IOP Publishing | Institute of Physics and Engineering in Medicine Physiological Measurement

Physiol. Meas. **35** (2014) 2543–2561 doi:10.1088/0967-3334/35/12/2543

# A combined segmenting and non-segmenting approach to signal quality estimation for ambulatory photoplethysmography

**J D Wander**[1,2] **and D Morris**[1]

[1] Microsoft Research, One Microsoft way, Redmond, WA 98052, USA
[2] Department of Bioengineering, University of Washington, 3720 15th Ave NE, Seattle, WA 98105, USA

E-mail: jdwander@uw.edu

Received 17 February 2014, revised 31 July 2014
Accepted for publication 8 August 2014
Published 19 November 2014

**Abstract**

Continuous cardiac monitoring of healthy and unhealthy patients can help us understand the progression of heart disease and enable early treatment. Optical pulse sensing is an excellent candidate for continuous mobile monitoring of cardiovascular health indicators, but optical pulse signals are susceptible to corruption from a number of noise sources, including motion artifact. Therefore, before higher-level health indicators can be reliably computed, corrupted data must be separated from valid data. This is an especially difficult task in the presence of artifact caused by ambulation (e.g. walking or jogging), which shares significant spectral energy with the true pulsatile signal. In this manuscript, we present a machine-learning-based system for automated estimation of signal quality of optical pulse signals that performs well in the presence of periodic artifact. We hypothesized that signal processing methods that identified individual heart beats (*segmenting* approaches) would be more error-prone than methods that did not (*non-segmenting* approaches) when applied to data contaminated by periodic artifact. We further hypothesized that a fusion of segmenting and non-segmenting approaches would outperform either approach alone. Therefore, we developed a novel *non-segmenting* approach to signal quality estimation that we then utilized in combination with a traditional *segmenting* approach. Using this system we were able to robustly detect differences in signal quality as labeled by expert human raters (Pearson's $r = 0.9263$). We then validated our original hypotheses by demonstrating that our *non-segmenting* approach outperformed the *segmenting* approach in the presence of contaminated signal, and that the combined system outperformed either individually. Lastly,

  

as an example, we demonstrated the utility of our signal quality estimation system in evaluating the trustworthiness of heart rate measurements derived from optical pulse signals.

Keywords: photoplethysmography, machine learning, signal quality index, heart rate, digital signal processing, pulse oximeter

(Some figures may appear in colour only in the online journal)

## 1. Introduction

Cardiovascular disease is one of the leading causes of mortality in both the United States (Hoyert *et al* 2012) and the European Union (Eurostat 2009). In the U.S. it accounts for over a half-million deaths and nearly a half-trillion dollars annually (CDC 2012). As physical activity (along with diet management) can dramatically reduce the risk of most cardiovascular conditions (CDC 2013), current guidelines from European health authorities stress prevention of these conditions before they become clinical emergencies (Mancia *et al* 2013). Monitoring of heart rate (HR) during and after exercise can allow for measurement of caloric expenditure (Ceesay *et al* 1989) and recovery rate (Cole *et al* 1999), and a number of cardiac rehabilitation protocols require consistent, dependable monitoring of heart-rate during exercise (Taylor *et al* 2004, Leon *et al* 2005). Most importantly, self-monitoring of health data such as HR is positively associated with improved healthy behavior (Fox and Duggan 2013). In addition to the individual health benefits, crowd-sourced collection of data from commercially available HR monitors presents a tremendous opportunity for epidemiologic study of risk factors, early symptoms, and progression of cardiovascular disease.

The current consumer standard for HR monitoring is the two-electrode chest strap, which is commonly used among fitness enthusiasts and elite athletes seeking to optimize training, but is too cumbersome and uncomfortable for full-time use. Recently, optical HR measurement at the wrist has emerged as an alternative. This technique leverages the fact that blood pulsing through the wrist changes the way light is absorbed by the tissue. Several commercial devices (MIO Alpha, *mioglobal.com*; Basis health tracker, *mybasis.com*) utilize this technology to provide continuous HR measurement. However, the accuracy of this technology relative to the gold standard of HR monitoring—the electrocardiogram—is still being evaluated (Schäfer and Vagedes 2013); ambient light and non-pulse-related blood movement tremendously distort the signal (Asada *et al* 2003). This is particularly true during motion (e.g. exercise), when artifact can make the pulse nearly indistinguishable from noise (Rhee *et al* 2001).

Use of optical pulse measurement—or the photoplethysmograph (PPG)—has until recently been limited to clinical environments where patients are generally sedentary, and trained personnel are on hand to manually determine the quality of the signals being recorded and adjust the recording equipment if necessary. This model of expert equipment and data management is unfortunately not scalable beyond the current application of PPG-based cardiac monitoring in staffed care centers, an obstacle that must be overcome for the development of ubiquitous PPG-based cardiac health monitoring. The high volumes of data generated in such a model cannot feasibly be hand-annotated for usable and unusable data, so a dependable measure of signal quality—a signal quality index (SQI)—is essential for assessing the trustworthiness of derived cardiac health metrics.

Furthermore, the nature of disturbances to the PPG signal changes dramatically when the devices are worn by mobile users. Repetitive motion artifact generated by ambulation, changes

in ambient light conditions, and environmental noise can all drastically impact quality of the recorded PPG. Computationally-determined SQIs have shown excellent promise in identifying contaminated periods in clinical PPG, but the majority of these approaches require the successful segmentation of the optical signal into individual beats (Weng *et al* 2005, Farooq *et al* 2010, Karlen *et al* 2012a 2012b). Such approaches may be confounded when the artifact to be identified is similar (quasi-periodic with a similar fundamental frequency) to the pulsatile signal of interest.

The primary aim of the work described below was to develop a system capable of successfully performing signal quality estimation in ambulatory settings. We hypothesized that such a system would benefit from employing both segmenting and non-segmenting approaches to signal quality estimation. Additionally, we wanted to understand which predictors of signal quality were most robust under ambulatory circumstances. Lastly, we wanted to validate our signal quality estimation algorithm for use in a common application: HR estimation.

## 2. Background and previous work

### *2.1. Photoplethysmography*

The term photoplethysmography (PPG) spans two classes of measurement devices—*transmissive* and *reflective*—both of which measure the quantity of light from one or more light emitting diodes that is not absorbed by the tissue and fluid through which the light passes. Transmissive PPG measures the light transmitted *through* a part of the body—typically a finger, toe, or earlobe—to a sensor on the other side of that tissue. Reflective PPG uses a photosensor that is co-located with the light source(s), and measures the light that is reflected back toward the source. Though transmissive PPG is much more common in clinical practice, all wrist-worn consumer-grade PPG-based HR monitors use the reflective approach. From a signal processing perspective, transmissive and reflective PPG are quite similar, though it has been observed that reflective PPG is more susceptible to motion artifact (Asada *et al* 2003).

The quantity of non-absorbed light that contributes to the PPG signal depends on the travel path, the optical density of the tissue, the volume of blood in the tissue, and the composition of the blood (Mannheimer 2007). Assuming travel path and optical tissue density to be constant, the PPG waveform can then be used to extract signals like HR and saturation of peripheral oxygen ($SpO_2$), as well as further derivative signals such as heart rate variability (HRV), respiratory rate (Nilsson *et al* 2000), and arterial wall stiffness (Smith *et al* 1999).

