August 5, 2005

To the Graduate School:

This thesis entitled "Real-time correction of heartbeat interbeat intervals" and written by Jeromie R. Rand is presented to the Graduate School of Clemson University. I recommend that it be accepted in partial fulfillment of the requirements for the degree of Master of Science with a major in Computer Engineering.

Adam Hoover, Advisor

We have reviewed this thesis
and recommend its acceptance:

Eric Muth

Stan Birchfield

Accepted for the Graduate School:

Real-time correction of heartbeat interbeat intervals

A Thesis

Presented to

the Graduate School of

Clemson University

In Partial Fulfillment

of the Requirements for the Degree

Master of Science

Computer Engineering

by

Jeromie R. Rand

August 2005

Advisor: Dr. Adam Hoover

ABSTRACT

This thesis considers the problem of monitoring a person's heartbeats using an electrocardiogram (ECG) and detecting and correcting misdetected beats in real-time. There are only two basic types of errors in interbeat interval data. An event that was present can be missed by the detector, or a spurious event can be added to the detection stream. This thesis considers the correction of these errors in a variety of combinations and from a variety of sources. It is important to detect and correct these errors because of the impact they can have on further analyses, especially in the field of heart rate variability (HRV).

Incoming interbeat interval (IBI) values are placed into a buffer as they are received. The corrections we consider involve a maximum of three IBI values. Our method buffers a minimum of six seconds of incoming data in order to provide enough context for evaluating corrections. The expected value of an incoming IBI is calculated as an acceptable deviation from the last known good value. If an IBI is flagged as a possible error, it is evaluated within its context by a series of rules designed to detemine the most likely error type that may have occurred.

Evaluation of the method was based on IBI data gathered from 18 healthy Clemson University students between the ages of 18 and 24. In all, fifty-four files containing 124,998 usable IBIs were collected and graded by two Clemson University graduate students. Six of the fifty-four files were used for training, while the others were reserved as a test set. This amounted to 15,095 IBI values to train on.

Results of analysis of the test set indicate a 97.4% agreement on classification in areas that humans deem to be correctable. Agreement on classification is significantly lower, but it follows the trend of human grader agreements. The method in its current state is probably not complete enough for use in sensitive clinical studies, but it may form the basis of a useful compoment of more complex systems designed for HRV analysis.

ACKNOWLEDGMENTS

I am indebted to Dr. Adam Hoover, my advisor, for his guidance throughout the term of this work. Without his helpful suggestions, words of motivation, and nearly unending patience this thesis would not exist.

I also owe a great debt to Dr. Eric Muth, whose insights into the psychological and physiological phenomena involved in my research provided a firm foundation for this work to build on.

I am grateful to Stephanie Jones, Jennifer Pappas, and Jason Moss for providing the data that made this thesis possible.

I would also like to thank Dr. Stan Birchfield for agreeing to serve on my Master's examination committee.

TABLE OF CONTENTS

List of Tables

Table Page

LIST OF FIGURES

Figure Page

Chapter 1

Introduction

1.1 Background

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 heart beats. The process of discretizing the raw electrical signal can introduce errors, however. Because the generated event series is often used as an input to other analyses in the realm of heart rate variability (HRV), it is essential that the errors in this data stream are kept to a minimum. Since it may also be useful to perform these analyses as the data are collected, it it also important that the data can be corrected as it is gathered. This thesis considers the problem of monitoring a person's heartbeats using an electrocardiogram (ECG) and detecting and correcting misdetected beats in real-time.

In order to understand the process of correcting interbeat interval data, it is useful to examine the underlying nature of the signal that is being sensed. The electrocardiogram, or ECG, is a measure of the average electrical activity in the heart at a particular moment in time. Because of the regular activity of the heart, the resulting waveform has a general shape that is repeated over time. This waveform has a series of peaks and troughs that correspond to significant cardiac events. Figure 1.1 shows an example of this waveform.

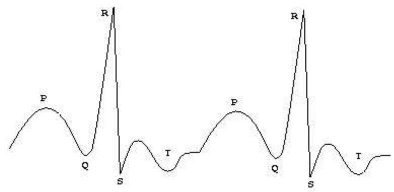


Figure 1.1: An example of ECG data.

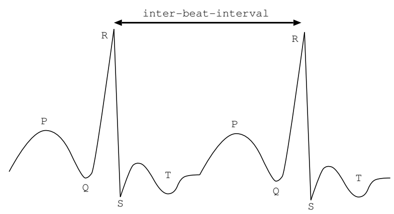


Figure 1.2: Definition of a interbeat interval (IBI) in ECG data.

There are three major sections of an ECG signal. The P wave is the beginning of a heartbeat and corresponds to atrial depolarization [9]. It is typically positive and rounded. The P wave usually lasts less than 120 milliseconds [7]. The QRS complex is a series of waveforms representing the ventricular muscle depolarization that are usually grouped together for analysis. The Q wave is the first negative portion of the ECG after the P wave. The R wave is a positive wave following the Q wave, and the S wave is another negative portion following the Q wave. Not all wave forms are necessarily present (the Q wave in particular may not be seen), but the grouping is known as the QRS complex even if some elements are missing. It is typically of shorter duration than the P wave, and the R portion has a relatively high magnitude [7]. The T wave is caused by the repolarization of the ventricular muscles. It is typically the first positive activity after the QRS complex.

While the P wave marks the beginning of heart activity in each cycle, the QRS complex is usually used to denote the time between cardiac events because its more prominant nature allows more reliable detection and better accuracy [14]. Figure 1.2 shows the typical definition of an interbeat interval (IBI), also known as an RR interval, in ECG data.

A number of different devices of varying quality can be used to gather IBI data, but they all work on a common principle. Electrodes in contact with the skin measure the electrical impulses of the heart. Some sort of timer runs until an event is triggered by the detection of a QRS complex. The time between events is reported as the length of the interbeat interval. There are two common ways of detecting the QRS complex. If implemented in hardware, the trigger typically utilizes a level detector, possibly in combination with a slope detector. This method introduces some errors. Because it is not guaranteed to record the peak of the R wave, there is some slight variation in the point at which the wave is recorded. The more accurate approach is to use a software detector that detects the inflection point of the R wave. This method is typically accurate to the nearest millisecond [27].

By definition, the detection of the QRS complex is a binary event: either the wave is present or it is not. Because of this, there are only two basic types of errors in interbeat interval data. An event that was present can be missed by the detector, or a spurious event can be added to the detection stream. All other errors are combinations of these two error types. There are two places artifacts can be introduced into the IBI stream. The ECG itself may become noisy or corrupt or the IBI detector may do a poor job of detecting the QRS complex. Problems in the ECG can arise from several reasons. Loose electrodes or broken wires will cause a wavering baseline with unusual waveforms throughout as seen in Figure 1.3. Muscular tremors, caused by tensed muscles, shivering, or health problems, will appear as an irregular baseline that obscures true waveforms. This is demonstrated in Figure 1.4. Figure 1.5 shows the effect of patient movements, which show up in a similar manner as muscle tremors. Electrical interference can be another cause of a poor ECG [7]. Figure 1.3 through Figure 1.5 are adapted from illustrations in Guide to ECG Analysis [7].

Physiological phenomena can also introduce errors or areas that appear to be errors into the ECG. Heart irregularities are an obvious source of unusual ECG data, but normal activity can also have an effect on the activity of the heart. In particular, the rate of breathing has a significant impact on heart activity [28]. Methods have been proposed for the removal

Figure 1.3: ECG data with loose electrodes[7].

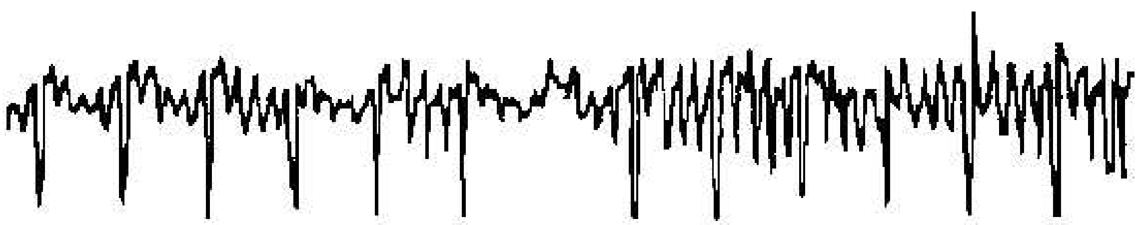
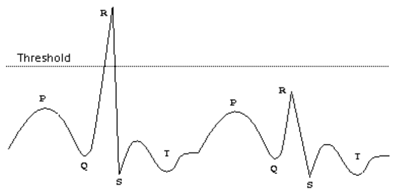


Figure 1.4: ECG data with muscle tremors[7].

