Deterioration of R-Wave Detection in Pathology and Noise: A Comprehensive Analysis Using Simultaneous Truth and Performance Level Estimation

Muhammad Kashif*, Stephan M. Jonas, and Thomas M. Deserno, *Senior Member, IEEE*

Abstract*— *Objective: **For long-term electrocardiography (ECG) recordings, accurate R-wave detection is essential. Several algorithms have been proposed but not yet compared on large, noisy, or pathological data, since manual ground-truth establishment is impossible on such large data.** ***Methods:*** **We apply the simultaneous truth and performance level estimation (STAPLE) method to ECG signals comparing nine R-wave detectors: Pan and Tompkins (1985), Chernenko (2007), Arzeno *et al.* (2008), Manikandan *et al.* (2012), Lentini *et al.* (2013), Sartor *et al.* (2014), Liu *et al.* (2014), Arteaga-Falconi *et al.* (2015), and Khamis *et al.* (2016). Experiments are performed on the MIT-BIH database, TELE database, PTB database, and 24/7 Holter recordings of 60 multimorbid subjects.** ***Results:*** **Existing approaches on R-wave detection perform excellently on healthy subjects (F-measure above 99% for most methods), but performance drops to a range of $F = 90.10\%$ (Khamis *et al.*) to $F = 30.10\%$ (Chernenko) when analyzing the 37 million R-waves of multimorbid subjects. STAPLE improves existing approaches ($\Delta F = 0.04$ for the MIT-BIH database and $\Delta F = 0.95$ for the TELE database) and yields a relative (not absolute) scale to compare algorithms' performances.** ***Conclusion:*** **More robust R-wave detection methods or flexible combinations are required to analyze continuous data captured from pathological subjects or that is recorded with dropouts and noise.** ***Significance:*** **STAPLE algorithm has been adopted from image to signal analysis to compare algorithms on large, incomplete, and noisy data without manual ground truth. Existing approaches on R-wave detection weakly perform on such data.**

***Index Terms*—Electrocardiography (ECG), multimorbid subjects, R-wave detection, simultaneous truth and performance level estimation (STAPLE).**

Manuscript received September 21, 2016; revised November 18, 2016; accepted November 21, 2016. Date of publication November 23, 2016; date of current version August 18, 2017. M. Kashif was supported by a joint German Academic Exchange Service (DAAD) and Higher Education Commission of Pakistan (HEC) scholarship program. *Asterisk indicates corresponding author.*

*M. Kashif is with the Department of Medical Informatics, RWTH Aachen University, 52057 Aachen, Germany, and also with the MIS Division, Pakistan Institute of Nuclear, Science and Technology, Nilore 45650, Islamabad, Pakistan (e-mail: muhammad.kashif@rwth-aachen.de).

S. M. Jonas and T. M. Deserno are with the Department of Medical Informatics, RWTH Aachen University.

Digital Object Identifier 10.1109/TBME.2016.2633277

I. Introduction

CARDIOVASCULAR diseases (CVDs) are the leading cause of death worldwide. According to the World Health Organization, approximately 17.3 million people died from CVDs in 2008, which is 30% of all global deaths. This number is increasing every year, expecting 23.3 million by 2030 [1], [2].

Electrocardiography (ECG) records the detailed electrical activity of the heart. It is, therefore, the most common clinical tool to diagnose CVDs and to mark risks. Usually, ECG is evaluated over short periods of time. Four to six cycles are plotted on scale paper and measured manually by the physician. In contrast, Holter monitoring is performed over 12 or 24 h on 3–12 leads. It aims to predict serious adverse events, such as the sudden cardiac death (SCD) [3], [4]. Today, advanced technology supports 12 lead monitoring over seven days and 24 h with a sampling rate of 1000 Hz, 10 bit per sample, collecting 600 000 to 900 000 cycles per subject. This large amount of data yields 15.6-km printout and cannot be read manually. It is parsed automatically for arrhythmia and other special patterns. This method of analysis is based on cycle decomposition, and requires robust R-wave detection. However, most of the recorded data remain uninspected.

Recently, wearable ECG devices (e.g., smart T-shirt,[1] portable ECG belt[2]) connected with smart devices (e.g., smartphone) have become available for ECG recordings [5]. In near future, ECG data will be recorded continuously while the subject is walking, driving, sleeping, and exercising. Such data are noisy and show temporary dropouts on one, many or all leads. Therefore, robust methods for R-wave detection are required to automatically analyze long-term recordings of noisy ECG data or data of morbid subjects.

Hence, automatic R-wave detection is a field of intensive research. Many methods have been proposed, which are based, for example, on the derivatives [6], [7], the wavelet transform [8], the Hilbert transform [9], [10], mathematical morphology [11], neural networks [12], and hybrid mixtures of approaches

[1]http://mobihealthnews.com/32774/healthwatch-seeks-fda-clearance-for-its-12-lead-ecg-tshirt

[2]https://www.getqardio.com/qardiocore-wearable-ecg-ekg-monitor-iphone

TABLE I
COMPARISON OF R-WAVE DETECTION METHODOLOGIES

Method	References	Year	Preprocessing	Feature Extraction	Peak Detection
Pan and Tompkins	[6]	1985	Bandpass filtering	Differentiation, squaring	Moving window integration, thresholding
Chernenko	[30]	2007	Fast Fourier transform (FFT)-based high-pass filtering	–	Window filter, thresholding.
Arzeno *et al.*	[9]	2008	Bandpass filtering	Differentiation, Hilbert transform,	Double thresholding
Manikandan *et al.*	[10]	2011	Chebyshev type I bandpass	Shannon energy envelope extraction	Hilbert transform, moving average filtering
Lentini *et al.*	[34]	2013	Gaussian bandpass	Adaptive matched filter	Thresholding
Sartor *et al.*	[35]	2014	Gaussian bandpass	VCG computation based on 6 leads (frontal plane), squaring, and summation	Thresholding
Liu *et al.*	[8]	2014	Wavelet transform, adaptive threshold	Energy window transform	Thresholding, energy window analysis
Arteaga-Falconi *et al.*	[7]	2015	Savitzky-Golay filter	Second derivative	Threshodling based on no. of samples in QRS
Khamis *et al.*	[38]	2016	Median and bandpass filtering	Combination of derivative and amplitude characteristics, QRS feature filtering	Thresholding, backtracking

[13]. So far, most of the methods have been evaluated on the MIT-BIH Arrhythmia Database [14], containing 48 excerpts of two lead ambulatory ECG recordings (each of 30 min length). Here, ground truth (GT) has been established manually by two or more cardiologists in consensus. On this data, some authors report accuracies of R-wave detection of more than 99%, but the existing approaches have limitations when applied to noisy signals [15] or recordings of multimorbid patients [16]. Therefore, it is necessary to evaluate the state-of-the-art R-wave detection methods on long-term data including drop-outs, noise, and massive pathological pattern. Since a 12 lead long-term recording yields 6–9 GB of uncompressed data, consensus GT cannot be established manually.

The lack of GT is a well-known problem in medical image segmentation. Here, the simultaneous truth and performance level estimation (STAPLE) method has been suggested to generate GT combining multiple observations, and to determine the best-performing observer [17], [18]. Disregarding whether the segmentation is done manually or automatically, STAPLE considers a data-point as either object or background by a majority vote, and iteratively updates the reliability of each observer or algorithm and the estimated GT by comparing the individual markings with the majority vote (GT). STAPLE has been used in various application domains, such as uterine cervix segmentation [19], brain segmentation in magnetic resonance imaging [20], [21], positron emission tomography [22], object reorganization [23], creating large-scale silver corpus [24], multiatlas segmentation [25], dental biofilm [26], and 3-D medical structures [27].

In this study, we aim to apply the STAPLE technique to R-wave detection in massive 1-D ECG signals, where each sample point is considered binary as R-wave peak or not, and to compare the performance of existing R-wave detection algorithms on large as well as noisy multiple-lead ECG data, that is not annotated manually (big data without GT).

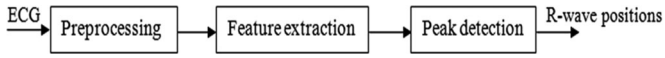

Fig. 1. Block diagram of the R-wave detection methodology.

II. MATERIAL AND METHODS

In this section, we describe the competing R-wave detection methods, the unification of R peak position, the mathematics of STAPLE, the experimental design, and the ECG databases used for evaluation.

A. R-Wave Detection Methods

Based on unstructured internet and literature research, we have identified 11 R-wave detection algorithms as the state-of-the-art. However, The method of Ferdi *et al.*, which is based on fractional digital differentiation [28], as well as the approach of Benmalek and Charef that uses digital differentiation and integration of fractional order for R-wave detection [29] were excluded as an implementation of these methods was not available and we were unable to implement the algorithms due to an incomplete description [28], [29]. The remaining nine methods are summarized in Table I.

