Real-Time Correction of Heart Interbeat Intervals

Jeromie Rand[1], Adam Hoover[1], Stephanie Fishel[2], Jason Moss[2], Jennifer Pappas[2], Eric Muth[2]
[1] Electrical and Computer Engineering Department
[2] Department of Psychology
Clemson University
Clemson, SC 29634
ahoover@clemson.edu, muth@clemson.edu

Abstract

Heartrate variability (HRV) is traditionally analyzed while a subject is in a controlled environment, such as at rest in a clinic, where it can be used as a medical indicator. This paper concerns analyzing HRV outside of controlled environments, such as on an actively moving person. We describe automated methods for inter-heartbeat interval (IBI) error detection and correction. We collected 124,998 IBIs from 18 subjects, undergoing a variety of active motions, for use in evaluating our methods. Two human graders manually labeled each IBI, evaluating 10IBIs as having an error, which is a far greater error percentage than has been examined in any previous study. Our automated method had a 96% agreement rate with the two human graders when they themselves agreed, with a 49% rate of matching specific error corrections and a 0.01% false alarm rate.

1 Introduction

The electrical activity of the heart, as measured by the electrocardiogram (ECG), can be used to construct an event series that indicates the time between individual heartbeats. Heartrate variability (HRV) analysis studies cyclical variations in a heartbeat series related to autonomic nervous system activity [13]. However, the process of discretizing the raw electrical signal and detecting individual heartbeat intervals is prone to errors. Even a small amount of mis-detected heartbeats is known to have a damaging impact upon any subsequent HRV analysis [10, 2]. This paper considers the problem of automatically detecting and correcting errors in real-time, to provide a cleaner series for subsequent HRV analysis.

In a clinical setting, ECG artifacts can be minimized by restricting subject motion, for example to bedrest, or by averaging signals from multiple ECG leads. Those errors that do appear are generally fixed by a human expert reviewing the data before proceeding to (off-line) HRV analysis. Our work is motivated by the potential for real-time HRV analysis in an unstructured environment. For example, a soldier could be equipped with an HRV monitor while on active duty [9]. Adaptive automation systems, such as a cruise control in an automobile, could be driven by an on-line HRV monitor [3]. During exercise, a person could monitor their own HRV in addition to the traditional heartrate. These types of systems must be able to work without a human expert in the loop. They must be able to work in real-time, to provide a useful feedback signal. They must be able to work in an environment where the subject is ambulatory, so that artifacts caused by muscle noise are the norm rather than the exception. In our data set (described later), we witnessed a 10% error rate in individual heartbeat detections, because of our active subjects. In prior studies, the error rate is generally found to be much less than 1%, because the data is captured on sedate subjects or in a controlled environment.

Most of the literature concerned with ECG errors looks at the problem of reducing noise within the ECG signal, or at building a better heartbeat interval detector [5, 7]. However, some amount of heartbeats are always going to be mis-detected, especially as the subject becomes active. This motivates the use of additional error detection after individual heartbeat detection. We use a small window of heart interbeat intervals (IBIs) to look for mis-detections. The context of IBIs immediately surrounding the one under consideration provides an indication as to the validity of the IBI. We call this process IBI error detection and correction.

Cheung [4] was perhaps the first to recognize the problem. He developed an algorithm to detect and correct errors by comparing an IBI with both its predecessor and its successor. Deviations beyond a certain percentage were considered errors. However, Cheung recognized that combinations of these errors or multiple errors found in series defeated his method, so that the method could only be applied to data containing limited errors.

Malik et. al. [8] explored comparing an IBI to the average IBI in an entire recording, and to the last accepted (assumed to not have an error) IBI. Evaluation was done on 24 hour recordings taken of subjects who had just undergone acute myocardial infarction, and normal subjects. Results were evaluated based upon distribution comparisons, rather than on manually graded IBIs. Therefore the percentage of IBIs containing errors is unknown, as is the actual success rate of detecting and correcting the errors.

Berntson et. al. [1] extended IBI error detection to include validation of the flagged error. The usual comparison of an IBI to its predecessor is used to detect

a potential error. Subsequently, the following IBI is compared to the supposed error to determine whether the detected problem was an actual error or a false alarm. A total of 15,360 IBIs were manually classified, with 33 being in error (about 0.002were also added to further test their methods, but these errors did not represent what could be expected from active real subjects.

