axis: $\mathbf{H}_m = \{\omega_i, \text{median}(\mathbf{H}_{p,i})\}_{i=1\ldots N_\omega}$. The peak with maximum amplitude in the median spectrum $\mathbf{H}_m$ is taken to correspond to the respiratory frequency.

IV. Materials and Methods

A. Data Collection

For the analysis described in this paper we used two independent, publically-available datasets: the CapnoBase benchmark dataset (available at www.capnobase.org), and a dataset extracted from the MIMIC-II waveform database (v3.0, derived from https://mimic.physionet.org/ and available at http://www.robots.ox.ac.uk/∼davidc).

1) The CapnoBase dataset: collected by Karlen et al. [36], this resource consists of PPG recordings and capnometry data, both recorded at sampling frequency $f_s = 300$ Hz, from 59 children (median age: 8.7, range: 0.8 - 16.5 years) and 35 adults (median age: 52.4, range: 26.2 - 75.6 years). The cases in the dataset were randomly selected by the authors from a larger collection of physiological signals collected during elective surgery and routine anaesthesia. In the work reported in [16], the CapnoBase dataset was divided into a test set consisting of 42 recordings of 8-minute duration (336 minutes in total), from 29 paediatric and 13 adult patients containing reliable recordings of spontaneous or controlled breathing, and a calibration set consisting of 124 recordings of 120 s (248 minutes) from the remaining 52 patients. As in [16], our results are reported using the test dataset of 42 recordings as defined above.

The capnometric waveform for each record was used as the reference "gold standard" recording for RR. Each breath in the capnogram in the database has been manually labelled by a research assistant, and the annotations were used to derive the reference RR values based on the average time between consecutive breaths.

2) The BIDMC Dataset: extracted from the MIMIC-II resource [37], this comprises PPG recordings and respiratory signals acquired using conventional impedance pneumography (IP), both sampled at $f_s = 125$ Hz, from 53 adult patients (median age: 64.81, range: 19-90+, 32 females). Those in the dataset were selected from a larger cohort of patients who

