R-peak detector stress test with a new noisy ECG database reveals significant performance differences amongst popular detectors

Bernd Porr*& Luis Howell†

Abstract

The R peak detection of an ECG signal is the basis of virtually any further processing and any error caused by this detection will propagate to further processing stages. Despite this, R peak detection algorithms and annotated databases often allow large error tolerances around 10%, masking any error introduced. In this paper we have revisited popular ECG R peak detection algorithms by applying sample precision error margins. For this purpose we have created a new open access ECG database with sample precision labelling of both standard Einthoven I, II, III leads and from a chest strap. 25 subjects were recorded and filmed while sitting, solving a maths test, operating a handbike, walking and jogging. Our results show that using an error margin with sample precision, common R peak detection algorithms perform much worse than previously reported. In addition, there are significant performance differences between detectors which can have detrimental effects on applications such as heartrate variability, thus leading to meaningless results.

1 Introduction

Heartbeat detection is the first processing step when calculating heart parameters. The simplest and most straightforward measure is heart rate which can be calculated by taking the difference between two R peak timestamps, mathematically the operation of a derivative. However, derivatives amplify

*School of Engineering University of Glasgow Glasgow, United Kingdom bernd.porr@glasgow.ac.uk

†School of Engineering University of Glasgow Glasgow, United Kingdom luisbhowell@gmail.com

noise and thus errors in the R-peak detection will lead to an even larger error in the heart rate. Such error is further amplified when taking the 2nd derivative between successive heart rate readings for calculating heart rate variability (HRV). This calls for narrow error margins. However, to our knowledge the impact on temporal precision has not been investigated systematically (or statistically) for ECG detection. In this paper we are going to address this given its impact on any further processing.

In terms of detectors, the earliest ones by Pan and Tompkins [1985], Englese and Zeelenberg [1979] used a combination of digital filters and differentiators. In the 1990's and early 2000's a variety of new techniques were utilised: filter banks [Afonso et al., 1999], the Hilbert transform [Benitez et al., 2000] and wavelet transforms [Cuiwei Li et al., 1995, Martinez et al., 2004]. Most recently, machine learning has begun to be used [Yildirim, 2018].

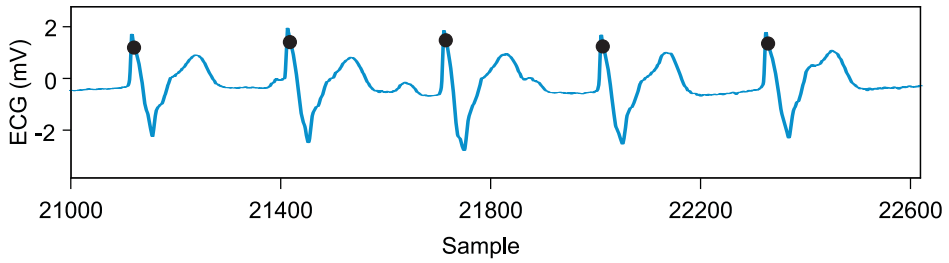

Figure 1: MITDB heartbeat annotations (dots) not located on R-peaks (record 107).

In order to benchmark a detector one needs an ECG database with annotated ECG R peaks. Almost all published research on heartbeat detection algorithms uses the MIT-BIH Arrhythmia Database (MITDB) [Goldberger et al., 2000, Moody and Mark, 2001] for testing. This database contains 48 ambulatory, 30-minute-long, annotated ECG recordings, 25 of which contain less common arrhythmias. The recordings have a sampling rate of 360 Hz with an 11 bit resolution over a 10 mV range. Although it has become the standard for detector evaluation, the almost exclusive use of this database poses an issue due to its two main shortcomings:

1. Very few examples of motion artefacts. The work by Benitez et al. [2000], Kalidas and Tamil [2017], Chan et al. [2017] tried to highlight noise resilience using only small sections ($3 - 10$ s) of a few select records. As these noisy sections make up such a small proportion of the database, they have very little impact on the overall results.

2. Inconsistent beat annotations which are often not located on the R-peak (Fig. 1). This poses a serious problem when benchmarking the

> detectors because it introduces a temporal jitter in the R-peak temporal time-stamps.

The gold standard of measuring R-peak detector performance is *sensitivity*. To calculate the sensitivity of the detectors one needs to specify a tolerance when comparing algorithm detections to the ground truth. The MIT Laboratory for Computational Physiology has a software library [Xie and Dubiel, 2018] for working with PhysioNet [Goldberger et al., 2000, Moody and Mark, 2001] databases such as the MITDB which internally specifies a tolerance and is widely used (often implicitly):

> The default tolerance is a tenth of the sampling rate.

At a sampling rate of 250 Hz this results in an error tolerance of 40 ms and at an average heart rate variability of 50 ms [Shaffer and Ginsberg, 2017], this renders any HRV readings meaningless.

Of the current thirty-three ECG databases on PhysioNet, two aim to provide examples of noisy ECG signals, however, both have flaws which limit their usefulness. The MIT-BIH Noise Stress Test Database (NSTDB) [Moody et al., 1984] features a 30-minute recording of noise typical of electrode motion artefacts and uses a script to add this on top of clean recordings from the MITDB. This has the advantage of being able to assess the effect of varying amounts of noise on the same signal. However, it has the disadvantages of using the inconsistent annotations of the MITDB and is not representative of a realistic recording as it is a *synthesised noisy signal* with a static level of noise. The other database, Motion Artefact Contaminated ECG Database [Behravan et al., 2015], consists of 27 recordings of a *single subject* standing, walking and performing a single jump. As this database features a small number of recordings from only *one subject* and is not annotated, it does not provide much use for detector evaluation.

In this paper we are presenting a new open access ECG database [Howell and Porr, 2018] from 25 subjects who performed different tasks such as sitting, performing a maths test, walking on a treadmill, using a hand-bike and jogging. ECGs were simultaneously recorded from both the Einthoven leads and from a chest strap. In addition, X/Y/Z acceleration and video footage was recorded to be able to attribute artefacts and which can potentially be used to counteract these. All ECGs were then annotated at sample precision allowing benchmarking of the different R-peak detectors at the highest possible precision. Having a database with increasing noise levels and strict timing requirements allows us to then determine which detector performs best and highlights the consequences of poor detection. This is particularly

relevant for applications such as heart rate variability where 2nd derivative quantities will be most susceptible to R peak jitter. This instructional and practical example will then inform us if sensitivity is actually a useful measure to evaluate R peak detectors.

The paper is structured as follows: first we describe the detectors, then we present the new database and finally we will benchmark the detectors against the different activities and recording techniques. As an application, we look into a measure of heart rate variability and determine how it is affected by two different detectors and recording techniques.

2 Methods

2.1 Glasgow University Database

The GUDB consists of two-minute ECG recordings from 25 subjects each performing five different tasks, for a total of 125 records. The tasks were chosen to be repeatable and representative of common, realistic scenarios. The tasks were as follows:

- sitting
- using a tablet to perform a maths test
- walking on a treadmill
- using a hand-bike
- jogging

Where the participant consented (24 out of the 25), a video synchronised to the data was recorded for each task. This video allows database users to see exactly how the movement was performed for each ECG recording and for any artefacts in the data to be identified. In addition the acceleration of the torso was recorded. With the exception of the most heavily noise contaminated records, all ECGs were annotated with high precision beat locations. The participants were all over the age of 18 and had no known cardiovascular conditions. The database is available through the University of Glasgow's open access research data repository [Howell and Porr, 2018]. The study was approved by the University of Glasgow ethics committee.