Figure 1.5: ECG data with patient movements[7].



(a) IBI error, missed (b) IBI error, split

Figure 1.6: Example data causing errors in IBI detection.

of noise in the ECG, but since they can introduce errors in the QRS area they may not be suitable for use in HRV analysis[31].

Figure 1.6 shows an example of a missed IBI and a spuriously detected IBI. In Figure 1.6(a), an R-wave is too low, causing the detector to miss and resulting in a doubled IBI value. In Figure 1.6(b), a P-wave is too high, crossing the threshold of the detector and causing it to return a false IBI that is a portion of the true IBI value. A variety of combinations of these two error types can occur (these are considered more fully in Chapter 2).

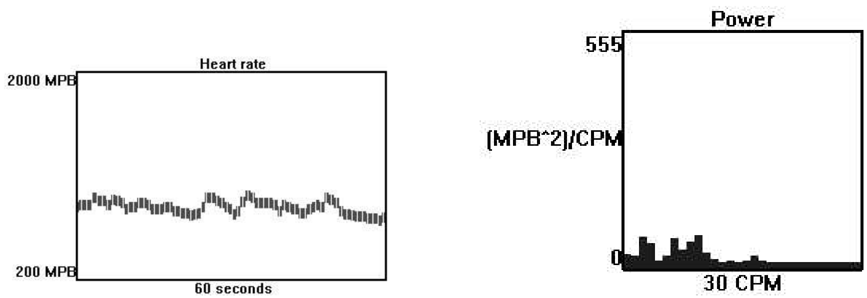
Heart rate variability is one of the primary applications of IBI data, and errors in an IBI stream can have a profound effect on this type of analysis [27]. There are three basic inputs that control heart rate variability. The sonoatrial node, also known as the pacemaker, has a tendency to keep the heart at a steady beat. Parasympathetic fibers have a tendency to increase IBIs. Activity in the parasympathetic system is affected by external stimuli and the sleep/wake cycle. Sympathetic nervous activity has a tendency to decrease IBIs. Activity in this system is strongly influenced by the environment [1].

Heart rate variability can be examined in the time domain or the frequency domain. Analyses in the time domain and frequency domain are both affected by artifacts. One common measure of heart rate variability in the time domain, the standard deviation of the differentiated RR (DRR) time series (rmSDD), is particularly susceptible to the effects of artifacts. Another common measure, the percentage of DRR values larger than 50ms (pNN50), is less susceptible to errors since every interval over 50ms is treated the same, but it has undesirable characteristics in the case of very low or very high HRV. García-

González and Pallàs-Areny [12] proposed a method that makes the rmSDD more robust by discarding outliers on the histogram of DRR values. This technique is primarily useful in patients with low heart rate variability, however, and does not supplant the need for artifact correction in all cases.

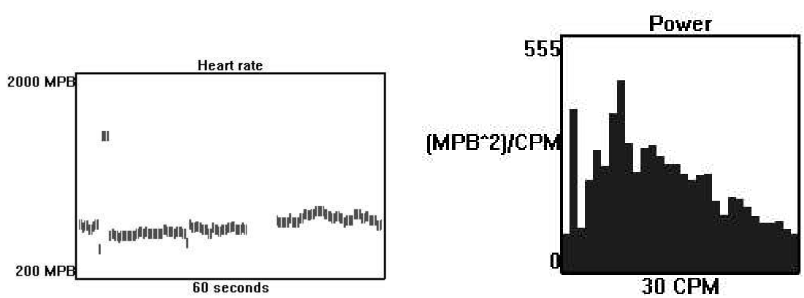
There are also several ways to examine HRV in the frequency domain. Autoregressive models and Fourier analysis are common choices. Autoregressive models tend to be better for short data segments because of the tendency for there to be some leakage between power segments with the fast Fourier transform (FFT) and the smaller resolution associated with using the FFT. The autoregressive model does have some disadvantages, such as the amplitude of the peaks not having a linear relationship with the sinusoidal power of the data in some models [29].

An example from Fourier analysis provides a good example of the effects of artifacts on HRV analysis. There are typically three components of the power spectrum of IBI data: a peak from 2-5 CPM, a peak centered around 6 CPM, and a peak from 9-30 CPM [17]. These peaks have been recognized in heart rate variability analysis since the early 1970s. The three peaks are related to respiratory frequency, arterial blood pressure control, and peripheral vasomotor regulation, respectively. The frequency and amplitude of the peaks is constantly changing in response to parasympathetic and sympathetic nervous activity [28]. Figure 1.7(a) shows an example of IBI data over a 60 second window, showing several cycles of activity at various frequency levels. Figure 1.7(b) shows the frequency space of a Fourier analysis of that data. The units on the x axis are cycles per minute (CPM). The units on the y axis are milliseconds per beat squared divided by cycles per minute. Notice the three peaks present in the data. Figure 1.8(a) shows an example of IBI data containing a missed beat error. Figure 1.8(b) shows how this affects the frequency space. Notice that all of the peaks are obscured by a huge influx of power across all frequencies. This is because an outlying IBI value severely distorts the appearance of all frequencies [30].



(a) Clean IBI series (b) Power analysis (Cycles Per Minute)

Figure 1.7: An example of power spectrum analysis of good IBI data.



(a) IBI series containing error (b) Power analysis

Figure 1.8: An example of power spectrum analysis of IBI data with errors.

1.2 Related Work

The study of heart rate variability is a field with a rich literature. There are a wide variety of applications for heart rate variability analysis. A number of studies have been done to explore the link between HRV analysis and autonomic nervous activity. A strong link has been established between HRV and parasympathetic nervous activity in the heart [2, 15, 16, 19, 23]. Abnormal parasympathetic nervous activity in the heart can be an element of diagnosing disease or other heart problems [5].

The examination of heart rate variability can play an important role in diagnosing the health of a patient [32]. Reduced heart rate variabilty can be an indication that patients are at a higher risk for mortality after acute myocardial infarction [10, 20, 21, 22, 24, 25]. This can provide a crucial element for the stratification of subjects after initial treatment. Heart rate variability can be an indicator of patients who may be at risk for a number of other health problems as well. Reduced HRV is an indicator of a higher general mortality risk in the elderly [33], and is a predictor of a number of possible heart problems in patients who otherwise appear to be healthy [34]. Several other health issues have links to heart rate variability. Studies have shown that particulate air pollution has a negative impact on heart rate variability [13, 26], and both depression [6] and panic disorder have been shown to be associated with abnormal HRV [11].

The activity of the autonomic nervous system can also be used to draw possible conclusions about the stress level or cognitive load of the subject [23]. Some researchers wish to use this information about the cognitive state of a person to create closed loop systems [17]. This could see a wide variety of applications in the field of augmented cognition.

While heavily edited data are not a replacement for good data and should only be used when no other option is available [18], the removal of some artifacts is usually necessary before subsequent analysis of the data is performed. The problem of IBI correction can be broken up into two parts: detection of erroneous IBI values and replacement of erroneous IBI values with a reasonable value [29]. There is no established method for automatically handling IBI error detection and correction. The correction of IBI data by hand is tedious and requires human graders to make subjective judgments in data with high variability. Some combination of automatic and manual correction is generally the preferred method, often allowing the machine to mark potential errors and letting the human decide if it is in fact an error and what correction to apply [3]. This work considers the problem of correcting IBI data without human intervention. Assuming a valid ECG signal, the only errors that can be introduced by an IBI detector are a falsely triggered threshold or a missed heartbeat. Cheung [8] shows that it is possible to develop an algorithm that can detect all errors of these types if some basic assumptions are met. Many errors that occur in IBI data are some combination of an IBI being split or combined with another IBI. Cheung's method is not guaranteed to succeed on these combinations, though the iterative nature of his algorithm allows it to correct many combinations of split and merge errors. Berntson et al [4]. describe another such method for the automatic correction of errors. It detects errors by evaluating an IBI population to determine the largest expected beat to beat difference for a normal IBI and the smallest expected difference for an error. A threshold is determined based on this information. If an IBI is above this threshold, it is further evaluated to see whether it matches a known error type. Based on the results of this evaluation, an error is corrected, marked as uncorrectable, or returned to a normal state. Berntson and his colleagues suggest an on-line implementation is possible with this algorithm examining a discrete window of accumulated statistics, but they do not provide further details of its implementation.