In general, all R-wave detectors follow a three-step pipeline, which is composed of preprocessing (filtering, etc.), feature extraction (differentiation, squaring, Hilbert transform, etc.), and peak detection (thresholding, etc.). This pipeline (see Fig. 1) can be applied while disregarding whether the method is operating on a single lead or on a combination of multiple leads (see Table I).

1) Pan and Tompkins: In 1985, Pan and Tompkins proposed an R-wave detection method based on the analysis of the slopes, amplitudes, and widths of QRS complexes [6]. The method includes a bandpass filter, a derivative, amplitude squar-

ing, a moving window integrator, adaptive thresholding, and a search procedure.

a) Preprocessing: A bandpass filter is composed by cascading low-pass and high-pass filters. It is applied on the ECG leads for noise attenuation.

b) Feature Extraction: The filtered signal is differentiated by using a five-point derivative operator. The derivative process provides the slopes of QRS complexes. Then, the signal is squared to emphasize higher frequencies from the QRS complexes and to ensure a positive signal. Also, small differences arising from P- and T-waves are suppressed. A moving window integrator is applied in order to produce the waveform that contains the information of the width of the QRS complex in addition to the slope of R-wave. The size of the window is important. For a large window size QRS complex and T-wave may merge together, while a very small window size may result in multiple peaks for a single QRS complex. It is suggested that the width of the window should be the same as the widest possible QRS complex. The rising edge of the resulting integration waveform corresponds to the QRS complex and the width of the QRS complex is equal to the time duration of the rising edge.

c) Peak Detection: A fiducial point of the QRS complex can be determined from the rising edge to be marked as the position of R-wave. Thresholds are adjusted automatically according to the characteristics of the signal.

d) Performance: Based on the MIT-BIH database [14], the authors report 507 false positive (FP) beats (0.437%) and 277 false negative (FN) beats (0.239%) for a total number of 116 137 beats. Hence, a sensitivity of 99.76% and a positive predictive value of 99.56% has been achieved.

e) Implementation: A MATLAB implementation of Pan and Tompkins' algorithm is available online.[3] We adopted the code such that the method takes ECG signal and sampling rate as input parameters and gave R-wave positions as output. No change in the code is made except the method unification (see Section II-B).

2) Chernenko: In 2007, Segey Chernenko has proposed a simple method to detect R-waves [30].

a) Preprocessing: In the preprocessing step, the fast Fourier transform (FFT) is applied on the ECG signal, low frequencies are removed, and the inverse FFT is applied to restore the filtered ECG signal.

b) Feature Extraction: The author did not apply any transform for feature extraction.

c) Peak Detection: The local maxima are found by applying a windowed filter. Initially, a window of default size is used. Then, a threshold is applied to remove the small peaks and to preserve the significant ones. These significant peaks were considered as the temporary R-waves. Then, the window size is optimized according to the RR distance and the process of window filtering is repeated. During this process, the final R-wave positions are obtained.

d) Performance: Author's experiments were performed on some private ECGs but evaluation on standard databases was not reported.

e) Implementation: The MATLAB code is available on author's webpage (http://www.librow.com/cases/case-2). No change in the code is made except the method unification (see Section II-B).

3) Arzeno et al.: In 2008, Arzeno *et al.* analyzed five R-wave detectors. They combined Hilbert transform with derivative, Hilbert transform with automatic and secondary thresholds and Hilbert transform with squaring function [9]. Since these algorithms are similar in functionality, we have chosen one algorithm from their work in which Hilbert transform is combined with derivative and with automatic and secondary thresholds to detect the R-waves.

a) Preprocessing: In the preprocessing step, a Kaiser window bandpass filter [31] is applied to the ECG data with a passband of 8–20 Hz to remove baseline wander and high-frequency noise.

b) Feature Extraction: Then, the signal is differentiated. Since the differentiation modifies the signal's phase and creates zero crossings in the location of the R-wave, Hilbert transform is applied in order to rectify the phase. The Hilbert transform yields prominent peaks as the location of R-waves, which corresponds to the zero-crossing of the differentiated ECG. The all-pass characteristics of the Hilbert transform prevents the signal distortion and preserves the necessary information of the QRS complex.

c) Peak Detection: Then, a variable threshold is applied, which is determined automatically based on the root mean squared value of the 1024-point data segment. The peaks are detected and the largest amplitude within a 200-ms window in the neighborhood of the identified peak is stored. If the current RR interval is increased 1.5 times the previous RR interval, a secondary threshold equal to 0.9 times of the current threshold is applied. The peaks are detected again after the secondary threshold. In the final stage, the real peaks are detected from the original ECG signal within ± 10 samples of the detected peaks in the transformed output.

d) Performance: The algorithm was tested on MIT-BIH database [14] and – according to the authors – achieved a sensitivity and a positive predictive value of 99.29% and 99.24%, respectively.

e) Implementation: In personal communication, Arzeno has provided us his MATLAB implementation. However, the Kaiser filter was not included in the code. On author's suggestion, the signal was preprocessed by the cascaded high- and low-pass filters, as already applied with the Pan and Tompkins method. No change in the code is made except the method unification (see Section II-B).

4) Manikandan et al.: In 2011, Manikandan *et al.* proposed a new method based on Shannon energy estimator and a peak-finding logic using Hilbert transform and moving average filter to overcome the detection problem of unusually shaped QRS complexes and noise [10].

[3]http://de.mathworks.com/matlabcentral/fileexchange/45840-complete-pan-tompkins-implementation-ecg-qrs-detector

a) Preprocessing: To reduce the noise, a fourth-order Chebyshev type I bandpass filter [32] for the bandwidth of 6–18 Hz is applied to the ECG signal.

b) Feature Extraction: A first-order differentiation is applied to obtain the slope information of the R-wave. The bandpass filtering and differentiation do not only reduce noise but also lower the amplitudes of P- and T-waves. The differentiated ECG signal then is normalized by dividing it by its absolute maximum value. The Shannon energy envelope [33] is applied to make the signal positive and to enhance the QRS complex. The Shannon energy envelope is computed using the following equation:

$$E_S\left[n\right] = -d^2\left[n\right]\ \log\left(d^2\left[n\right]\right) \tag{1}$$

where E_S is the Shannon energy envelope and d is the normalized differentiated ECG signal.

The Shannon energy envelope reduces the effects of low-value noise components and produces smooth sharp local maxima. These local maxima in the Shannon energy envelope indicate the approximate locations of the R-peaks. Then, the zero-phase filtering is applied to obtain the sharp peaks around the QRS complex. For this, a rectangular impulse response of length L is applied on the signal in both the forward and backward direction. The length L is approximately the same as the possible width of QRS complex.

c) Peak Detection: The peak finding logic is based on the Hilbert transform and zero-crossing point detection. While Arzeno *et al.* have applied the Hilbert transform on the differentiated ECG signal to shift the zero-crossings onto the locations of the R-peaks, it is used here for the opposite: it is applied to locate the R-peaks from the Shannon energy envelope and operates on the zero-crossing of the output signal. However, the small amplitude R-waves may not be located at the zero-crossing points. For this, a moving average filter is applied to remove the low-frequency drift. After that, the negative to positive zero-crossing points are detected, which correspond to the locations of the R-waves. Finally, the true R-wave positions were detected from the original ECG by searching the maximum amplitude within $\pm$ 25 samples of the candidate R-wave.

d) Performance: Like the previous methods, the algorithm was validated using the ECG records of the MIT-BIH database [14]. The authors report a high performance with sensitivity and positive predictive value of 99.93%, and 99.88%, respectively.

e) Implementation: We have implemented the algorithm in MATLAB from the original paper. The length of the impulse response is set to $L = 55$, 77, and 154 samples for sampling rate $r = 360$, 500, and 1000 Hz, respectively. Similarly, the length of moving average filter is set to $l = 900$, 1250, and 2500 samples for sampling rates of $r = 360$, 500, and 1000 Hz, respectively.

5) Lentini et al.: In 2013, Lentini *et al.* in cooperation with TMD have developed an adaptive template-based R-wave detection method. On different time-scales, the filter shape adapts such that the sum of all its coefficients is zero [34].