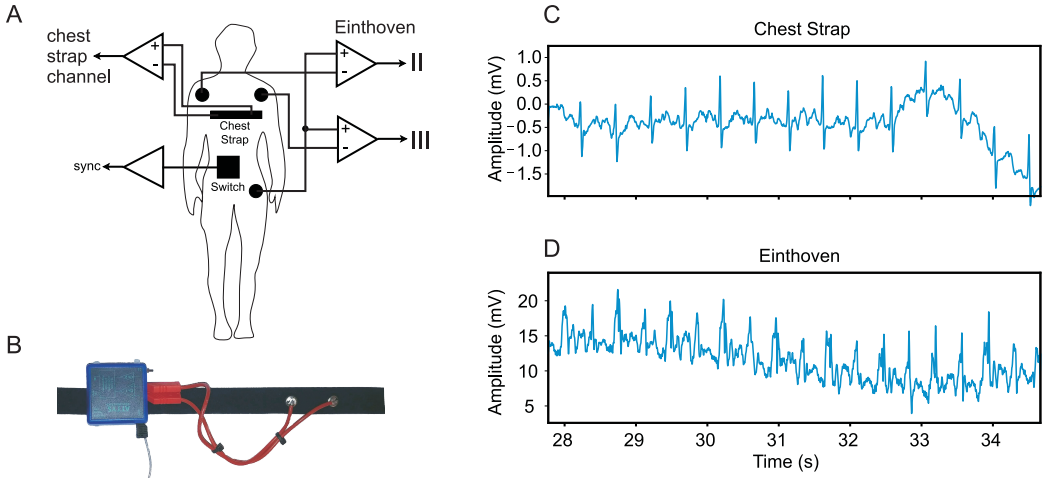

Figure 2: A) Wiring of the two two channel wireless biosignal amplifiers. B) Photo of the chest strap and wiring of the amplifier. The GND was placed on the shoulder. Comparison between chest strap (C) and Einthoven recording (D) while subject was jogging.

2.1.1 Equipment

The ECG signals were recorded using an Attys Bluetooth data acquisition board (Glasgow Neuro LTD, Glasgow). This board has a sampling rate of 250 Hz and a resolution of 24-bit over a range of ± 2.42 V. As this device is wireless, it increases electrical isolation and allows a moving subject to be recorded easily without the need of a cumbersome tether. The Attys features two analogue recording channels, one through an amplifier and the other through a differential amplifier. Two Attys were used at the same time to record different configurations representing a best and worst-case recording setup. This allows the impact of recording setup on signal noise to be investigated.

The best-case setup uses an Attys mounted on an elastic electrode chest strap (Amazon, UK), connected with short cables zip tied together (Fig. 2A,B). This configuration minimises the effect of cable movement artefacts as much as possible and is worn tightly on the subject to prevent the electrodes from moving. As the chest strap is worn high around the chest, the electrodes are approximately in the same location as V1 and V2 in the standard six electrode chest configuration [Macfarlane and Coleman, 1995]. The left electrode on the chest strap is connected to the positive terminal of the differential amplifier and the right electrode is connected to the negative. The GND terminal was connected to a silver chloride electrode (Pulse Medical, UK) on the right shoulder. The second channel of the Attys records the switch signal

used to synchronise the data with the video. The switch (Fig. 2A) is worn on a belt around the waist and when switched produces an audible click and shorts channel two to GND. The circuit diagram configuration can be seen in Fig. 2A.

The worse-case configuration uses the second Attys connected to standard ECG electrodes (Pulse Medical, UK) with loose cables. The positive terminal of the differential amplifier is connected to the left hip, the negative terminal to the right shoulder and the GND terminal to the left shoulder. The Attys is put into ECG mode where the positive terminal of the CH1 differential amplifier is connected internally to the CH2 amplifier input. This configuration allows two ECG signals to be recorded using only three cables. CH1 records Einthoven II between the left hip and right shoulder and CH2 records Einthoven III between the left hip and GND on the left shoulder. The circuit diagram for this configuration can be seen in Figure 2A.

Having introduced the the best and worse-case measurement situations, we show the corresponding raw signals as shown in Fig. 2C, D when the subject is jogging. The chest strap recording remains largely noise free while the Einthoven signal has significant noise contamination.

2.1.2 Protocol

Before the experiment begins, the participant will read the information sheet and sign the consent form (see Howell and Porr 2018), opting in or out of the video recording. The recording equipment is then connected to the participant and the experiment runs as follows:

1. 120 second ECG recording, sitting down
2. 120 second ECG recording, timed maths questions on a tablet
3. 120 second break
4. 120 second ECG recording, walking on a treadmill at 2 kph
5. 120 second break
6. 120 second ECG recording, using a hand bike
7. 120 second break
8. 120 second ECG recording, jogging on a treadmill at 7 kph
9. Electrodes and chest strap will be removed from participant

2.1.3 Post-processing

To annotate the data with heartbeat locations, a Python script was created which uses a Matplotlib [Hunter, 2007] interactive plot. An ECG data file is loaded into the plot and ran through a heartbeat detection algorithm (engzee segmenter from the BioSPPy library [Carreiras et al., 2018]) to provide an initial estimation of R-peak locations. This estimation is then manually inspected to remove any false positives and add any missing R-peaks. Where there was too much noise to reliably annotate the entire recording, no annotation file was made. Of the 125 recordings, 2 chest strap and 19 loose cable recordings were unable to be annotated, this mostly occurred in the jogging scenario. The annotation sample locations are saved to a .tsv file when the plot is closed. This is performed for both the chest strap ECG and the Einthoven II loose cable recording.

2.2 MITDB

The MITDB records are loaded using the WFDB Python library [Xie and Dubiel, 2018], only the first channel is used which is modified limb lead II (MLII). The MITDB annotations contain both beat labels and rhythm labels, using the PhysioBank annotation codes [PhysioNet, 2016], the rhythm labels are removed.

2.3 Detector Software Implementation

The algorithms chosen represent popular heartbeat detectors as well as a range of different techniques. The main criteria for selection was that the algorithm could be implemented in a real time system. All algorithms were implemented in Python, code can be found at Howell and Porr [2019].

2.3.1 Pan and Tompkins

The algorithm by Pan and Tompkins [1985] pre-processes the ECG signal before peak detection to reduce noise and emphasise the QRS complex (Fig. 3A-E). The first step is a bandpass filter with a passband of $5 - 15$ Hz, this will remove the DC offset, baseline wander, 50 Hz power line interference and reduce the amplitude of T-waves and some movement noise. This filter is implemented as a first order Butterworth IIR filter. After the bandpass filter, the signal is then differentiated, effectively a high-pass filter to highlight the sharp slopes of the QRS complex. To further emphasise the QRS complex, the signal is then squared, this also has the effect of making all points positive. The final processing step is a moving window average with a window of

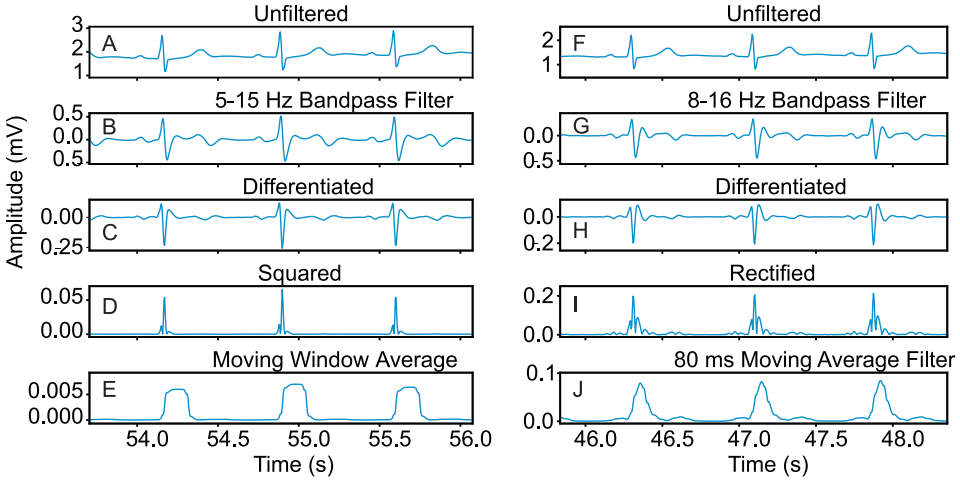

Figure 3: ECG processing steps for the Pan and Tompkins algorithm (A-E) and the Hamilton algorithm (F-J).

150 ms. This window was chosen to match the width of the widest possible QRS complex.