### *2.2. Artifact in the PPG signal*

PPG is susceptible to contamination from multiple sources. If the device is not being worn correctly, or there is poor physical contact between the photosensor and the wearer's tissue, ambient light will contaminate the PPG and in some cases saturate the sensor. Additionally, even if the interface between the photosensor and tissue is good, fluid flow in the tissue associated with movement or pressure changes (i.e. not associated with the heart's pumping of blood) will also change the observed PPG.

A number of studies have leveraged correlations between motion and PPG contaminants in an effort to mitigate artifact (Asada *et al* 2004, Gibbs *et al* 2005, Wood and Asada 2006). It is worth noting, however, that the nature of this motion and corresponding PPG artifact can vary greatly between clinical and ambulatory use cases. In the former, signal contaminants have traditionally been thought of as discrete artifact events, such as a single motion from

the wearer or removal of the sensor (Silva *et al* 2012, Karlen *et al* 2012b). However, during ambulation and exercise, signal contamination must be thought of as continuous and periodic, with significant spectral energy in the same frequencies as the physiological signals of interest ($0.5\,\text{Hz} < f < 3\,\text{Hz}$).

### *2.3. Signal quality estimation*

Because the PPG is subject to contamination, a number of algorithms have been suggested to either detect reduction in signal quality or to reduce the presence of these contaminants (referred to as *signal quality estimation* and *signal conditioning*, respectively). This manuscript focuses primarily on the problem of signal quality estimation. The majority of signal quality estimation algorithms rely on the quasi-periodic nature of the PPG signal to allow for *segmentation* of the time series into individual beats (Weng *et al* 2005, Farooq *et al* 2010, Karlen *et al* 2012a 2012b). Many of these algorithms use the derivative of the PPG (dPPG) for beat segmentation (Weng *et al* 2005, Farooq *et al* 2010). Assuming that beats can be successfully segmented, there have been a number of approaches to signal quality estimation that leverage the fact that beat morphology is fairly consistent over short periods (Weng *et al* 2005, Sukor *et al* 2011, Karlen *et al* 2012b). Recognizing that there are multiple different signal contaminants, and that different predictors may be specific to as few as one of these contaminants, there has been a valuable effort to fuse multiple signal quality predictors in to a single quality metric (Clifford *et al* 2012, Li and Clifford 2012).

One limitation of the above-mentioned signal quality estimation approaches for the purposes of consumer HR monitoring is that they typically utilize either clinical data from one or more publicly available databases or data collected in short segments from stationary users. In some cases, researchers developed devices that could be used during ambulatory behavior (Asada *et al* 2003, Karlen *et al* 2013), but did not stress-test signal quality estimation algorithms by providing them with a wide variety of artifact types. The nature of signal contamination in all of these datasets was discrete and aperiodic in nature, a condition that may be met in clinical settings, but does not represent the types of contamination observed during ambulatory use.

## 3. Methods

We employed a supervised learning approach to training of our algorithm, thus it was necessary to collect optical HR data and manually label these data for time periods where quality was good/poor. Manual quality ratings (MQRs) were then used as ground-truth labels for subsequent ML methods. Additionally, accelerometry data and chest-strap HR were collected for comparative analyses.

### *3.1. Data collection*

Data were gathered from 11 subjects (3 female) with no known history of cardiovascular illness. The recordings obtained were as follows: (1) two channels of reflective PPG recorded from the lateral surface of the left wrist using custom hardware, (2) 3-axis accelerometry (ADXL 327EB, Analog Devices Inc., Norwood, MA) recorded from the same location, and (3) inter-beat intervals and HR from a chest-strap HR monitor (HRM3, Garmin Ltd, Olathe, KS). PPG and accelerometer data were sampled at 1000 Hz (NI-9206, National Instruments Corp., Austin, TX). All behavioral cueing and data collection were performed using custom

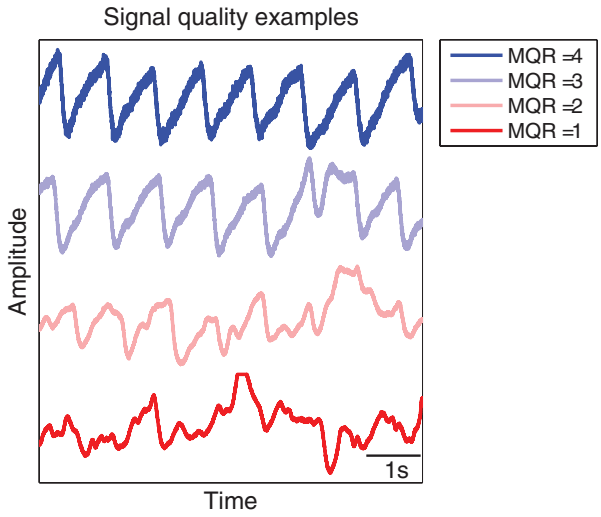


**Figure 1.** Examples of the four classes of manually-rated signal quality. Note the pulse waveform is entirely preserved when MQR = 4 and is preserved to lessening degrees as signal quality decreases.

C# and MATLAB software (MathWorks Inc., Natick, MA). Chest-strap data from a single subject were corrupted, resulting in the exclusion of that subject from the HR application portion of the analyses below.

### *3.2. Experimental protocol*

While data were being recorded, subjects performed three behavioral tasks: standing, walking in place, and jogging in place. Subjects were informed of the current behavioral task on a computer monitor placed in front of and approximately 50 cm away. Transitions in behavior were cued with an audible tone. The recording session began with 30 s of standing still. Then subjects performed three blocks of the three behaviors, randomized within-block. Each behavior lasted for 80 s, resulting in a total session time of 12.5 min.

### *3.3. Manual quality annotation*

Data from each channel were divided into seven-second windows, randomized and presented to expert raters in conjunction with accelerometer data from the same time period. Raters were instructed to rate the signal quality on a scale from one to four, using the following guidelines: (4) excellent signal quality, all beats can be easily visually identified; (3) good signal quality, fewer than all but more than half of beats can be visually identified; (2) mediocre signal quality, fewer than half but more than one beat can be visually identified; (1) poor signal quality, one or fewer beats can be visually identified. Accelerometer measurements were included in the rating process to lessen the likelihood that manual raters mistook periodic artifact for pulse. Figure 1 gives examples of data windows corresponding to the four manual quality ratings (MQR). Data were manually rated by three experts, two of whom only rated every other window, resulting in a total of two ratings for each window. Final MQR was taken as the mean