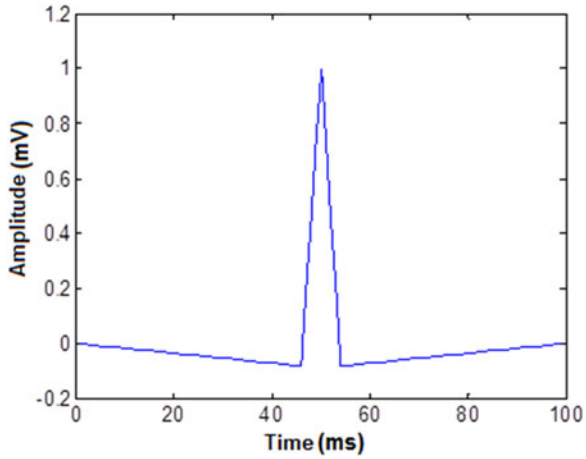

Fig. 2. Adaptive matched filter.

a) Preprocessing: A bandpass filter is composed by subtracting two low-pass Gaussian filters. It is applied on the ECG leads for noise attenuation.

b) Feature Extraction: The signal is filtered with an adaptive matched filter that was designed to emphasize the R-wave by exploiting the shape of the QRS complex (see Fig. 2). The area under the filter always sums to zero disregarding its peak width, which is adopted to match the width of the QRS complex (R-peak width of 10 ms).

c) Peak Detection: A local threshold is estimated within a window of length 1440, 2000, and 4000 samples for sampling rate of $r = 360$, 500, and 1000 Hz, respectively. Based on the threshold, the R-waves are detected and the maximum within the R-wave is used as the position of the R-wave.

d) Performance: The method was not yet applied to any standard ECG database for quantitative performance evaluation.

e) Implementation: A MATLAB implementation is available from the authors and has been integrated in our framework for ECG analysis. The value of alpha is set as $\alpha = 9$ and $\alpha = 4$, for the two low-pass Gaussian filters in the preprocessing step.

6) Sartor et al.: In addition, Sartor *et al.* in cooperation with SMJ and TMD have applied the idea of R-wave detection on the vector cardiogram (VCG) [35]. The VCG is based on all the six frontal leads (I, II, III, aVL, aVR, aVF) out of any 12 lead ECG recording. In VCG, not individual channels, but the magnitude of the heart's electrical activity is plotted with respect to its direction, i.e., the angle the activity vector is pointing to. Therefore, the electrical activity is displayed in a two-dimensional space for the frontal plane.

a) Preprocessing: All six relevant leads are filtered with a Gaussian bandpass, which is identical to the one used in the approach of Lentini *et al.*

b) Feature Extraction: A squared magnitude measure of the electrical activity is calculated as follows:

$$M\left(t\right) = \mathrm{I}^2\left(t\right) + \mathrm{II}^2\left(t\right) + \mathrm{III}^2\left(t\right) + \mathrm{aVL}^2\left(t\right) + \mathrm{aVR}^2\left(t\right) + \mathrm{aVF}^2\left(t\right). \tag{2}$$

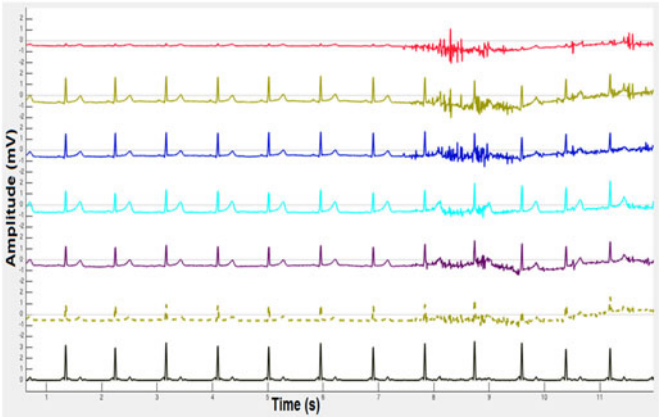

Fig. 3. From top to bottom, Leads I, II, III, aVR, aVL, and AVF are displayed. The bottom most signal is *M*(t), see (2). Adding the linear dependent leads 1 to 6 reduces noise and hence, emphasizes the R-waves.

M yields a positive signal with emphasized R-waves (see Fig. 3). Depending on the database, some summands in (2) might be unavailable, i.e., set to zero. For instance, the MIT-BIH database provides just one of the required channels (lead II), and (2) yields

$$M_{\mathrm{MIT-BIH}}(t) = \mathrm{II}^2(t).$$

The squared magnitude *M* is filtered with a Gaussian low-pass filter G_{low} again to reduce high-frequency noise as follows:

$$\hat{M}(t) = M(t) * G_{\mathrm{low}}. \tag{3}$$

After that t_{peak} is calculated

$$t_{\mathrm{peak}} = \mathrm{argmax}\left[\hat{M}(t)\right]. \tag{4}$$

Here, t_{peak} corresponds to the maximum excitation amplitude of the cycle and determines the R-wave position.

c) Peak Detection: A local threshold is calculated within a window of length 720, 1000, and 2000 samples for sampling rate of $r = 360$, 500, and 1000 Hz, respectively, and the position of the R-wave is calculated.

d) Performance: In the paper of Sartor *et al.*, the method is not evaluated quantitatively. According to the authors, a database with GT on 12 lead recording was unavailable [35].

e) Implementation: In our MATLAB implementation, the width of the Gaussian window is set as $w = 37$, 51, and 101 samples for sampling rate of $r = 360$, 500, and 1000 Hz, respectively, and alpha is set as $\alpha = 4$.

7) Liu et al.: In 2014, Liu *et al.* proposed another method for R-wave detection based on the wavelet transform, an energy window analysis, and a peak detection logic [8].

a) Preprocessing: To simulate a rather unified noisy condition, Gaussian white noise is added to the ECG signal. Then, the signal is decomposed into eight scales depending on the frequency using the wavelet transform. In this case, both the signal and noise are decomposed into different scale. Then, the adopted threshold is applied on each scale to isolate the ECG signal from the noise. The QRS complexes are located in several scales. For sampling rate of $r = 360$ and 500 Hz, the signal is reconstructed using the scales 3–5. If $r = 1000$ Hz, then the signal is reconstructed using scales 4–7

$$S = dh_3 + dh_4 + dh_5 + dh_6 + dh_7 \tag{5}$$

where dh_3, dh_4, dh_5, dh_6, and dh_7 denote the denoised signals at the scales 3–7, respectively. *S* is the reconstructed signal. If $r = 360$ or 500 Hz, then $dh_6 = 0$ and $dh_7 = 0$ and if $r = 1000$ Hz, then $dh_3 = 0$.

b) Feature Extraction: The advantage of wavelet reconstruction is that the signal mainly contains the information of QRS complex. The P- and T-waves are removed during the process of wavelet reconstruction. After reconstruction, an energy window transform is applied to enhance the QRS complex. The energy window transform can be represented as

$$E_n = \sum_{n-\frac{N}{2}+1}^{n+\frac{N}{2}} S^2[n] \tag{6}$$

and

$$n = \frac{N}{2}, \frac{N}{2}+, \ldots, m - \frac{N}{2} \tag{7}$$

where E_n is the energy, *N* is the length of the slide window, and *m* is the length of sample data.

c) Peak Detection: A threshold is applied to remove the small energy peaks. Then, a group of energy peak points are obtained through energy window analysis. A refining strategy is applied to avoid the missing or redundant peaks. In the end, the maximum amplitudes in a certain range (± 20 samples) around the energy peak points are searched in the original ECG signal which are treated as R-peaks.

d) Performance: The MIT-BIH database was used to validate the effectiveness of this method [14]. However, the results were not reported for all records, since some have been excluded from the experiments. On the remaining data, sensitivity and positive predictive values of 98.65% and 99.44% were reported, respectively.

e) Implementation: We have implemented the code from the original paper in MATLAB. The chosen length of the slide window is 36, 50, and 100 samples for $r = 360$, 500, and 1000 Hz, respectively, and the value of threshold T_{h} in peak detection is set to be

$$T_h = 0.3 \times \mathrm{med}(E_n) \tag{8}$$

where $\mathrm{med}(E_n)$ is the median of E_n.

8) Arteaga–Falconi et al.: In 2015, Arteaga–Falconi *et al.* presented an R-wave detection method based on the second derivative [7].

a) Preprocessing: Since the method was proposed for filtered ECG, no preprocessing was done. However, in personal communication the author suggested to use a Savitzky–Golay filter [36]: a built-in MATLAB function (sgolayfilt) with polynomial order $K = 1$ as well as frame size $F_r = 11, 17,$ and 33 for $r = 360$, 500, and 1000 Hz, respectively.

b) Feature Extraction: A second derivative of the filtered ECG is calculated and the output series is inverted. The inverted second derivative series is sorted by amplitude in descending order, where all values in the beginning of the series belong to the QRS complexes. However, it is uncertain how many values belong to the QRS complexes. To resolve this issue, an index of the last possible QRS complex value is calculated based on the sampling rate, time length of the signal, and the maximum possible heart rate (which is constant and approximated as $HR_{\max} = 220$ (bpm) [37]) as follows:

$$L_R = HR_{\max} \times t_f \times s_{\text{qrs}} \tag{9}$$

and

$$S_{\text{qrs}} = k \times r \tag{10}$$

where L_R is the last possible index, $HR_{\max}$ is the maximum possible heart rate, t_f is the time length of the ECG signal (in seconds), S_{qrs} is the possible number of samples in QRS complex, k is the proportionality constant, and r is sampling rate. All entries beyond the index are eliminated. Then, the values in the sorted array are overlaid to their original location.

c) Peak Detection: A threshold based on the possible number of samples in QRS complex is applied to identify the R-waves. In the end, the maximum value corresponding to the QRS complex in the original ECG signal is considered as the position of R-wave.

d) Performance: The algorithm was evaluated on selected subset of the MIT-BIH database [14], and the authors report sensitivity and positive predictive values of 99.43% and 99.22%, respectively.

e) Implementation: We have received the code from the author. According to the authors, the constant parameter is set as $k = 0.019843$ [7].