Sapoznikov et. al. [12] explored the idea of filtering an IBI series using a polynomial smoothing function. A large degree-of-freedom polynomial was fit to a window of IBI data, and then subtracted from the IBI series, in order to eliminate errors that resemble spikes. Subsequently, an error is detected by comparing an IBI with an updated mean IBI from the window, and the last accepted (assumed to not have an error) IBI. Visual evaluation was performed by two experts on data from 3 subjects during 6 hour sleep periods. Because the IBIs were not manually graded, the overall percentage of errors in the data is unknown. However, it can be assumed to be small because the subjects were sleeping.

Compared to all the previously discussed methods, we extend the handling of IBI error detection and correction in several important aspects. First, we make no assumptions about the motions of the subject. The subject may be at rest, normally moving (e.g. in an office environment), or fully active (e.g. in exercise or combat). Second, we assume that the process must run on-line and in real-time. Therefore it cannot rely upon statistics of entire recordings, but must calculate any needed statistics on the fly as it runs. Third, we extend the previously discussed methods of IBI error detection, correction, and validation to a more comprehensive rule set. This rule set covers combinations of error types, as well as sequences of multiple errors. Finally, we quantitatively evaluate our methods on over 100,000 manually graded IBIs, containing a far larger percentage of errors than in previously reported studies.

2 Methods

Figure 1 shows an overview of our methods. In this view we may consider IBIs a commodity that are produced by an IBI detector, cleaned by our "engine", and consumed by an IBI analysis method. The IBI detector is assumed to be running independent of our engine. Heartbeats occur asynchronously, so of course the input to our engine is asynchronous. It is assumed that an IBI value is given to our engine as soon as the heartbeat ending the interval is detected. Inside the engine, a buffer of recent IBI values is maintained. Error detection and correction is performed somewhere in the middle of this buffer (discussed in detail below), so that there are always a set of preceding IBIs that have already been validated, and a set of succeeding IBIs waiting to be validated. All these surrounding IBIs serve to provide context for the IBI under consideration. As time proceeds, IBIs move from right to left in the buffer. The engine provides a synchronous output, essentially an oversampled but time-delta consistent estimate of the IBI series, in addition to the corrected asynchronous IBI series. Both outputs are delayed approximately 6 seconds past the original IBI detector (discussed more below).

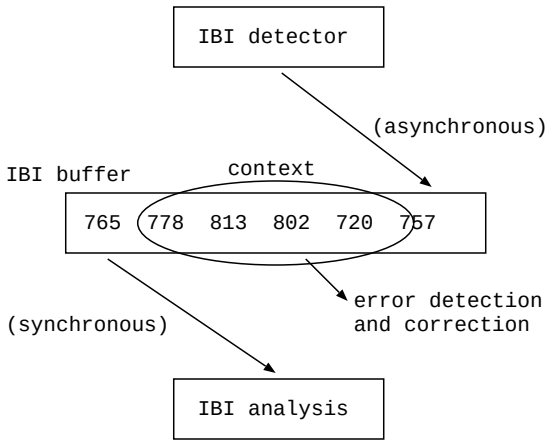

Figure 1: Process view of IBI detection, error analysis, and IBI analysis.

Section 2.1 describes our methods for automated IBI error detection and correction, which we call the analysis engine. Section 2.2 describes the data set, consisting of 124,998 IBIs, used to evaluate our methods. Our engine uses several parameters that can be adjusted to obtain optimal performance. Section 2.3 describes the parameters and the methods we used to choose the best set.

2.1 Engine

Figure 2 shows a detailed view of the heart of the engine, which is a buffer of IBI values. The engine maintains a pointer to the position in the buffer of the IBI under consideration. IBIs ahead of the IBI under consideration (left of it in the figure) have already been tested for errors, and possibly replaced with corrected values. IBIs behind the IBI under consideration have been detected but are awaiting consideration, and are referred to as incoming values. The buffer must be of sufficient size to allow all the necessary information contained in the sequence of IBIs to be used in

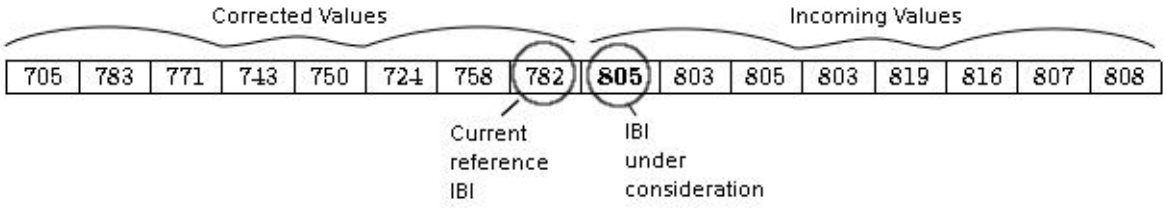

Figure 2: A snapshot of the buffer used for correction context.