Sapoznikov et al [29]. describe several methods for detecting erroneous IBIs. First they tried removing values based on mean heart rate. Using a global heart rate was completely unacceptable for identifying errors. If a threshold was set that would detect all (or a reasonable number) of errors, it would cause a high rate of false positives. A moving updated mean performed better, but still had a high rate of false positives when heart rate underwent changes. Another method was to fit the HR data to a polynomial and mark values that deviated by a certain threshold as being in error. The changes in the polynomial caused by the errors causes problems, especially in areas with multiple errors. A third method is to examine the differences between consecutive beats. This method works best if the beats are compared to the last known good value, but fails in areas where overall changes in heart rate are contaminated with errors. After comparing these various solutions, the recommendation is to use both the updated mean and the last normal HR value as benchmarks. When tested on sleep data, using both values failed to detect 6% of artifacts and mislabeled a small number of correct values as artifacts.

1.3 Overview

Some of the above methods assume that the entire IBI trace has been collected and is available during error correction. Others collected data in carefully controlled clinical settings with limited activity that reduced the type and number of errors present. We consider the case where IBI errors must be detected and corrected in real-time. Such is the case, for example, if heart rate variability analysis is to be used for real-time monitoring or feedback. The applications of such a device are numerous. It could be used for alertness monitoring in long term vigilance tasks, increasing the safety of critical alertness activities such as truck driving. It could also be used to close the loop and allow computer systems that can respond to a user's physiological state. A system such as this would be desirable, for instance, in managing cognitive load for optimal performance during crucial tasks. Researchers desire to enhance soldiers of the twenty first century with such computer assisted cognitive capabilities [17].

We also consider the possibility of additional types of errors. While any errors derived from a valid ECG signal will result in combinations of simple split or combine solutions, errors caused by an inability to properly sense this signal or by faulty equipment are equally important to detect, even if they can not be corrected.

To our knowledge, this work is the first to consider IBI error detection and correction in real-time on mobile subjects and with multiple devices. This thesis describes a method for correcting IBI data as it is gathered that is designed to run online. It introduces minimal lag into the IBI signal and is capable of detecting and classifying a wide variety of errors. Section 2.1 describes the actual correction process. It is a rule based approach that works within the bounds of a contextual buffer. Section 2.2 describes our data set of 124,998 usable IBI values collected from 18 subjects. We evaluate the performance of the method against that of a set of human graders using the train and test paradigm. Results are presented in Chapter 3.

Chapter 2

Methods

Section 2.1 describes our engine for automated IBI error detection and correction. Section 2.2 describes the data set, consisting of 124,998 IBIs, used to evaluate our method. Our engine accepts several adjustable parameters that can be modified for optimum performance. Section 2.3 describes the parameters and the process used to choose the best set.

2.1 Engine

Incoming IBI values are placed into a buffer, illustrated in Figure 2.1, as they are received. This buffer provides context for the two questions that must be asked as an IBI is evaluated. The first is, should this IBI be marked as an error? The second is, if this IBI is an error, what is the most appropriate correction? This buffer must be of sufficient size to allow all the necessary information contained in the sequence of IBIs to be used in the decision making process. Past research has shown that temporally related heartbeats contain more information about the predicted state of an IBI value under consideration. [4] This indicates that the buffer does not need to be very large to make an appropriate evaluation of the state of a heart beat. The possible corrections applied provide another constraint on the size of the buffer. The corrections we consider (see Table 2.1) involve a maximum of three IBI values. The maximum value for an IBI is approximately 1500ms, and they are significantly

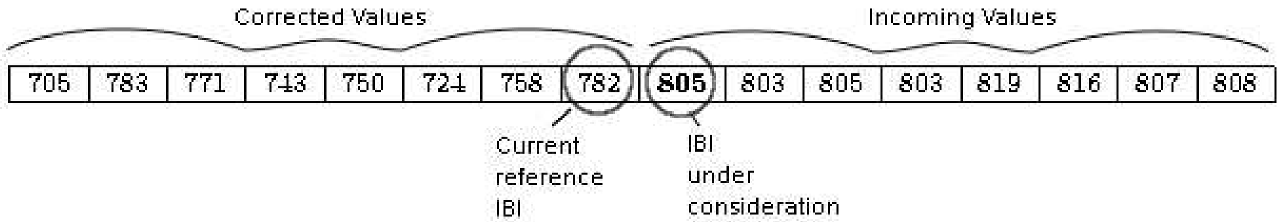


Figure 2.1: A snapshot of the buffer used for correction context.

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 any corrections we may wish to try. Our method buffers a minimum of six seconds of incoming data in order to provide enough context for evaluating corrections. Since IBI data arrive asynchronously, in practice we are unable to guarantee a fixed amount of time in the buffer. We try to keep the time in the buffer close to six seconds to minimize lag introduced into the system. The reason our buffer maintains a fixed amount of time instead of simply working on a fixed number of IBI values is primarily to facilitate the resampling of IBI values into synchronous data for consumer processes that may need to work on IBI values in this format. Specifically, spectral analysis through the use of the Fourier transform requires synchronously sampled data to provide meaningful results. As data arrive and are added to the buffer, a timestamp is also associated with each value. If the device generating the IBI values has a separate clock from that on the recording system, this can provide some measure of useful redundancy.

Past history is not necessary for feeding a synchronously sampled consumer, which is reflected in our implementation. As values are deemed to be correct or to have a correction applied, they are moved into a portion of the buffer that contains a fixed number of values rather than a certain time period. The number of values contained in this section of the buffer was a variable we trained to and is more fully examined in Section 2.3. Statistics on the values in this section of the buffer are used to predict the expected value of the next IBI.

The expected value is calculated as an acceptable deviation from the last known good value. Equation 2.1 shows how the threshold for determining if a new value has exceeded this deviation is calculated. In the equation, T is the threshold, N is the number of values in the past buffer, pb represents the past buffer, and m is a multiplier. N and m are fixed at runtime, but were tried with several different reasonable values. Results from this portion of the experiment can be found in Section 2.3. In plain language, the mean successive difference of the trusted values in the buffer is multiplied by some value to generate the threshold. The generated value is limited to a predefined range of possible values, also determined at runtime and also trained for as described in Section 2.3. If the buffer is not yet full, a reasonable estimate is used until the buffer is full and meaningful statistics can be calculated. During the experiments described in this thesis, a default value of 100ms was used. An IBI may also be marked as an error if the timestamp data have a significant deviation from the recorded value. This is most useful for catching errors where the data were corrupted between the IBI generating device and the recording system.

The portion of the engine that detects errors can be temporarily suspended if too many errors occur in a particular neighborhood. This capability was introduced to minimize the effect of the engine generating a sequence of data that appears to be valid but in fact deviates from the original data. In practice, an observation of human graders indicates that more than 2 or 3 consecutive corrections are rarely, if ever, used. If corrections are occurring with 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 OK. If the counter exceeds a value of 3, the next IBI is not examined for potential errors and the counter is again decremented. The only exception to this rule is for IBI values at or exceeding the maximum reportable hardware value. These IBIs are replaced with a valid timestamp or marked as uncorrectable even if the counter is at its maximum value.

$$T = \frac{\sum_{i=0}^{N-2} |pv_i - pv_{i+1}|}{N-1} * m \tag{2.1}$$

There are at least two possible approaches to correcting IBIs. In the first, the sequence of values could be viewed as a continuous function over time with some kind of smoothing function applied to deviations. Alternatively, since we have already established that the only types of errors that can occur in a stream of IBI data are the omission of an event and the addition of a spurious event, we could try to reconstruct the series of events that may have caused a particular error or sequence of errors from the surrounding good IBI values. From this we could make a reasonable approximation of the original values of the erroneous IBIs. Our method takes the second approach, using a set of rules to try to classify the error sequence that led to an abnormal IBI.

A human considering IBIs can frequently discern errors where IBIs have been split or combined due to some faulty detection. These errors become apparent when IBI values are plotted in sequence with the value on the Y axis. Figure 2.2 demonstrates the appearance of several common error types when plotted in this manner. When a specific error pattern has been recognized, it allows the human grader to determine the proper correction to apply. Human experts can also recognize faulty IBIs if the reported values are outside of the physiologically expected range. Additionally, most IBI detecting hardware has a maximum value it will report. Any IBIs at this maximum value can be quickly recognized as errors.

