2021 43rd Annual International Conference of the
IEEE Engineering in Medicine & Biology Society (EMBC)
Oct 31 - Nov 4, 2021. Virtual Conference

Optimal Preprocessing of Raw Signals from Reflective Mode Photoplethysmography in Wearable Devices

Florian Wolling*[1], Sudam Maduranga Wasala*, and Kristof Van Laerhoven*

Abstract— **The optical measurement principle photoplethysmography has emerged in today's wearable devices as the standard to monitor the wearer's heart rate in everyday life. This cost-effective and easy-to-integrate technique has transformed from the original transmission mode pulse oximetry for clinical settings to the reflective mode of modern ambulatory, wrist-worn devices. Numerous proposed algorithms aim at the efficient heart rate measurement and accurate detection of the consecutive pulses for the derivation of secondary features from the heart rate variability. Most, however, have been evaluated either on own, closed recordings or on public datasets that often stem from clinical pulse oximeters in transmission instead of wearables' reflective mode. Signals tend furthermore to be preprocessed with filters, which are rarely documented and unintentionally fitted to the available and applied signals. We investigate the influence of preprocessing on the peak positions and present the benchmark of two cutting-edge pulse detection algorithms on actual *raw* measurements from *reflective* mode photoplethysmography. Based on 21806 pulse labels, our evaluation shows that the most suitable but still universal filter passband is located at 0.5 to 15.0 Hz since it preserves the required harmonics to shape the peak positions.**

I. INTRODUCTION

Wearable devices have become increasingly popular, particularly in the form factor of wrist-worn fitness trackers and smartwatches. At the same time, photoplethysmography (PPG) has been established as the standard technique for monitoring the wearer's heart rate (HR), one of the human's primary vital signs. Originally introduced by Hertzman [1] in 1937, the simple optical measurement principle enables the non-invasive measurement of HR and peripheral oxygen saturation. Since then, pulse oximeters are present at regular wards in clinical settings and apply *transmission* mode PPG, usually at the fingertip or earlobe, which emits light at one side of the perfused tissue and measures the amount of transmitted light at the opposite side. In contrast, the *reflective* mode, utilized in today's wearable devices, detects the non-absorbed but scattered light from the superficial layers of the skin. In both modes, the signal directly obtained from the sensor is inversely proportional to the captured blood volume changes in the skin. It is, however, common practice to invert the signal amplitude during preprocessing, to be consistent with the associated arterial blood pressure (ABP), which regularly leads to confusion [2], [3], [4].

In research, and especially in field studies with numerous devices, the Empatica E4 [5] has been established as a popular and commercially available tool for the monitoring of vital signs over long term [6]. Besides the early detection and diagnosis of heart diseases in medical studies, secondary features from the heart rate variability (HRV), derived from the pseudo-periodic heartbeat, have shown to be linked to the wearer's emotions and affective state [7]. The evidence of the findings is, however, biased and limited to signals from specific devices such as the aforementioned E4.

For researchers, it is comfortable to obtain the measurements from such embedded sensors: The signals are usually straightforward to interpret and analyze, since they come already conditioned and preprocessed (Fig. 3, middle). To the inexperienced observer from disciplines other than signal processing, the sensing devices seem to deliver proper *raw* signals because they come directly from them. However, the embedded software of commercial wearables is usually closed, not adaptable, and hence limits the signal's information content as well as possible applications. The use of actual *raw* sensor data would demand for more knowledge and effort from the researcher, but also allows for customization to meet individual, research-specific requirements.

In context of the extraction of the respiration rate from PPG signals, Pimentel et al. [8] emphasize that "Future studies should concentrate on the use of [...] raw data sources as a benchmark for comparison". However, in a review of public datasets [4], we have revealed that, although advertised as such, most datasets do not actually contain *raw* but conspicuously filtered signals. For this reason, most large and promising datasets are not suitable to benchmark available algorithms or even to determine optimal preprocessing parameters. Likewise, Reiss et al. [9] state that "State-of-the-art publications rely mostly on the two datasets introduced for the IEEE Signal Processing Cup" [10], [11] (Fig. 3, top), which do not contain actual *raw* PPG signals [4], and found that "existing approaches are highly parametrised and optimised for specific scenarios of small, public datasets".

This study aims at deeper understanding of the *raw* PPG signal's characteristics, directly obtained from the sensor, to pave the way for more universal, reliable, and accurate sensor-integrated algorithms, running on wearable devices.

In this paper, we make the following contributions:

- We provide 21806 labels, manually set and validated by an expert rater, for the public dataset of Biagetti et al. [12] that contains 286 minutes of *raw*, *reflective* mode PPG recordings from seven subjects during three different exercises.
- We benchmark the two popular, cutting-edge algorithms of Karlen et al. [13] (2012) and van Gent et al. [14] (2019).
- We investigate the influence of preprocessing on the peak positions and hence the performance of these algorithms.

*Ubiquitous Computing, University of Siegen, Germany.
[1]Corresponding author: florian.wolling@uni-siegen.de

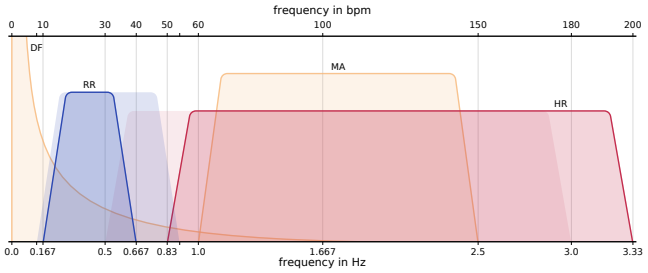

Fig. 1: Frequency components present in *raw* PPG signals. Frequency bands: fundamental frequencies of heart rate (HR) and respiration rate (RR) according to [15], [16]; DC offset and $1/f$ fluctuations [17] (DF); general noise and disturbances through daily motion [18] (MA). Natural limits of HR and RR change with the age from infants and children to adults (solid lines). Overlapping areas are critical since they are hard to distinguish and to assign [18]. Y-axis unspecified.