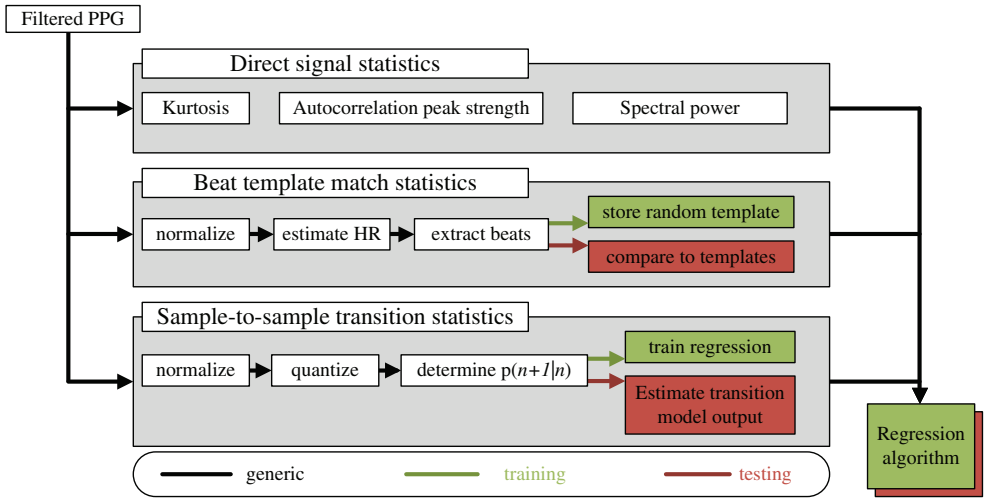


**Figure 2.** Overview of system architecture. After pre-processing, a number of signal statistics or features are derived from each data window of PPG. The majority of these features are used directly by the final regression model, with the exception of the sample-to-sample transition statistics, which first become consolidated in to a single feature using a support-vector regression model.

the two ratings for each window, resulting in seven possible quality ratings for each window (1, 1.5, 2, 2.5, 3, 3.5, or 4).

The goal of our feature extraction and supervised machine learning will be to predict these quality labels on new data windows.

### *3.4. Algorithm description*

We sought to design an algorithm that could perform robustly in the presence of the large motion artifact present in PPG when subjects were performing the behavioral tasks listed above. To perform well in these varied use conditions, the algorithm utilizes a hybrid approach to signal quality estimation, incorporating information from a variety of independent predictors, all used as features in a supervised learning approach to deriving an SQI. The features are organized in to three subsets: *direct signal statistics*, *beat template match statistics*, and *sample-to-sample transition statistics*, each of which is discussed separately below. Figure 2 outlines the algorithm architecture.

Raw PPG signals were band-passed with a zero-phase digital filter (4th order Butterworth, $f_{LP} = 0.5$ Hz, $f_{HP} = 50$ Hz).

Both filtered PPG signals and accelerometer signals were divided into seven-second windows, with no overlap. Subsequent signal quality estimation was performed and evaluated on these windows.

For each seven-second window of 3-axis accelerometer data, an average accelerometer power was calculated by removing the offset from each channel, full-wave rectifying all samples, and taking the average across all samples and the three channels. This value was then log-transformed to be quasi-normally distributed and to allow for use of standard statistical methods. From this point forward, the value resulting from these operations is referred to as *log accelerometer power*.

The following sections provide detailed feature extraction methods for the three feature subsets: *direct signal statistics*, *beat template match statistics*, and *sample-to-sample transition statistics*.

*3.4.1. Direct signal statistics.* We extracted three types of direct statistics from each 7 s window: kurtosis, autocorrelation peak strength, and spectral power.

*Kurtosis* of pulsatile signals has been used previously as a predictor of signal quality of the ECG waveform (Li *et al* 2008), leveraging the tendency of uncorrelated (i.e. noisy) data to be Gaussian distributed. For each seven second window of PPG data ($x = [x_1, x_2, \ldots x_M]$ ), we calculated the estimate of *kurtosis* ($\widehat{K}$) of that window as follows:

$$\hat{k} = \frac{1}{M}\sum_{i=1}^{M}\left[\frac{\mathrm{x}_i - \hat{\mu}}{\hat{\sigma}}\right]^4 \tag{1}$$

Where $M$ is the number of samples in the data window and $\hat{\mu}$ and $\hat{\sigma}$ represent the sample mean and sample standard deviation respectively. Finally, *kurtosis* values for each window were then log-transformed to de-emphasize the impact of outliers in our linear models.

Clean PPG signals are quasi-periodic and thus strongly self-correlated at lags related to their periodicity. Correspondingly, the second signal statistic we derived from each window was a feature from the autocorrelation function, which can be calculated with the following equation:

$$R_{xx}(m) = \sum\nolimits_n \mathrm{x}_n \mathrm{x}^*_{n-m} \tag{2}$$

This result can be simplified to not include the complex conjugate as $x$ is real-valued, and subsequently normalized such that $R_{xx}(0) = 1$. Auto-correlations were calculated at lags from 0 to 3000 samples (0 to 3 s). We then smoothed the resultant autocorrelation function with a rectangular filter that was 50 samples (0.05 s) wide. From the smoothed autocorrelation function, we then calculated peak-to-trough distances for all peaks not located at a lag of zero. In cases where peaks had troughs on either side, trough height was calculated as the average of the two troughs. The largest of these peak-to-trough distances was reserved as a feature that will subsequently be referred to as *autocorrelation peak strength*.

Especially in the case of electromagnetic (EM) interference, spectral characteristics of the PPG signal are correlated with signal quality. Thus the third set of features we extracted was the *spectral power* at a number of frequencies (1, 3, 5, 7, 9, 13, 17, 21, 25, and 29 Hz). These were estimated using Welch's method of spectral density estimation with a window width of 1024 samples (1.024 s) with overlap of 512 samples (0.512 s).

*3.4.2. Beat template-match statistics.* Individual beats are highly stereotyped in the PPG signal. Morphological characteristics of the heartbeat waveform can be learned and potentially used to differentiate high-quality PPG signal from periodic artifact (e.g. artifact associated with repetitive movements like walking) by comparing a test signal to a dictionary of previously observed high-quality beats (Karlen *et al* 2012b). To perform this comparison it is first necessary to attempt to segment the continuous PPG waveform into individual beats. We achieved this using an adaptation of the method of repeated Gaussian filters proposed by Karlen *et al* (Karlen *et al* 2012b), wherein HR frequency and phase candidates for a given PPG window are evaluated by constructing windows of repeated Gaussian waveforms corresponding to each combination of phase and frequency and correlating those windows with a transformed version of the dPPG. The frequency/phase combination that produces the highest correlation with the dPPG is subsequently used for beat segmentation. The method employed in this manuscript is similar, with two differences. First, instead of using the

discrete cosine transform (DCT) for HR estimation, we extract potential HRs by finding peaks in the autocorrelation of the first derivative of the PPG waveform. This allows for finer resolution of HR than methods based on the FFT or DCT. Second, when selecting the estimated HR for a given window, any candidate HRs that are determined to be factors of other candidate HRs are penalized. This penalty lowers the overall selection score for these candidate heart rates by 50%.

During the process of beat segmentation, three signal quality features are extracted. First, the *autocorrelation HR score* is calculated as the maximum peak-to-nearest-trough difference in the autocorrelation function of the data window. Second, the *Gaussian correlation score* is the correlation coefficient of the repeated Gaussian filters with the dPPG. Lastly, the *HR estimate score*, which reflects the accuracy of the HR estimate, is calculated as the product of the *autocorrelation HR score*, the *Gaussian correlation score*, and the candidate heart rate penalty (if applicable).

During initial training of the beat template-match algorithm, a template dictionary was populated with 20 templates representing high-quality pulse waveforms that were randomly selected from training data such that all subjects in the training set contributed an equal number of templates to the dictionary. Only data windows taken when the subjects were standing and with MQR of 4 contributed seed templates to the template dictionary.