Our automated correction procedure attemps to classify errors in a manner similar to human graders. If an IBI has been flagged as a possible error in the detection phase, the IBI is evaluated within its context by a series of rules designed to determine the most likely error type that may have occurred. Each rule represents a possible error type. The rules, in the order they are applied, are listed in Table 2.1. The order in which possibilities

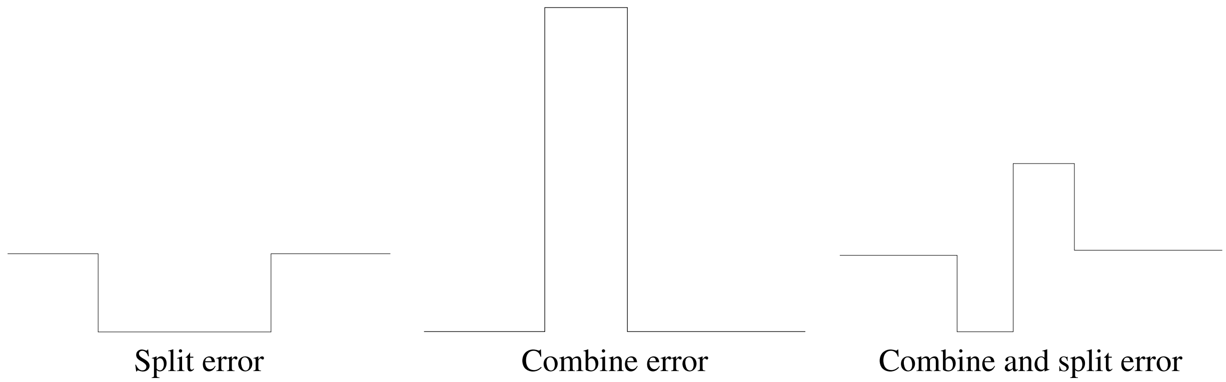


Figure 2.2: Appearance of errors in IBI data

are inspected is significant, as the first classification that is accepted terminates the search process. In order to be deemed an acceptable error, the corrected value must lie with some threshold of the reference IBI and the IBI value following the inspected IBI. This threshold is generated in a manner identical to the threshold for detection described in Equation 2.1, but the values for the multiplier, threshold minimum, and threshold maximum may differ from the values used for detection. Like the values used in the detection process, these were trained for optimum performance as described in Section 2.3.

2.2 Data

Evaluation of the method 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 motions designed to simulate active conditions and have a high probability of inducing errors. Participants completed 2 sets of the following tasks followed by an 8 minute baseline period: punching arms, jumping jacks, running in place, and crunches. Each subject repeated the experiment three times with a different device each time. These were the Polar S810 (Lake Success, NY), the Biolog 3991 (UFI Corp., Morro Bay, CA) with standard electrodes, and the Biolog 3991 with fetrodes. In all, fifty-four files containing 124,998 usable IBIs were collected and graded by human raters. Two Clemson University graduates students were

Correction	Description	Cause
Hardware trigger	replace value outside of maximum or minimum values allowed by hardware	transmission error
Split	Divide an IBI into two equal values	Missed heartbeat
Split 3	Divide an IBI into three equal values	Missed two heartbeats
Combine	Add two IBIs together	False trigger
Combine 2 / Split 2	Replace two IBI values with their average	Combination of missed heartbeat and false trigger
Combine 2 / Split 3	Get three new IBI values by adding two and dividing them by three	Combination of missed heartbeat and false trigger
Combine 3 / Split 3	Replace three IBI values with their average	Combination of missed heartbeat and false trigger
Physiological trigger	Replace value outside maximum or minimum physiological value	Probably some uncaught error from above
Uncorrectable	Could not apply any rule, but IBI appears faulty	

Table 2.1: Possible corrections applied by program

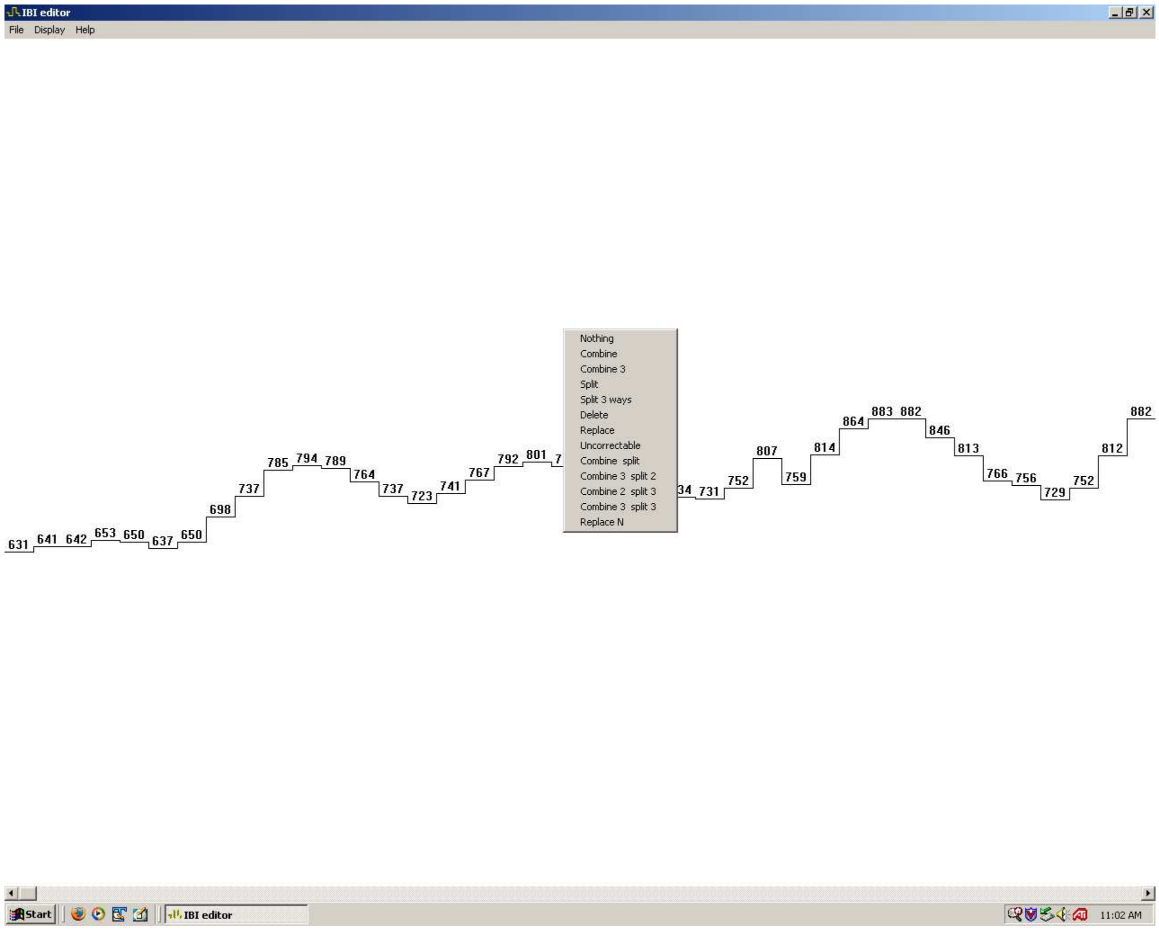


Figure 2.3: IBIedit program

given basic 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 corrections the human graders agreed on. Decisions the human graders did not agree on were deemed uncorrectable and not used for training purposes.

The human graders corrected the files using a program developed specifically for that purpose. The program (see Figure 2.3) allowed them to view the IBI data at multiple resolutions and apply any of the corrections allowed by our method. Human graders were also allowed to add their own custom correction. The program generated a list of corrections applied on each IBI present in the original file. This list was designed to allow easy automated comparison between the human graders and the automated method to facilitate the training of our method as described in Section 2.3.

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 all deal with the threshold generation process. Table 2.3 shows a summary of the variables along with the values used for testing.

Six of the fifty-four files were used for training. This amounted to 15,095 total IBI values to train on. We tried several different sets of parameters until one was found that seemed stable. For each training pass, we used five values for each of the seven possible parameters. Each of the six files was run through the engine with every possible combination of values. For each file at every combination we used the sum of the IBI values that both humans and the computer agreed were correct and those that all agreed were a specific error as a metric for the performance of the engine with that parameter set. The values for each file were added and stored as a total metric for the performance with that particular parameter set. The maximum value in this N-cube consisting of 5^7 values was deemed the best result.