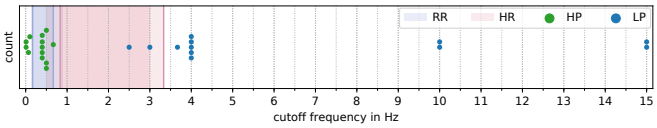

Fig. 2: Review of 14 publications. Distribution of applied filter cutoff frequencies: lower corner (high-pass, HP) and upper corner (low-pass, LP), in respect of respiratory (RR) and cardiac (HR) bands (see Fig. 1). Ref.: [11], [19], [20], [21], [22], [23], [24], [25], [26], [27], [28], [29], [30], [31].

II. RELATED WORK

We first introduce the most common methods to detect the pulse in PPG signals. Afterwards, we briefly survey applied preprocessing strategies before we lead over to a review and discussion of previous research on the optimal conditioning and filtering of *raw* PPG measurements.

A. Methods for Heart Rate Monitoring

In general, most available algorithms for HR determination from PPG signals can be divided into two major categories: time domain and frequency domain approaches. In time domain, the individual, consecutive pulses of the heartbeat are identified by means of significant fiducial points. The diastolic pulse onset is the most common one, but also other characteristic points have been subject to evaluation and show slightly different applicability and accuracy [32], [33]. In the end, the HR can be determined either by counting and averaging the number of peaks per unit time, or by directly calculating the individual reciprocal of the inter-beat interval (IBI) for an instantaneous measure [34]. In contrast, approaches in frequency domain often aim for the application in resource-constrained systems. They aggregate and characterize a longer signal period of several seconds to minutes by means of a decomposition method or transformation function [35], [36]. Besides advanced techniques for spectral estimation [11], [25], the most frequently applied ones are the fast Fourier transformation (FFT) and Welch's method for a smoothed periodogram. The fundamental frequency, associated with the predominant (not average) HR, is then identified within the signal's frequency spectral representation and validated through different heuristics. This shortcut prevents these approaches, however, from the derivation of secondary information such as the heart rate variability (HRV) metrics [37], [38], the tachogram, or the interval function [39]. Since modern wearable applications demand for these measures, this research concentrates on HR tracking algorithms applied in time domain. Before the pulse feature detection and validation takes place, usually preprocessing and motion artifact removal stages are applied [19], [40].

Preprocessing: As illustrated in Fig. 1, the spectrum of *raw* PPG signals is composed of diverse superimposing frequency components. Although the bands of heart rate (HR) and respiration rate (RR) are limited by nature, it is not advisable to use filters with fixed passband limits to extract the components [18]. The plausibility of frequencies' occurrence highly depends on the individual and there is no consensus on optimal, generalized ranges [41]. For adults, the bands typically range from about 0.833 to $3.333\,\mathrm{Hz}$ (50–$200\,\mathrm{bpm}$) for HR and about 0.133 to $0.667\,\mathrm{Hz}$ (8–$40\,\mathrm{bpm}$) for RR. In the age of infants to young adults, the spectra for HR and RR range from 0.5 to $3.0\,\mathrm{Hz}$ (30–$180\,\mathrm{bpm}$) and 0.667 to $0.9\,\mathrm{Hz}$ (40–$54\,\mathrm{bpm}$) respectively [15], [16]. Consequently, the universal cardiac and respiratory frequency bands, for both infants and adults, overlap. The separation of the desired signal components from in-band noise, especially motion artifacts from daily activities such as walking and jogging (1.0 to $2.5\,\mathrm{Hz}$), becomes even more challenging [18], [40].

The artifact-free *raw* PPG signal (Fig. 3, bottom) is dominated by a large DC offset while the AC signal amplitude comprises only about 1–$10\,\%$ of the total scope [42], [43]. Depending on the ADC's resolution, typically $\geq 16\,\mathrm{bit}$, the digital representation and storage of measurements requires a lot of memory. The simple elimination of the DC offset, often taken as unnecessary, easily reduces the extent.

The *raw* signal contains also other frequency components that are usually not of interest and removed by signal conditioning and preprocessing techniques. It is general standard to limit the signal's spectral bandwidth by any type of bandpass filter, of which the Butterworth is the most common one. The passband's lower $f_{c,hp}$ and upper $f_{c,lp}$ cutoff frequencies are defined through successive high-pass and low-pass filter stages (depicted in Fig. 5). The low-pass stage rejects noise at higher frequencies which hinder the accurate detection and localization of the small AC pulse peaks. At the same time, the natural baseline wander of the physiological signal contains low-frequency components [17], [37], [39] which blur and smear the pulses along steep and large slopes. Thus, the high-pass filter stage is applied to detrend the signal and to remove e.g. the RR fluctuations [8], [15], [22]. Particularly the first ten harmonics of the fundamental HR shape the pulse waveform and need to be preserved [20], [44]. Fig. 2 illustrates the widespread distribution of utilized cutoff frequencies from 14 publications. In previous research, we have reviewed commonly applied conditioning and preprocessing strategies [4] and have further investigated the influence of the sensor's sampling rate on the HR determination [35].

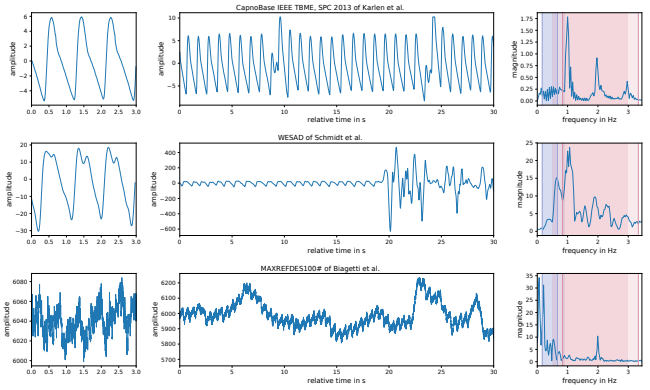

Fig. 3: Typical excerpts from CapnoBase [10], [45] (top), WESAD [7] using the popular Empatica E4 [5] (middle), and the MAXREFDES100# [12], [46] (bottom). Close-up (left), 30 s window (middle), respective frequency spectrum (right). Preprocessed, smoothed, and zero-centered (top and middle). Actual *raw* signals from *reflective* mode PPG (bottom).