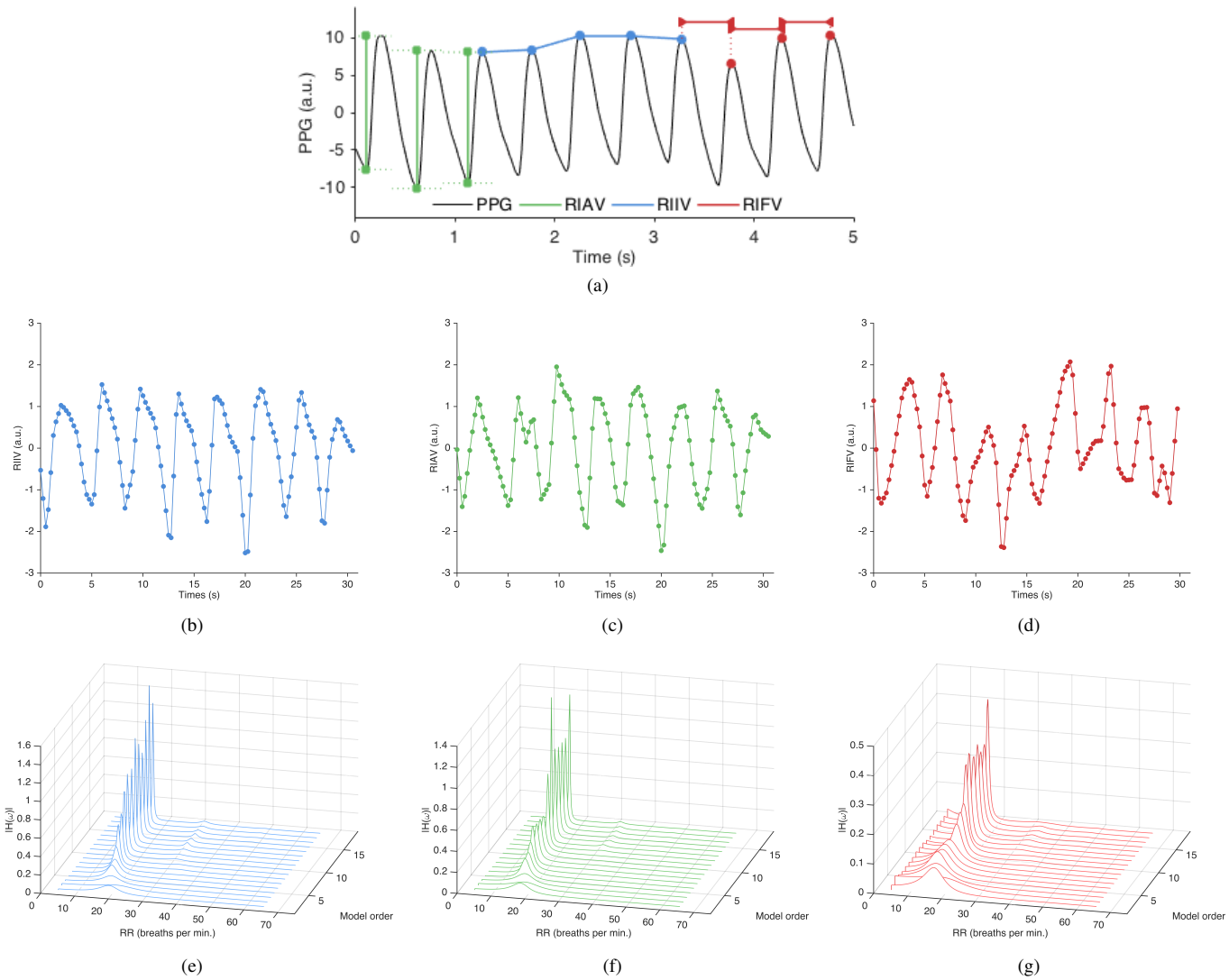


Fig. 2. (a) Representation of the extraction method of the respiratory-induced variations from the PPG. (b)-(c)-(d) Respiratory-induced variations (normalised) extracted from an example 32 s PPG sliding window used for RR estimation: RIIV, RIAV and RIFV. (e)-(f)-(g) Corresponding AR spectra computed for different model orders. We observe a clear peak at around 18 breaths per min. for all respiratory-induced variations. Nevertheless, the peak is more distinct (higher magnitude) for the RIIV and RIAV spectra (e), (f), as the signals from which they are derived (b), (c), appear to be less artefactual than the other signal (d). Therefore, the AR spectra represented in (g) have a "lower weight" in the resulting median spectrum.

were admitted to medical and surgical intensive care units at the Beth Israel Deaconess Medical Center (BIDMC), Boston, USA. 53 recordings of 8-minute duration were randomly selected as the test set for this database.

The IP waveform for each record was used as the reference recording for RR. Each breath in the IP signals was manually (independently) annotated by two research assistants, and both sets of annotations were used to derive the reference RR values. For each set of annotations, the RR value was determined based on the average time between consecutive breaths within a given window; only those windows of data for which the agreement between both estimates was within 2 breaths per min were retained, and the mean value of the two estimates was taken as the reference RR. As a result, using a window size of 32 s, for example, resulted in 97.5% of all available reference windows being deemed to be "valid" according to our criterion.

B. Methods evaluation

We evaluated the performance of our method for two window sizes (32-second and 64-second duration), with successive windows having 29 and 58 seconds overlap; i.e., a new estimate is computed every 3 and 6 seconds, respectively. The window sizes were selected as they did not need zero padding (for frequency-based analysis) and were within reasonable physiological and clinical limits, as discussed in [16]. In this study, RR was estimated within the plausible range of respiratory frequencies set to 4 to 65 breaths per minute. The signal quality metric (SQI) described above was used to identify and discard windows of PPG data that are artefactual. For windows of data in which SQI < 0.90, an estimate of RR was not produced. Therefore, only "valid" windows are retained for the analysis. The value for the SQI was selected heuristically.

The estimated RR values obtained by each method (for each

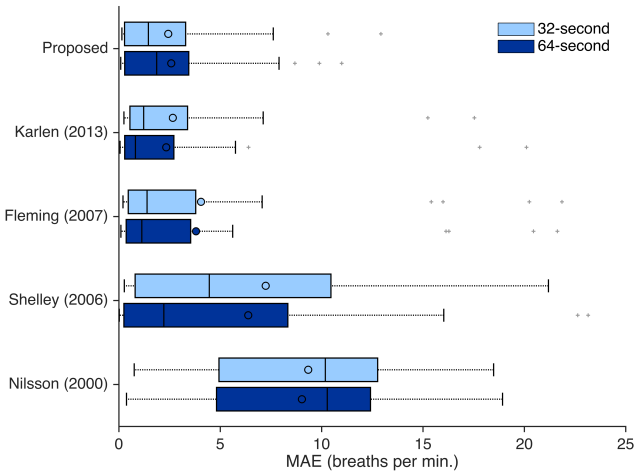
window size) from the PPG recordings were compared with the reference RR obtained from the reference gold standard recordings in each database (as detailed above). Performance was assessed on each dataset by calculating the mean absolute error (MAE) in breaths per minute for each record, defined as

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|\hat{y}_i - y_{\text{ref},i}\right| \quad (2)$$

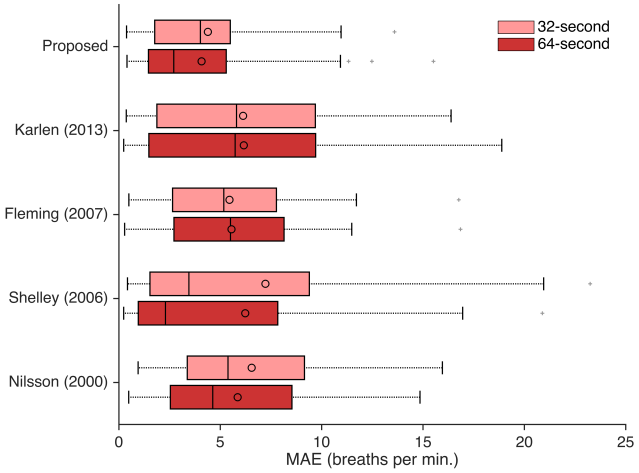
where n is the number of observations, $\hat{y}_i$ is the estimated respiratory rate and $y_{\text{ref},i}$ is the reference respiratory rate for observation i. In addition, we determined the number of valid windows retained by our SQI.