Several different N-cubes were used. We sought one that had most of the values in the middle instead of the edges of the cube. The final N-cube used for the results presented in this thesis was generated using the parameters in table 2.3. The parameters in thie N-cube determined to give the best results were pb=5, m1=10, t1min=50, t1max=200, m2=25, t2min=10, and t2max=100. It should be noted that throughout this training process, an offline version of the engine was used so that we could use the same set of files repeatedly. The offline version maintains an identical amount of context to the online version and is functionally identical.

Parameter	Description	Possible values
past buffer size	The number of ibi values that are put in the buffer used to determine the threshold for future error detection.	3, 5, 10, 15, 25
multiplier 1	Multiplier used in generating threshold for error detection	1, 5, 10, 25, 50
threshold 1 min	Minimum threshold generated when creating threshold for error detection	10, 50, 70, 100, 150
threshold 1 max	Maximum threshold generated when creating threshold for error detection	10, 50, 100, 150, 200
multiplier 2	Multiplier used in generating threshold for correction acceptance	1, 5, 10, 25, 50
threshold 2 min	Minimum threshold generated when creating threshold for correction acceptance	10, 50, 70, 100, 150
threshold 2 max	Maximum threshold generated when creating threshold for correction acceptance	10, 50, 100, 150, 200

Table 2.2: Parameters used in training.

Chapter 3

Results

In order to provide a ground truth for testing the method, the human graders were first evaluated against one another. We computed the rate of agreement for each error type as

$$Rate = \frac{Agree}{Human1 + Human2 - Agree} \tag{3.1}$$

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. Results can be seen in Table 3.1.

Several important observations can be made from the results of the human versus human data. First, there are a few corrections that dominate the decisions made by the human graders. 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. In areas the human graders agreed upon, these three corrections made up 89% of the total corrections. There are several factors that may contribute to these errors occurring more frequently than others. The motions used in this study may be prone to causing these types of errors. The prominence of certain errors may be related to the conditions under which data was collected. In errors caused by motion, there is a high chance that many of the errors are caused by insufficient contact of the electrode with the subject. This would be likely to cause missed beats, which would be fixed by splitting an IBI if it occurred in isolation or by marking an area as uncorrectable if it occurred frequently enough to cause the data to be impossible to reconstruct. Another factor that may contribute to the high usage and high rate of agreement of the split error would be that it is one of the easiest to identify. It stands out plainly in a series of data and is simple enough to be the obviously correct solution if found in isolation. Similar factors may contribute to the high use of the combine and split operation. Because it is essentially a smoothing operation, some of its use may be due to over smoothing the original data.

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 corrections should probably not be part of any automated algorithm because of their extremely low rate of use. It should be noted that delete and replace N are not included in the current version of the algorithm, and replace is only used if the external timestamp seems to be valid when the received IBI value is not.

Another important observation 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 area was in need of correcting 96% of the time, the best rate of agreement for choosing what 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. It also suggests that the human graders require additional training. If the correct data stream is not obvious, then the area should be marked as uncorrectable.

The number of corrections that the human graders agreed upon was used as the ground truth for calculating the computer's rate of success. Areas that either human grader marked

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 3.1: Human vs human agreement rates by correction type.

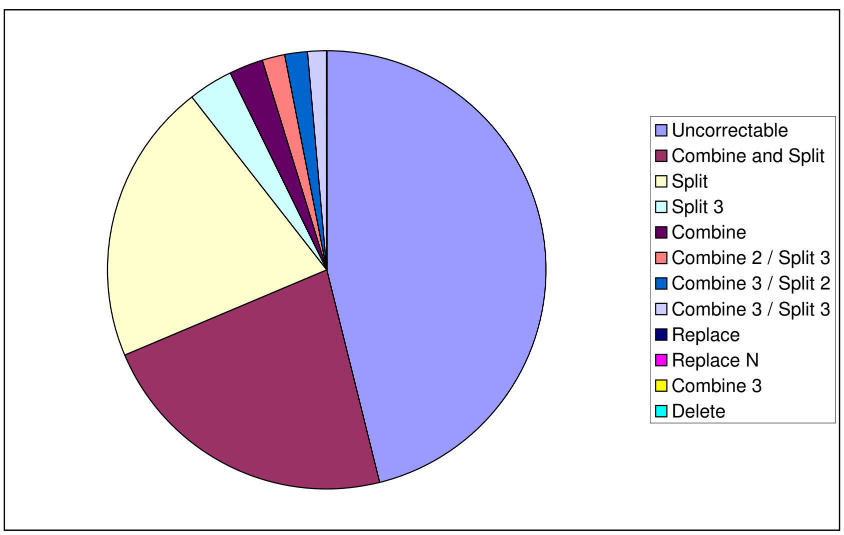


Figure 3.1: Percent correction application by human graders in areas the human graders agreed on the correction used

as uncorrectable or that the human graders disagreed on were not used in the training process.

As mentioned in Chapter 1, the problem of correcting ibi data can be broken up into the two distinct areas of detection and classification. Table 3.2 shows the computer's detection performance on the training data in correctable areas the human graders agreed on. The table is broken down by file to demonstrate how the results varied across different subjects. Some files, such as Sub09Polar, demonstrated a very high rate of agreement on handling errors, while others, such as Sub08Electrode, had a noticably poorer performance. This indicates that certain types of areas with errors are handled better than others by the automated method. Overall, the humans and computer agreed on the classification of 99.0% (97.0 + 2.0) of the data. The computer made the conservative choice to leave an area alone on 60% of the erroneously labeled IBI values, leaving improperly applied corrections on only 0.4% of the data. It is usually preferable to miss some corrections instead of over correcting the data [8], so the tendency towards leaving unsure areas alone is promising. Results were similar on the test set, seen in Table 3.3, though performance was not as good. The humans and computer agreed on 97.4% (95.6 + 1.8) of the labels. The computer made erroneous corrections on 58.3% of mislabeled data (1.4 / (1.4 + 1.0)), and 1.4% of the data was labeled with an erroneous correction. This shows that the disposition towards conservative correction did not hold in the test set. More complex training may be able to change this trend in future experiments. The difference between classification perfromance on the test and training sets is illustrated in Figures 3.2 and 3.3. The area with at least one entity (humans or computer) marking an error is broken out to show the significant change in the treatment of erroneous portions of the data.

Table 3.4 illustrates the computer's performance at classification for the combined test and training set. The computer's rate of agreement with the combined human graders is similar to the human graders' rate of agreement with one another. Figure 3.5 shows the percent agreement of the human graders with one another for each correction along

File Name	H = OK C = OK (# of IBIs)	H = OK C = Error (# of IBIs)	H = Error C = OK (# of IBIs)	H = Error C = Error (# of IBIs)	Total (# of IBIs)
Sub08Electrode	2353	17	32	33	2435
Sub08Fetrode	2785	1	4	12	2802
Sub08Polar	2480	9	4	43	2536
Sub09Electrode	1643	19	28	61	1751
Sub09Fetrode	1662	14	6	74	1756
Sub09Polar	2550	2	7	55	2614
Totals	13473	62	81	278	13894
Percent	97.0%	0.4%	0.6%	2.0%	100%

Table 3.2: Classification performance on training set

File	H = OK C = OK (# of IBIs)	H = OK C = Error (# of IBIs)	H = Error C = OK (# of IBIs)	H =Error C = Error (# of IBIs)	Total (# of IBIs)
Totals	98347	1409	1071	1835	102662
Percent	95.6%	1.4%	1.0%	1.8%	100%