Peaks at least 300 ms apart are identified in the pre-processed signal (Fig. 3E) and classified as noise or a QRS complex depending on an adaptive threshold. This threshold is calculated as

$$NPKI = 0.125 \cdot PEAKI + 0.875 \cdot NPKI \quad (1)$$

$$SPKI = 0.125 \cdot PEAKI + 0.875 \cdot SPKI \quad (2)$$

$$THRESHOLD = NPKI + 0.25 \cdot (SPKI - NPKI) \quad (3)$$

where $PEAKI$ is the most recently detected peak of that type. $NPKI$ is an estimate of the current noise peak and $SPKI$ is an estimate of the current QRS peak. A peak is classified as a QRS complex if its amplitude is greater than that of the adaptive threshold otherwise it is classified as noise. To identify any QRS complexes which have potentially been missed, the average interval from the last nine QRS peaks is calculated. If the interval of a detected QRS peak is greater than 166 % of the average interval, it is assumed a QRS peak has been missed. To find this peak, the interval is scanned again with the threshold halved. If the newly detected peak exceeds the minimum interval criteria, it is added to the QRS peaks.

2.3.2 Hamilton

The method by Hamilton [2002] is based on the work by Pan and Tompkins [1985] (section 2.3.1) but makes alterations to the pre-processing stage

(Figure 3F-J). It differs by using a passband of $8 - 16$ Hz, rectifying the differentiated signal instead of squaring it and using an 80 ms moving average window over 150 ms.

QRS detection is based on three rules: Peaks must be at least 300 ms from the last detected R-peak, if a peak amplitude is above the detection threshold it is classified as a QRS complex and if the interval between two detected peaks exceeds 1.5 times the average RR interval, a QRS peak has been missed. Missed peaks must be at least half the detection threshold and occur at least 360 ms after the last detection to classify as a QRS complex. When a peak is detected, if its amplitude is greater than that of the detection threshold, its value is added to a buffer containing the last 8 QRS peak amplitudes. If its amplitude is lower, its value is added to a buffer containing the last 8 noise peak values. The mean of these buffers gives the average QRS peak and the average noise peak. Using these average peak values, the detection threshold for an R peak *thres* is calculated as:

$$thres = avg\ noise\ peak + 0.45 \cdot (avg\ QRS\ peak - avg\ noise\ peak) \quad (4)$$

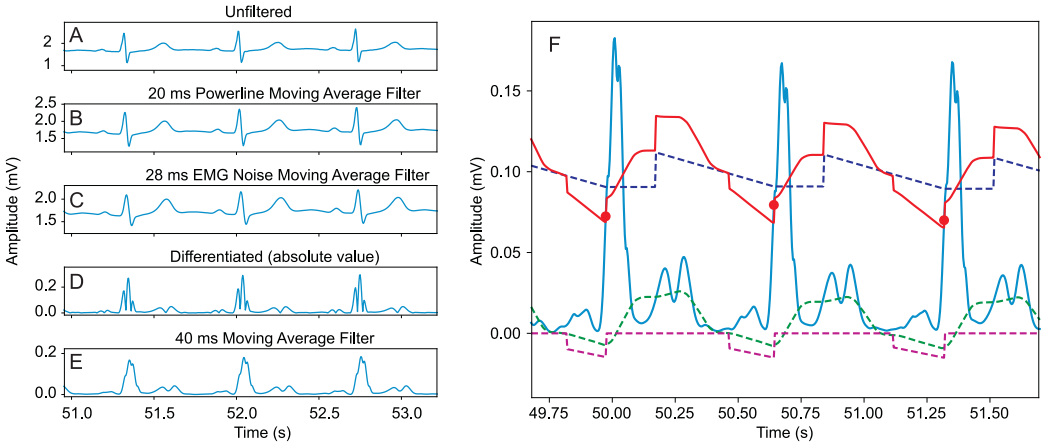

Figure 4: ECG processing steps for the Christov algorithm (A-E). F) illustrates the adaptive thresholds for the Christov algorithm on the pre-processed signal. Solid red: overall threshold MFR, dashed blue: steep-slope threshold M, dashed green: integrating threshold F, dashed magenta: beat expectation threshold R, red dot: QRS detection.

2.3.3 Christov

The method by Christov [2004] filters the ECG through a series of moving averages to pre-process the signal (Figure 4A-E), a combination of adaptive

thresholds is then used for QRS detection. The first moving average filter is used to remove any power-line interference. It is implemented as an FIR filter where the number of coefficients is equal to the number of samples in a 50 Hz period (20 ms). Each coefficient has the value of $\frac{1}{\text{number of coefficients}}$. The second moving average is implemented in the same way but with a period of 28 ms to reduce EMG noise. To emphasise the QRS complexes, the absolute value of the differential is then taken. A final moving average filter is used at a period of 40 ms to reduce noise amplified by the differentiation process.

Central to this algorithm is a sophisticated interplay of three thresholds: to identify QRS complexes, the pre-processed signal Y is compared to a threshold comprised of three independent adaptive thresholds: a steep-slope threshold M, an integrating threshold F and a beat expectation threshold R. We now define the different thresholds.

For the first five seconds of the signal, the steep-slope threshold, M, is calculated as $M = 0.6 \cdot max(Y)$. During this period multiple QRS complexes should be detected. After this initial period, a new steep-slope threshold is calculated as $0.6 \cdot max(Y)$ in the 200 ms after a QRS detection. In the 200 ms after a QRS detection no beat can be detected. This new value is added to a buffer of 5 values and the M threshold is calculated as the average of the buffer. In the interval $200 - 1200$ ms after a QRS complex has been detected, M is linearly decreased reaching a final value of 60 % of its starting value.

The purpose of the integrating threshold, F, is to increase the overall threshold when EMG noise is present in the signal, thereby reducing false positive detections. The integrating threshold is updated for every sample by adding the difference between the maximum value of Y in the most recent 50 ms and the maximum value of Y in the earliest 50 ms of the past 350 ms. A weighting coefficient of $\frac{1}{150}$ is used:

$$F = F + \frac{max(Y_{\text{latest 50 ms in 350 ms interv}}) - max(Y_{\text{earliest 50 ms in 350 ms interv}})}{150} \quad (5)$$

The beat expectation threshold, R, is intended to predict when a beat is going to occur and lower the threshold accordingly. When a beat is detected, the beat interval is stored in a buffer with four previous intervals and the buffer average, R_m, is calculated. In the interval between a detected beat and $+\frac{2}{3}R_m$, the beat expectation threshold is equal to zero. As a beat is expected in the period $\frac{2}{3}R_m$ to R_m, the R threshold decreases by 1.4 times less than the M threshold reduction. After the period R_m, the decrease stops.

The overall threshold, MFR, is calculated as the sum of the three independent thresholds, $MFR = M + F + R$ (Figure 4F).

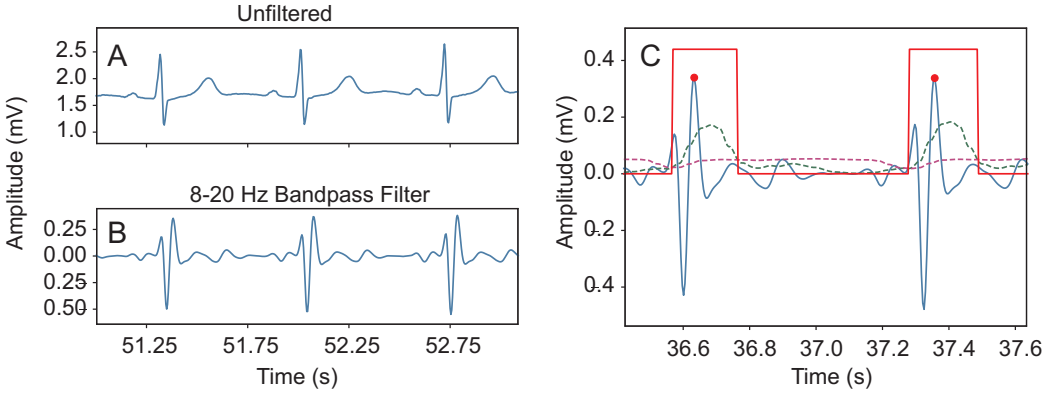

Figure 5: A) ECG processing steps of the algorithm by Elgendi et al. [2010]. B) R-peak detection using the two moving average detectors.