The performance of our method was compared with that of previously-proposed approaches, using both datasets: Karlen (2013) [16], Fleming (2007) [25], Shelley (2006) [24], and Nilsson (2000) [23]. These methods were selected as being representative of the state-of-the-art or as being key studies that we considered "benchmark" investigations. For all methods, RR estimations were obtained only for valid windows as identified by our SQI; i.e., windows of PPG data what were deemed artefactual (windows for which $SQI < 0.9$) were not considered in the comparison of the different methods. We note that in the case of the first method, Karlen (2013), the fusion approach may discard additional RR estimations. For a fair comparison with this approach, we used our own PPG segmentation algorithm for obtaining the respiratory-induced variations, and SQI to discard artefactual windows of PPG data.

A Kruskal-Wallis test, a non-parametric version of the classical one-way analysis of variance (ANOVA) that uses the chi-square statistic (χ^2), was performed on the errors from the different methods for each dataset (and window size), in order to compare the median of the MAEs of the different methods and determine if the distribution of the errors are the same. Results from Bonferroni post-hoc pairwise comparisons (a total of 10 pairwise comparisons, at a significance level of $p = 0.05$) are reported. The 95% confidence intervals for the difference of two medians were determined based on the method described in [38]. We also compared, for each dataset, the error obtained between the two window sizes across the different methods by performing a Kruskal-Wallis group analysis.



(a)



(b)

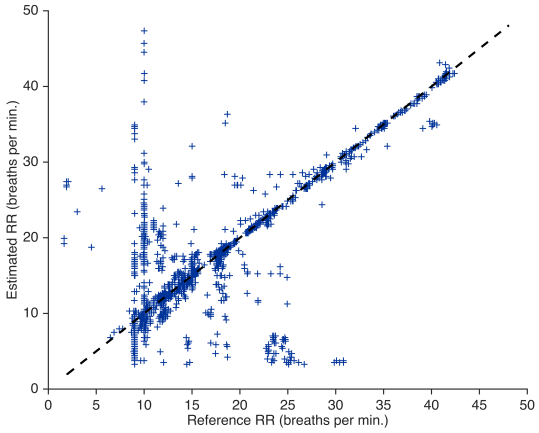
Fig. 3. Comparison of the mean absolute error (MAE) for all methods using (a) the CapnoBase dataset, and (b) the BIDMC dataset, for both window sizes. The boxplot shows distributions of MAEs, with lower quartile, median and upper quartile values displayed as left, middle, and right horizontal lines of the boxes. Whiskers are used to represent the most extreme values within 1.5 times the interquartile range from the central box. Outliers (data with values beyond the ends of the whiskers) are displayed as crosses. Circles represent the mean values.

V. Results

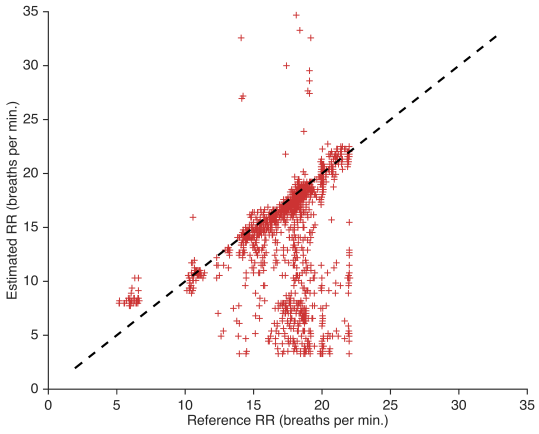
Figure 3 shows the results of applying the various methods to the CapnoBase and BIDMC datasets, showing the distributions of MAEs for window sizes of 32 and 64 secs.

The different window sizes used for estimating RR did not show a substantial difference in error for both CapnoBase ($\chi^2 = 2.61$, $p = 0.106$) and BIDMC ($\chi^2 = 2.33$, $p = 0.127$) datasets. In general, however, a trend for lower error rates when larger windows are used can be observed for all methods (Figure 3). Window sizes from 16 to 128 seconds have been used in previous studies [16], [22], [30]. On the one hand, choosing smaller window sizes would be ideal as they yield shorter computation and processing times at each time step. On the other hand, larger window sizes may improve the accuracy of the estimate and decreases the lowest detectable RR. For example, if we consider that two respiration cycles (two periods) are necessary for obtaining an accurate estimation, the lowest detectable RR for a 32-s window is 3.75 breaths per min [16]. However, it is important to note that most methods for estimating RR assume that a respiratory-induced variation signal is governed by a single, or dominant, frequency, which is expected to correspond to RR. Therefore, choosing larger window sizes may lead to an increased number of "violations" of this assumption, as there is more room for variability in the RR. This may subsequently lead to increased errors. Therefore, our analysis focused on window sizes of 32 and 64 seconds.