9) Khamis et al.: In 2016, Khamis *et al.* proposed a new method for R-wave detection by generating a QRS feature containing information about derivative and amplitude characteristics of the filtered ECG signal [38]. The algorithm is not only developed for clean clinical data but also designed for poor quality telehealth signals.

a) Preprocessing: At first, the ECG signal is detrended to reduce transient start up effects of any later filtering. Then, a median filter is applied to minimize any baseline drift. After that, the signal is bandpass filtered (0.7–20 Hz) to retain peak frequency components of the QRS complex. The filtered ECG signal is denoted as $x[n]$.

b) Feature Extraction: The QRS feature is generated by combining the derivative and the amplitude characteristics of the filtered ECG signal $x[n]$. The absolute value of the derivative $d_x[n]$ is obtained. Then, an amplitude envelope $a_{\text{x}}[\text{n}]$ of $x[n]$ is computed using maximum and minimum filters (window width 0.1 s). The upper and lower envelopes are denoted as $x_u[n]$ and $x_l[n]$, respectively,

$$a_x[n] = x_u[n] - x_l[n]. \tag{11}$$

Then, the feature $z_{\text{raw}}[n]$ is obtained

$$z_{\text{raw}}[n] = a_x[n] \times d_x[n]. \tag{12}$$

The multiplication in (12) suppresses P- and T-waves, and the QRS feature filtering smoothes the feature and removes spurious peaks, which occur due to local variation in the ECG signal. For this, a low-pass filter with an adaptive cutoff frequency is applied on $z_{\text{raw}}[n]$, yielding the final QRS feature signal $z[n]$.

c) Peak Detection: A peak to peak amplitude signal $a_z[n]$ is calculated as $a_z[n] = z_u[n] - z_l[n]$, where $z_u[n]$ and $z_l[n]$ are the upper and lower signal envelopes, respectively. A threshold $\lambda = 0.2 \cdot \text{median}(a_z[n])$ is computed. Then, the peaks are searched by following the feature signal $z[n]$, starting from the first sample of the signal $z[0]$. The local maxima and the local minima are searched when $z[n]$ rises above $z[0] + \lambda$, or falls below $z[0] - \lambda$. In the end, backtracking is performed to correct the obviously missed or incorrectly selected R-waves.

d) Performance: The algorithm was evaluated on the MIT-BIH and the TELE databases [14], [38], and the authors reported the sensitivity 99.76% and 98.05% and the positive predictive value 99.80% and 95.75%, for the MIT-BIH database and the TELE database, respectively.

e) Implementation: The MATLAB code is available online [38]. We have not made any change in the code except the method unification (see Section II-B). Since the implementation fails for sampling rate $r > 600$ Hz, we have down-sampled data with $r = 1000$ Hz to $r' = 500$ Hz.

B. Method Unification

Most of the methods in Section II-A determine the maximum point in the QRS complex as the position of the R-wave. The approaches of Pan and Tompkins and Sartor *et al.*, however, do not seek the maximum in the signal's domain. For these methods, we performed an additional maximum detection to ensure that there is a certain interval framing the R-peak, where no additional R-waves are allowed, and that different methods do not end up with displacements of one or two sample points. In this way and for all methods, each same point is considered binary as a position of R-wave or not.

C. STAPLE Method

Originally, STAPLE has been proposed for medical image segmentation. A structure is segmented in an image by indicating its presence or absence at each pixel or voxel. STAPLE is based on several segmentations and computes a probabilistic estimate of the GT segmentation and a measure of the performance level achieved by each of the compared methods, which may be a human rater or an automatic segmentation algorithm [17], [18].

The principle of STAPLE is summarized in [39]. The algorithm takes a set of segmentations from J experts as input. The labeling of each voxel, in an image of I voxels, provided by the expert is referred as segmentation decisions d_{ij}, indicating the label given by each expert j for voxel i, $i \in [1, \ldots, I]$. The goal of STAPLE is then to estimate both a reference standard segmentation T, and performance parameters $\theta = \theta_1, \ldots, \theta_j, \ldots, \theta_J$ describing the agreement between each expert and the reference standard by computing the sensitivity and specificity. Here, $\theta_{\text{jśs}}$ is the probability that the expert

j gave the label $\acute{s}$ to a voxel i when the reference standard label is s, i.e., $\theta j\acute{s}s = \mathrm{P_r}(\mathrm{d_{ij}} = \acute{s}|\mathrm{T_i} = \mathrm{s})$.

Since, reference standard is unknown, an Expectation–Maximization approach [40] is used to estimate T and the expert performance parameters through the maximization of the expected value of the complete data log likelihood $\mathrm{Q}(\theta|\theta^{(k)})$

$$\mathrm{Q}\left(\theta/\theta^{(k)}\right) = \sum_i \sum_j \sum_s W_{\mathrm{si}} \log(\theta_j d_{\mathrm{ijs}}) \qquad (13)$$

where W_{si} denotes the posterior probability of T for label s: $P_r(Ti = s|D, \theta^{(k)})$. The EM algorithm, which is guaranteed to converge to a local maximum, proceeds to identify the optimal estimate θ' by iterating two steps:

1) *E-step*: Compute $\mathrm{Q}(\theta|\theta^{(k)})$, the expected value of the complete data log likelihood given the current estimates of the expert parameters at iteration k: $\theta^{(k)}$.
2) *M-step*: Estimate new performance parameters $\theta^{(k+1)}$ by maximizing $\mathrm{Q}(\theta|\theta^{k})$.

In this study, we substitute images with ECG signals and pixels with sample points. The R-wave detection methods are considered as experts. The output of each R-wave detection method is transformed into a binary vector such that 1 represents R-wave position in the vector and 0 for all other places (representing no R-wave). These vectors are given then as an input to the binary STAPLE method [17]. Initially, STAPLE assigns the same weights (value of the performance parameters) to all methods and computes the true R-wave positions based on the initial weights and the inputs from all R-wave detection methods. After computing the initial GT, STAPLE estimates the performance of all R-wave detection methods with reference to the initial GT and assign them new weights based on their performance. The process is iterated until the weights stay stable. Then, STAPLE gives the GT positions of R-waves and the performance of each R-wave detection method.

D. Evaluation

In this section, we describe the hypothesis to approve or disapprove, according experiments, appropriate metrics, and databases used.

1) Hypotheses: In this hypothesis-driven research, all experiments are designed and performed in order to confirm particular assumptions. In total, we have worded five hypotheses.

1) H1: Our MATLAB implementations of R-wave detection algorithms as well as the code that we have received from the authors directly or from the Internet yield equal performance as originally reported by the authors. Such "equality" is given if the F-measures deviates less than 1.0 percentage point.
2) H2: STAPLE is a reliable means of determining GT for various databases.
3) H3: STAPLE improves R-wave detection compared to any of the individual detection algorithms.
4) H4: The authors of previous work probably have tuned their algorithms to just one database, yielding overfitting. We assume that when the algorithms are applied to different data, the performance drops down. Again, a deviation of more than 1.0 percentage point is regarded to approve the hypothesis.
5) H5: There is an additional drop in performance when the methods are applied to long-term and noisy ECG recordings of subjects suffering from severe pathologies. On the other hand, we expect improved performance when the algorithms are applied to recordings of healthy subjects only.

2) Experiments: To prove (or disapprove) the hypotheses, we have designed according experiments:

1) H1: The F-measure is computed on the data the authors have used when publishing their work, and the obtained values are compared with the ones published by the authors.
2) H2: The STAPLE approach is applied to annotated databases, and compared with the manual GT.
3) H3: The performance of STAPLE and the nine methods individually are compared using databases with manual GT.
4) H4: All methods are applied to a different but comparable database. To make the computations comparable, GT is generated using the STAPLE approach, and any GT that comes with the data is disregarded.
5) H5: All methods are applied to further databases, and GT is generated again using the STAPLE approach. Here, data for healthy as well as pathologic subjects are required.

3) Metrics: Quantitative evaluation requires the use of metrics. According to the previous publications [8]–[10], we apply precision (also called positive predictive value) and recall (sensitivity). Specificity is not used, because of the high values for true negatives (TN): only one out of 1000 ECG sample points yields an R-wave (on $r = 1000$ Hz). True positive (TP), FN, and FP determine the correctly detected R-wave, a missed R-wave, and a noisy spike not being an R-wave or P- or T-waves falsely assumed as R-wave, respectively. F-measure is computed as the harmonic mean of precision and recall to provide a unique figure allowing to rank and compare the methods. Precision P, Recall R, and F-measure F were computed by using following equations:

$$P = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \qquad (14)$$

$$R = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \qquad (15)$$

$$F = 2 \times \frac{P \times R}{P + R}. \qquad (16)$$

4) ECG Databases: Several databases and data repositories are meanwhile available. For instance, the Telemetric and Holter ECG Warehouse (THEW) project has collected several recordings from different clinical trials [41]. However, access to THEW data is not granted without a written research proposal. Further important selection criteria are the sampling rate, the number of leads, the duration of recordings, and the grade of pathological pattern contained in the databases (see Table II).