2.3.4 Elgendi et al

Elgendi et al. [2010] uses a bandpass filter and two moving averages to segment the ECG signal into blocks containing potential QRS complexes (Fig. 5). The ECG signal is filtered using a second order Butterworth IIR filter with a passband of $8 - 20$ Hz (Fig. 5B). The first moving average has a window of 120 ms to match the approximate duration of a QRS complex. A wider window of 600 ms is used for the second moving average to match the approximate duration of a complete heartbeat. Both moving averages are performed on the rectified bandpass filtered signal. Sections of the filtered ECG where the amplitude of the first moving average is higher than that of the second are marked as blocks containing a potential heartbeat (Figure 5C, red square wave). Blocks with a width of less than 80 ms are ignored as this is smaller than a QRS complex. The maximum value of the filtered ECG in each block is then stored as a detected QRS. Detections which follow the previous one by less than 300 ms are removed.

2.3.5 Kalidas and Tamil

The method by Kalidas and Tamil 2017, Fig. 6 is based on the work by Pan and Tompkins [1985] (section 2.3.1) but uses the Stationary Wavelet Transform (SWT) to remove noise and emphasise the QRS complexes instead of a bandpass filter (Fig. 6B). The Stationary Wavelet Transform is a method of decomposing a signal into uniform frequency bands using a mother wavelet. In this algorithm, the SWT is performed on the ECG signal using the Daubechies 3 wavelet (Fig. 6F). A SWT level of three was selected based on experimental detection results. As shown by the Fourier transform of the

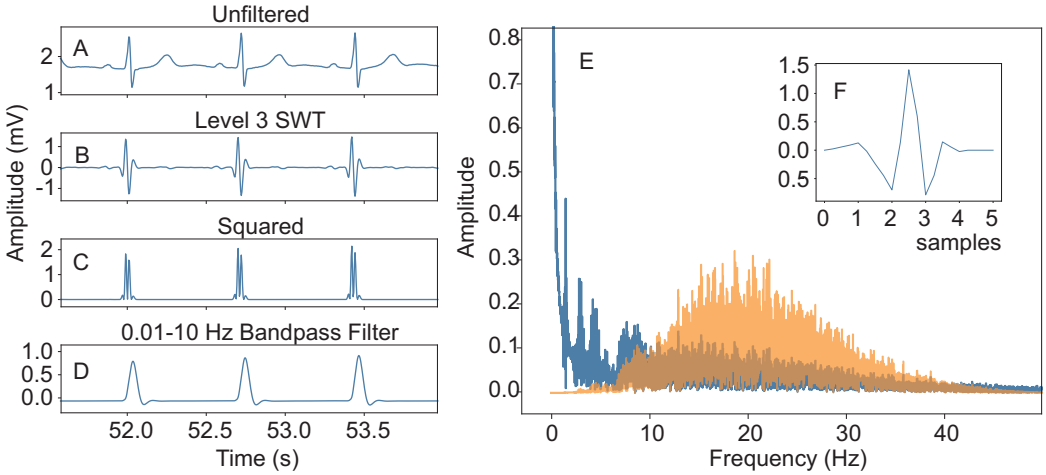

Figure 6: A-D) ECG processing steps for the algorithm by Kalidas and Tamil [2017] E) Frequency spectrum of 250 Hz sampling rate unfiltered ECG (blue) and ECG after level 3 SWT (orange). F) Daubechies 3 wavelet.

level 3 SWT (Fig. 6E), for a sampling rate of 250 Hz the frequency spectrum is centred around 20 Hz with a width of approximately 20 Hz, roughly matching the QRS frequencies. After the SWT has been calculated, detailed coefficients are extracted and then squared. In the original paper, a moving average was performed on the squared signal, however, it was found that using a bandpass filter instead significantly increased sensitivity and accuracy. Peak detection on the bandpass signal is then identical to that of Pan and Tompkins [1985].

2.3.6 Engelse and Zeelenberg with modifications by Lourenco et al (Engzee Mod)

Lourenço et al. [2012] expanded on the work by Englese and Zeelenberg [1979] by adapting their algorithm to work in real time and replacing the fixed threshold with an adaptive one. The first pre-processing step is an IIR band-stop filter at $48 - 52$ Hz to remove any powerline interference. The signal is then differentiated

$$y[n] = x[n] - x[n-4] \tag{6}$$

and passed through a five tap FIR windowed smoothing filter with the coefficients $h = [1, 4, 6, 4, 1]$.

QRS detection is based on two conditions, the first of which is peak detection. The adaptive threshold is based on the work by Christov [2004] and uses the steep-slope threshold, M. Refer to section 2.3.3 for how this threshold

is obtained. The second condition is that within 160 ms of a peak being detected there are at least 10 ms of consecutive points with an amplitude of less than $-M$. If this condition is met, the unfiltered signal is scanned in a window between where the peak was detected and after the 10 ms of consecutive points. The maximum value in this window is determined to be the R-peak.

2.3.7 Matched Filter

This method uses a matched filter to match an ECG signal to a QRS template, outputting a pulse for a detection. In our case a QRS template at both 250 and 360 Hz is chosen from motion free ECG recordings. The ECG recording is filtered before template selection to remove the DC offset and power-line interference. Each template is saved in a csv file.

For detection, the ECG signal is first filtered using a fourth order IIR bandpass filter with a passband of $0.1-48$ Hz to remove DC and 50 Hz power-line noise. Based on the sampling rate, the corresponding template is loaded and then time reversed. Using the time reversed template as coefficients, the pre-filtered ECG signal is processed using an FIR filter. The output of this FIR filter is then squared to increase the signal to noise ratio. The method by Pan and Tompkins [1985] (section 2.3.1) is performed on the squared signal to detect the R-peaks.

2.4 Evaluating Detectors

To evaluate the detectors, a script was used to compare each algorithm's detected R-peaks to the annotation locations. For the MITDB, all 48 records were tested. For the Glasgow University Database, all recordings with annotations were tested: 123 for the chest strap and 106 for the loose cable setup. A *detection tolerance of zero* was used for both databases, where only the exact annotation location was accepted as a true positive detection. Note that all detectors *delay* the signals so that their R-peak timestamps will be always later. In order to be able to do sample precision evaluation we subtracted the median delay from the true R peak. The median delay was calculated separately for every subject, activity and measurement protocol as this will be different each time. The performance of the detector algorithms was compared using sensitivity which is the proportion of the total number of R-peaks that were correctly detected:

$$Se = \frac{TP}{TP + FN} \tag{7}$$

where TP is the number of true R peak detections and FN is the number of false negative ones.

The sensitivities for sitting, jogging, Einthoven, chest strap and the different detectors are now compared with the help of the Wilcoxon rank test. Valid p values need be based on at least 20 different sensitivity pairs comparing two conditions. Alpha level is set to 0.05.

2.4.1 Heart rate variability

To have a practical application for precise R peak detection and to test how the results above impact on a real application we calculated the heart rate variability for the conditions "sitting" and "maths test". We test if the maths test significantly impacts on the normalised root mean square differences of successive heart rates (RMSSD). This is a popular measure in HRV analysis and is based on the 2nd derivative of the R peak time stamps and thus will heavily rely on a precise timing with any jitter changing the result. We use the normalised RMSSD (nRMSSD) to eliminate its dependency on the heartrate. We test "sitting" against "maths" nRMSSD with the help of the Wilcoxon rank test for the three conditions:

1. Ground truth: nRMSSD based on labelled R peaks
2. Wavelet detector [Kalidas and Tamil, 2017] based R timings and resulting nRMSSD
3. Lourenço et al. [2012] detector derived R timings and resulting nRMSSD

We apply these two detectors to both chest strap and Einthoven so that we have five different comparisons where an alpha level of 5% indicates which ones are significant. The statistical test is the Wilcoxon rank test which counts the occurrences when the nRMSSD is larger or smaller between the nRMSSD readings from the different subjects. Note that this exercise is intended to show different noise levels in a realistic experimental setting where sitting still will introduce less noise than doing the maths test on the tablet. If maths really stresses out a person or rather sitting still in a lab environment is not of relevance here.