The Kruskal-Wallis test showed that there were significant differences between the RR estimation methods, using a win-



(a)



(b)

Fig. 4. Scatter plot comparing the reference RR with the RR estimates from the PPG using the proposed method, using (a) the CapnoBase dataset, and (b) the BIDMC dataset, for 32-second windows.

dow size of 32 secs, for the CapnoBase dataset ($\chi^2 = 46.26$, $p < 0.001$). The post-hoc multicomparison tests according to the Bonferroni method revealed that the performance of the proposed method was significantly different from that of the approaches of Shelley (2006) [24] ($p = 0.016$) and Nilsson (2000) [23] ($p < 0.001$). Comparable MAEs between our proposed method and the approaches by Karlen (2013) [16] ($p = 0.974$) and Fleming (2007) [25] ($p = 0.897$) were observed (see Table I). For the BIDMC dataset, the Kruskall-Wallis test showed that there was no statistical difference between the errors of the methods tested at a level of significance of 5% ($\chi^2 = 8.75$, $p = 0.068$).

Using a window size of 64 secs, the Kruskal-Wallis test showed that there were significant differences between the RR estimation methods for both the CapnoBase ($\chi^2 = 38.12$, $p < 0.001$) and BIDMC ($\chi^2 = 10.87$, $p = 0.028$) datasets. Our method was found to have a significantly different performance than the approach proposed by Nilsson ($p < 0.001$ and $p = 0.041$ for CapnoBase and BIDMC datasets, respectively), but comparable performances to the other methods, using the Bonferroni multicomparison tests.

Figure 4 shows that the majority of windows in the CapnoBase dataset have reference RR values around 10 - 20 breaths per min, with a tail that extends to approximately 45 breaths per min. In the BIDMC dataset, it may be seen that the distribution of RR values falls into the range 5 - 25 breaths per min. We note that, using the CapnoBase dataset, the largest estimation errors of the proposed method occurred for lower RR values. Using the BIDMC dataset, we note a bias on some of the estimation errors; our method underestimated RR for a substantial proportion of windows through the range of 12 to 25 breaths per min for this dataset (Figure 4). This may be caused by low-frequency, non-respiratory processes that were not removed during our procedure for extracting the respiratory-induced variations, and hence, were detected by the multiple spectra approach.

While the approaches by Fleming (2007), Shelley (2006) and Nilsson (2000) take into account a single respiratory-induced variation (which corresponds to the RIIV component), the method described by Karlen et al. [16] and the proposed method described in this paper combine the estimations of three derived respiratory-induced variations using different strategies (as described above). The proposed approach retains more estimations than that by Karlen (2013) in both datasets (Table I). Specifically, we note that using the BIDMC dataset, for a window size of 32 secs, in average, 65% of the windows in each record were discarded using the fusion approach proposed by Karlen et al. in [16], while a substantially reduced number of estimations (6%) were discarded by the approach proposed here.

VI. Discussion

The robust estimation of RR in a number of important healthcare-related applications (e-health, m-health, wellness) is a topic in which a substantial amount of work has been done in the last decade. Given the importance of estimating RR in a robust manner, this paper set out to understand the reasons for why estimation of RR in a clinical setting remains an "unsolved" problem by proposing a novel method for estimating RR from PPG recordings.

The method proposed in this paper combines the RR estimations from the three derived respiratory-induced variations by "fusing" the corresponding AR spectra computed over several model orders, and selecting the dominant frequency of the resulting spectrum as that of the respiratory rate. A single RR estimation from the three sources is, therefore, generated at each time step. An SQI based on the agreement of two different beat detectors (combined with a flat-line detector) is used as a quality indicator of periods of the PPG recording and which discard the RR estimation at each time step if the window of data is deemed to be of "poor" quality.

A different approach for combining the results of three respiratory-induced variations has been proposed by Karlen et al. [16]. This method involves not generating an estimate of RR if the standard deviation of those three estimates exceeds 4 breaths per min; if the three estimates are similar, an estimate of RR is generated using the mean of the available values. As noted by the authors, this approach introduces a major

TABLE I
PERFORMANCE OF RR ESTIMATION METHODS (FOR BOTH WINDOW SIZES): PERCENTAGE OF WINDOWS RETAINED (N) PRESENTED WITH MEAN AND STANDARD DEVIATION (SD) PER RECORD; MEAN ABSOLUTE ERROR (MAE) PRESENTED AS MEDIAN AND INTER-QUARTILE RANGE ($25^{TH} - 75^{TH}$ PERCENTILES); AND MEAN DIFFERENCE BETWEEN THE ERROR OBTAINED WITH EACH METHOD AND THAT OF THE PROPOSED METHOD (95% CONFIDENCE INTERVAL, CI).