TABLE II
DESCRIPTION OF THE DATABASES IN USE

Database	Leads	Subjects	Records	Sampling rate [Hz]	Length [min]	Amplitude [mV]	Quantization [bit]	Manual GT	Pathology	Total length [hr]	Total length [beats]
MIT-BIH	2	47	48	360	30 +/− 0	±5	11	Yes	Yes	24	109404
TELE	1	120	250	500	0.48 +/− 0.24	±5	12	Yes	Yes	2	6708
PTB (complete)	12	290	549	1000	1.81 +/− 0.39	±16	16	No	Yes	16.5	72586
PTB (healthy only)	12	52	80	1000	1.97 +/− 0.06	±16	16	No	No	2.62	10655
ESRD	12	60	60	1000	9559 +/− 1734	±6	10	No	Yes	9559	37 mill

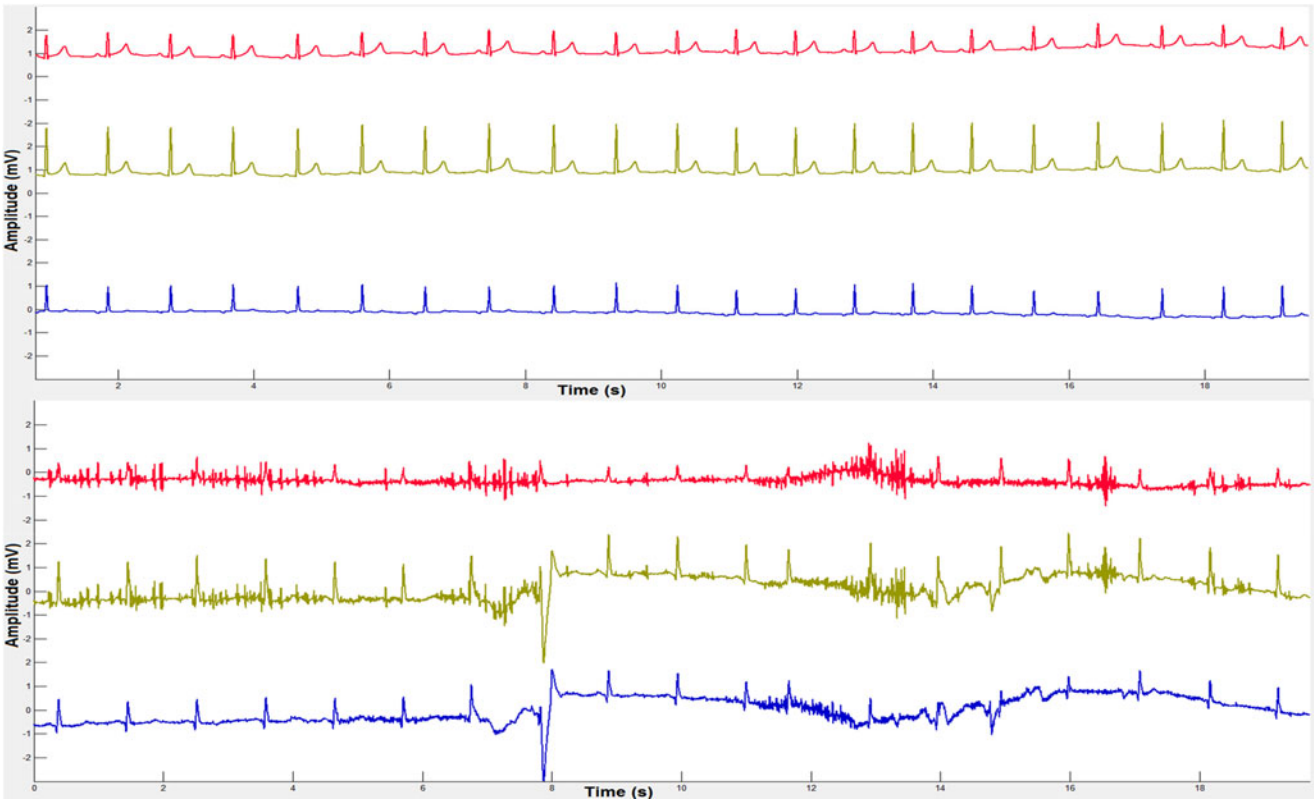

Fig. 4. Leads I, II, and III extracted from a healthy subject of the PTB database (top) and of a multi-morbid subject of the ESRD database (bottom).

The MIT-BIH Arrhythmia database [14] has been used by many of the authors of previous papers, and hence we use it for experiments H1–H5. It contains 48 excerpts of two lead ambulatory ECG recordings, each of 30 min length, which are obtained from 47 subjects. The recordings have been digitized at 360 Hz per channel on 11 bit resolution over a 10-mV range. The recordings contain normal beats as well as pathological patterns. Two or more cardiologists independently annotated each record; disagreements were resolved to obtain the computer-readable reference annotations for each beat (approximately 110 000 annotations in total).

A publically available TELE database [38] comprises 250 single lead-I ECG signals recorded using a remote monitoring system called the TeleMedCare Health Monitor (TeleMedCare Pty., Ltd., Sydney, N.S.W., Australia). The ECG is sampled at a rate of 500 Hz using dry metal Ag/AgCl plate electrodes. ECG measurements were recorded by the patients who were trained to use the device [42]. Three scorers annotated the data independently. After that, the scorers annotated the signals as a group to reconcile the individual annotations. Segments of artifacts were identified and masked. QRS annotations in the masked regions were discarded [38]. To remove artifacts from the signals, we applied visual mask which is available with the TELE database. The database is used for experiment H2–H5.

The PTB Diagnostic ECG database is available freely to the public. It contains 549 records from 290 subjects, out of which 52 are healthy controls (18%) [43]. The sampling rate is 1000 Hz. Each record contains 15 signals including conventional 12 leads signals. The total recording duration for all 549 records is 16.5 h. Manual GT is unavailable. However, PTB is composed of clean clinical data similar to the MIB-BIH database. Therefore, we have selected the PTB database for experiments H4 and H5.

TABLE III
AUTHOR-REPORTED PERFORMANCES VERSUS OUR COMPUTATIONS ON THE MIT-BIH DATABASE

METHOD	Reported by the Authors			Reproduced Results			ΔF
	P [%]	R [%]	F [%]	P [%]	R [%]	F [%]	
Pan and Tompkins	99.56	99.76	99.66	99.03	98.94	98.99	–0.67
Chernenko*	—	—	—	99.31	86.25	92.32	—
Arzeno *et al.*	99.24	99.29	99.26	98.78	98.67	98.72	–0.54
Manikandan *et al.*	99.88	99.93	99.90	99.24	99.03	99.13	–0.77
Lentini *et al.**	—	—	—	96.89	94.94	95.90	—
Sartor *et al.**	—	—	—	94.14	50.62	65.84	—
Liu *et al.***	99.44	98.65	99.04	97.71	98.83	98.27	–0.77
Arteaga-Falconi *et al.***	99.22	99.43	99.32	97.31	95.76	96.53	–2.79
Khamis *et al.*	99.80	99.76	99.78	99.73	99.66	99.69	–0.09

*The method has not been evaluated by its authors on MIT-BIH database.
**The authors' original results have been computed on a reduced dataset.

In a controlled clinical trial (Clinical-Trials.gov: NTC02001480) on surrogates for SCD in patients with diabetes mellitus and end stage renal disease (ESRD), a large database of 12 lead ECG has been recorded from 60 subjects.[4] This collection ensures frequent occurrence of pathological patterns, movement noise, and temporarily dropouts on one to many leads (see Fig. 4). All subjects have been recorded continuously for a complete week (24/7) using a Holter monitor (FD12 plus, Schiller, Austria) in its scientific mode (r = 1000 Hz) yielding 600 000–900 000 heart cycles per person. A total of 10 080 hours of ECG was recorded for analysis (approximately 42 million cycles, an equivalent to 1000 km of printed ECG). There is no manually determined GT. Here, it is used for experiment H5.

III. RESULTS

This section is ordered according to the experiments that have been described in Section II-D2.