the decision making process. However, we desire to keep the buffer as small as possible, to minimize the delay between IBI detection and cleaned IBI output. Note that output occurs at the "current reference IBI" in Figure 2. The additional corrected IBIs in the buffer are maintained only to provide better information on how to analyze the IBI under consideration.

To arrive at a buffer size, we consider the buffer in two parts. The input size of the buffer must be large enough to accommodate the maximum number of IBIs involved in a correction. For our rule set (described below), this number is three. The maximum expected value for an IBI is 1500ms (they are of course much shorter under most circumstances). It is also necessary to have at least one IBI received after the currently evaluated IBI in order to decide on the appropriateness of an attempted correction. Therefore our method buffers a minimum of $(3+1) \times 1500 =$ six seconds of incoming data in order to provide the minimum context for evaluating corrections. The second part of the buffer, on the output side, maintains a "past history" of corrected IBI values. Statistics of these data are used during the analysis of the IBI under consideration. A longer past history does not increase the lag of the system, therefore the actual length of the past history is a parameter we train in order to determine optimal performance.

Since IBI data arrive asynchronously, in practice we are unable to guarantee a fixed amount of time in the buffer, so we simply try to keep the time in the buffer as close to six seconds as possible. For example, Figure 2 shows 6.76 seconds of incoming data, including the IBI under consideration. The reason our buffer maintains a fixed amount of time instead of working on a fixed number of IBI values is primarily to facilitate the resampling of IBI values into synchronous data for HRV analysis. Specifically, spectral analysis through the use of the Fourier transform requires synchronously sampled data to provide meaningful results [6].

The first task of the engine is to determine whether or not the IBI under consideration is an error. The engine uses an adaptive threshold to compare the IBI under consideration against the current reference IBI. The threshold is calculated anew for each IBI under consideration according to Equation 1.

$$T = \frac{\sum_{i=0}^{N-2} |b_i - b_{i+1}|}{N-1} m \qquad (1)$$

In the equation, T is the threshold, N is the number of values maintained in the past buffer, b_i represents the values in the past buffer, and m is a multiplier. During operation of the engine, N and m are fixed; the specific values used during our experiments are discussed in Section 2.3. In plain language, the mean successive difference of the trusted values in the past buffer is multiplied by a constant to generate a threshold. That threshold is limited to a predefined range of possible values, also fixed at runtime and also trained for as described in Section 2.3. If the buffer is not yet full, a reasonable estimate for T is used until the buffer is full and meaningful statistics can be calculated. During the experiments described in this paper, a default value of 100ms was used.

The portion of the engine that detects errors can be temporarily suspended if too many errors occur in a particular neighborhood. This capability is needed to minimize the potential for the engine to generate a sequence of data that appears to be valid but in fact has locked into an error pattern that deviates from the true IBI series. In practice, an observation of human graders correcting IBI data indicated that more than 2 or 3 consecutive corrections are rarely, if ever, used. If corrections are occurring at a higher frequency, the area is almost certainly uncorrectable. In order to establish this behavior, a counter is incremented each time a correction is applied and decremented each time an IBI is marked as valid. If the counter exceeds a value of 3, the next IBI is not examined for potential errors and the counter is again decremented.

The second task of the engine is to correct an IBI under consideration marked as an error. Our methods use a set of rules. The rules are designed to emulate how a human expert would manually grade the

Correction	Description
Hardware trigger	Replace value outside of allowed hardware range
Split	Missed heartbeat; divide IBI into two equal values
Split 3	Two missed hearbeats; split IBI into three equal values
Combine	False trigger; add two IBIs together
Combine 2 / Split 2	Replace two IBI values with their average
Combine 2 / Split 3	Get three new IBI values as average of two
Combine 3 / Split 3	Replace three IBI values with their average
Physiological trigger	Replace value outside of allowed physiological range
Uncorrectable	Could not apply any rule, but IBI appears faulty

Table 1: Possible corrections applied by program.