	Window size: 32 secs					
	CapnoBase dataset			BIDMC dataset		
Method	N (SD)	MAE	Mean difference [95% CI]	N (SD)	MAE	Mean difference [95% CI]
Proposed	92 (18)	1.5 (0.3 – 3.3)	–	94 (16)	4.0 (1.8 – 5.5)	–
Karlen (2013)	61 (28)	1.2 (0.5 – 3.4)	−0.3 [−1.6 to 1.2]	35 (21)	5.8 (1.9 – 9.7)	1.8 [−0.8 to 4.4]
Fleming (2007)	92 (18)	1.4 (0.5 – 3.8)	−0.1 [−1.4 to 1.3]	94 (16)	5.2 (2.6 – 7.7)	1.2 [−0.6 to 2.9]
Shelley (2016)	92 (18)	4.5 (0.8 – 10.5)	3.0 [0.2 to 5.8]	94 (16)	3.5 (1.5 – 9.4)	−0.5 [−3.2 to 2.1]
Nilsson (2000)	92 (18)	10.5 (4.9 – 12.7)	8.7 [6.9 to 10.6]	94 (16)	5.4 (3.4 – 9.2)	1.4 [−0.2 to 2.9]
	Window size: 64 secs					
Proposed	92 (19)	1.9 (0.3 – 3.4)	–	94 (18)	2.7 (1.5 – 5.3)	–
Karlen (2013)	64 (29)	0.8 (0.3 – 2.7)	−1.1 [−2.2 to 0.2]	34 (24)	5.7 (1.5 – 9.7)	3.0 [0.1 to 6.0]
Fleming (2007)	92 (19)	1.1 (0.4 – 3.5)	−0.7 [−1.9 to 0.5]	94 (18)	5.5 (2.7 – 8.1)	2.8 [0.7 to 4.9]
Shelley (2016)	92 (19)	2.2 (0.2 – 8.3)	0.4 [−3.0 to 3.7]	94 (18)	2.3 (0.9 – 7.9)	−0.4 [−2.1 to 1.3]
Nilsson (2000)	92 (19)	10.2 (4.8 – 12.4)	8.4 [6.5 to 10.3]	94 (18)	4.6 (2.5 – 8.5)	1.9 [0.2 to 3.7]

limitation: "*an RR estimation is only available for periods of data that do not contain artifacts or have an agreement between the three estimations*" [16]. In fact, using the dataset, an average of 36% of the windows in each record were eliminated due to the disagreement of the three estimates (Table I). This effect is exacerbated when the approach is used with the PPG recordings of the BIDMC dataset, where more than 60% of the windows (on average) in each recording were deemed to be low-quality estimations due to this fusion approach. We also note that discarding data due to the agreement of the three respiratory-induced variations may not produce results with an improved performance. While this strategy led to a good performance (compared to that of the proposed method) using CapnoBase recordings, it generated larger estimations errors with the BIDMC dataset (Figure 3, Table I).

With patients who are elderly and/or unwell, or who are undertaking treatments with intake of certain drugs, it is likely that one or more of these three sources may be consistently unrepresentative of respiration, but the estimations from the other respiratory-induced variations may still be considered for generating reliable RR estimations. This effect is most obviously noticed, for example, in the decrease in the RSA phenomenon in the elderly [15], [26], corresponding to non-robust estimates of RR from RIFV. An effective "fusion" approach should be able to cope with one or more of the respiratory signals consistently showing a lack of respiratory-related information. The method proposed in this paper copes with this effect by investigating the magnitude of the spectra corresponding to the three respiratory-induced variations. The magnitude of the spectrum corresponding to a noisy waveform is lower than that corresponding to a less noisier waveform (Figure 2). Therefore, the presence of artifacts in the respiratory-induced variations is encoded in the magnitude of the corresponding AR spectra, which works as a weighting factor for determining the "fused" spectrum, from which the single RR estimation is computed. Another method that combines all three respiratory-induced variations, derived using continuous wavelet decomposition, has been used to estimate RR from the PPG [39]. However, the approach used an undisclosed algorithm, so that it could not be easily reproduced here.

We investigated the use of a new strategy that takes advantage of the notion of "model fusion", such that model complexity may be determined automatically, in an unsupervised manner, leading to more robust estimations of RR based on a "committee" of models of varying complexity (using different model orders). As noted before, the selection of the AR model order may be problematic, as it typically depends

on the data and type of dataset used [25]. We observe that the approach described by Fleming et al. [25] was implemented with a fixed model order of 11, which was seen to achieve the best performance on both datasets from all other model orders and strategies for determining a single best model order (such as the AIC and BIC)[1]. Hence, it is important to note that it may achieve a different performance in a third independent dataset. The proposed method performed well with both datasets (compared to the other methods), suggesting that the approach described in this paper is a promising means of coping with different types of datasets and overcome the problem of model selection.