Motion Artifacts: In any application for the conscious human, especially during physical activity, the loosely attached sensor captures motion-induced distortions which result from hemodynamic effects, slight sensor displacements, and tissue deformation [47]. The occurring interference and artifacts, typically observable as abrupt changes, affect the desired pulsatile signal and impede the application of simple algorithms: threshold-based approaches using derivatives, moving averages, or the slope sum function [21] and relics from ECG analysis [48]. The identification, exclusion, or even removal of motion artifacts remains the biggest challenge and has most recently been surveyed by Ismail et al. [40]. Besides approaches assessing the pulse morphology and applying adaptive filters in time domain [13], others analyze signals from auxiliary sensors, mostly accelerometers [23], [31] but even optical flow sensors known from computer mouses [49].

B. Studies on Optimal Preprocessing and Filtering

In recent years, there has already been research aiming at the identification of optimal techniques and filter parameters for the preprocessing of PPG signals. Those studies did, however, not lead to a clear consensus. They either targeted specific applications or comprise weaknesses which we now intend to address with our research.

In 2012, Stuban et al. [44] evaluated the optimal filter bandwidth for pulse oximetry (SpO2), which traditionally applies transmission mode PPG. The research concentrated on the estimation of the arterial oxygen saturation from the ratio of measurements at two different wavelengths, red and infrared light, sampled at 40 Hz. They concluded that the "harmonics of the pulse signal do not contribute to the accuracy of pulse oximetry" and consequently "filtering out the harmonics [...] does not degrade the accuracy". The DC and very low frequency components have been removed by a 2$^{\text{nd}}$ order infinite impulse response (IIR) high-pass filter. The lower cutoff frequency was set to 0.1 Hz and "must be lower than the fundamental frequency of the pulse". Noise and harmonics of the pulse have been removed by a 100$^{\text{th}}$ order finite impulse response (FIR) low-pass filter and five upper cutoff frequencies at 0.66, 1.0, 1.5, 3.0, or 15.0 Hz.

In 2018, Liang et al. [50] published an impressively large dataset of 657 PPG snippets, captured at the left index fingers of 219 subjects. Since the recordings are very short, just 2.1 s long, the dataset's applicability is limited. Based on a selection of 219 pulses, classified as "excellent", "acceptable", and "unfit", they determined the 4$^{\text{th}}$ order Chebyshev II to be the optimal filter technique – at least for these short signals.

Most recently, in 2020, Cassani et al. [51] analyzed the spectral coherence and the signal-to-noise ratio between "isolated" and the original, "raw" pulses. They determined the optimal filter passband to be 0.6 to 3.3 Hz for adults and 1.0 to 2.7 Hz for children. The spectral analysis showed a half-power bandwidth of 0.8 to 2.4 Hz for adults and of 0.9 to 2.7 Hz for children. The study analyzed 27000 pulses from the well-known CapnoBase IEEE TBME [10], [45] (Fig. 3, top) dataset containing signals from a fingertip pulse oximeter, but not *raw*, *reflective* mode PPG signals [4].

In 2019, Bastos et al. [52] investigated the optimal parameters for Butterworth and maximal overlap discrete wavelet transform (MODWT) filters which are "widely employed" in resource-constrained wearables. Considering very few cutoff frequencies, again a dataset from [45] (Fig. 3, top) and the MIMIC-II BIDMC [8] were applied, both unfortunately not containing actual *raw* signals from *reflective* mode PPG [4].

III. METHODS AND MATERIALS

We first introduce the used dataset and highlight the preparation of ground truth. Subsequently, we investigate the general pulse peak displacement due to filtering and then benchmark two popular algorithms on filtered time series.

A. Dataset

Public datasets of actual *raw* signals from *reflective* mode PPG sensors are scarce [4]. The adequate benchmark of available algorithms and preprocessing techniques requires, however, large datasets of such kind. In this research, we decided for the very recent dataset of Biagetti et al. [12] from 2020. It is originally intended for the application of machine learning techniques in human activity recognition. With in total 286 minutes of *raw* PPG measurements (Fig. 3, bottom), it provides a set of 105 recordings from 7 subjects wearing the MAXREFDES100# [46], a commercially available reference design. Simultaneous PPG and acceleration signals, sampled at a rate f_s of 400.0 Hz, are provided for the subjects performing three exercises: *rest*, *squat*, and *step*. Hence, the time series are not entirely clean but contain motion artifacts which do affect the applied algorithms.

An ideal dataset would uniformly cover the entire range of the natural HR (30–200 bpm [15], [16]). This requirement would, however, hardly be possible without a health risk for the volunteers. As illustrated in Fig. 4, the used dataset covers a broad HR spectrum of at least 40 to 160 bpm, with a strong core area ranging from 50 at rest to 110 bpm at light exercise. The three exercises show the mean HR of 72.9 ± 11.1 bpm (1.22 ± 0.18 Hz) for *rest*, 98.6 ± 16.0 bpm (1.64 ± 0.27 Hz) for *squat*, and 106.2 ± 21.2 bpm (1.77 ± 0.35 Hz) for *step*.

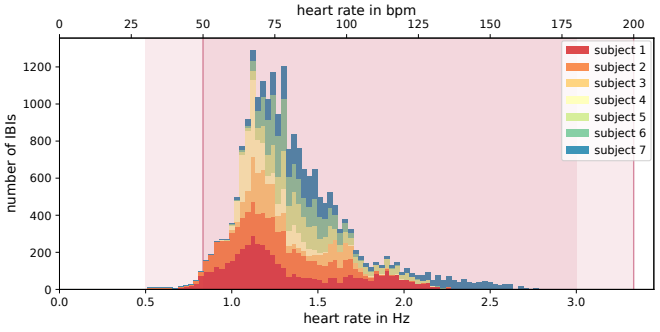

Fig. 4: Distribution of the seven subjects' instantaneous heart rate, derived from the inter-beat intervals (IBI) of manually labeled pulses. Limited cardiac frequency band (red, Fig. 1).