3 Results

3.1 Glasgow University Database

Figure 7 shows the results of the University of Glasgow Database for sitting (A) and jogging (B) comparing the different detectors and the two electrode

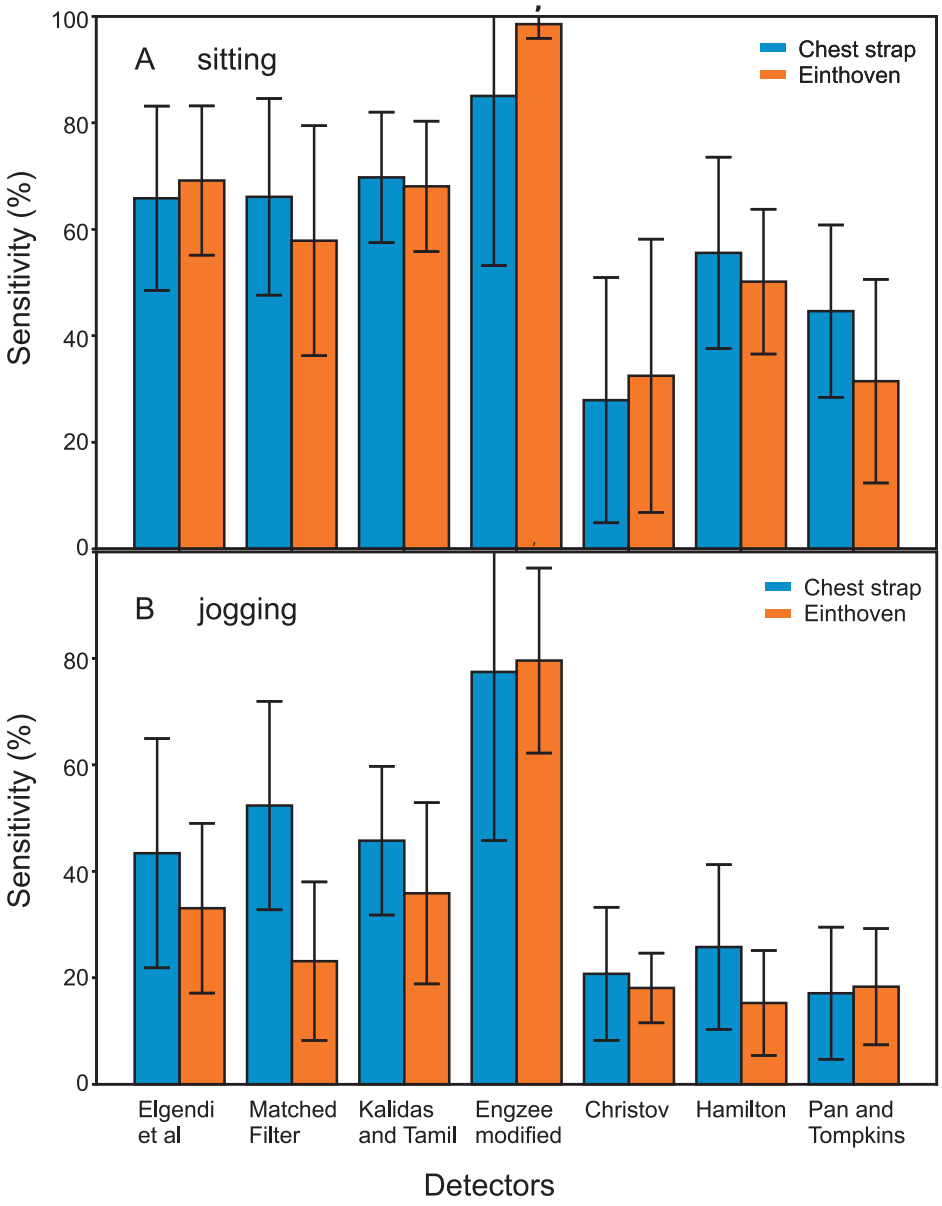

Figure 7: Detector results on the Glasgow University Database for chest strap and Einthoven recordings while sitting (A) and jogging (B).

configurations: chest strap with V2-V1 and Einthoven II. Qualitatively it can be seen that the detectors by Engzee, Elgendi, Kalidas and the matched filter perform best while Christov, Hamilton and Pan Tompkins perform worst. Amongst the better performing detectors the Engzee detector stands out with close to 100 % sensitivity. This is the only detector which performs a peak search on the ECG signal and is thus more computationally expensive

than the other ones. We are now investigating the different detectors by comparing their sensitivities to find significant differences. The following statistical tests have been done by comparing pairs of sensitivities with the Wilcoxon rank test where the minimal number of pairs had to be 20. Under Einthoven/jogging this condition could not be reached and the results have been excluded. Similarly the detector by Christov has been performing so badly that not enough sensitivity pairs were available. A p-value of less than 0.05 indicates that the sensitivities are significantly different. All the values in the tables below have been obtained with the Wilcoxon rank test.

3.1.1 Sitting vs jogging

As a first test we had a look at how noise from jogging compared to sitting degrades the sensitivity of the detector. Because Einthoven was too noisy we only compare the results from the chest strap. These are the p-values testing for a significant differences between sitting vs jogging:

Elgendi	Matched	Kalidas	Engzee	Christov	Hamilton	Pan
0.00*	0.03*	0.00*	0.02*	—	0.00*	0.00*

Overall there is a significant difference between just sitting and jogging. All detectors perform significantly worse while jogging. Note that the Christov detector has performed so badly that sensitivity readings could not be calculated for most subjects.

3.1.2 Sitting

The lowest noise levels are expected to be while sitting on a chair. We are going to first compare the Einthoven leads with the chest strap and then compare the detectors for both Einthoven and the chest strap.

Einthoven vs chest strap As a first step we investigate how the electrode configurations (Einthoven vs chest strap) alter the sensitivity and check which detectors are most susceptible to it:

Elgendi	Matched	Kalidas	Engzee	Christov	Hamilton	Pan
0.57	0.17	0.64	0.06	—	0.08	0.01*

The p-values above show that there is no significant difference between an Einthoven recording and that from a chest strap – apart from the detector by Pan and Tompkins [1985]. This is expected as during sitting the muscle and movement artefacts will be low and thus even the Einthoven recording with its loose cables has no negative impact.

Einthoven II Let's now compare the different detectors against each other when recording Einthoven II while sitting:

	Elgendi	Matched	Kalidas	Engzee	Christov	Hamilton	Pan
Elgendi	—	0.07	0.75	0.00*	—	0.00*	0.00*
Matched	0.07	—	0.10	0.00*	—	0.01*	0.00*
Kalidas	0.75	0.10	—	0.00*	—	0.00*	0.00*
Engzee	0.00*	0.00*	0.00*	—	—	0.00*	0.00*
Christov	—	—	—	—	—	—	—
Hamilton	0.00*	0.01*	0.00*	0.00*	—	—	0.07
Pan	0.00*	0.00*	0.00*	0.00*	—	0.07	—

The above p-values confirm what has been already observed qualitatively: the detectors by Pan and Tompkins [1985] and Hamilton [2002] are significantly worse than our best performers, in particular Engzee [Lourenço et al., 2012] but also Elgendi et al. [2010], machted filter and Kalidas and Tamil [2017].

Chest strap Instead of using the standard Einthoven leads one can use the chest strap offering less noise, even while sitting.

	Elgendi	Matched	Kalidas	Engzee	Christov	Hamilton	Pan
Elgendi	—	0.84	0.38	0.03*	—	0.07	0.00*
Matched	0.84	—	0.35	0.02*	—	0.10	0.00*
Kalidas	0.38	0.35	—	0.02*	—	0.00*	0.00*
Engzee	0.03*	0.02*	0.02*	—	—	0.00*	0.00*
Christov	—	—	—	—	—	—	—
Hamilton	0.07	0.10	0.00*	0.00*	—	—	0.07
Pan	0.00*	0.00*	0.00*	0.00*	—	0.07	—

The p-values above reveal that at a sensitivity of just 80 % the detector by Engzee [Lourenço et al., 2012] is still significantly better than any other detector. Less well performing at about 65 % are Elgendi et al. [2010], the matched filter and Kalidas and Tamil [2017]. There is no significant difference between these three.

3.2 Jogging

We now move on to jogging and investigate how noise impacts on the performance of the detectors.

3.2.1 Einthoven

Most of the Einthoven recordings while jogging are too noisy to allow any labelling of the R peaks and, thus, we have no statistical results. Note that we could have taken the labelling from the chest strap as a fall-back however we decided against it because of the substantially different shapes of the R-peaks between chest and Einthoven leads.