Table 3.3: Classification performance on test set

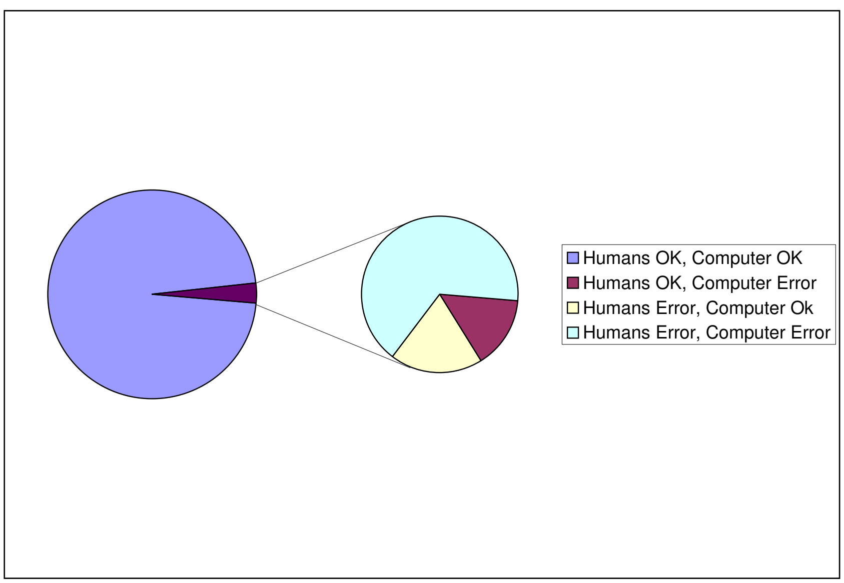


Figure 3.2: Classification performance on training set

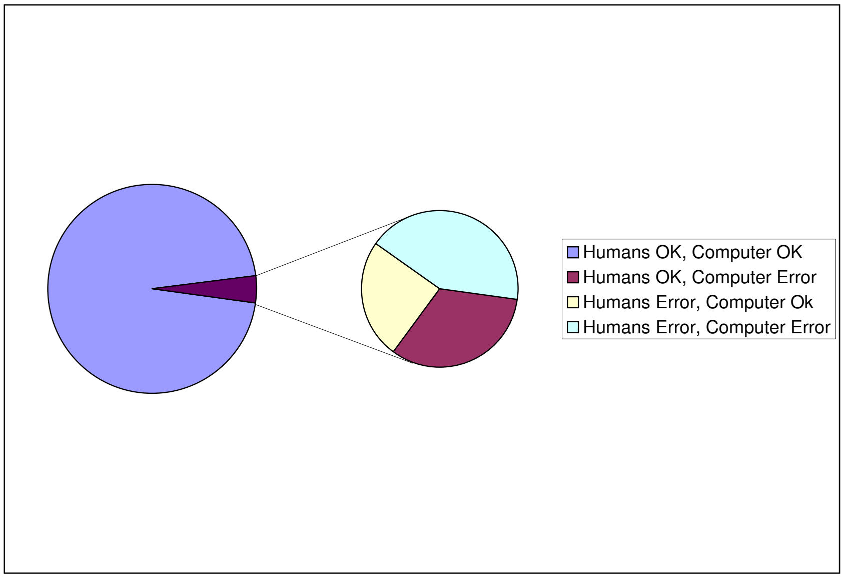


Figure 3.3: Classification performance on test set

with the agreement rate for the computer with both humans. The results appear to be highly correlated, suggesting that areas that the humans are more sure about provide more positive feedback to the algorithm. A training set with a higher rate of agreement between the human graders may be able to enhance performance of the algorithm.

Figure 3.4 shows the usage of various error types by the automated method. It is interesting to note the difference in distribution between the automated method and the human graders, whose error usage is shown in Figure 3.1. This would seem to indicate that the automated method needs improvement in the classification stage of error detection. Ideally, the automated method would apply corrections with the same frequency as human graders.

Figures 3.6 through 3.29 demonstrate the results of corrections on each of the files in the training set. Areas that drop to 0 in the corrected data are spots that have been marked as uncorrectable. Notice that the computer handles most of the isolated errors, like those in figure 3.10, quite well and has difficulty handling areas of concentrated errors like those

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 3.4: Computer vs human agreement rates by correction type.

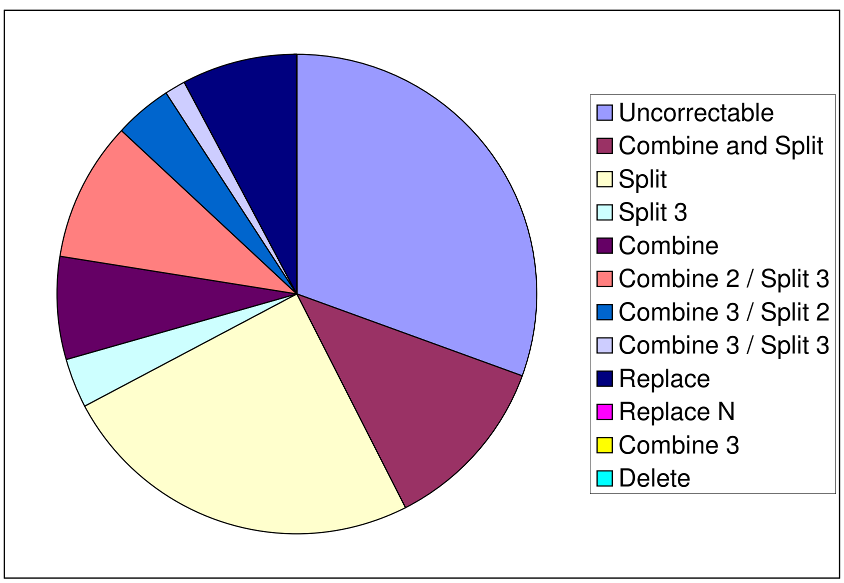


Figure 3.4: Percent correction application by the automated method

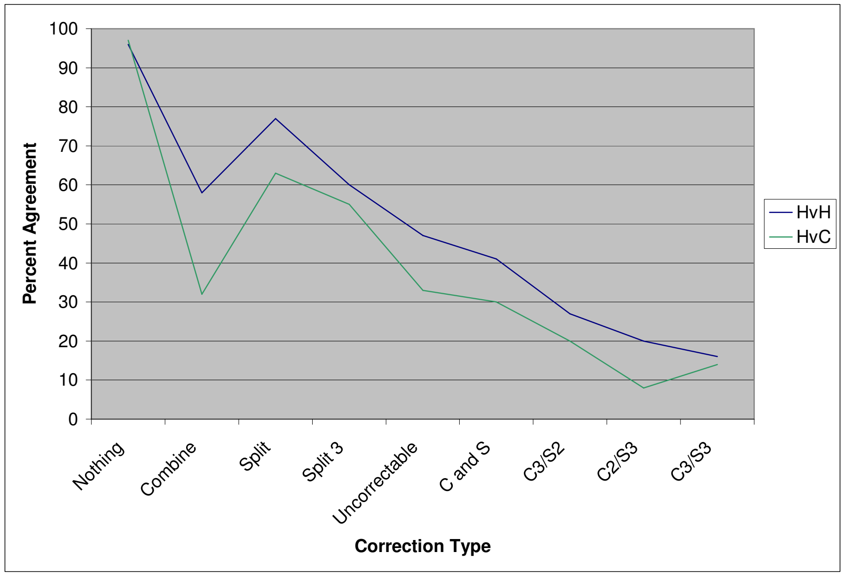


Figure 3.5: Human vs human and humans vs computer agreement by error type

in Figure 3.18. These areas of concentrated error show the need to mark an entire area as uncorrectable instead of considering each IBI individually.