B. Ground Truth

For every recording in daily life, the supply of ground truth tends to be the major issue. Usually, a second sensing device or even a second sensing modality is used to provide the information with, at best, a higher degree of reliability and precision. In case of the HR, wearable long-term ECG devices are mostly employed since the electrodes are directly attached to the skin and hence enable the reliable and robust measurement, at cost of comfort. Unfortunately, the selected dataset does not provide ECG as ground truth. The PPG signal has thus manually been analyzed and annotated by a human expert rater with year-long experience. Consequently, the performance of the applied algorithms is not compared against a reference device but the labels accurately set by the expert. We hence demonstrate the theoretical limits by means of the human and their ability to interpret PPG signals.

Data Annotation: For the purpose of the comfortable and reliable annotation, we have developed a graphical tool which allowed the expert rater to label the pulse onsets within the raw time series. We consciously decided against an automatic labeling or preselection to avoid the expert being influenced and biased in their decision. Without a doubt, this decision resulted in more and monotonous manual work – reams of clicks to select the in total 21806 peaks. To prevent faults due to fatigue, the expert has split the work up into one subject per day, first the larger but easier-to-label data of *rest* and, after a break, the shorter but ambitious recordings of the *squat* and *step* exercises. In total 104 of 105 time series, 278 of 286 min (97.3 %) are annotated with 21 806 peak labels, only 7.73 min do not contain distinguishable signals or are considerably affected by motion artifacts and hence excluded. 88 of 105 time series are entirely labeled. The subset *squat 3* of *subject 5* was rejected since it does not contain any clearly distinguishable pulses.

While the ECG's well-known R spike is pointed and hence relatively 'easy' to identify, even in noisy signals, the typical PPG waveform is rather smooth and round. Recorded at a higher sampling rate [35], the *raw* PPG signal also shows a large portion of noise and baseline wander which blurs the optimal pulse peak and makes the identification of its exact position ambiguous (see Fig. 5). Consequently, the very top of the pulse is not always distinct but often subject to interpretation. In contrast to deterministic algorithms, the expert has, however, intuition, grounded in experience, to 'see' which tiny wave is an actual pulse onset and which one is just negligible noise or motion-induced distortion.

To exclude the influence of filtering from the beginning, the labels have been set within the *raw* signal before applying any filter. A second panel allowed the expert, however, to glance at the detrended and smoothed signal for orientation and validation, to avoid the selection of any invalid pulses. A 4^{th} order ($2 \times 2^{\text{nd}}$) *filtfilt* forward-backward zero-phase band-pass filter, passband 0.5 to 30.0 Hz, was applied. Unnaturally short or long IBIs have automatically been labeled as invalid and were subject to revision by the expert. The remaining uncertain intervals were finally excluded from the studies.

C. Study I: Peak Displacement

Filtering considerably changes the trend and shape of the *raw* PPG signal. By narrowing down the passband, the peak positions are conspicuously affected, smeared, and blurred (see Fig. 5). In the first study, we investigated the influence of filtering on the positions of pulse onsets [32], [33], the maximum peaks in *raw* signals respectively. We applied a 4^{th} order ($2 \times 2^{\text{nd}}$) Butterworth band-pass filter with 40×40 non-equidistant lower $f_{c,hp}$ and upper $f_{c,lp}$ cutoff frequencies: $\{/, 0.005, \ldots, 2.5\,\text{Hz}\} \times \{/, 199.0, \ldots, 2.5\,\text{Hz}\}$. The applied *filtfilt* forward-backward filter method with zero phase allows to protect and preserve the signal's original phase. To track the peak displacement $\varepsilon_d = |p_0 - \hat{p}|$, we applied a simple hill climbing algorithm to follow the original position p_0 up to the closest local maximum of the filtered signal at $\hat{p}$, implemented as a function $\hat{p}(\, f_{c,hp}\,,\, f_{c,lp}\,)$. Any $\varepsilon_d > 250\,\text{ms}$ has been excluded as a slipped outlier.

Fig. 6 illustrates the mean ε_d results for the individual exercises (left) which sum up to the averaged overall results (right). The boundaries of the minimum error plateaus with $\varepsilon_d \leq 0.5\,\text{samples}$ (blue) are shifted (red arrows) due to the increasing HR (vertical line) from *rest* (1.22 Hz) via *squat* (1.64 Hz) to *step* (1.77 Hz). Accordingly, an extended $f_{c,lp}$ (y-axis) is required to cover a sufficient number of harmonics, for an adequate contour and peak reconstruction, but it also allows for a higher $f_{c,hp}$ (x-axis).

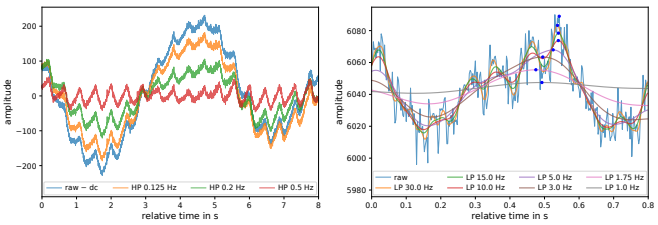

Fig. 5: Left: Effect of diverse high-pass filters for detrending. Right: Illustration of the pulse peak's position (blue) displacement due to the application of diverse low-pass filters. Conspicuously affected, smeared, and blurred pulse contour, vanishing with the baseline wander, due to the elimination of the fundamental HR's higher harmonics.