With the seed dictionary populated, quality features for all data windows were extracted by correlating each beat within a window to all 20 templates. To length-match test beats with templates from the dictionary, the two were aligned at peaks in their respective derivatives and the longer beat was truncated to be the length of the shorter. The final four quality features derived from this process were the *mean beat correlation*, *median beat correlation*, *minimum beat correlation*, and *maximum beat correlation*.

Pulse wave morphology can differ from subject to subject. To account for this possibility, and improve the predictive capability of the template matching model, we allowed the template dictionary to converge to a user-specific pulse morphology in an unsupervised fashion. On a subject-by-subject basis, initial seed templates were incrementally replaced by subject specific templates that had sufficiently high correlation scores. The correlation requirement for beat replacement was increased from an initial threshold of 0.89 to a final threshold of 0.94 in increments of 0.0025 with each subsequent dictionary replacement. Once all 20 templates had been replaced by user-specific templates, this updating process was halted.

*3.4.3. Sample-to-sample transition statistics.* A major susceptibility of the template matching approach is that it relies on its ability to successfully segment the continuous PPG waveform in to individual beats. In an effort to allow for signal quality estimation in cases where beat segmentation was not reliable, we developed a novel, non-segmenting approach that utilizes the quasi-periodic, highly stereotyped nature of the PPG waveform without the need for beat segmentation.

For each filtered PPG data window, we downsampled from 1000 to 15 Hz, normalized the window to be on the interval [0, 1], and quantized the signal to a bit depth of 4. This means that any given sample within the window could then take on 1 of 16 discrete values. Next, using all samples within the data window, we determined the transition matrix from sample $n$ to sample $n+1$ which was then normalized to become the probability distribution $p(x_{n+1} \mid x_n)$. As can be seen in figure 3, an example transition matrix for high-quality signal is qualitatively quite different from the transition matrix for low-quality signal.

Because of the high dimensionality of this feature subset (256 features) relative to the other features extracted from each data window, we employed a late fusion approach and trained a linear support vector machine (SVM) for regression (SVR) on these transition-statistics

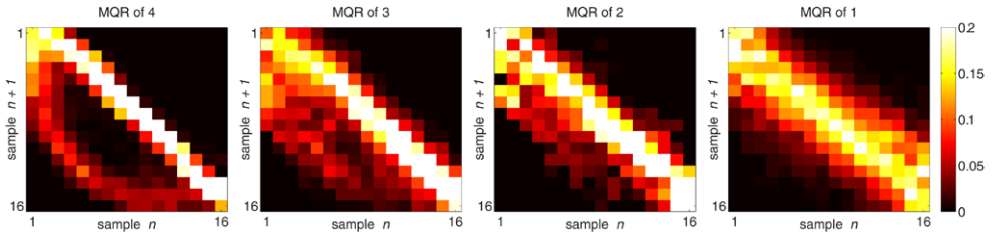


**Figure 3.** Example transition probability distribution functions for various MQRs. Warm colors represent higher probabilities. Color scaling stops at a maximum probability of $p = 0.20$ to enhance off-diagonal components.

features using MQRs as labels. This SVM was implemented using the LIBLINEAR library (http://www.csie.ntu.edu.tw/~cjlin/liblinear/) and utilized L2-loss regularization. The prediction output of the SVR was then fused with the remaining feature set for signal quality prediction (section 3.5). In each fold, the SVR was trained on training data alone, but features were extracted for all training and test data to allow for training of the 2nd-stage ML algorithm.

### *3.5. Signal quality prediction*

We trained a classification and regression tree (CART; (Breiman 1993)), using the features derived above, to predict quality labels on the same four-point scale used for manual quality labeling.

Model training and testing were performed using leave-one-out cross validation. In each fold an entire subject's data trace was withheld from the training set for use as test data. This resulted in 11 folds total. To encourage model sparseness and reduce the chance of overfitting, the minimum number of observations per tree leaf was set to 30. For error analyses, the test results from each fold were combined to allow for assessment of the algorithm's performance across all folds.

A quality prediction was generated for every seven-second window in the test recording for each fold.

### *3.6. Heart rate estimation*

In the example application of our signal quality estimation framework, we estimate HR for individual data windows. This is done to allow for comparison of error in HR estimation as a function of SQI, to demonstrate the utility of such a measure.

There are myriad methods for heart rate estimation, and comparison of these methods is not the focus of this manuscript. Thus for the purposes of this comparison, we extract potential HRs using the method described above in section 3.4.2. The HR chosen for each window was the beat frequency with the highest overall *HR estimate score*.

### *3.7. Statistical comparisons*

In section 4.2, on a subject-by-subject basis, log accelerometer power and MQR values were averaged across all observations for a given behavior to remove statistical bias due to multiple correlated observations. Significance of an overall interaction between behavioral state and log accelerometer power, as well as between behavioral state and MQR were evaluated with

a one-way, repeated measures ANOVA. For these same two outcome measures, individual relationships between pairs of behavioral states were evaluated using two-sided paired-sample student's *t*-tests.

Similarly, in section 4.4, statistical assessment of consistency of SQI prediction error across folds was performed using a one-way ANOVA.

In section 4.5, comparison of segmenting and non-segmenting approaches for low-MQR windows, as well as comparison of segmenting, non-segmenting and combined approaches for all windows was performed using two-sided paired-sample student's *t*-tests.

Lastly, in section 4.6, statistical comparison of error in HR estimation as a function of HR stratum was performed on a pairwise basis (six possible pairs of strata) using two-sided Wilcoxon rank-sum tests that were subsequently Bonferroni corrected for multiple comparisons.

## 4. Results

The results section is organized as follows:

- Sections 4.1 and 4.2 present a brief summary of the data collected as well as a confirmation that subjects were performing the behavioral task as instructed. This serves to verify the impact of ambulation on MQR.
- Section 4.3 provides an assessment of the predictive capability of individual features with respect to MQR, demonstrating specifically that, for our ambulatory dataset, our non-segmenting approach correlates with MQR better than the other features that were evaluated.
- Sections 4.4 and 4.5 present an cross-validated assessment of our ML-based signal quality estimation framework, with a specific focus on SQI error as a function of both MQR and artifact periodicity. The latter comparison serves to demonstrate the relative contributions of the non-segmenting and segmenting approaches in the presence of periodic artifact.
- Lastly, section 4.6 provides an example application of the utility of our framework in the task of HR-estimation.

### *4.1. Data summary*

After discarding data windows that bridged behavioral states, the mean (± standard deviation) number of seven-second data windows among subjects was 112.7 ± 1.42. The mean number data windows for standing, walking in place, and jogging in place were 40.3 ± 0.65, 36.4 ± 0.81 and 36.1 ± 0.83, respectively.