3.2.2 Chest strap

As shown in the introduction, the chest strap offers much less noise and the p-values comparing the different detectors are:

	Elgendi	Matched	Kalidas	Engzee	Christov	Hamilton	Pan
Elgendi	—	0.01*	0.36	0.00*	—	0.00*	0.00*
Matched	0.01*	—	0.00*	0.01*	—	0.00*	0.00*
Kalidas	0.36	0.00*	—	0.00*	—	0.01*	0.00*
Engzee	0.00*	0.01*	0.00*	—	—	0.00*	0.00*
Christov	—	—	—	—	—	—	—
Hamilton	0.00*	0.00*	0.01*	0.00*	—	—	0.04*
Pan	0.00*	0.00*	0.00*	0.00*	—	0.04*	—

While the detector by Engzee [Lourenço et al., 2012] is significantly better than any other detector it shows that with a standard deviation of 50 % it is not usable for serious applications. The next section about heart rate variability will test the difference between Engzee [Lourenço et al., 2012] and Kalidas and Tamil [2017] in a real application.

Summary of the GUDB results In summary, the Engzee detector [Lourenço et al., 2012] performs significantly better than the other detectors. In the mid-field we have Elgendi et al. [2010], Kalidas and Tamil [2017] and the matched filter. The detectors by Pan and Tompkins [1985] and Hamilton [2002] are significantly worse compared to the other ones. The stellar performance of the Engzee detector [Lourenço et al., 2012] might have its origin in its precise R peak search which is good in noise free environments but risky in noisy ones. However, the mid field seems to be less susceptible to noise because they bandpass filter the ECG prior to detection.

Comparing the chest strap and loose cable results, it becomes clear that the recording setup is as important as the detector algorithm used. If the subject is going to remain stationary for the entirety of the recording, the Einthoven II setup will provide good results. However, if any movement is planned, it is crucial to choose a setup which reduces cable movement, an electrode chest strap has been shown to be very effective in this regard.

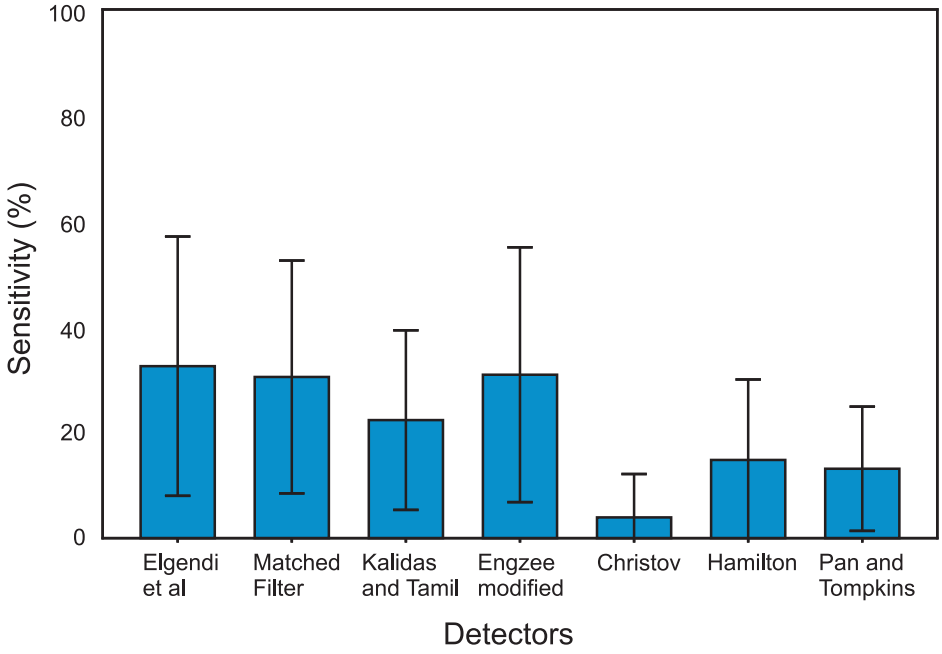

Figure 8: Detector sensitivities on the MITDB.

3.3 MITDB

As predicted, when using an evaluation tolerance of 0, the inaccurate annotations of the MITDB cause a reduction in sensitivity (Figure 8). However, the size of this reduction was far greater than expected with no detector achieving over 32% sensitivity.

A detailed comparison between the detectors reveals consistently with our database that Elgendi et al. [2010] and the matched filter perform significantly best. The wavelet detector [Kalidas and Tamil, 2017] here is significantly worse than Elgendi et al. [2010] but not compared to the matched filter:

	Elgendi	Matched	Kalidas	Engzee	Christov	Hamilton	Pan
Elgendi	—	0.39	0.03*	0.91	0.00*	0.00*	0.00*
Matched	0.39	—	0.17	0.84	0.00*	0.00*	0.00*
Kalidas	0.03*	0.17	—	0.04*	0.00*	0.03*	0.01*
Engzee	0.91	0.84	0.04*	—	0.00*	0.00*	0.00*
Christov	0.00*	0.00*	0.00*	0.00*	—	0.02*	0.03*
Hamilton	0.00*	0.00*	0.03*	0.00*	0.02*	—	0.91
Pan	0.00*	0.00*	0.01*	0.00*	0.03*	0.91	—

Overall the results from the MIT database confirm our results although on a much lower sensitivity level, probably because of the erroneous annotations.

However, overall Elgendi et al. [2010], the matched filter, the stationary wavelet detector [Kalidas and Tamil, 2017] and Engzee [Lourenço et al., 2012] appear again as the best performing detectors.

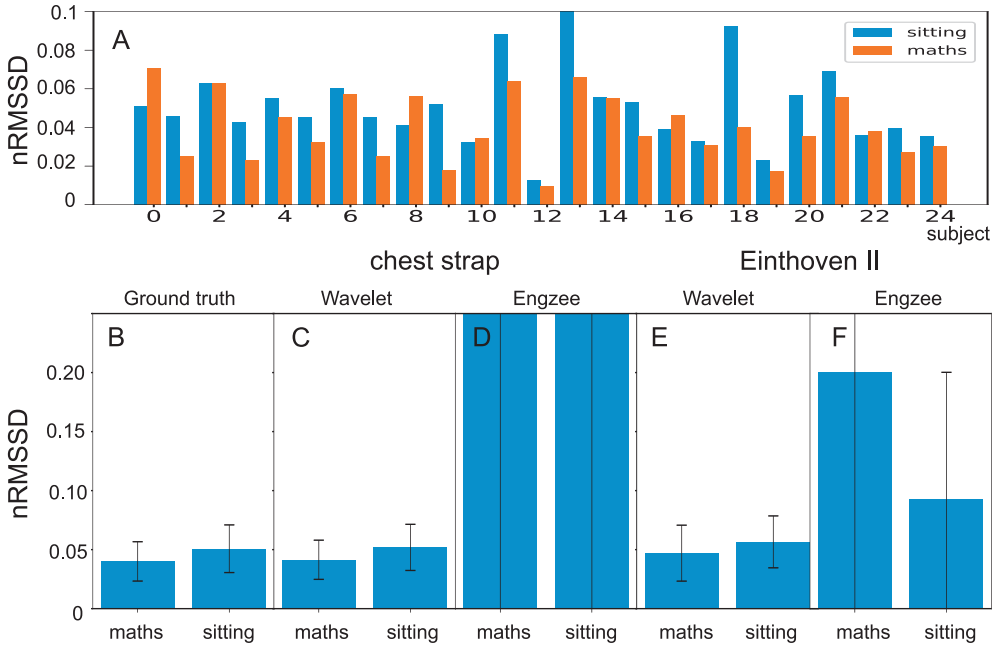

Figure 9: Heartrate variability test between sitting and maths test using different detectors and both the chest leads and Einthoven. A) normalised RMSSD for every subject for the maths test and just sitting B) Ground truth between maths test and sitting. C) Chest strap ECG analysed with the wavelet detector. D) Chest strap ECG analysed with the Pan Tompkins detector. E) Einthoven II analysed with the wavelet detector. F) Einthoven II analysed with the Pan Tompkins detector. None of the maths vs sitting tests are significant but the nRMSSDs of the Pan Tompkins based results are significantly different to both the wavelet detector results and the ground truth.