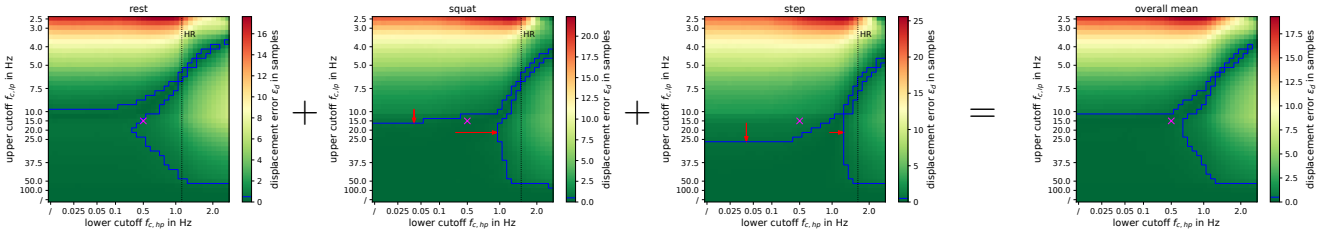

Fig. 6: Evaluation results of the displacement error ε_d due to filtering at diverse lower $f_{c,hp}$ (x-axis) and upper $f_{c,lp}$ (y-axis). Left to right: results of subsets *rest*, *squad*, *step*, and their overall mean. Shifting (red arrows) boundaries of minimum error plateau $\varepsilon_d \leq 0.5\,\text{samples}$ (blue) due to increasing HR (vertical). Proposed filter passband (magenta mark): 0.5 to 15.0 Hz.

D. Study II: Benchmark of Algorithms

The results of the previous *Study I* serve as the upper boundary of maximal achievable accuracy from filtered PPG signals. To benchmark available algorithms and to investigate the effect of preprocessing on their performance, we applied two popular algorithms on the filtered time series: 1) Karlen et al. [13] from 2012 and 2) van Gent et al. [14] from 2019. Those are based on two fundamentally different principles to identify the pulse peaks in time domain.

To assess the algorithms' performance, every position p_0 from the manual annotations is assigned to its closest counterpart $\hat{p}_a$ from the detected pulse peaks. If the error distance $|p_0 - \hat{p}_a| \leq 250\,\text{ms}$, the pair $(p_0, \hat{p}_a)$ is classified as true positive (*TP*). All missed p_0 without a counterpart $\hat{p}_a$ within reach are classified as false negative (*FN*) while the surplus of erroneously detected peaks falls into false positive (*FP*). This approach enables to apply the popular *F1*-score (1), the harmonic mean of *precision* (PPV) and *recall* (TPR) (2), to measure the peak detection performance. For all pairs in *TP*, the average error distance $\varepsilon_a = |p_0 - \hat{p}_a|$ is determined analog to the displacement error ε_d in *Study I*.

$$F1 := 2 \cdot \frac{PPV \cdot TPR}{PPV + TPR} \tag{1}$$

$$PPV := \frac{TP}{TP + FP}\,,\quad TPR := \frac{TP}{TP + FN} \tag{2}$$

In case of an algorithm's ideal performance, the *F1*-score would hit 1.0 and ε_a would match the theoretical limit ε_d. Since the local optima of *F1*-score and ε_a can be conflicting (see Fig. 7–8), a simple parameter optimization is applied to find a trade-off by multiplying the two normalized metrics.

1) Karlen et al. [13]: Intended for usage on resource-constrained devices, the algorithm consists of two stages. First, the incremental-merge segmentation (IMS) algorithm extracts the morphological features by segmenting and compressing the signal into straight lines. It is implemented as a sliding window of size m, which is the only parameter that requires tuning, but also depends on f_s. With a larger m the algorithm is faster and less susceptible to noise, but the determined peak positions are also less precise. Subsequently, the extracted lines with positive gradient are classified as *artifact* or *valid* pulse using simple adaptive thresholds. The authors state that "No other filtering than the standard band-pass filter applied by pulse oximeter manufacturers to remove the DC component [...] is necessary", but they do not specify proven values. Before, the algorithm has been evaluated using two datasets of which one is from CapnoBase [45], from *transmission* mode pulse oximeters. Since the algorithm regards the inverted pulse direction, consistent with the ABP, the time series have been flipped before its application.

2) van Gent et al. [14]: The open-source *HeartPy* toolkit aims for the computational efficient but particularly reliable pulse detection independent from the utilized sensor. Besides a comfortable Python library, an implementation for embedded devices in C is also available. First, a moving average, with a default window size w of 750 ms (300 samples at f_s), is used to identify local maxima as a first selection of *candidate peaks*. Since an excessive or missing single peak significantly increases the standard deviation of successive differences (SDSD), this measure is combined with the constraint of the natural HR limits (40–180 bpm by default) to stepwise adjust the threshold and hence to find the optimal peak selection of minimal SDSD. To compare the algorithms' performance, its validation heuristic is set to the previously discussed natural HR limits of 30 to 200 bpm [15], [16].

IV. Results & Discussion

The optimal cutoff frequencies largely depend on the subjects' HR. At rest it is low while the frequency band tends to be narrow (*rest*: 1.22 ± 0.18 Hz). With increasing activity, the HR increases and the frequency band widens (*squat*: 1.64 ± 0.27 Hz; *step*: 1.77 ± 0.35 Hz).

Accordingly, *Study I* demonstrates that the most universal and effectual filter passband ranges from the theoretical minimum 0.5 Hz of the natural HR to appropriate 15.0 Hz. As illustrated in Fig. 6, it is applicable for HR at *rest* as well as during exercise such as *squat* and *step*. While the lower $f_{c,hp}$ can be 'easily' estimated and fixed to the minimum HR to be expected, the upper $f_{c,lp}$ is more critical and difficult to specify. A generous $f_{c,lp}$ allows to cover more harmonics of the fundamental HR, which eventually refine the pulse contour. At a HR of 0.5 Hz (30 bpm), the covered 29th harmonic is, however, not very gainful. Nevertheless, the upper 15.0 Hz cutoff is required to cover at least 3 harmonics of a HR at 3.3 Hz (200 bpm) – 10 harmonics would, however, be ideal [20], [44]. A wider passband of up to 25.0 Hz would result in slightly more pointed and accurate peak contours but also gives unnecessary space for high-frequency noise.