We also observe that the proposed method underestimates RR for many windows of the BIDMC dataset (Figure 4). This is caused by the presence of low-frequency, non-respiratory processes that our method was not able to distinguish from the true underlying respiratory process. As noted in [40], there are baseline fluctuations in the PPG waveform that are independent of respiration and which are part of a separate vascular response to the sympathetic nervous system. These low-frequency fluctuations ($\sim$0.12 Hz) are often referred to as Mayer waves [41], and are thought to represent the baroreflex mediated oscillation of arterial blood pressure. Additionally, very low frequency ($\sim$0.05 Hz) sympathetically mediated variations in the baseline may also be apparent as a vascular response in the regulation of body temperature [15], [42]. Therefore, an additional step may be required to better select which frequency corresponds to RR by discarding unwanted frequencies. Furthermore, we observe that our method considers all windows from one record to be independent; i.e., the RR estimate obtained at a certain time step does not influence the value of the RR estimate of the subsequent time step. It is straightforward to imagine a set of time-based rules that would avoid abrupt changes in the RR estimations from one window to the next, and, hence, reduce estimation errors. Kalman filtering has been used in order to combine RR values estimated from different sources and produce a smoother time-series of RR estimations [43]. Nevertheless, we designed (and evaluated) this approach for applications in which the duration of the PPG recording is not necessarily very long, which precludes the use of Kalman filtering and time-averaging techniques.

The proposed method comprises the use of two segmentation algorithms and the computation of AR models of different model orders, which may represent a substantial computational load. We note that the method was implemented and tested using the MATLAB software framework, v.R2012b (Mathworks, Natick, MA, USA), and it was designed to be used on each single window (independently of the others). The average processing time for a record of 8 minutes (150 estimations) is 1.6 secs on a single processor thread on a 2.4 GHz PC. The number of AR models to be calculated is proportional to the frequency at which estimations are displayed. We note that the optimisation tools of the AR processes may be modified in order to reduce the computational load of the proposed method without compromising the accuracy of its RR estimates. Also, we note that other, more efficient, segmentation algorithms may be used for extracting the respiratory-induced variation time-series and, hence, reduce the computational load of the proposed method.

Finally, one of the major reasons that we identified for not translating RR estimation algorithms into clinical practice was the lack of large-scale validation studies using datasets that match the conditions under which a system would be used in practice. In this study, data collected from subjects with a wide range of ages, from paediatric to elderly patients (the only significant age cohort omitted was the neonatal cohort), under controlled ventilation or spontaneously breathing were used. As with previous studies [16], [44], no significant differences in the performance of the proposed method between controlled and spontaneous breathing subjects were observed (see Supplementary Material). This study is limited, however, by not including data from patients outside of the hospital setting, and by using data acquired from patients whilst being stationary, rather than truly ambulatory. We used the largest publicly-available database from the literature (CapnoBase), and augmented it with a database derived from a well-understood, and well-investigated repository (Physionet)[2]. There are two major limitations with the latter. The first limitation is the absence of a capnometric waveform, from which reference RR values may be extracted for comparison. Nevertheless, we overcome this problem by using the IP waveform as the respiration source and the annotations performed by two independent research assistants that allowed the extraction of reliable reference RR values to which RR estimations from the PPG may be compared. The second limitation is the patient population and range of RR values. We note that these recordings are solely from adults, and that RR values in this dataset are concentrated in the range between 5 and 25 breaths per min. Therefore, the use of this single dataset may preclude the evaluation of algorithms in the full interval of RR values.

VII. Conclusion

We have presented the development of a novel algorithm for estimating RR from the PPG. The algorithm fuses the estimates of three respiratory-induced variations (RIFV, RIIV, and RIAV) using the corresponding spectra computed using multiple autoregressive AR models of different orders. The method was evaluated in two independent test sets including recordings from pediatric and adult in-hospital patients. Our analysis demonstrated the importance of using alternative datasets for evaluating the performance and generalisation ability of proposed methods. Future studies should concentrate on the use of these (and additional) raw data sources as a benchmark for comparison of new RR estimation approaches.

Data access statement

This manuscript is in compliance with the UK Research Councils Common Principles on Research Data Policy, and the data used in this research are openly available from public sources as described in the text.

[1]The results of this analysis are not included in this manuscript.

[2]The dataset used for this analysis will be made available at http://www.robots.ox.ac.uk/$\sim$davidc

REFERENCES

[1] G. Smith *et al.*, "Review and performance evaluation of aggregate "track and trigger" systems," *Resuscitation*, vol. 77, pp. 170–179, 2008.

[2] C. Alvarez *et al.*, "Predicting out of intensive care unit cardiopulmonary arrest or death using electronic medical record data," *BMC Medical Informatics and Decision Making*, vol. 13, no. 28, 2013.

[3] L. Tarassenko *et al.*, "Centile-based early warning scores derived from statistical distributions of vital signs," *Resuscitation*, vol. 82, no. 8, pp. 1013–1018, 2011.

[4] National Institute for Clinical Excellence, "Guideline CG50 - acutely ill patient in hospital: Recognition of and response to acute illness in adults in hospital," Technical Report, 2007.

[5] Royal College of Physicians, "National early warning scores (NEWS): Standardising the assessment of acute-illness severity in the NHS," Royal College of Physicians, Tech. Rep., 2012.