IBI data. Each rule represents a possible error type. The rules, in the order they are applied, are listed in Table 1. For each rule, the correction is applied, and the result is evaluated against the context of surrounding IBIs in the buffer. The order in which possibilities are inspected is significant, as the first classification that is accepted terminates the search process. In order to be deemed an acceptable correction, the corrected value must lie within some threshold of both the current reference IBI and the IBI value following the set of IBIs used for the correction. This threshold is generated in a manner identical to the threshold for detection described in Equation 1. The actual values for the multiplier, threshold minimum, and threshold maximum used for correction acceptance may differ from the values used for detection. Like the values used in the detection process, these were trained for optimal performance as described in Section 2.3.

2.2 Data

Evaluation of our methods was based on IBI data gathered from 18 healthy Clemson University students between the ages of 18 and 24. The subjects performed a series of active motions that have a high probability of inducing errors. Participants completed 2 sets of the following tasks, each task lasting 2 minutes: punching arms, jumping jacks, running in place, and crunches. Between each task was a 1 minute rest period. Prior to all these tasks was a 10 minute baseline, during which the subject sat at rest, and a second 10 minute baseline, in which a person engaged in reading and light sitting activities. Following all the tasks was an 8 minute cooldown period, where again the subject sat at rest.

Each subject repeated the set of tasks three times using a different IBI detecting device each time. This was done to test the variance in the percentage of IBI errors produced by different devices. The devices were the Polar S810 (Lake Success, NY), the Biolog 3991 (UFI Corp., Morro Bay, CA) with standard electrodes, and the Biolog 3991 with fetrodes, a sensor patented and sold by UFI.

The data were recorded for purposes of training our method and evaluating its performance against two human evaluators. Although the data were analyzed from stored recordings, it is important to note that the data were processed by our methods as though they were appearing during live, real-time operation. No advantage was gained by operating on pre-recorded series of data.

In all, 54 recordings containing 124,998 IBIs were collected and evaluated by human graders. Two Clemson University graduates students were given training on the correction of IBI data and asked to independently inspect the data for errors. Their decisions were combined into a single ground truth by accepting all error *detections* the human graders agreed upon (even if they did not necessarily agree upon the correction to apply to the error). In total, the human graders agreed that an IBI was either valid or in error on 119,346 IBIs. Only those decisions for error detection that the human graders agreed upon were used to train our methods, and to evaluate the performance of our methods on identifying actual error type.

2.3 Training

In order to achieve the best results, we used a portion of the data set to train the adjustable parameters introduced in Section 2.1 according to the train and test paradigm. The parameters trained all deal with the threshold generation process. As explained in Section 2.1, the threshold for detecting an error is computed anew for each IBI, and a second threshold is computed for accepting the computed correction. Both of these thresholds are limited to minimum and maximum values, which are constant while the engine is running. Table 2 shows a summary of the seven variables controlling this process, along with the range of values used for training.

Data from six of the 54 recordings were used for training, providing 15,095 total IBI values. All these IBIs were run through the engine with every possible

Parameter	Possible values
N	3, 5, 10, 15, 25
m_{detect}	1, 5, 10, 25, 50
min(m_{detect})	10, 50, 70, 100, 150
max(m_{detect})	10, 50, 100, 150, 200
m_{accept}	1, 5, 10, 25, 50
min(m_{accept})	10, 50, 70, 100, 150
max(m_{accept})	10, 50, 100, 150, 200

Table 2: Parameters trained for engine.

combination of parameter values listed in Table 2. For each set of parameter values, we calculated the sum of the IBI values that both humans and the computer agreed were correct and those that all agreed were a specific error. The maximum sum out of all possible 5^7 sets of parameter values was deemed the best result.

In deciding what range of values to search for each parameter, we performed several combinatorial searches similar to the one just described. We sought a set of parameter ranges that had most of the largest sums of IBI agreement in the middle of the 7D space, instead of near the edges. The parameter ranges listed in Table 2 are the final 7D space used to train our method. The parameters in that particular 7D space determined to give the best results were $N = 5$, $m_{detect} = 10$, $min(m_{detect}) = 50$, $max(m_{detect}) = 200$, $m_{accept} = 25$, $min(m_{accept}) = 10$, and $max(m_{accept}) = 100$.