### *4.2. Confirmation of behavioral compliance*

We determined that there was a significant effect of behavioral state on log accelerometer power ($F(2, 20) = 392.34$, $p < 0.0001$), indicating that the magnitude of mechanical forces acting on the body and sensor varied as a function of behavioral state. Similarly, there was a significant effect of behavioral state on MQR ($F(2, 20) = 200$, $p < 0.0001$). Individual *t*-tests between behavioral conditions determined that both log accelerometer power and MQR were significant between all condition pairs ($N = 11$, $p < 0.0001$ in all cases). Figure 4 shows these behavioral results. Though the first of these results seems quite obvious, it serves to confirm that subjects were, in fact, performing the task as instructed. The second demonstrates the strong interaction between behavior (standing, walking, and running) and the amount of

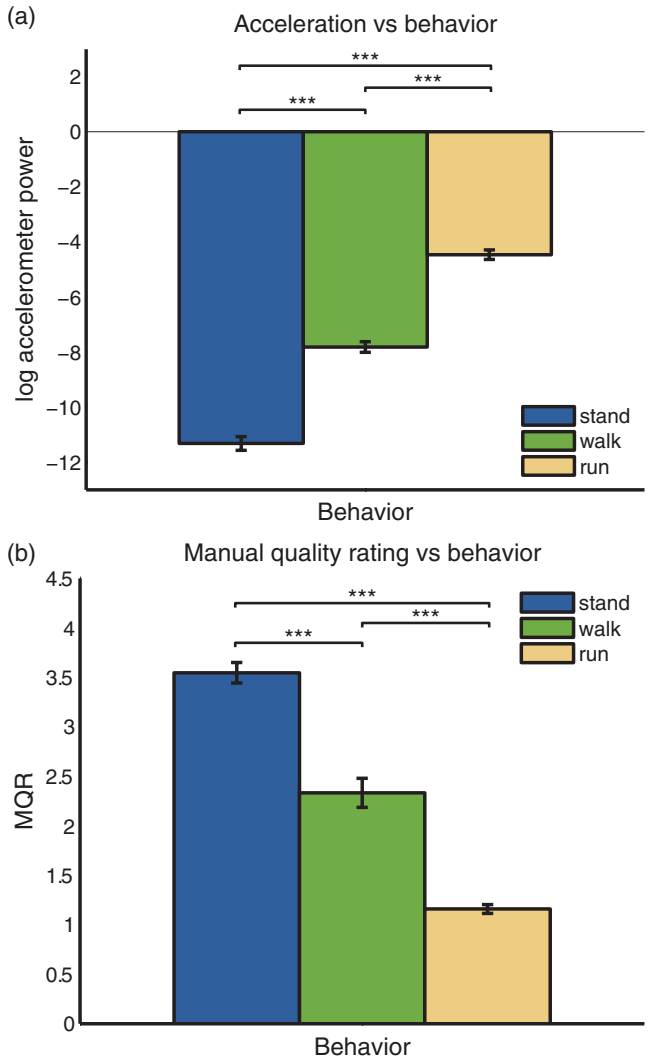


**Figure 4.** Confirmation of behavioral compliance. Subplot (*a*) shows the relationship between log accelerometer power and behavior, with the three behavioral states shown in different colors. Subplot (*b*) shows the relationship between MQR and behavior, utilizing the same color scheme. Three stars ('***') above significance bars represent $p$ values of less than 0.0001.

artifact present in the PPG signal. This serves as additional motivation of the importance of signal quality estimation during ambulatory activity.

### *4.3. Individual quality estimators*

In mobile heart rate monitoring applications, computational power and energy use limitations result in significant engineering constraints. Thus, heart rate monitoring device firmware will need to be as efficient as possible; performing on-device calculation of all of the signal quality features described above may not always be feasible. With that in mind, we sought to

**Table 1.** Results of linear regression between individual features/feature groups and MQR. Pearson's r values not shown for multiple regressions. Regression using the sample-to-sample transition features and log accelerometer power (both non-segmenting approaches) resulted in the highest correlations and lowest residual error. Feature subgroups and their corresponding Pearson's r values and $\sqrt{MSE}$ values are shown in **bold** type. Higher Pearson's r or lower $\sqrt{MSE}$ signify that a given feature is a better predictor of MQR.

| Feature (feature type) | Pearson's r | $\sqrt{MSE}$ |
|---|---|---|
| ***Log accelerometer power (non-segmenting)*** | **–0.8211** | **0.6924** |
| ***Sample-to-sample transition statistics (non-segmenting)*** | **0.898** | **0.5337** |
| ***Direct signal statistics (non-segmenting)*** | | **0.7528** |
| Log kurtosis | –0.4674 | 1.0724 |
| Autocorrelation peak strength | 0.5707 | 0.9961 |
| Spectral power | | 0.9238 |
| ***Beat template match statistics (segmenting)*** | | **0.6979** |
| Autocorrelation HR score | 0.5809 | 0.9874 |
| Gaussian corr. Score | 0.5904 | 0.9790 |
| HR est. score | 0.6498 | 0.9221 |
| Mean template correlation | 0.7595 | 0.7890 |
| Med. template correlation | 0.6813 | 0.888 |
| Max. template correlation | 0.4442 | 1.0868 |
| Min. template correlation | 0.7395 | 0.8166 |

understand the relationship between individual quality features and MQR, specifically seeking out features that were most strongly correlated with signal quality.

We performed regression analyses on a feature-by-feature basis as well as on the feature subsets described above, which are reported in table 1. It is important to note that these analyses were not performed using a complete cross-validation approach, only cross-validated training for the *sample-to-sample transition* SVR. Furthermore, initial beat templates were contributed to the *template-match* feature algorithm by all subjects and the *sample-to-sample transition* SVR was trained using training data from all subjects (Pearson's *r* and $\sqrt{MSE}$ values reported only on test data). This analysis was conducted with the goal of understanding how well different features and feature sets individually correlated with MQR. Though all features were significantly correlated with MQR, regression using the *sample-to-sample transition* features and *log accelerometer power* resulted in the highest correlations and lowest residual error (calculated as the square root of mean squared error [$\sqrt{MSE}$] between MQR and SQI; Pearson's *r* of 0.8980 and –0.8211, and $\sqrt{MSE}$ of 0.5337 and 0.6924, respectively), meaning that these features would likely be the most robust predictors of MQR. See table 1 for complete results.

### *4.4. Estimation of signal quality*

We then trained a CART using all features and employing the leave-one-out validation approach described above. During testing of all folds we achieved a mean correlation of 0.9263 and an $\sqrt{MSE}$ of 0.4627. Additionally, we trained individual CART models for each feature group to directly compare predictive performance of each feature group. Results are summarized in table 2. Observing that in some cases the CART was capable of estimating

**Table 2.** Results of leave-one-out CART-based regression analyses using all features and feature subsets.

| Feature group | Pearson's r | $\sqrt{MSE}$ |
|---|---|---|
| **All Features** | **0.9263** | **0.4627** |
| *Log accelerometer power (non-segmenting)* | 0.8283 | 0.713 |
| *Sample-to-sample transition statistics (non-segmenting)* | 0.9134 | 0.4932 |
| *Direct signal statistics (non-segmenting)* | 0.7373 | 0.8355 |
| *Beat template match statistics (segmenting)* | 0.8343 | 0.6692 |

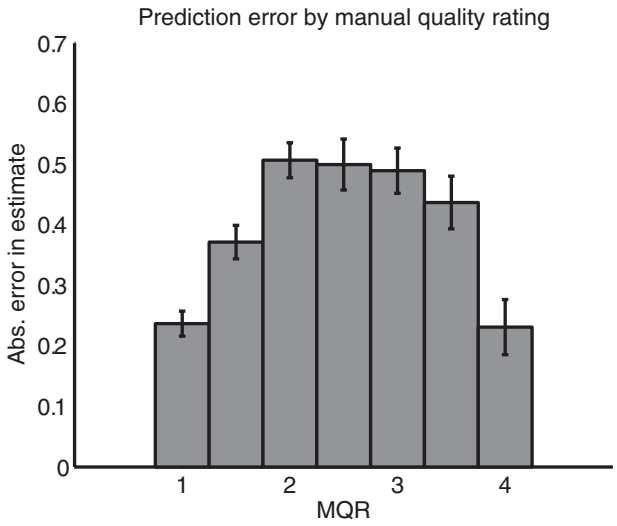


**Figure 5.** Signal quality estimate error magnitude as a function of MQR. Error magnitude was calculated as the absolute value of the difference between MQR and SQI for each window. Note decreased error rates at the extrema of manual quality ratings.