3.4 Heart rate variability

Fig. 9 shows the results of the heart rate variability test where panel A shows the individual nRMSSD readings from the different subjects for both sitting and maths. These were calculated using the annotated sample-precise time stamps and are thus the ground truth which will be compared to the different nRMSSDs obtained from different detectors. At an alpha of 5%, this ground truth reveals a significant difference of the nRMSSDs between sitting and maths test (Fig. 9B).

Now let's look at the results using the actual detectors. When using a chest strap (Fig. 9C-D), the wavelet detector [Kalidas and Tamil, 2017] reproduces the slight decrease of the nRMSSD during the maths test and this is also significant. Also the results of the wavelet detector based nRMSSDs are significantly identical to the ground truth. However, the Engzee detector [Lourenço et al., 2012] has significant problems against the ground truth and generates very large nRMSSD values and thus cannot be recommended. This is most likely due to the way the Engzee detector determines the final R peak position namely trying to find the true maximum in the ECG signal, most likely ending up in local maxima within its search window.

When using the Einthoven leads, the wavelet detector [Kalidas and Tamil, 2017] performs perfectly by also being significant between maths test and sitting and generating significantly identical nRMSSD readings for both sitting and maths test. However, the Engzee detector [Lourenço et al., 2012] fails again.

This shows that sensitivity readings themselves need to be taken with caution: in the previous section the Engzee detector was clearly standing out as the best detector at about 99 % sensitivity. However, here we see that the detector using the stationary wavelet transform with about 60 % sensitivity provides much better performance. Most likely the reason is that the wavelet transform *very effectively removes noise* because of optimal bandpass filtering while the Engzee algorithm operates on the unfiltered ECG to detect the R peak and thus is highly susceptible to noise.

4 Discussion

This work was undertaken in response to the apparent lack of research into how realistic movement noise effects the performance of ECG heartbeat detection algorithms. To provide a dataset for evaluating the detectors and a resource for future research, an open access database of ECG recordings was created [Howell and Porr, 2018]. This database consists of 125 recordings of realistic scenarios, the majority of which have been videoed and annotated with high precision. Seven real-time heartbeat detection algorithms were implemented in Python to be evaluated, representing a range of popular techniques [Howell and Porr, 2019].

In addition to evaluating these algorithms on the new database, they were also tested on the most popular existing database, the MIT-BIH Arrhythmia Database [Goldberger et al., 2000, Moody and Mark, 2001]. This database has two main deficiencies: very few examples of noisy ECGs and inaccurate annotations. These inaccurate annotations produced very low sensitivity

results, below 35% when using a detection tolerance of 0. This is particularly troubling when almost all research papers on the subject report sensitivities upwards of 99%. Almost perfect results like these using the MIT database were all obtained using large tolerances, producing misleading and unreliable results. This, and the lack of noisy ECGs make a strong argument for the need for a new standard ECG database.

When the same detectors were tested on the new database, the majority were able to achieve above 60% sensitivity as a result of the high precision annotations. A comparison of the two recording setups used in the new database showed that an electrode chest strap is a very effective method of reducing noise while the subject is moving. Using the chest strap, the reduction in performance as a result of movement noise is minor.

Evaluating the algorithms across the two databases revealed that when the ECG signal is ideal, there is very little difference in performance between the detectors. The performance of the detectors then drops and diverges as the ECG signal becomes noisy. This highlights the significant opportunity for development and further research in this area. It is hoped that the database created here can help spur this future development by providing a high-quality dataset for testing.

Table 1: Comparison of claimed and tested sensitivities for the MITDB

Detector	Cited Tolerance (ms)	Stat. tests	Cited MITDB Sens-itivity (%)	Our MITDB Sens-itivity (%)
Pan and Tompkins [1985]	Not stated	No	99.30	12.85
Hamilton [2002]	Not stated	No	99.80	13.86
Elgendi et al. [2010]	Not stated	No	98.31	30.50
Matched Filter	N/A	N/A	N/A	27.95
Lourenço et al. [2012]	10% of RR	No	96.50*	28.92
Christov [2004]	60+	No	99.69	3.92
Kalidas and Tamil [2017]	Not stated	No	99.88	21.86

*own ECG database and performance not measured against annotations

Table 1 shows the results reported by different original publications of R peak detectors. All but one have used the MIT-BIH database for benchmarking. We will quickly revisit the different results.

Pan and Tompkins [1985] were one of the first who developed a real-time

QRS detection algorithm which was written in assembly language on a Z80 microprocessor to be able to cope with the filtering demands. The performance was analysed by playing tapes of ECG recordings from the MIT/BIH database which were turned into digital signals, processed in the Z80 processor and then turned back into analogue signals for comparison to the annotations. Coincidence between the detected R peak and the annotations was done by visual inspection so that a precise jitter tolerance couldn't be given. Overall Pan and Tompkins [1985] reports a sensitivity of 99.3 % which is in stark contrast to our findings of 13.13 %.

Hamilton [2002] had in mind an open source R peak detector with a high sensitivity which is a further development of the detector proposed by Pan and Tompkins [1985]. In order to be able to run it even on a microcontroller, Hamilton [2002] also presented a stripped down version of their full detector implementation, omitting certain QRS detection rules. For their original detector they report a sensitivity of 99.74 % and for the microcontroller version one of 99.80 %. The sensitivities were not compared to other detectors nor a statistical analysis has been performed. The authors state that the detectors perform comparably. The jitter tolerance is not mentioned in the paper. However we report a sensitivity of just 14.79 % which is virtually identical to the original algorithm by Pan and Tompkins [1985].

The work by Elgendi et al. [2010] aimed to optimise the bandpass filter frequencies used for the R-peak detection. Different frequency bands were benchmarked and the 8-20 Hz band was chosen to be optimal having the highest sensitivity. However all sensitivities, even the sub-optimal ones vary only between 92.66% and 98.31% where some deviate less than one percent. There is no statistical analysis in the paper stating which band is significantly better than another one. A temporal tolerance is not given. However, compared to our result at zero jitter tolerance at 32.53 % their tolerance used must have been substantially larger reporting such high sensitivity values. Given the lack of statistical tests in the paper and the omission of the tolerance render the recommendations towards cut-off frequencies questionable.

Christov [2004] starts off not just with Einthoven I but creates a complex lead combining the derivatives from all standard 12 leads. This is then sent through various thresholds to detect the R peaks. Christov [2004] achieved a sensitivity of 99.96 % at a tolerance of 60 ms with larger temporal differences allowed – if approved by an independent expert. However our results at sample precision tolerance yields just 3.92 %. The article has no discussion section and thus does not compare its results to other approaches and has no statistical evaluations. In the conclusion section the author claims that: "The statistical indices are higher than, or comparable to those, cited in the scientific literature." [Christov, 2004].

Similar to Christov [2004], taking the derivative of the ECG signal is the central idea by Englese and Zeelenberg [1979] which is then turned into a real-time version by Lourenço et al. [2012] using an adaptive threshold. The paper then compares the real-time version with the previous off-line version and the detector by Christov [2004]. This paper does not use the MIT-DB database because the authors recorded their own ECGs. They also devised a different performance measure: the deviation from the mean RR interval needs to be less than 10 % for individual RR pairs or in other words: the heartrate needs to be within normal limits of the resting heartrate variability [Shaffer and Ginsberg, 2017]. Their online algorithms lead to average valid RR intervals between 84.5 % and 96.5 % depending on on- or offline algorithms, use of electrodes, the algorithm itself and filtering. These high readings are expected because their criterion will most likely detect only crude deviations from RR intervals, for example a missed beat which then results in twice the length or an additional spurious detection which results in a very short RR interval. Even more important is that this measure does not compare against the ground truth of RR intervals and only looks at the self consistency of the RR intervals – which might have been wrong in relation to the annotations in the 1st place. However, sensitivity actually compares ground truth with the detector output. Here, we have a similar sensitivity to other detectors at 28.92 %. There is no statistical test which of these results differ significantly but could have been easily performed with the reasonably large number of subjects ($N = 62$).