3 Results

In order to provide a ground truth for testing our methods, the human graders were first evaluated against one another. As mentioned in Section 2.2, the human graders agreed that an IBI was either correct or in error 119,346 out of 124,998 cases. We further broke down the 119,346 agreed detection cases by the error correction applied. Table 3 presents those results. The percentage rate of agreement for each error type was computed as

$$Rate = \frac{Agree}{Human1 + Human2 - Agree} \quad (2)$$

where $Agree$ is the number of values the human graders agreed upon for each correction type, $Human1$ is the number of times the first human grader used that correction type, and $Human2$ is the number of times the second human grader used that correction type.

Several important observations can be made from the results of the human versus human comparison. First, out of 124,998 original IBIs, the human graders agreed on how to correct 119,346 total IBIs, or 95% of the cases[1]. Out of these agreements, 113,291 IBIs were agreed to be error-free, and 6,055 (5%) were agreed to exhibit a specific error or be uncorrectable. These numbers show a far higher percentage of errors than have been previously reported in the literature, largely due to the active motions of our subjects. Putting the disagreements together with those agreed as having a specific error, we are faced with a data set containing 10% errors. This suggests a strong need for a reliable error detection and correction method if HRV is to be successfully used in an automated fashion on active subjects.

Another important observation from Table 3 is which error correction types dominated the decisions made by the human graders. Excluding normal IBIs, the most common decision was to label an IBI or sequence of IBIs as uncorrectable. This was followed by the combine and split correction and then by the split correction. For erroneous IBIs that the human graders agreed upon, these three corrections made up 89% of the total corrections. Several of the corrections were used an insignificant portion of the time. Combine 3, delete, replace, and replace N were all used on less than 0.1% of the total data set, and combine 3 and delete were never agreed upon by the human graders. These results suggest that an automated correction method does not need a large set of options to determine how to correct an error.

Another important observation from Table 3 is that the agreement rate between the human graders was relatively low when deciding upon what correction to apply. While they were able to agree if an IBI was correct 96% of the time (correction type is “do nothing”), the best rate of agreement for choosing a correction to apply was the split correction at 77%. There is a rapid fall off in the rate of agreement after this, with only 2 of the other 12 possible corrections having a rate of agreement over 50%. The low agreement rates imply that the problem is difficult, and suggests that corrective measures should err more on the side of caution; if the correct data stream is not obvious, then the area should be marked as uncorrectable.

The 119,346 IBIs for which the human graders agreed upon a correction were used to analyze the performance of our automated method. Table 4 shows

[1]We stressed to the human graders to make every possible attempt to correct an IBI, before labeling it as uncorrectable, because we wanted to correct as many as possible. This likely led to the low 95% agreement rate.

Correction Type	Human 1 (# of IBIs)	Human 2 (# of IBIs)	Agree (# of IBIs)	Percent Agreement
Nothing	115843	115015	113291	96%
Combine	268	175	162	58%
Combine 3	0	0	0	–
Split	1488	1390	1256	77%
Split 3	253	266	195	60%
Delete	2	11	0	0%
Replace	54	63	4	4%
Uncorrectable	3510	5258	2790	47%
Combine and Split	2653	2099	1370	41%
Combine 3 / Split 2	319	165	102	27%
Combine 2 / Split 3	301	313	104	20%
Combine 3 / Split 3	275	223	69	16%
Replace N	32	20	3	6%

Table 3: Human vs human agreement rates by correction type.

our method's performance at classification for these data. In this table our method is labeled "computer". Overall, our method agreed with the human graders in 114,761 out of 119,396 IBIs, or 96% of the cases. Breaking that down, out of 113,291 IBIs where the humans agreed the IBI was not in need of correction, our method applied some correction in $113,291 - 111,820 = 1,471$ cases, or $1,471/113,291 = 0.01\%$ cases. Out of the remaining 6,055 IBIs where the humans agreed on the error correction, our method identified the same correction in 2,941 (49%) cases. These numbers show a very low false alarm rate for our method (0.01%) but that there is a great deal of improvement possible in correcting errors.

Additional analysis of our results may be found in [11].

4 Conclusions

This paper presents a method for the automatic correction of IBI data. This automated method was evaluated against a pair of human graders using 124,998 IBIs, about 5% of which the human graders did not agree upon, and about 5% of which the human graders agreed upon a specific error correction. The results show a 96% overall agreement rate for IBIs on which the human graders themselves agreed. The automated method identified the same correction as the humans on 49% of the erroneous IBIs, with a 0.01% false alarm rate, on this challenging data set. These rates are certainly not high enough for use in sensitive clinical studies, but we believe they are a step in the right direction for use in automated, high activity tasks.