[6] D. R. Prytherch *et al.*, "ViEWS - towards a national early warning score for detecting adult inpatient deterioration," *Resuscitation*, vol. 81, no. 8, pp. 932–937, 2010.

[7] G. Smith *et al.*, "The ability of the national early warning score (NEWS) to discriminate patients at risk of early cardiac arrest, unanticipated intensive care unit admission, and death," *Resuscitation*, vol. 84, no. 4, pp. 465–470, 2013.

[8] L. Tarassenko and D. Clifton, "Semiconductor wireless technology for chronic disease management," *Electronics Letters*, vol. S30, pp. 30–32, 2011.

[9] G. Clifford and D. Clifton, "Annual review: Wireless technology in disease state management and medicine," *Annual Review of Medicine*, vol. 63, pp. 479–492, 2012.

[10] L. Clifton *et al.*, "Predictive monitoring of mobile patients by combining clinical observations with data from wearable sensors," *IEEE Journal of Biomedical and Health Informatics*, vol. 18, no. 3, pp. 722–730, 2014.

[11] F. Hardinge *et al.*, "Using a mobile health application to support self-management in COPD - development alert thresholds derived from variability in self-reported and measured clinical variables," *American Journal of Respiratory and Critical Care Medicine*, p. A1396, 2014.

[12] V. W *et al.*, "Remote plethysmographic imaging using ambient light," *Optics Express*, vol. 16, no. 26, pp. 21 434–21 445, 2008.

[13] M. Z. Poh *et al.*, "Advancements in noncontact, multiparameter physiological measurements using a webcam," *IEEE Transactions on Biomedical Engineering*, vol. 58, no. 1, pp. 7–11, 2011.

[14] L. Tarassenko *et al.*, "Non-contact video-based vital sign monitoring using ambient light and auto-regressive models," *Physiological Measurement*, vol. 35, pp. 807–831, 2014.

[15] D. Meredith *et al.*, "Photoplethysmographic derivation of respiratory rate: A review of relevant respiratory and circulatory physiology," *Journal of Medical Engineering and Technology*, vol. 36, no. 1, pp. 60–66, 2012.

[16] W. Karlen *et al.*, "Multiparameter respiratory rate estimation from the photoplethysmogram," *IEEE Transactions on Biomedical Engineering*, vol. 60, no. 7, pp. 1946–1953, 2013.

[17] C. Orphanidou *et al.*, "Signal quality indices for the electrocardiogram and photoplethysmogram: Derivation and applications to wireless monitoring DOI 10.1109/JBHI.2014.2338351," *IEEE Journal of Biomedical and Health Informatics*, 2015.

[18] B. N. Li *et al.*, "On an automatic delineator for arterial blood pressure waveforms," *Biomedical Signal Processing and Control*, vol. 5, no. 1, pp. 76–81, 2010.

[19] W. Zong *et al.*, "An open-source algorithm to detect onset of arterial blood pressure pulses," in *Computers in Cardiology, 2003*, Sept 2003, pp. 259–262.

[20] R. A. Cernat *et al.*, "Real-time extraction of the respiratory rate from photoplethysmographic signals using wearable devices," in *Proc. European Conference on Ambient Intelligence, Eindhoven, Netherlands*, 2014, pp. 1–17.

[21] J. Lázaro *et al.*, "Deriving respiration from photoplethysmographic pulse width," *Medical & Biological Engineering & Computing*, vol. 51, no. 1, pp. 233–242, 2013.

[22] P. H. Charlton *et al.*, "An assessment of algorithms to estimate respiratory rate from the electrocardiogram and photoplethysmogram," *Physiological Measurement*, 2016.

[23] L. Nilsson *et al.*, "Monitoring of respiratory rate in postoperative care using a new photoplethysmographic technique," *Journal of Clinical Monitoring and Computing*, vol. 16, no. 4, pp. 309–315, 2000.

[24] K. H. Shelley *et al.*, "The use of joint time frequency analysis to quantify the effect of ventilation on the pulse oximeter waveform," *Journal of Clinical Monitoring and Computing*, vol. 20, no. 2, pp. 81–87, 2006.

[25] S. G. Fleming and L. Tarassenko, "A comparison of signal processing techniques for the extraction of breathing rate from the photoplethysmogram," *International Journal of Biological and Medical Sciences*, vol. 2, no. 4, pp. 232–236, 2007.

[26] C. Orphanidou *et al.*, "Data fusion for estimating respiratory rate from a single-lead ECG," *Biomedical Signal Processing and Control*, vol. 8, no. 1, pp. 98–105, 2013.

[27] P. Leonard *et al.*, "An algorithm for the detection of individual breaths from the pulse oximeter waveform," *Journal of clinical monitoring and computing*, vol. 18, no. 5-6, pp. 309–312, 2004.

[28] M. Pimentel *et al.*, "Probabilistic estimation of respiratory rate using gaussian processes," in *Conference proceedings: Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Jul. 2013, pp. 2902–2905.