Study II demonstrates the very different character of the two applied algorithms: 1) The algorithm of Karlen et al. [13]

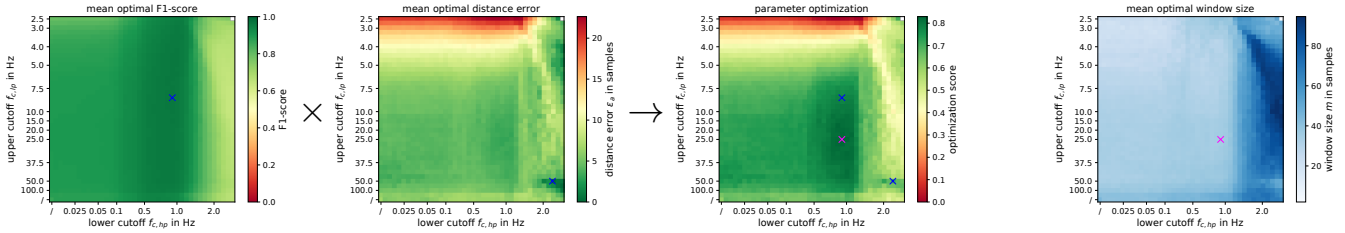

Fig. 7: Evaluation chain and results for the algorithm of Karlen et al. [13]. Left to right: mean optima of *F1*-score, distance error ε_a, fused parameter optimization, and window size m versus lower $f_{c,hp}$ and upper $f_{c,lp}$. Local optima of *F1*-score and ε_a (blue). Optimal configuration (magenta): $f_{c,hp}$ of 0.9375 Hz, $f_{c,lp}$ of 25.0 Hz, and m of 34.236 samples resulting in an *F1*-score of 0.958 and ε_a of 3.037 samples. Relatively homogeneous plateau of possible parameters of similar quality.

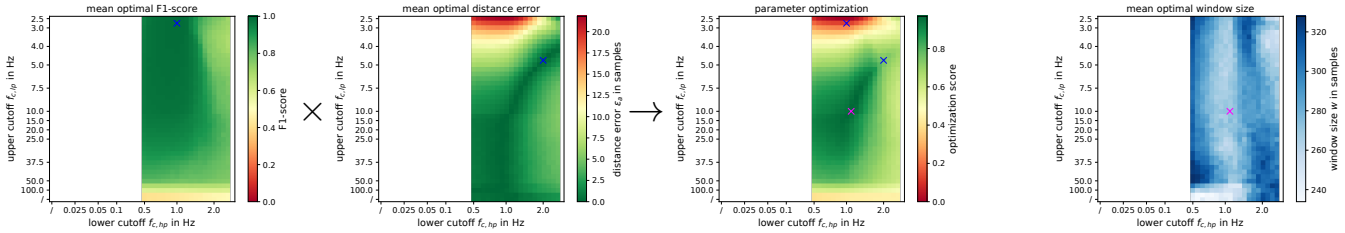

Fig. 8: Evaluation chain and results for the algorithm of van Gent et al. [14]. Left to right: mean optima of *F1*-score, distance error ε_a, fused parameter optimization, and window size m versus lower $f_{c,hp}$ and upper $f_{c,lp}$. Local optima of *F1*-score and ε_a (blue). Optimal configuration (magenta): $f_{c,hp}$ of 1.125 Hz, $f_{c,lp}$ of 10.0 Hz, and m of 270.426 samples resulting in an *F1*-score of 0.970 and ε_a of 0.051 samples. Rather complex texture with abruptly falling peak of optimal parameters.

performed best with a small window m of 34.236 samples (85.589 ms) and a filter passband from 0.9375 to 25.0 Hz. The configuration results in an *F1*-score of 0.958 and an ε_a of 3.037 samples (7.594 ms). Fig. 7 shows that the *F1*-score is optimal close to the fundamental HR but remains constant along a varied upper $f_{c,lp}$. The ε_a stays relatively steady along the lower $f_{c,hp}$ until it passes the HR but increases considerably along with a decreasing upper $f_{c,lp}$. The m is homogeneous and plane until it passes the HR along the lower $f_{c,hp}$. 2) The algorithm of van Gent et al. [14], in contrast, performed best with a large window w of 270.426 samples (676.065 ms) and a narrower passband from 1.125 to 10.0 Hz. The configuration results in an *F1*-score of 0.970 and an impressively small ε_a error of 0.051 samples (0.127 µs). Due to the immense increase of peak candidates in raw and noisy signals, accompanied by increasing processing efforts, the evaluation in Fig. 8 is limited to $f_{c,hp} \geq 0.5$ Hz.

Limitations: Because the used dataset does not provide ground truth information, we had to rely on an experienced expert rater to provide the annotations for the PPG data afterwards, inadvertently being subject to some bias as well. In case of peaks vanishing with the baseline, the identification can be rather subject to interpretation than a distinct recognition. Inaccuracies due to imperfect label placement are, however, statistically compensated through the large number of peak labels. The use of an ECG reference channel would, without question, be expedient. Very low as well as very high HR are underrepresented in this dataset. Follow-up studies should focus on a broader HR diversity that spans the entire range of the natural HR from 0.5 to 3.3 Hz (30–200 bpm [15], [16]). Also, since PPG at the wrist shows a location-specific composition, the harmonics may contribute differently to the pulse peak at other measurement locations.

V. CONCLUSIONS

Photoplethysmography is an emerging optical measurement principle which *reflective* mode is the standard technique to monitor the wearer's heart rate in wearable devices. The resource constraints of these demand for high efficiency while the applications require reliable and accurate pulse detection for HRV measurements. Most available algorithms have, however, been evaluated on just few publicly available datasets of conspicuously filtered signals. In this research, we highlight the importance of benchmarking on actual *raw* PPG signals. Based on the dataset of Biagetti et al. [12] and 21806 peak labels, manually annotated by an expert rater, the impact of preprocessing on pulse peak positions and the performance of peak detection algorithms has been evaluated. Applying 40×40 filter configurations, two popular algorithms are benchmarked: 1) Karlen et al. [13] from 2012 and 2) van Gent et al. [14] from 2019. In summary, algorithm 2) is more complex than 1) but, in absence of low-frequency baseline wonder, its concept results in a significantly higher precision. In general, the filter passband of 0.5 Hz to 15.0 Hz showed the best universality by preserving the heart rate's harmonics for distinct and precise pulse peak positions.