MQR in cross-validation analyses better than our correlation analyses above suggests that non-linear models may better capture the relationships between those feature sets and MQR.

It is interesting and important to note that the greatest errors in SQI (calculated as the magnitude of the difference between SQI and MQR for a given window) occurred for windows where MQR was intermediate (i.e. not 1 or 4). This can be seen in figure 5. Mean magnitude of prediction error was consistent across all folds, with no significant effect of fold on SQI error ($F(10, 66) = 0.94$, $p = 0.5074$) suggesting that the results were consistent across all subjects.

### *4.5. Signal quality estimation in the presence of periodic artifact*

We performed additional analyses using the two CART models trained using segmenting (*beat template match*) features and our non-segmenting (*sample-to-sample transition*) features. In order to test our hypothesis that *non-segmenting* signal quality algorithms are advantageous to their segmenting counterparts specifically when applied to PPG signal that has been contaminated by periodic artifact, we first looked only at data windows where MQR was low (less than or equal to 1.5, implying at least one rater scored the window as a 1), comparing the

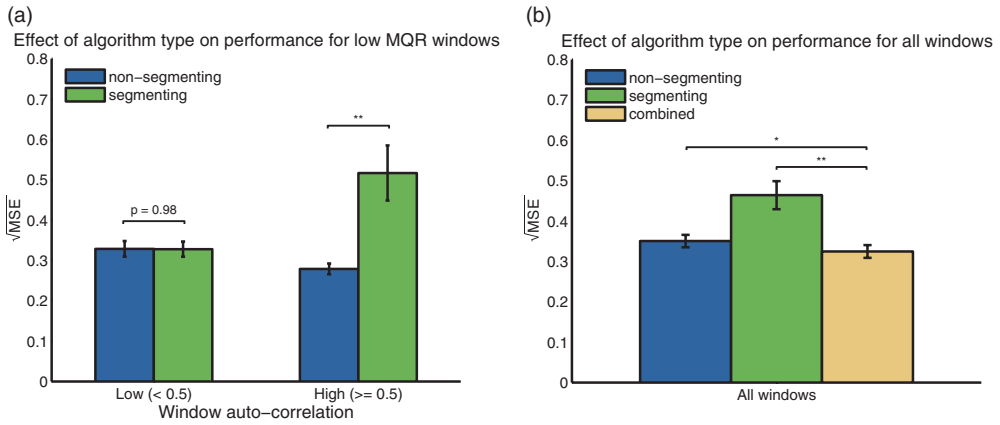


**Figure 6.** Subplot (*a*) depicts the differences in performance (as measured by $\sqrt{MSE}$) of segmenting and non-segmenting algorithms in cases where signal quality is low and the signal is either very autocorrelated or not well autocorrelated. Note that the segmenting algorithm performed worse than the non-segmenting algorithm in cases where the signal was highly autocorrelated. Subplot (*b*) considers all windows, and demonstrates that a combined approach that utilizes both segmenting and non-segmenting methods out performs either approach alone. One star ('*') above significance bars represents a p value of less than 0.05, but greater than 0.01. Two stars ('**') above significance bars represent a p value of less than 0.01.

performance of these two regression models in the case where the signal was highly autocorrelated (*autocorrelation peak strength* was greater than or equal to 0.5) to the case where it was not (*autocorrelation peak strength* was less than 0.5). Low-quality regions that are highly periodic are very likely contaminated with periodic motion artifact. As would be expected from their relative regression strengths outlined in table 1, the model trained on *sample-to-sample transition* (non-segmenting) features out-performed the model trained on *beat template match* (segmenting) features ($N = 11$, $p = 0.0341$). Breaking this down by window type, we see that non-segmenting and segmenting approaches demonstrated equivalent performance on non-autocorrelated windows ($N = 11$, $p = 0.9765$), but the segmenting approach significantly outperformed the non-segmenting approach on the autocorrelated windows ($N = 11$, $p = 0.0037$). This result demonstrates that the non-segmenting approach shows improved performance relative to the segmenting approach specifically during periods where the PPG signal was contaminated with strongly periodic artifact, such as during walking or running.

In order to address our second hypothesis that a combined approach would outperform either of these individual approaches we then considered all data windows for each subject, comparing the performance of these two individual CART models to the performance of the model trained on all features. The combined model outperformed both the model trained only on segmenting features ($N = 11$, $p = 0.0034$) and the model trained on non-segmenting features ($N = 11$, $p = 0.0116$).

Both of these results are summarized in figure 6.

### *4.6. Signal quality prediction applied to HR estimation*

The notion of signal quality is, in reality, application-dependent. Features relevant to successful estimation of signal quality may change depending on what physiological parameters are going to be extracted from the signal downstream in the processing pipeline. In the same

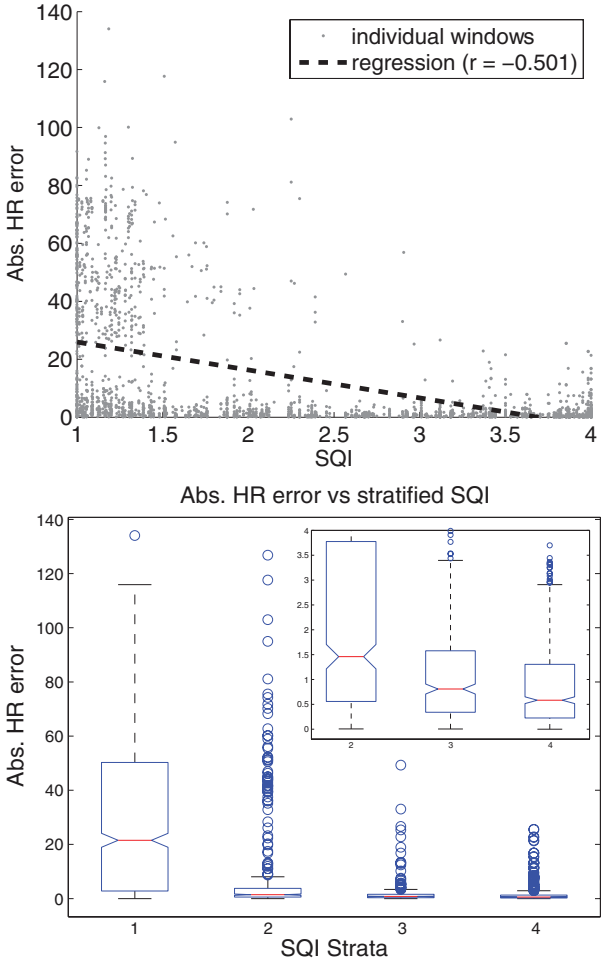


**Figure 7.** Results from HR analysis. Subplot (*a*) shows absolute error in HR estimate as a function of automatically determined signal quality for data windows from the 10 subjects included in the HR analysis. Regressing absolute HR error against signal quality resulted in a Pearson's r-value of −0.501. Regression fit is shown as a dashed black line. Subplot (*b*) presents the same data, where signal quality has been stratified in to integer values from one to four. With reference to each box plot the horizontal red line depicts the median, the extents of the notch surrounding the median are the 95% confidence intervals of this estimate, the extents of the box are the upper and lower quartiles, the whiskers extend to the greatest values not considered outliers, and outliers, which are considered as any values more than 1.5 times the inter quartile range above or below their respective quartlies, are shown as circles. The inset plot shows the distributions of the three largest strata in more detail.