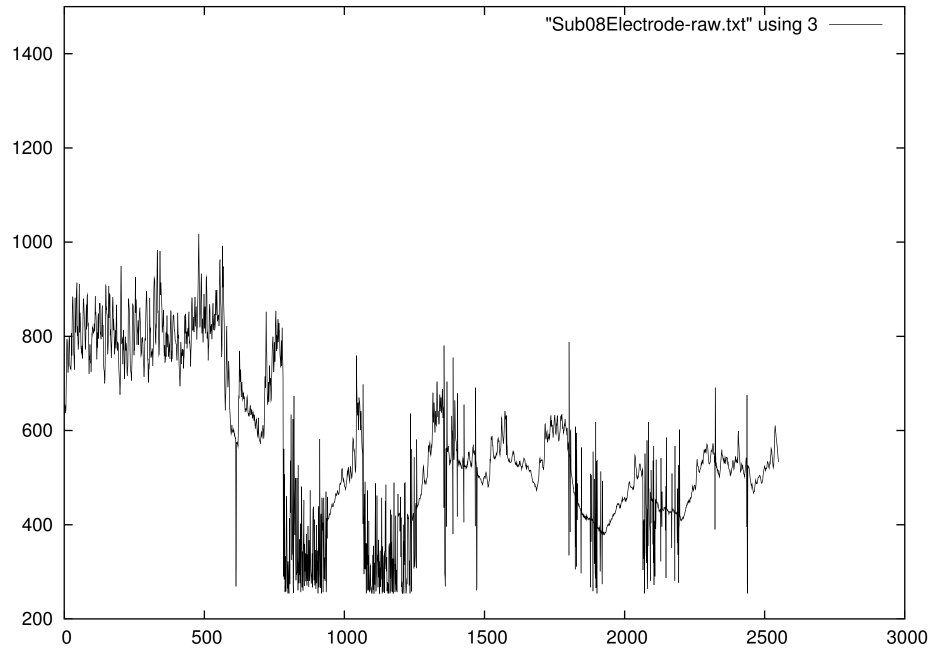


Figure 3.6: Subject 8 Electrode uncorrected

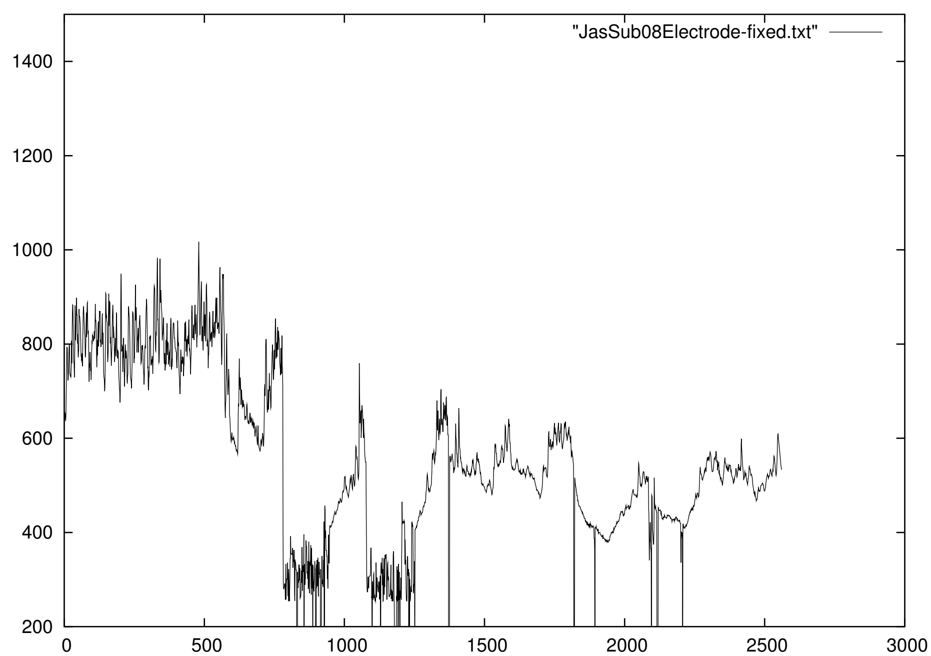


Figure 3.7: Subject 8 Electrode corrected by Human 1

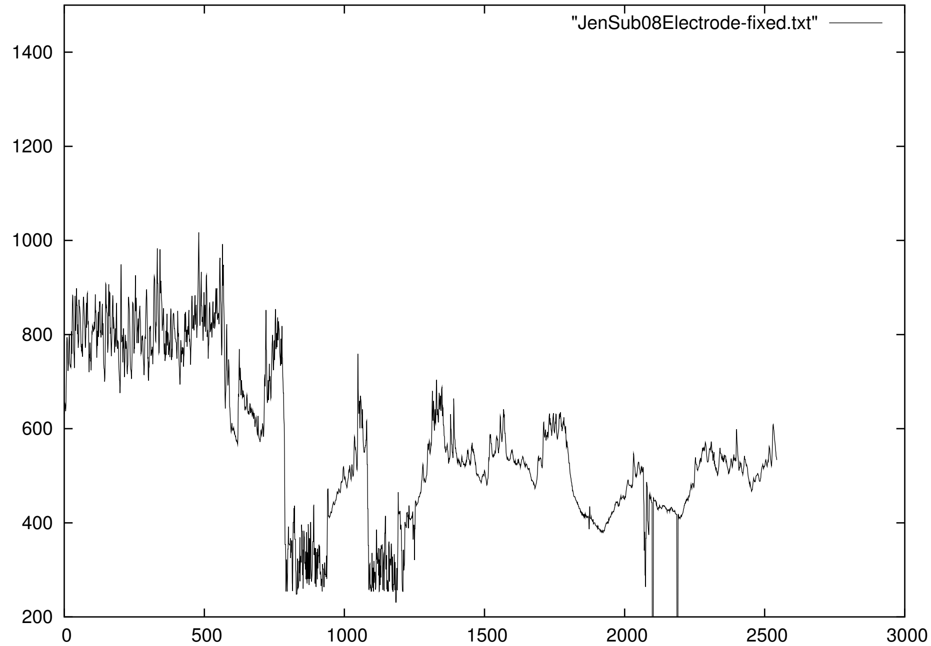


Figure 3.8: Subject 8 Electrode corrected by Human 2

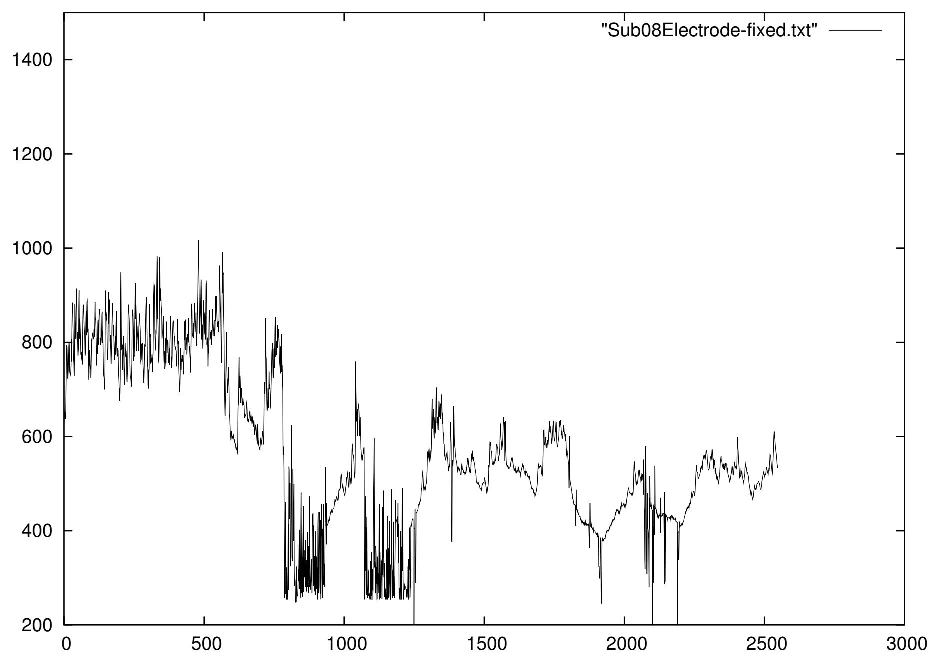


Figure 3.9: Subject 8 Electrode corrected by Computer

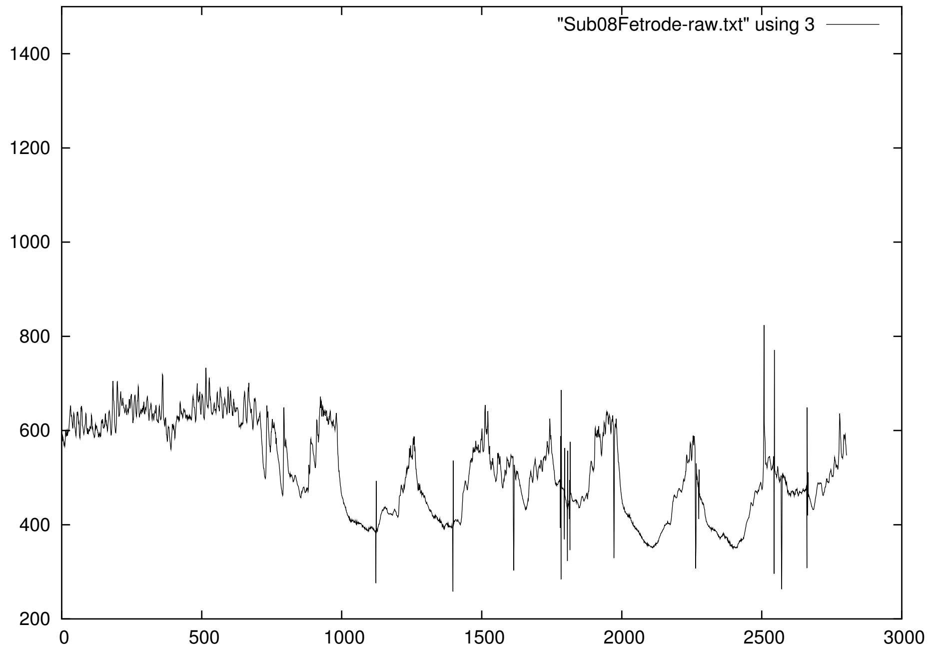