We encourage researchers to use the publicly available dataset of Biagetti et al. [12] in combination with the supplementary annotations from this research to benchmark their own algorithms as well as machine learning approaches. The annotation files, provided as `*.pkl` and `*.csv`, of the 21806 diastolic pulse onset labels are available for download from: `https://ubicomp.eti.uni-siegen.de/home/datasets/embc21/`

ACKNOWLEDGMENT

The large amount of data has been processed by the OMNI cluster at the University of Siegen in Germany.

REFERENCES

[1] A. B. Hertzman, "Photoelectric Plethysmography of the Fingers and Toes in Man," *Experimental Biology and Medicine*, vol. 37, no. 3, pp. 529–534, 1937.

[2] T. Y. Abay, "Reflectance Photoplethysmography for Non-invasive Monitoring of Tissue Perfusion," Doctoral Thesis, University of London, 2016.

[3] C. Choi, B.-H. Ko *et al.*, "PPG pulse direction determination algorithm for PPG waveform inversion by wrist rotation," *IEEE EMBC*, vol. 2017, pp. 4090–4093, 2017.

[4] F. Wolling and K. Van Laerhoven, "The Quest for Raw Signals: A Quality Review of Publicly Available Photoplethysmography Datasets," in *DATA '20*. ACM, 2020.

[5] Empatica Inc., "Empatica E4 Wristband," https://www.empatica.com/research/e4/, accessed: 2021-03-31.

[6] C. McCarthy, N. Pradhan *et al.*, "Validation of the Empatica E4 Wristband," in *IEEE EMBS ISC*, 2016, pp. 1–4.

[7] P. Schmidt, A. Reiss *et al.*, "Introducing WESAD, a Multimodal Dataset for Wearable Stress and Affect Detection," in *ICMI '18*. ACM, 2018, p. 400–408.

[8] M. A. F. Pimentel, A. E. W. Johnson *et al.*, "Toward a Robust Estimation of Respiratory Rate From Pulse Oximeters," *IEEE TBME*, vol. 64, no. 8, pp. 1914–1923, 2017.

[9] A. Reiss, I. Indlekofer *et al.*, "Deep PPG: Large-Scale Heart Rate Estimation with Convolutional Neural Networks," *Sensors*, vol. 19, no. 14, 2019.

[10] W. Karlen, S. Raman *et al.*, "Multiparameter Respiratory Rate Estimation from the Photoplethysmogram," *IEEE TBME*, vol. 60, no. 7, pp. 1946–1953, 2013.

[11] Z. Zhang, Z. Pi, and B. Liu, "TROIKA: A General Framework for Heart Rate Monitoring Using Wrist-Type Photoplethysmographic Signals During Intensive Physical Exercise," *IEEE TBME*, vol. 62, no. 2, pp. 522–531, 2015.

[12] G. Biagetti, P. Crippa *et al.*, "Dataset from PPG wireless sensor for activity monitoring," *Data in Brief*, vol. 29, 2020.

[13] W. Karlen, J. M. Ansermino, and G. Dumont, "Adaptive Pulse Segmentation and Artifact Detection in Photoplethysmography for Mobile Applications," *IEEE EMBS*, vol. 2012, pp. 3131–3134, 2012.

[14] P. van Gent, H. Farah *et al.*, "Analysing Noisy Driver Physiology Real-Time Using Off-the-Shelf Sensors: Heart Rate Analysis Software from the Taking the Fast Lane Project," *JORS*, vol. 7, 2019.

[15] P. Dehkordi, A. Garde *et al.*, "Extracting Instantaneous Respiratory Rate From Multiple Photoplethysmogram Respiratory-Induced Variations," *Frontiers in Physiology*, vol. 9, p. 948, 2018.

[16] S. Fleming, M. Thompson *et al.*, "Normal ranges of heart rate and respiratory rate in children from birth to 18 years of age: a systematic review of observational studies," *The Lancet*, vol. 377, no. 9770, pp. 1011–1018, 2011.

[17] M. Kobayashi and T. Musha, "1/f fluctuation of heartbeat period," *IEEE TBME*, vol. 29, no. 6, pp. 456–457, 1982.

[18] T. Tamura, Y. Maeda *et al.*, "Wearable Photoplethysmographic Sensors—Past and Present," *Electronics*, pp. 282–302, 2014.

[19] C. Fischer, B. Domer *et al.*, "An Algorithm for Real-Time Pulse Waveform Segmentation and Artifact Detection in Photoplethysmograms," *IEEE J-BHI*, vol. 21, no. 2, pp. 372–381, 2017.

[20] A. Kamal, J. B. Harness *et al.*, "Skin photoplethysmography — a review," *Comp. Meth. Prog. Biomed.*, vol. 28, pp. 257–269, 1989.

[21] W. Zong, T. Heldt *et al.*, "An Open-Source Algorithm to Detect Onset of Arterial Blood Pressure Pulses," in *Computers in Cardiology*. IEEE, 2003, pp. 259–262.

[22] P. H. Charlton, T. Bonnici *et al.*, "An assessment of algorithms to estimate respiratory rate from the electrocardiogram and photoplethysmogram," *Physiol. Meas.*, vol. 37, no. 4, pp. 610–626, 2016.

[23] E. De Giovanni, S. Murali *et al.*, "Ultra-Low Power Estimation of Heart Rate Under Physical Activity Using a Wearable Photoplethysmographic System," in *DSD 2016*. IEEE, 2016, pp. 553–560.

[24] S. Lee, H. Shin, and C. Hahm, "Effective PPG sensor placement for reflected red and green light, and infrared wristband-type photoplethysmography." IEEE ICACT, 2016, pp. 556–558.

[25] D. Dao, S. M. A. Salehizadeh *et al.*, "A Robust Motion Artifact Detection Algorithm for Accurate Detection of Heart Rates From Photoplethysmographic Signals Using Time-Frequency Spectral Features," *IEEE J-BHI*, vol. 21, no. 5, pp. 1242–1253, 2017.