way, thresholds for minimum allowable signal quality may also change. With this in mind, we sought to evaluate our signal quality estimation algorithm in the context of HR monitoring, a common application of the PPG signal. Using all signal quality features, and comparing HR estimated using the same algorithm that was used to estimate HR during beat segmentation to ground truth HR obtained from an ECG chest strap, we found that HR estimation

error magnitude was significantly anti-correlated with SQI (Pearson's $r = -0.501$, $p < 0.0001$). Stratifying quality values as $SQI_1 < 1.5$, $1.5 <= SQI_2 < 2.5$, $2.5 <= SQI_3 < 3.5$, and $SQI_4 <= 3.5$, we found significant pair-wise relationships between all strata ($p < 0.01$ for $SQI_3$-$SQI_4$ and $p < 0.0001$ for all other pairs, Wilcoxon rank-sum test). In other words, our signal quality prediction approach is generally able to predict scenarios where a real application—HR estimation—would likely be unreliable.

Both the linear trend and HR estimation error magnitude as a function of stratified SQI can be seen in figure 7. Note that these analyses were only performed on data from 10 of the 11 subjects, as the HR chest strap data were corrupt for the 11th subject.

## 5. Discussion

In this study we have expanded the paradigm of signal quality estimation beyond the clinical environment to a model that could be applied to continuous data collection using a non-invasive, wearable biosensor. We have developed and validated a signal quality estimation system that can estimate graded signal quality and demonstrated the utility of signal quality estimation in an example cardiac output metric. Specifically, we have proposed a novel signal quality feature set that does not rely on successful segmentation of PPG waveforms in to individual beats, and demonstrated the circumstances under which this feature set is superior to alternatives that do rely on beat segmentation. These findings equip us with an important tool as we move toward the eventual goal full-time monitoring of cardiac health outside of the clinical environment.

The majority of segmenting approaches to signal quality estimation leverage the fact that beat morphology is fairly similar from beat to beat, and beats that deviate greatly from the current model of beat morphology are likely artifactual. However, it is difficult for segmenting approaches to differentiate periodicity in the PPG due to beating of the heart from periodicity due to ambulatory movement. Therefore, these methods are susceptible to falsely developing a self-reinforcing model of pulse morphology that is based solely on motion-induced PPG changes, which could lead to performance differences as were demonstrated in figure 6. The *sample-to-sample transition statistics* that we proposed in this manuscript leverage the differences between pulse wave morphology of uncontaminated PPG and PPG corrupted by quasi-periodic motion; specifically, the asymmetry intrinsic to clean PPG. When fused with features that are sensitive to discrete motion events (e.g. autocorrelation metrics), we demonstrated that both periodic and aperiodic signal contaminants can be robustly identified.

There are two primary strategies that are applied to the problem of contaminated PPG signal. The first, which was the focus of this manuscript, is the estimation of an SQI that can be utilized in downstream processes to choose to include or exclude data segments from subsequent analyses. The second, which is in some ways a more difficult problem, is the estimation and subsequent removal of components of the PPG signal that are attributable to motion of the limb or digit where the device is being worn. Though these two techniques are by no means mutually exclusive, the majority of investigations focus on one or the other (see (Krishnan *et al* 2010) for an exception). That being said, we see a tremendous opportunity in the coupling of signal conditioning with signal quality estimation, specifically the use of an SQI to provide information about the success of an applied signal conditioning approach and to determine whether the procedure resulted in a reliable signal. As we observed in this manuscript, PPG signal can be contaminated by a variety of different sources, and different signal conditioning approaches may be more or less successful depending on the source contaminant.

In addition to software-based approaches for improving PPG signal quality, many groups are exploring hardware and interface-based changes to PPG acquisition systems that can make them more robust to motion artifact. Approaches involving multiple wavelengths of transmitted light (Patterson and Yang 2011, Yousefi *et al* 2014), modulated LED brightness (Patterson and Yang 2011), and deliberate control of sensor-to-skin interface pressure (Asada *et al* 2003) are all being considered. Robust SQI frameworks provide an objective mechanism by which to compare PPG signals recorded using these different systems.

An optical pulse monitor that is intended to be worn all day by an individual who is going about their normal routine will, by the very nature of its use, be subjected to a variety of different signal contaminants. An important follow-up study to the work presented here will be to evaluate contaminant nature and corresponding signal quality estimation of optical signals during more natural use cases. Our experimental conditions reflected only a subset of behaviors that, though they did effectively reflect changes in signal quality due to exercise conditions, certainly did not span the breadth of activities that a wearer would undergo throughout the course of a day. As such, we expect that ideal signal quality assessment systems will employ a variety of strategies for quality estimation, possibly even modifying their strategy dynamically based on estimation of the user's state (e.g. sedentary, exercising, etc.) and environmental conditions. It is important to note that, even in our specific set of experimental conditions, algorithm performance was maximal when features derived using both segmenting and non-segmenting methods used together.

An additional valuable point of discussion concerns how we define high and low signal quality. As has been mentioned previously, systems that are trained to identify signal quality, necessary quality acceptance thresholds, and remediation strategies in the case of low signal quality are all highly application-dependent. In this study, in an effort to achieve consistency, raters were trained to rate signal quality using a specific rubric. However, this rubric was not designed with a specific application in mind. Though we presented the use of signal quality estimation in a HR application, our raters evaluated PPG signal quality based on their ability to visually identify individual beats, which is a metric that may be most suitable for applications that require resolving individual beats in the PPG signal such as HRV or pulse transit time (Smith *et al* 1999). Furthermore, though raters were instructed to use a specific rubric when assessing the quality of PPG data, we observed substantial variability between raters, especially at intermediate values of MQR. This may partially explain why signal quality estimation errors were highest for intermediate MQR values (see figure 5); however, it also speaks to the robustness of our algorithm to noise in the ground-truth labels. In the absence of a large number of ratings from multiple raters, such variability is effectively additive noise that weakens the effectiveness of training for a supervised ML algorithm. We find it promising that automated signal quality detection is robust even in the presence of noise in the data labels, resulting in excellent predictive capability.

Taken together, these points punctuate the notion that signal quality in general, and specifically systems that are trained to programmatically estimate signal quality, must be considered in the context of the application for which they are intended. As an example, in the development of a PPG-based HR monitor, instead of labeling data windows with manual quality ratings, it may make more sense to train signal quality estimators on errors in HR estimation as compared to ground truth HR derived from an electrocardiogram (ECG), or alternatively to set an error threshold and train a quality estimator to predict whether signal quality is sufficient to estimate HR with sufficient accuracy to be within that threshold. Fortunately, modern ML methods such as those used in this study will function equally well in this application-specific case, and may even identify meaningful features that are specifically related to the ability of a HR estimation algorithm to correctly extract HR. We recognize that the HR application we

demonstrated could be expanded significantly by employing more sophisticated approaches to HR estimation, such a Kalman filter that takes in to account previous heart rate estimates and prior knowledge about the rate at which HR can change (Li *et al* 2008), however, such work is beyond the scope of this manuscript. Our example served to illustrate the utility of PPG signal quality estimation in a model example.