This study adds to the body of literature on the automated correction of heart interbeat intervals. In particular, this is the only study we are aware of that attempts to automatically correct heavily corrupted real IBI data. Our methods work on the data as it is acquired, so it is available for real-time feedback; this could find use in research problems where an offline correction process is not feasible. One improvement that could be made to our methods would be to evaluate subsequences of IBIs as groups. Currently, although our methods use neighboring IBIs for context, each IBI is still evaluated individually. Evaluating a neighborhood of IBIs might be particularly helpful when processing an uncorrectable area. Further study is required.

Acknowledgments

This work was partially supported by an Office of Naval Research grant and by Honeywell Corporation.

References

[1] G. Berntson, K. Quigley, J. Jang and S. Boysen, "An Approach to Artifact Identification: Application to Heart Period Data", in *Psychophysiology*, vol. 27 no. 5, 1990, pp. 586-598.

[2] G. Berntson and J. Stowell, "ECG ertifacts and heart period variability: Don't miss a beat!", in *Psychophysiology*, vol. 35, 1998, pp. 127-132.

Correction Type	**Computer** (# of IBIs)	**Humans** (# of IBIs)	**Agree** (# of IBIs)	**Percent Agreement**
Nothing	113986	113291	111820	97%
Combine	383	162	132	32%
Combine 3	0	0	0	–
Split	1324	1256	993	63%
Split 3	170	195	130	55%
Delete	0	0	0	–
Replace	416	4	0	0%
Uncorrectable	1636	2790	1101	33%
Combine and split	645	1370	469	30%
Combine 3 / Split 2	212	102	52	20%
Combine 2 / Split 3	495	104	46	8%
Combine 3 / Split 3	79	69	18	14%
Replace N	0	3	0	0%

Table 4: Computer vs human agreement rates by correction type.

[3] W. Boucsein and S. Miyake, chairs, "Real-Time Psychophysiological Measures for Adaptive Automation Systems", abstracts from symposium session in *Psychophysiology*, vol. 42 issue s1, Sep 2005, pp. S25-S26.

[4] M. Cheung, "Detection of and Recovery from Errors in Cardiac Interbeat Intervals", in *Psychophsyiology*, vol. 18 no. 3, 1981, pp. 341-346.

[5] G. Friesen, T. Jannett, M. Jadallah, S. Yates, S. Quint and H. Nagle, "A comparison of the noise sensitivity of nine QRS detection algorithms", in *IEEE Trans. on Biomedical Engineering*, vol. 37 no. 1, Jan 1990, pp. 85-98.

[6] A. Hoover and E. Muth, "A Real-Time Index of Vagal Activity", in *Int'l Journal of Human-Computer Interaction*, vol. 17 no. 2, 2004, pp. 197-209.

[7] B.-U. Kohler, C. Hennig and R. Orglmeister, "The Principles of Software QRS Detection", in *IEEE Engineering in Medicine and Biology Magazine*, Jan/Feb 2002, pp. 42-57.

[8] M. Malik, T. Cripps, T. Farrell and A. Camm, "Prognostic value of heart rate variability after myocardial infarction. A comparison of different data-procesisng methods", in *Medical & Biological Engineering & Computing*, Nov 1989, pp. 603-611.

[9] E. Muth, A. Kruse, A. Hoover and D. Schmorrow, "Augmented Cognition: Aiding the soldier in high and low workload environments through closed-loop human-machine interactions", Chapter 6 in *Military Life: The Psychology of Serving in Peace and Combat*, edited by T. Britt, C. Castro and A. Adler, Praeger Security International, 2006, pp. 108-127.

[10] S. Porges and E. Byrne, "Research methods for measurement of heart rate and respiration", in *Biological Psychology*, vol. 34, Nov 1992, pp. 93-130.

[11] J. Rand, "Real-Time Correction of Heart Interbeat Intervals", master's thesis, Electrical & Computer Engineering Dept., Clemson University, 2005.

[12] D. Sapoznikov, M. Luria, Y. Mahler and M. Gotsman, "Computer processing of artifact and arrhythmias in heart rate variability analysis", in *Computer Methods and Programs in Biomedicine*, vol. 39, 1992, pp. 75-84.

[13] Task Force of the European Society of Cardiology and the North American Society of Pacing and Electrophysiology, "Heart rate variability: standards of measurement, physiological interpretation, and clinical use", in *Circulation*, vol. 93 no. 5, March 1996, pp. 1043-1065.