In contrast to any other paper above Kalidas and Tamil [2017] employed the stationary wavelet transform to filter the ECG which resulted in a high sensitivity in their paper of 99.88 %. However, also the cited sensitivities of their competitors are in the 99 % band and thus render the comparison useless. Neither a statistical significance test nor the detection tolerance is published. However, given our calculated sensitivity of 22.34 % compared to their 99.88 %, it suggests that again a large temporal jitter was permitted, however the exact margin is not mentioned in the text.

Overall, with virtually every paper reporting very high sensitivities of 98 % or more is not helpful at all and have only been possible because of high temporal tolerances of probably 100 ms or even more. In addition the sole use of the MIT-BIH database with mostly artefact free ECGs again overestimate the performance of these algorithms. These factors combined grossly overestimate the performances of the algorithms and thus making a comparison impossible.

The results from the heartrate variability example show that sensitivity itself is a poor measure of performance because it is just a binary value if an R-peak has been detected within a time frame or not. However, given

that both heartrate and heartrate variability operate on the 1st and 2nd derivatives of the R peak timestamps a new performance measure should reflect the *temporal jitter* between the R peak time stamps, for example by measuring the standard deviation between the true R peak positions and the actual R peak positions.

References

V.X. Afonso, W.J. Tompkins, T.Q. Nguyen, and Shen Luo. ECG beat detection using filter banks. *IEEE Transactions on Biomedical Engineering*, 46(2):192–202, 1999. ISSN 00189294. doi: 10.1109/10.740882. URL http://ieeexplore.ieee.org/document/740882/.

Vahid Behravan, Neil E. Glover, Rutger Farry, Patrick Y. Chiang, and Mohammed Shoaib. Rate-adaptive compressed-sensing and sparsity variance of biomedical signals. In *2015 IEEE 12th International Conference on Wearable and Implantable Body Sensor Networks (BSN)*, pages 1–6. IEEE, jun 2015. ISBN 978-1-4673-7201-5. doi: 10.1109/BSN.2015.7299419.

D.S. Benitez, P.A. Gaydecki, A. Zaidi, and A.P. Fitzpatrick. A new QRS detection algorithm based on the Hilbert transform. *Computers in Cardiology*, 27:379–382, 2000. doi: 10.1109/CIC.2000.898536. URL http://ieeexplore.ieee.org/document/898536/.

C Carreiras, AP Alves, A Lourenço, F Canento, H Silva, and A Fred. BioSPPy - Biosignal Processing in Python, 2018. URL https://github.com/PIA-Group/BioSPPy/.

Hsiao-Lung Chan, Fu-Tai Wang, Yi-Sheng Lee, and Chun-Li Wang. Heartbeat Detection Using Oscillatory Envelope Pattern in Noisy Electrocardiogram. *Computing in Cardiology*, pages 1–4, sep 2017. doi: 10.22489/CinC.2017.086-353. URL http://www.cinc.org/archives/2017/pdf/086-353.pdf.

Ivaylo I Christov. Real time electrocardiogram QRS detection using combined adaptive threshold. *BioMedical Engineering OnLine*, 3(1): 28, aug 2004. ISSN 1475925X. doi: 10.1186/1475-925X-3-28. URL http://biomedical-engineering-online.biomedcentral.com/articles/10.1186/1475-925X-3-28.

Cuiwei Li, Chongxun Zheng, and Changfeng Tai. Detection of ECG characteristic points using wavelet transforms. *IEEE Transactions on Biomedical Engineering*, 42(1):21–28, 1995. ISSN 00189294. doi: 10.1109/10.362922. URL http://ieeexplore.ieee.org/document/362922/.

Mohamed Elgendi, Mirjam Jonkman, and Friso Deboer. Frequency Bands Effects on QRS Detection. In *Biomedical Engineering Systems and Technologies*, pages 428–431. Springer, 2010. URL http://www.elgendi.net/papers/QRS{_}Bands{_}final.pdf.

W Englese and C Zeelenberg. A single scan algorithm for QRS-detection and feature extraction. *Computers in Cardiology*, 6:37–42, 1979.

A L Goldberger, L A Amaral, L Glass, J M Hausdorff, P C Ivanov, R G Mark, J E Mietus, G B Moody, C K Peng, and H E Stanley. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. *Circulation*, 101(23):E215–20, jun 2000. ISSN 1524-4539.

P. Hamilton. Open source ECG analysis. In *Computers in Cardiology*, pages 101–104, Memphis, TN, USA, 2002. IEEE. ISBN 0-7803-7735-4. doi: 10.1109/CIC.2002.1166717. URL http://ieeexplore.ieee.org/document/1166717/.

Luis Howell and Bernd Porr. High precision ECG Database with annotated R peaks, recorded and filmed under realistic conditions, 2018. URL http://researchdata.gla.ac.uk/716/.

Luis Howell and Bernd Porr. Popular ECG R peak detectors written in python, August 2019. URL https://doi.org/10.5281/zenodo.3357365.

John D. Hunter. Matplotlib: A 2D Graphics Environment. *Computing in Science & Engineering*, 9(3):90–95, 2007. ISSN 1521-9615. doi: 10.1109/MCSE.2007.55. URL http://ieeexplore.ieee.org/document/4160265/.

Vignesh Kalidas and Lakshman Tamil. Real-time QRS detector using Stationary Wavelet Transform for Automated ECG Analysis. In *2017 IEEE 17th International Conference on Bioinformatics and Bioengineering (BIBE)*, pages 457–461. IEEE, oct 2017. ISBN 978-1-5386-1324-5. doi: 10.1109/BIBE.2017.00-12. URL http://ieeexplore.ieee.org/document/8251332/.

André Lourenço, Hugo Silva, Paulo Leite, Renato Lourenço, and Ana Fred. Real Time Electrocardiogram Segmentation for Finger Based ECG Biometrics. In *Biosignals 2012*, pages 49–54, 2012.

P W Macfarlane and E N Coleman. RESTING 12-LEAD ECG ELECTRODE PLACEMENT AND ASSOCIATED PROBLEMS. Technical report, The Society for Cardiological Science and Technology, 1995. URL http://www.scst.org.uk/resources/RESTING{_}12.pdf.

J.P. Martinez, R. Almeida, S. Olmos, A.P. Rocha, and P. Laguna. A Wavelet-Based ECG Delineator: Evaluation on Standard Databases. *IEEE Transactions on Biomedical Engineering*, 51(4):570–581, apr 2004. ISSN 0018-9294. doi: 10.1109/TBME.2003.821031. URL http://ieeexplore.ieee.org/document/1275572/.

G. B. Moody and R. G. Mark. The impact of the mit-bih arrhythmia database. *IEEE engineering in medicine and biology magazine : the quarterly magazine of the Engineering in Medicine & Biology Society*, 20(3):45–50, May/Jun 2001. ISSN 0739-5175. URL http://www.ncbi.nlm.nih.gov/pubmed/11446209.

GB Moody, WE Muldrow, and RG Mark. A noise stress test for arrhythmia detectors. *Computers in Cardiology*, 11:381–384, 1984. URL https://physionet.org/physiobank/database/nstdb/.

Jiapu Pan and Willis J. Tompkins. A Real-Time QRS Detection Algorithm. *IEEE Transactions on Biomedical Engineering*, BME-32(3):230–236, mar 1985. ISSN 0018-9294. doi: 10.1109/TBME.1985.325532. URL http://ieeexplore.ieee.org/document/4122029/.

PhysioNet. PhysioBank Annotations, 2016. URL https://www.physionet.org/physiobank/annotations.shtml.

Fred Shaffer and J. P. Ginsberg. An overview of heart rate variability metrics and norms. *Frontiers in public health*, 5:258, Sep 2017. ISSN 2296-2565. doi: 10.3389/fpubh.2017.00258. URL http://www.ncbi.nlm.nih.gov/pubmed/29034226.

Chen Xie and Julien Dubiel. WFDB Python, 2018. URL https://github.com/MIT-LCP/wfdb-python.

Özal Yildirim. A novel wavelet sequence based on deep bidirectional lstm network model for ecg signal classification. *Computers in biology and medicine*, 96:189–202, May 2018. ISSN 1879-0534. doi: 10.1016/j.compbiomed.2018.03.016. URL http://www.ncbi.nlm.nih.gov/pubmed/29614430.