1) H1: A comparison was possible for six out of the nine R-wave detection methods. The smallest and largest deviations in F-measure are observed with the methods of Khamis *et al.* and Arteaga–Falconi *et al.*, yielding $\Delta F = -0.09$, and $\Delta F = -2.79$, respectively (see Table III). Since Arteaga–Falconi *et al.* have computed their results on a subset of the MIT-BIH database that figure cannot be used for the hypothesis evaluation. All other values of ΔF are below the given 1.0 threshold, whether we consider our implementations as "equal".
2) H2: Comparing STAPLE with manual GT on the MIT-BIH database, a performance of $F = 99.73\%$ is obtained. However, the TELE database yields $F = 97.60\%$ (see Table IV). Here, $\Delta F > 1.0\%$ and STAPLE cannot be considered as reliable GT (H2 does not hold).
3) H3: For the MIT-BIH database, the top three methods: Khamis *et al.*, Manikandan *et al.*, and Pan and Tompkins obtained the F-measures 99.69%, 99.13%, and 98.99%, respectively, and STAPLE improves all the individual results by obtaining F-measure 99.73%. Similarly for the TELE database, the top three methods: Khamis *et al.*, Arzeno *et al.*, and Pan and Tompkins obtained the F-measure 96.65%, 92.22%, and 92.21%, respectively, while STAPLE obtained the F-measure 97.60% improving all individual results (see Table IV), which confirms our Hypothesis H3.
4) H4: Most authors have evaluated their methods on the MIT-BIH database (see Table III). Here, we applied the methods on the TELE and the PTB (complete) datasets as well. Based on the manual GT, the overall performance drops from MIT-BIH to TELE, and $\Delta F < -1.0$ for all methods. Using STAPLE-based GT mean-based ΔF for MIT-BIH to TELE again is negative for all methods (see Tables V and VI). However, for MIT-BIH to PTB (complete) $\Delta F = +24.02$ and $\Delta F = +3.05$ for Sartor *et al.* and Chemenko, respectively. The performance gain of the Sartor methods can be explained by the fact that Sartor is based on six leads, but MIT-BIH as well as TELE only provides two and one leads, respectively. For the method of Chemenko, the author stated explicitly that he has not tuned the parameters on the dataset. With all other values being certainly below the threshold $\Delta F < -1.0$, we have confirmed the supposed overfitting (H4 is true).
5) H5: On the pathological ESRD data, STAPLE has generated in total 37 006 460 R-waves. The performance ranges from $\Delta F = 90.10\%$ (Khamis *et al.*) to $\Delta F = 30.10\%$ (Chemenko). Comparing the mean of MIT-BIH and TELE with ERSD, there is a drastic change in the performance ranging from $\Delta F = -0.62$ to $\Delta F = -61.23$ for Liu *et al.* and Chemenko, respectively (see Table VI). Comparing the baseline with PTB (healthy only), $\Delta F > 1.0$ for all methods, except Arenzo *et al.* ($\Delta F = -1.31$). Again, the results of Sartor method are covered by the six lead effects. However, we can conclude that the state of the art in R-wave detection is not performing robustly on continuous data recordings and noisy data (H5 approved).

IV. DISCUSSION

R-wave detection is an important step in automatic ECG analysis, and a large number of methods have been proposed in the scientific literature. Although authors of such approaches usually report a performance above 99% in terms of accuracy or F-measure, novel approaches still are being developed [38], clearly indicating a lack of performance with existing approaches. Indeed, most of the current R-wave detection methods are not tested comprehensively on several ECG databases. This limits the performance assessment in terms of parameter choice, robustness to noise, and robustness to different type of ECG recordings [2]. Even when using the same database, that is the MIT-BIH, many authors have excluded records [7], [8] or segments with ventricular flutter [9], [10] to improve performance. In addition, the recording devices are changing rapidly, and in

[4]ERSD database is intended to be made publicly available after the clinical trial has been completed.

TABLE IV
OVERALL PERFORMANCE OF THE *R*-WAVE DETECTION METHODS ON THE STAPLE GT AS COMPARED TO THE MANUAL GT

Database	MIT-BIH							TELE						
GT	Manual			STAPLE				Manual			STAPLE			
Method	*P* [%]	*R* [%]	*F* [%]	*P* [%]	*R* [%]	*F* [%]	ΔF	*P* [%]	*R* [%]	*F* [%]	*P* [%]	*R* [%]	*F* [%]	ΔF
Pan and Tompkins	99.03	98.94	98.99	98.42	99.67	99.04	+0.05	88.53	96.20	92.21	91.19	97.40	94.19	+1.98
Chernenko	99.31	86.25	92.32	99.34	87.47	93.03	+0.71	86.84	94.31	90.42	88.70	94.89	91.69	+1.27
Arzeno *et al.*	98.78	98.67	98.72	98.68	99.33	99.00	+0.28	90.6	93.90	92.22	92.67	94.60	93.62	+1.40
Manikandan *et al.*	99.24	99.03	99.13	98.32	99.47	98.89	-0.26	78.22	88.18	82.90	79.9	88.57	84.01	+1.11
Lentini *et al.*	96.89	94.94	95.90	96.82	96.19	96.51	+0.61	92.68	89.46	91.04	95.05	90.39	92.66	+1.62
Sartor *et al.*	94.14	50.62	65.84	94.07	51.28	66.38	+0.54	99.15	20.81	34.40	99.51	20.96	34.63	+0.23
Liu *et al.*	97.71	98.83	98.27	96.86	99.27	98.05	-0.22	68.21	80.83	73.99	69.75	81.38	75.12	+1.13
Arteaga-Falconi *et al.*	97.31	95.76	96.53	97.00	96.76	96.88	+0.35	96.48	48.28	64.36	97.46	48.44	64.72	+0.36
Khamis *et al.*	99.73	99.66	**99.69**	99.07	99.37	99.22	-0.47	95.80	97.53	**96.65**	97.01	97.30	97.15	+0.50
STAPLE	99.85	99.60	**99.73**	100	100	100	+0.27	96.89	98.33	**97.60**	100	100	100	+2.40
Improvement to best method			**+0.04**							**+0.95**				

TABLE V
PERFORMANCE (F-MEASURE [%]) OF *R*-WAVE DETECTION METHODS ON VARIOUS DATABASES BASED ON STAPLE-GENERATED GT, RANK IS THE AVERAGE OF INDIVIDUAL RANKS OF THE METHODS FOR ALL DATABASES

METHOD	PTB (healthy only)		MIT-BIH		PTB (complete)		TELE		ESRD		Rank
	Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std	
Pan and Tompkins	99.55	3.81	99.05	2.35	94.76	17.14	93.50	12.38	82.71	16.17	3
Chernenko	99.57	1.50	91.64	14.09	94.69	18.25	91.01	10.96	30.10	30.50	6
Arzeno *et al.*	93.17	20.04	98.97	1.86	84.99	28.16	89.99	20.55	69.21	24.12	6
Manikandan *et al.*	99.60	1.83	98.85	2.26	95.96	15.98	83.42	16.90	87.28	18.36	3
Lentini *et al.*	99.34	1.47	96.47	6.25	92.27	20.53	91.79	9.24	74.04	27.71	5
Sartor *et al.*	97.79	11.66	61.15	35.11	85.17	31.38	26.89	36.51	69.72	33.12	8
Liu *et al.*	99.04	4.03	97.89	3.08	93.89	17.67	74.38	13.89	85.51	17.94	5
Arteaga-Falconi *et al.*	86.98	30.98	96.15	9.68	51.60	44.86	57.14	38.28	67.62	32.33	8
Khamis *et al.*	99.73	1.46	99.27	3.12	96.01	15.90	95.64	11.70	90.10	17.03	1

TABLE VI
PERFORMANCE DIFFERENCES ΔF OF THE *R*-WAVE DETECTION METHODS ON VARIOUS DATABASES, HERE, ΔF IS COMPUTED BASED ON STAPLE-GENERATED GT EXCEPT FOR THE RIGHT MOST COLUMN [MIT-BIH TO TELE (ON MANUAL GT)]

METHOD	Mean (MIT-BIH and TELE) to PTB (Healthy)	MIT-BIH to PTB (complete)	MIT-BIH to TELE	Mean (MIT-BIH and TELE) to ESRD	MIT-BIH to TELE (on Manual GT)
Pan and Tompkins	+3.27	−4.29	−5.55	−13.57	−6.78
Chernenko	+8.24	+3.05	−0.63	−61.23	−1.90
Arzeno *et al.*	−1.31	−13.98	−8.98	−25.27	−6.50
Manikandan *et al.*	+8.47	−2.89	−15.43	−3.85	−16.23
Lentini *et al.*	+5.21	−4.20	−4.68	−20.09	−4.86
Sartor *et al.*	+53.77	+24.02	−34.26	+25.70	−31.44
Liu *et al.*	+12.91	−4.00	−23.51	−0.62	−24.28
Arteaga-Falconi *et al.*	+10.34	−44.55	−39.01	−9.03	−32.17
Khamis *et al.*	+2.28	−3.26	−3.63	−7.36	−3.04

the near future, contactless ECG recording [44], smart textiles [45], and other portable devices with nonstandard limb electrode positions [46] may allow lifetime ECG monitoring and will result in even more varying quality and nature of ECG data.

Therefore, we have analyzed the performance of several methods on prior reported and on new, more challenging data. Our implementations have been proven to perform equally to the prior reports. However, on more challenging data, the performance of all methods significantly drops as expected and might be insufficient for trend prediction and alarming [3].