[26] A. Temko, "Accurate Heart Rate Monitoring During Physical Exercises Using PPG," *IEEE TBME*, vol. 64, no. 9, pp. 2016–2024, 2017.

[27] A. Chatterjee and U. K. Roy, "PPG Based Heart Rate Algorithm Improvement with Butterworth IIR Filter and Savitzky-Golay FIR Filter," in *IEEE IEMENTech*. IEEE, 2018, pp. 1–6.

[28] Q. Xie, Q. Zhang *et al.*, "Combining Adaptive Filter and Phase Vocoder for Heart Rate Monitoring Using Photoplethysmography During Physical Exercise," *IEEE EMBC*, pp. 3568–3571, 2018.

[29] H. Chung, H. Lee, and J. Lee, "Finite State Machine Framework for Instantaneous Heart Rate Validation Using Wearable Photoplethysmography During Intensive Exercise," *IEEE J-BHI*, vol. 23, no. 4, pp. 1595–1606, 2019.

[30] N. Huang and N. Selvaraj, "Robust PPG-based Ambulatory Heart Rate Tracking Algorithm," *IEEE EMBC*, vol. 2020, pp. 5929–5934, 2020.

[31] M. Wójcikowski and B. Pankiewicz, "Photoplethysmographic Time-Domain Heart Rate Measurement Algorithm for Resource-Constrained Wearable Devices and its Implementation," *Sensors*, vol. 20, no. 6, p. 1783, 2020.

[32] H. F. Posada-Quintero, D. Delisle-Rodríguez *et al.*, "Evaluation of pulse rate variability obtained by the pulse onsets of the photoplethysmographic signal," *Physiol. Meas.*, vol. 34, no. 2, pp. 179–187, 2013.

[33] E. Peralta Calvo, R. Bailón Luesma *et al.*, "Optimal Fiducial Points for Pulse Rate Variability Analysis from Forehead and Finger PPG Signals," *Physiol. Meas.*, vol. 40, no. 2, 2019.

[34] D. Chabot, M. Bayer, and A. de Roos, "Instantaneous heart rates and other techniques introducing errors in the calculation of heart rate," *Canadian Journal of Zoology*, vol. 69, no. 4, pp. 1117–1120, 1991.

[35] F. Wolling and K. Van Laerhoven, "Fewer Samples for a Longer Life Span: Towards Long-Term Wearable PPG Analysis," in *iWOAR '18*. ACM, 2018, pp. 1–10.

[36] D. Biswas, N. Simoes-Capela *et al.*, "Heart Rate Estimation From Wrist-Worn Photoplethysmography: A Review," *IEEE Sensors Journal*, vol. 19, no. 16, pp. 6560–6570, 2019.

[37] G. E. Billman, "Heart rate variability - a historical perspective," *Frontiers in Physiology*, vol. 2, p. 86, 2011.

[38] F. Shaffer and J. P. Ginsberg, "An Overview of Heart Rate Variability Metrics and Norms," *Frontiers in public health*, vol. 5, p. 258, 2017.

[39] G. Baselli, S. Cerutti *et al.*, "Heart rate variability signal processing: A quantitative approach as an aid to diagnosis in cardiovascular pathologies," *Int. J. Bio. Med. Comput.*, vol. 20, pp. 51–70, 1987.

[40] S. Ismail, U. Akram, and I. Siddiqi, "Heart rate tracking in photoplethysmography signals affected by motion artifacts: a review," *EURASIP*, vol. 2021, no. 1, pp. 1–27, 2021.

[41] P. H. Charlton, D. A. Birrenkott *et al.*, "Breathing Rate Estimation from the Electrocardiogram and Photoplethysmogram: A Review," *IEEE RBME*, vol. 11, pp. 2–20, 2018.

[42] A. A. Kamshilin and N. B. Margaryants, "Origin of Photoplethysmographic Waveform at Green Light," *Physics Procedia*, vol. 86, pp. 72–80, 2017.

[43] Y.-H. Kao, P. C.-P. Chao, and C.-L. Wey, "Design and Validation of a New PPG Module to Acquire High-Quality Physiological Signals for High-Accuracy Biomedical Sensing," *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 25, no. 1, pp. 1–10, 2019.

[44] N. Stuban and M. Niwayama, "Optimal filter bandwidth for pulse oximetry," *Rev. Sci. Instrum.*, vol. 83, no. 10, p. 104708, 2012.

[45] W. Karlen, M. Turner *et al.*, "CapnoBase: Signal database and tools to collect, share and annotate respiratory signals." STA, 2010, p. 25.

[46] Maxim Integrated, "MAXREFDES100#: Health Sensor Platform," https://www.maximintegrated.com/en/design/reference-design-center/system-board/6312.html, accessed: 2021-04-30.

[47] R. W. C. G. R. Wijshoff, M. Mischi *et al.*, "Reducing motion artifacts in photoplethysmograms by using relative sensor motion: phantom study," *Journal of Biomedical Optics*, vol. 17, no. 11, p. 117007, 2012.

[48] J. Pan and W. J. Tompkins, "A Real-Time QRS Detection Algorithm," *IEEE TBME*, vol. 32, no. 3, pp. 230–236, 1985.

[49] N. D. P. Ferreira, C. Gehin, and B. Massot, "Optical flow sensor as a reference for reduction of motion artefacts in photoplethysmographic measurements," *IEEE EMBC*, vol. 2020, pp. 4421–4424, 2020.

[50] Y. Liang, Z. Chen *et al.*, "A new, short-recorded photoplethysmogram dataset for blood pressure monitoring in China," *Scientific Data*, vol. 5, no. 180020, 2018.

[51] R. Cassani, A. Tiwari, and T. H. Falk, "Optimal filter characterization for photoplethysmography-based pulse rate and pulse power spectrum estimation," *IEEE EMBC*, vol. 2020, pp. 914–917, 2020.

[52] L. Bastos, D. Rosario, E. Cerqueira *et al.*, "Filtering Parameters Selection Method and Peaks Extraction for ECG and PPG Signals," in *IEEE LATINCOM*, 2019, pp. 1–6.