[29] M. A. F. Pimentel *et al.*, "Probabilistic estimation of respiratory rate from wearable sensors," in *Wearable Electronics Sensors*. Springer, 2015, pp. 241–262.

[30] A. Garde *et al.*, "Estimating respiratory and heart rates from the correntropy spectral density of the photoplethysmogram," *PloS one*, vol. 9, no. 1, p. e86427, 2014.

[31] P. Leonard *et al.*, "Standard pulse oximeters can be used to monitor respiratory rate," *Emergency Medicine Journal*, vol. 20, no. 6, pp. 524–525, 2003.

[32] S. A. Shah *et al.*, "Respiratory rate estimation during triage of children in hospitals," *Journal of Medical Engineering & Technology*, vol. 39, no. 8, pp. 514–524, 2015, pMID: 26548638.

[33] W. Karlen *et al.*, "Adaptive pulse segmentation and artifact detection in photoplethysmography for mobile applications," in *IEEE Engineering in Medicine and Biology Society*, 2012, pp. 3131–3134.

[34] M. A. F. Pimentel *et al.*, "Heart beat detection in multimodal physiological data using a hidden semi-markov model and signal quality indices," *Physiological Measurement*, vol. 36, no. 8, p. 1717, 2015.

[35] R. Takolo *et al.*, "Tutorial on univariate autoregressive spectral analysis," *Journal of Clinical Monitoring and Computing*, vol. 19, pp. 401–410, 2005.

[36] W. Karlen *et al.*, "Capnobase: Signal database and tools to collect, share and annotate respiratory signals," in *Annual Meeting of the Society for Technology in Anesthesia*, 2010.

[37] M. Saeed *et al.*, "Multiparameter intelligent monitoring in intensive care II (MIMIC-II): A public-access intensive care unit database," *Critical Care Medicine*, vol. 39, pp. 952–960, May 2011.

[38] D. G. Bonett and R. M. Price, "Statistical inference for a linear function of medians: confidence intervals, hypothesis testing, and sample size requirements." *Psychological methods*, vol. 7, no. 3, p. 370, 2002.

[39] P. S. Addison *et al.*, "Developing an algorithm for pulse oximetry derived respiratory rate (rr_{oxi}): A healthy volunteer study," *Journal of Clinical Monitoring and Computing*, vol. 26, no. 1, pp. 45–51, 2012.

[40] B. Khanoka *et al.*, "Sympathetically induced spontaneous fluctuations of the photoplethysmographic signal," *Medical and Biological Engineering and Computing*, vol. 42, no. 1, pp. 80–85, 2004.

[41] C. Julien, "The enigma of mayer waves: facts and models," *Cardiovascular research*, vol. 70, no. 1, pp. 12–21, 2006.

[42] P. D. Larsen *et al.*, "Spectral analysis of ac and dc components of the pulse photoplethysmograph at rest and during induction of anaesthesia," *International journal of clinical monitoring and computing*, vol. 14, no. 2, pp. 89–95, 1997.

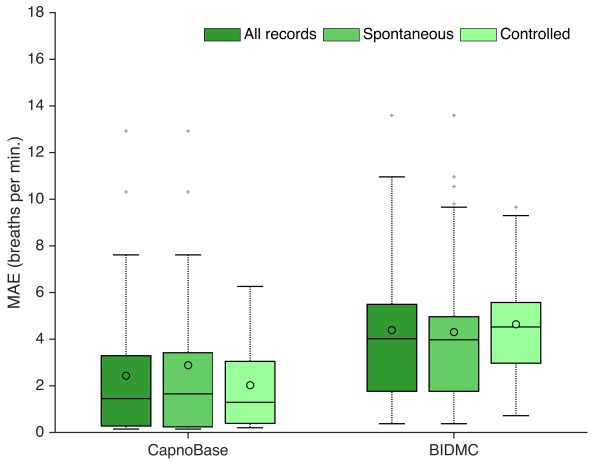
[43] S. Nemati *et al.*, "Data fusion for improved respiration rate estimation," *EURASIP J. Adv. Signal Process*, vol. 2010, pp. 10:1–10:10, Feb. 2010.

[44] L. Nilsson *et al.*, "Respiration can be monitored by photoplethysmography with high sensitivity and specificity regardless of anaesthesia and ventilatory mode," *Acta anaesthesiologica scandinavica*, vol. 49, no. 8, pp. 1157–1162, 2005.

SUPPLEMENTARY MATERIAL

The performance of the proposed method was compared between subjects under controlled ventilation and subjects spontaneously breathing. There are 22 (52%) records corresponding to subjects under controlled breathing in CapnoBase, and 14 (26%) records corresponding to subjects that were under mechanical ventilation (during the recording) in BIDMC. The information for latter was extracted using the MIMIC-II clinical database [37]. Figure 5 shows the results of applying the proposed method to the CapnoBase and BIDMC datasets

showing the distribution of MAEs for a window size of 32 secs.



(a)

Fig. 5. Results for both datasets split up into controlled and spontaneous breathing. The boxplots show the MAE for the proposed method using a window size of 32 secs.