The effect of performance loss has been reported previously in the literature [9], [33]. Reasons may include overfitting (tuning of parameters) to the database that has been used for evaluation and, thus, not allowing generalization of the algorithm, different tolerances for counting a detection as a TP or FP [38], or different strategy of comparison with the GT. Since we have only implemented the code for Manikandan *et al.* and Liu *et al.*, and both methods are performing superior to the others (except Khamis *et al.*) in big data (see Table V), we assume that our implementations are correct. The original authors supplied other methods.

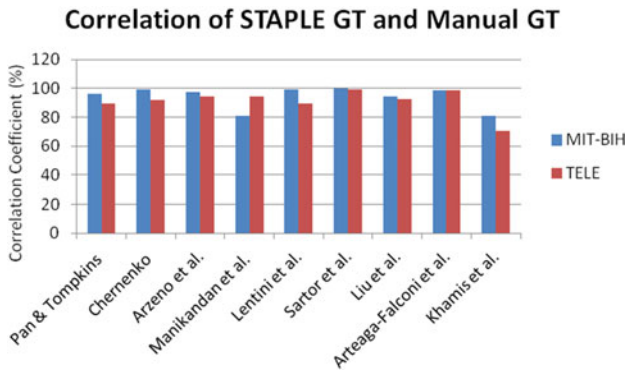

Fig. 5. Correlation of F-measures obtained by the R-wave detection methods for the STAPLE GT and the manual GT on the MIT-BIH database and on the TELE database.

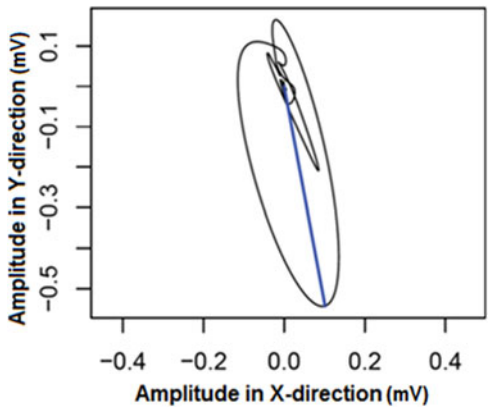

Fig. 6. Two-dimensional vector plot of an ECG cycle. *X*-direction corresponds to the lead I and *Y*-direction corresponds to the lead aVF. The blue line depicts the vector in the maximum direction corresponding to t_{peak}.

Most of the authors reported overall performance. Therefore, in order to compare with the prior reported results, we have reported overall performance in Tables III and IV. However, for further analysis, we have reported mean and standard deviation of F-measure in Tables V, which is more meaningful, and shows the variation in performance on per recording basis.

In this paper, the STAPLE method has been adopted for R-wave detection in ECG. To the best of our knowledge, STAPLE has never been applied in signal analysis before. With our experiments, we have not been able to show that STAPLE allows reliable GT establishment. However, STAPLE can be used to compare methods and to improve individual methods in a combined method.

In order to observe the correlation of performances of the R-wave detection methods for two different GTs of the MIT-BIH database, we computed the correlation coefficient of F-measures obtained by the R-wave detection methods for the manual GT and the STAPLE GT over the records of the MIT-BIH database. All methods show a strong positive correlation for the two GTs (see Fig. 5). The average correlation coefficient of 0.94 indicates that the R-wave detection methods have similar performance on the STAPLE-based and manual GT. Likewise, the correlation coefficient is calculated for the STAPLE GT and manual GT over 250 records of the TELE database. Again all methods show a strong positive correlation (see Fig. 5), yielding an average correlation coefficient of 0.91. This again is indicating that the R-wave detection methods can be evaluated using STAPLE in big data for which manual GT determination is not feasible.

We have observed a clear tendency of performance loss from PTB (healthy subjects), MIT-BIH, PTB (all subjects), TELE toward ESRD (see Table V). This is in line with the findings of Elgendi *et al.* who revisited R-wave detection methodologies and concluded that current R-wave detection algorithms are not suitable for long-term ECG recordings, or for portable and wearable ECG systems [2].

Most of the authors are restricted to the MIT-BIH database for the development and testing of the R-wave detection methods [7]–[10], but recently Khamis *et al.* have evaluated their method on telehealth ECG recordings [38]. Other studies addressing the usefulness in clinical routine are rarely found [2].

So far, the lacking GT has hindered the development of methods being more robust to large noisy data. Therefore, we adopted the STAPLE algorithm to ECG signal analysis. In line with Dewalle–Vignion *et al.* [22], we found that STAPLE on ECG data is combining the strengths of the individual algorithms, while eliminating their weaknesses. Meyer *et al.* have combined two state-of-the-art methods for QRS complex detection and proved that the combined approach outperforms both individual algorithms [13]. However, applying STAPLE is more than just a combination of approaches. STAPLE is a reliable iterative process toward an optimum on a per-recording level, thus more flexible than most other combination methods. However, there is also a limitation with the STAPLE approach. When majority of the methods perform alike, STAPLE gives the similar results. This holds for FNs as well as FPs. Hence, the necessity of developing new and diverse as well as robust R-wave detection methods is emphasized again.

Comparing the performance of the individual methods, Khamis *et al.* performed best among all. It consistently has given the best performance on all databases. The authors have evaluated their method for multiple databases, which makes it more effective compared to the other methods.

The methods Chernenko and Arteaga-Falconi *et al.* show disparate behavior. For instance, Chernenko performs well ($F = 94.69\%$) on the PTB database. However, its performance on the ESRD database is worst ($F = 30.10\%$). This is because Chernenko applied a fixed global threshold based on the higher amplitudes of the temporarily detected R-waves. Whenever there is more variation in amplitudes of the R-waves of the ECG recordings, the R-waves with lower amplitudes are eliminated in final detection. This is why it performs third best ($F = 99.57\%$) on healthy subjects of the PTB database which contain normal beats, but failed completely on more challenging ECG recordings of multimorbid subjects.

Similarly, Arteaga-Falconi *et al.* yield a larger standard deviation for all databases, which shows more variation over various signals. Arteaga-Falconi *et al.* is the only method performing better ($F = 67.62\%$) on the ESRD database as compared to the PTB database ($F = 51.60\%$). This is because Arteaga-Falconi

et al. use a fixed threshold based on the possible number of samples in QRS complex to identify the R-waves, where the threshold is computed from sampling rate. Even for the same sampling rate, the number of samples in QRS complexes may differ significantly for various types of ECG recordings. Therefore, this method cannot detect R-waves when the number of samples in the QRS complexes is below the threshold.

Comprehensive evaluation also showed that our method (Sartor *et al.*) cannot compete with other approaches. This method is the only one combining several leads, and hence, should have a higher potential. However, the simple squared linear combination of different measures (2) seems insufficient, and further filtering may improve the concept. Also, an analysis of the directions in terms of the vectorial components of M may improve the method. Fig. 6 shows the VCG in the (x,y) plane. It can be seen clearly that the correct R-wave must not only yield certain amplitude but must also be located in a certain section of the plane.

In the current study, most of the R-wave detection algorithms are numerically efficient and take a fraction of second to process a 30-min ECG recording of the MIT-BIH database. Any combination of algorithms, such as STAPLE, takes more time. We are using nine algorithms, where Pan and Tompkins and Liu *et al.* are computationally more expensive, yielding 4.25–5.00 s in total. STAPLE itself, since computed with 2-bit logical data type takes only 0.13 s. Hence, it is feasible to implement STAPLE-based R-wave detection on PCs and on portable devices such as tablets and smartphones.

V. Conclusion

In this paper, we have confirmed that current R-wave detection algorithms are suitable for clear data recorded from healthy subjects, but their performance drops significantly when applied to signals derived from long-term Holter recordings of multimorbid subjects. The STAPLE method is suitable to compare algorithms on large but unannotated databases. It improves the R-wave detection performance of an individual algorithm for clean clinical as well as noisy telehealth ECG data. Nonetheless, more robust R-wave detection methods and adaptive combination schemes are required for continuous ECG monitoring or data of multimorbid subjects.

References

[1] A. Alwan, "Global status report on noncommunicable diseases 2010," *World Health Organ*., 1st ed. Geneva, Switzerland: WHO, 2011.

[2] M. Elgendi *et al.*, "Revisiting QRS detection methodologies for portable, wearable, battery-operated, and wireless ECG systems," *PloS One*, vol. 9, p. e84018, Jan. 2014.

[3] T. M. Deserno and N. Marx, "Computational electrocardiography: Revisiting Holter ECG monitoring," *Methods Inf. Med.* vol. 2, p. 55, 2016.

[4] C. W. Israel, "Mechanisms of sudden cardiac death," *Indian Heart J.*, vol. 66, pp. S10–S17, Feb. 2014.

[5] Y. D. Lee and W. Y. Chung, "Wireless sensor network based wearable smart shirt for ubiquitous health and activity monitoring," *Sens. Actuators B Chem.*, vol. 140, no. 2, pp. 390–395, Jul. 2009.