Though we applied our algorithm exclusively to PPG signals, we expect that they may be equally applicable to signal quality detection during direct measurement of the pulsatile pressure wave. Pressure wave signals are morphologically quite similar to the PPG waveform. Applanation tonometry (AT) of the radial artery has been proposed for estimation of aortic pressure waveform, brachial blood pressure, and response to hypertensive treatment (Nelson *et al* 2010). Derivation of these metrics, however, requires observation of the accurate pulse pressure wave form, and correspondingly high signal quality. The AT signal is equally, if not more so, susceptible to motion artifact, punctuating the need for robust signal quality estimation.

The work discussed in this manuscript represents a significant step forward toward ubiquitous, non-invasive heart health sensing. With the continued development of wearable biosensors come many opportunities for reduction in morbidity of major cardiac diseases. As the quantity of data collected from these sensors continues to increase, the importance of effective automated signal quality estimation will increase correspondingly.

## References


Asada H *et al* 2003 Mobile monitoring with wearable photoplethysmographic biosensors *IEEE Eng. Med. Biol.* **22** 28–40

Asada H, Jiang H-H and Gibbs P 2004 Active noise cancellation using MEMS accelerometers for motion-tolerant wearable bio-sensors *Conf. Proc. IEEE EMBS* **3** 2157–60

Breiman L (ed) 1984 *Classification and Regression Trees* (Boca Raton, FL: CRC Press)

CDC 2012 Heart disease and stroke prevention *Centers for Disease Control and Prevention* (www.cdc.gov/chronicdisease/resources/publications/AAG/dhdsp.htm)

CDC 2013 Physical activity recommendations for adults *Centers for Disease Control and Prevention* (www.cdc.gov/physicalactivity/everyone/guidelines/adults.html)

Ceesay S, Prentice A and Day K 1989 The use of heart rate monitoring in the estimation of energy expenditure: a validation study using indirect whole-body calorimetry *Br. J. Nutr* **61** 175–86

Clifford G D *et al* 2012 Signal quality indices and data fusion for determining clinical acceptability of electrocardiograms *Physiol. Meas.* **33** 1419–33

Cole C R *et al* 1999 Heart-rate recovery immediately after exercise as a predictor of mortality *N. Engl. J. Med.* **341** 1351–7

Eurostat 2009 *Health Statistics—Atlas on Mortality in the European Union* (Luxembourg: Office for Official Publications of the European Communities)

Farooq U *et al* 2010 PPG delineator for real-time ubiquitous applications. *Conf. Proc. IEEE EMBS* **2010** 4582–5

Fox S and Duggan M 2013 *Tracking for Health* (Washington, DC: Pew Research Center)

Gibbs P T, Wood L B and Asada H 2005 Active motion artifact cancellation for wearable health monitoring sensors using collocated MEMS accelerometers *Smart Structures and Materials Inter. Soc. Opt. Photon.* pp 811–9

Hoyert D L and Xu J 2012 *Deaths: Preliminary Data for 2011: National Vital Statistics Reports* (Hyattsville, MD: National Center for Health Statistics) **61**

Karlen W, Ansermino J M and Dumont G 2012a Adaptive pulse segmentation and artifact detection in photoplethysmography for mobile applications *Conf. Proc. IEEE EMBS* **2012** 3131–4

Karlen W *et al* 2012b Photoplethysmogram signal quality estimation using repeated Gaussian filters and cross-correlation *Physiol. Meas.* **33** 1617–29

Karlen W, Petersen C and Gow J 2013 Photoplethysmogram processing using an adaptive single frequency phase vocoder algorithm. *Biomed. Eng. Syst. Technol. Commun. Comput. Inform. Sci.* **273** 31–42

Krishnan R, Natarajan B B and Warren S 2010 Two-stage approach for detection and reduction of motion artifacts in photoplethysmographic data *IEEE Trans. Biomed. Eng.* **57** 1867–76

Leon A S *et al* 2005 Cardiac rehabilitation and secondary prevention of coronary heart disease *Circulation* **111** 369–76

Li Q and Clifford G D 2012 Dynamic time warping and machine learning for signal quality assessment of pulsatile signals *Physiol. Meas* **33** 1491–501

Li Q, Mark R and Clifford G 2008 Robust heart rate estimation from multiple asynchronous noisy sources using signal quality indices and a Kalman filter *Physiol. Meas.* **29** 15–32

Mancia G *et al* 2013 2013 ESH/ESC guidelines for the management of arterial hypertension *Eur. Heart J.* **34** 2159–219

Mannheimer P D 2007 The light-tissue interaction of pulse oximetry *Anesth. Analg.* **105** S10–7

Nelson M R *et al* 2010 Noninvasive measurement of central vascular pressures with arterial tonometry: clinical revival of the pulse pressure waveform? *Mayo Clin. Proc. Mayo Clin.* **85** 460–72

Nilsson L, Johansson A and Kalman S 2000 Monitoring of respiratory rate in postoperative care using a new photoplethysmographic technique *J. Clin. Monit. Comput.* **16** 309–15

Patterson J A C and Yang G-Z 2011 Ratiometric artifact reduction in low power reflective photoplethysmography *IEEE Trans. Biomed. Circuits Syst.* **5** 330–8

Rhee S, Yang B and Asada H 2001 Artifact-resistant power-efficient design of finger-ring plethysmographic sensors *IEEE Trans. Biomed. Eng.* **48** 795–805

Schäfer A and Vagedes J 2013 How accurate is pulse rate variability as an estimate of heart rate variability? A review on studies comparing photoplethysmographic technology with an electrocardiogram *Int. J. Cardiol.* **166** 15–29.

Silva I, Lee J and Mark R G 2012 Signal quality estimation with multichannel adaptive filtering in intensive care settings *IEEE Trans. Biomed. Eng.* **59** 2476–85

Smith R P *et al* 1999 Pulse transit time: an appraisal of potential clinical applications *Thorax* **54** 452–7

Sukor J, Redmond S and Lovell N 2011 Signal quality measures for pulse oximetry through waveform morphology analysis *Phys. Meas.* **32** 369

Taylor R S *et al* 2004 Exercise-based rehabilitation for patients with coronary heart disease: systematic review and meta-analysis of randomized controlled trials *Am. J. Med.* **116** 682–92

Weng J, Ye Z and Weng J 2005 An improved pre-processing approach for photoplethysmographic signal *Conf. Proc. IEEE EMBS* **1** 41–4

Wood L B and Asada H 2006 Noise cancellation model validation for reduced motion artifact wearable PPG sensors using MEMS accelerometers *Conf. Proc. IEEE EMBS* **1** 3525–8

Yousefi R *et al* 2014 A motion-tolerant adaptive algorithm for wearable photoplethysmographic biosensors *IEEE J. Biomed. Health Inform.* **18** 670–81