[6] J. Pan and W. J. Tompkins, "A real-time QRS detection algorithm," *IEEE Trans. Biomed. Eng.*, vol. BME-32, no. 3, pp. 230–236, Mar. 1985.

[7] J. S. Arteaga-Falconi *et al.*, "R-peak detection algorithm based on differentiation," in *Proc. IEEE 9th Int. Symp. Intell. Signal Process.*, May 2015, pp. 1–4.

[8] X. Liu *et al.*, "A novel R-peak detection method combining energy and wavelet transform in electrocardiogram signal," *Biomed. Eng.*, vol. 26, no. 01, p. 1450007, Feb. 2014.

[9] N. M. Arzeno *et al.*, "Analysis of first-derivative based QRS detection algorithms," *IEEE Trans. Biomed. Eng.*, vol. 55, no. 2, pp. 478–484, Feb. 2008.

[10] M. S. Manikandan and K. P. Soman, "A novel method for detecting R-peaks in electrocardiogram (ECG) signal," *Biomed. Signal Process. Control*, vol. 7, no. 2, pp. 118–128, Mar. 2012.

[11] F. Zhang and Y. Lian, "QRS detection based on multiscale mathematical morphology for wearable ECG devices in body area networks," *IEEE Trans. Biomed. Circuits Syst.*, vol. 3, no. 4, pp. 220–228, Aug. 2009.

[12] B. Abibullaev and H. D. Seo, "A new QRS detection method using wavelets and artificial neural networks," *J. Med. Syst.*, vol. 35, no. 4, pp. 683–691, Aug. 2011.

[13] C. Meyer *et al.*, "Combining algorithms in automatic detection of QRS complexes in ECG signals," *IEEE Trans. Inf. Technol. Biomed.*, vol. 10, no. 3, pp. 468–475, Jul. 2006.

[14] G. B. Moody and R. G. Mark, "The impact of the MIT-BIH arrhythmia database," *IEEE Eng. Med. Biol. Mag.*, vol. 20, no. 3, pp. 45–50, May 2001.

[15] S. Kharabian *et al.*, "Fetal R-wave detection from multichannel abdominal ECG recordings in low SNR," in *Proc. IEEE Annu. Int. Conf. Eng. Med. Biol. Soc.*, 2009, pp. 344–347.

[16] N. T. Liu *et al.*, "Development and validation of a novel fusion algorithm for continuous, accurate, and automated R-wave detection and calculation of signal-derived metrics," *J. Critical Care*, vol. 28, no. 5, pp. 885-e9–885-e18, Oct. 2013.

[17] S. K. Warfield *et al.*, "Validation of image segmentation and expert quality with an expectation-maximization algorithm," *in Proc. 5th Int. Conf. Med. Image Comput . Comput.-Assisted Intervention*, 2002, pp. 298–306.

[18] S. K. Warfield *et al.*, "Simultaneous truth and performance level estimation (STAPLE): An algorithm for the validation of image segmentation," *IEEE Trans. Med. Imag.*, vol. 23, no. 7, pp. 903–921. Jul. 2004.

[19] S. Gordon *et al.*, " Evaluation of uterine cervix segmentations using ground truth from multiple experts," *Comput. Med. Imag. Graph.*, vol. 33, no. 3. pp. 205–216, Apr. 2009.

[20] M. B. Cuadra *et al.*, "Comparison and validation of tissue modelization and statistical classification methods in T1-weighted MR brain images," *IEEE Trans. Med. Imag.*, vol. 24, no. 12, pp. 1548–1565, Dec. 2005.

[21] T. Rohlfing *et al.*, "Extraction and application of expert priors to combine multiple segmentations of human brain tissue," in *Proc. 6th Int. Conf. Med. Image Comput . Comput.-Assisted Intervention*, 2003, pp. 578–585.

[22] A. S. Dewalle-Vignion *et al.*, "Is STAPLE algorithm confident to assess segmentation methods in PET imaging?" *Phys. Med. Biol.*, vol. 60, no. 24, pp. 9473–9491, Nov. 2015.

[23] F. Mattern *et al.*, "Adaptive performance-based classifier combination for generic object recognition," in *Proc. Int. Fall Workshop Vis., Modeling Vis.*, 2005, pp. 139–146.

[24] M. Krenn *et al.*, "Creating a large-scale silver corpus from multiple algorithmic segmentations," in *Proc. Med. Comput. Vis. Workshop*, 2015, pp. 103–115.

[25] H. Wang *et al.*, "Multi-atlas segmentation with joint label fusion," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 3, pp. 611–623, Mar. 2013.

[26] A. Mansoor *et al.*, "A statistical modeling approach to computer-aided quantification of dental biofilm," *IEEE J. Biomed. Health Informat.*, vol. 19, no. 1, pp. 358–366, Jan. 2015.

[27] J. E. Cates *et al.*, "GIST: An interactive, GPU-based level set segmentation tool for 3D medical images," *Med. Image Anal.*, vol. 8, no. 3, pp. 217–231, Sep. 2004.

[28] Y. Ferdi *et al.*, "R wave detection using fractional digital differentiation," *ITBM-RBM*, vol. 24, no. 5, pp. 273–280, Dec. 2003.

[29] M. Benmalek and A. Charef, "Digital fractional order operators for R-wave detection in electrocardiogram signal," *IET Signal Process*., vol. 3, no. 5, pp. 381–391, Sep. 2009.

[30] S. Chernenko. (2012). ECG Processing. R-peaks detection [Online]. Available: http://www.librow.com/cases/case-2

[31] J. F. Kaiser, "Nonrecursive digital filter design using the I_0-sinh window function," in *Proc. IEEE Int. Symp. Circuits Syst.*, 1974, pp. 20–23.

[32] A. B. Williams and F. J. Taylor, *Electronic Filter Design Handbook*, 4th ed. New York, NY, USA: McGraw-Hill, 2006.

[33] H. Liang *et al.*, "Heart sound segmentation algorithm based on heart sound envelogram," in *Proc. IEEE Comput . Cardiol*, 1997, pp. 105–108.
[34] M. Lentini *et al.*, "Long ECG and pattern extraction," in *Proc. Int. Conf. Appl. Math . Inform.*, 2013.
[35] M. Sartor *et al.*, "Nicht-lineare zeitnormierung im langzeit-EKG," in *Bildverarbeitung Für Die Medizin* . Berlin, Germany: Springer-Verlag, 2014, pp. 300–305.
[36] A. Savitzky and M. J. Golay, "Smoothing and differentiation of data by simplified least squares procedures," *Anal. Chem.*, vol. 36, no. 8, pp. 1627–1639, Jul. 1964.
[37] R. A. Robergs and R. Landwehr, "The surprising history of the 'HRmax= 220-age' equation," *J. Exercise Physiol.*, vol. 5, no. 2, pp. 1–10, May 2002.
[38] H. Khamis *et al.*, "QRS detection algorithm for telehealth electrocardiogram recordings," *IEEE Trans. Biomed. Eng.*, vol. 63, no. 7, pp. 1377–1388, Jul. 2016.
[39] O. Commowick and S. K. Warfield, "Incorporating priors on expert performance parameters for segmentation validation and label fusion: A maximum a posteriori STAPLE," in *Proc. Med. Image Comput. Comput.-Assisted Intervention*, 2010, pp. 25–32.
[40] A. P. Dempster *et al.*, "Maximum likelihood from incomplete data via the EM algorithm," *J. Roy. Statist. Soc. Series B Statist. Methodol.*, vol. 1, pp. 1–38, Jan. 1977.
[41] J. P. Couderc, "The telemetric and Holter ECG warehouse initiative (THEW): A data repository for the design, implementation and validation of ECG-related technologies," in *Proc. Annu. Int. Conf. IEEE Eng. Med . Biol. Soc.*, 2010, pp. 6252–6255.
[42] S. J. Redmond *et al.*, "Electrocardiogram signal quality measures for unsupervised telehealth environments," *Physiol. Meas.*, vol. 33, no. 9, pp. 1517–1533, Aug. 2012.
[43] A. L. Goldberger *et al.*, "Physiobank, physiotoolkit, and physionet components of a new research resource for complex physiologic signals," *Circulation*, vol. 101, no. 23, pp. 215–220, Jun. 2000.
[44] A. Cordes *et al.*, "A portable magnetic induction measurement system (PIMS)," *Biomed. Techn.*, vol. 57, no. 2, pp. 131–138, Apr. 2012.
[45] M. M. Baig *et al.*, "A comprehensive survey of wearable and wireless ECG monitoring systems for older adults," *Med. Biol. Eng. Comput.*, vol. 51, no. 5, pp. 485–495, May 2013.
[46] R. Bond *et al.*, " Data driven computer simulation to analyse an ECG limb lead system used in connected health environments," *Methods Inf. Med.*, vol. 55, no. 3, pp. 258–265, 2016.

Authors' photographs and biographies not available at the time of publication.