Figure 3.10: Subject 8 Fetrode uncorrected

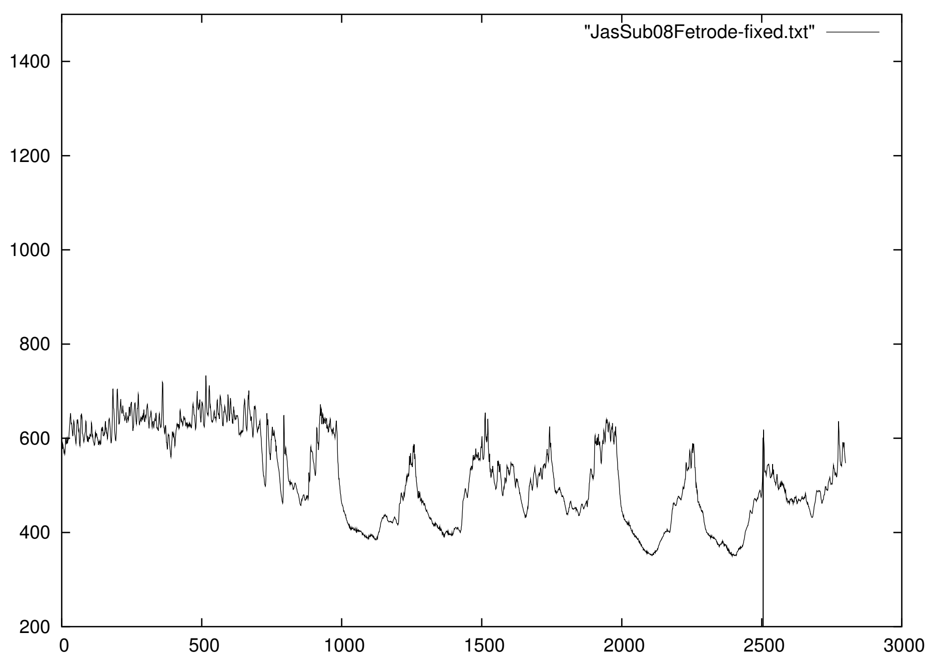


Figure 3.11: Subject 8 Fetrode corrected by Human 1

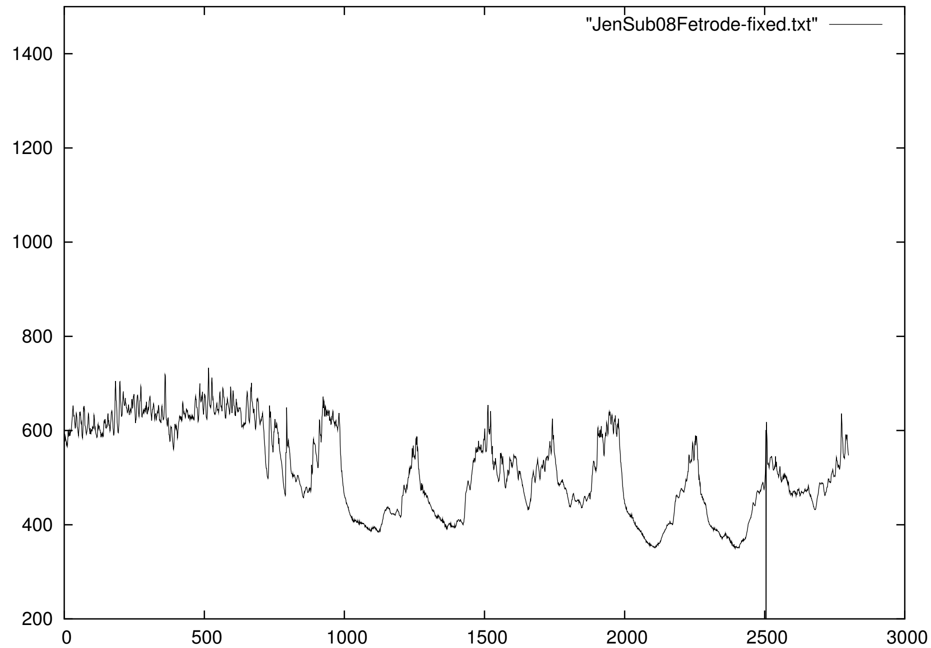


Figure 3.12: Subject 8 Fetrode corrected by Human 2

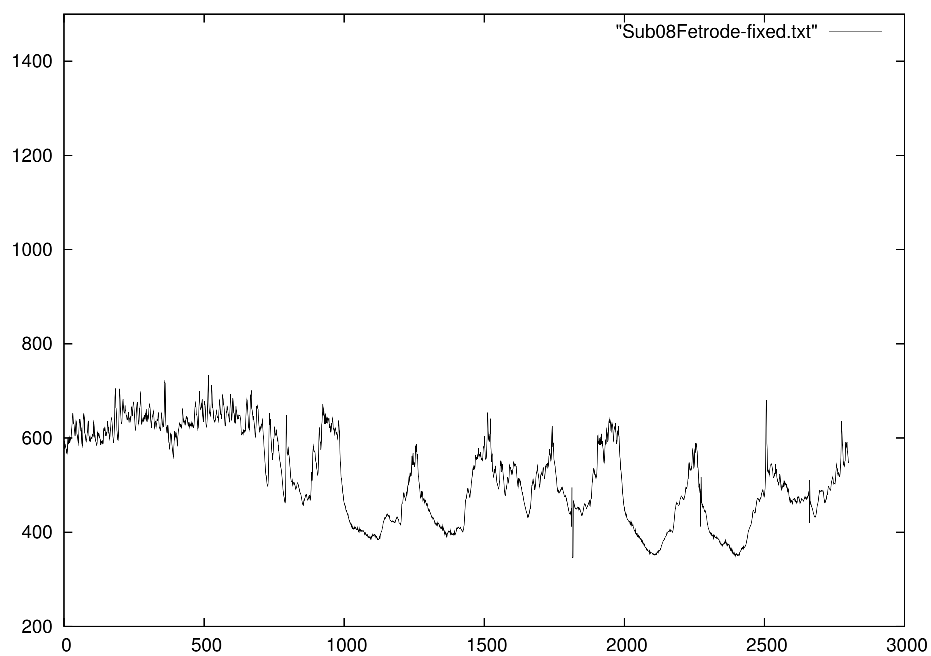


Figure 3.13: Subject 8 Fetrode corrected by Computer

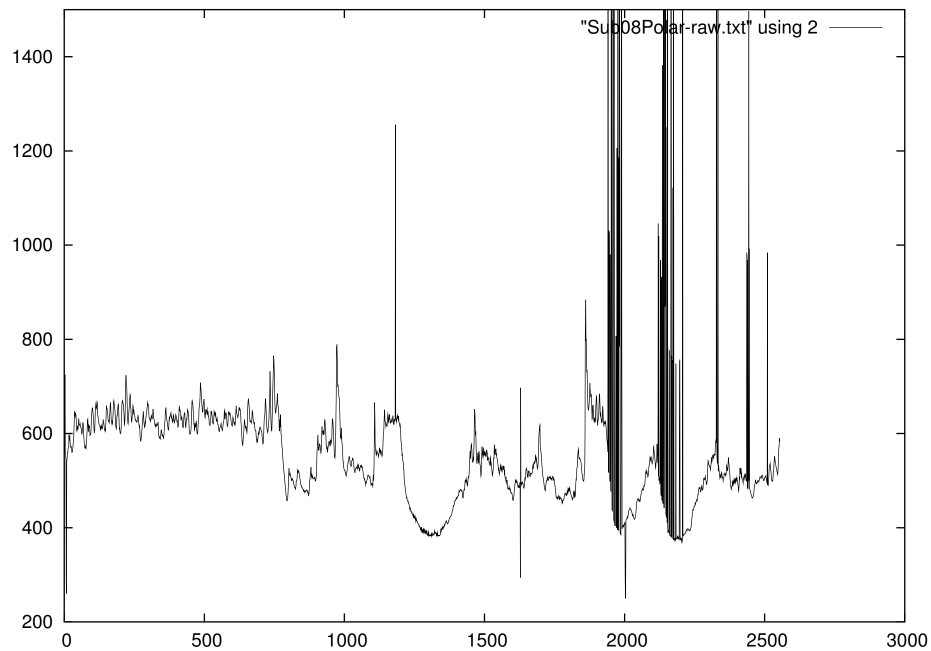


Figure 3.14: Subject 8 Polar uncorrected

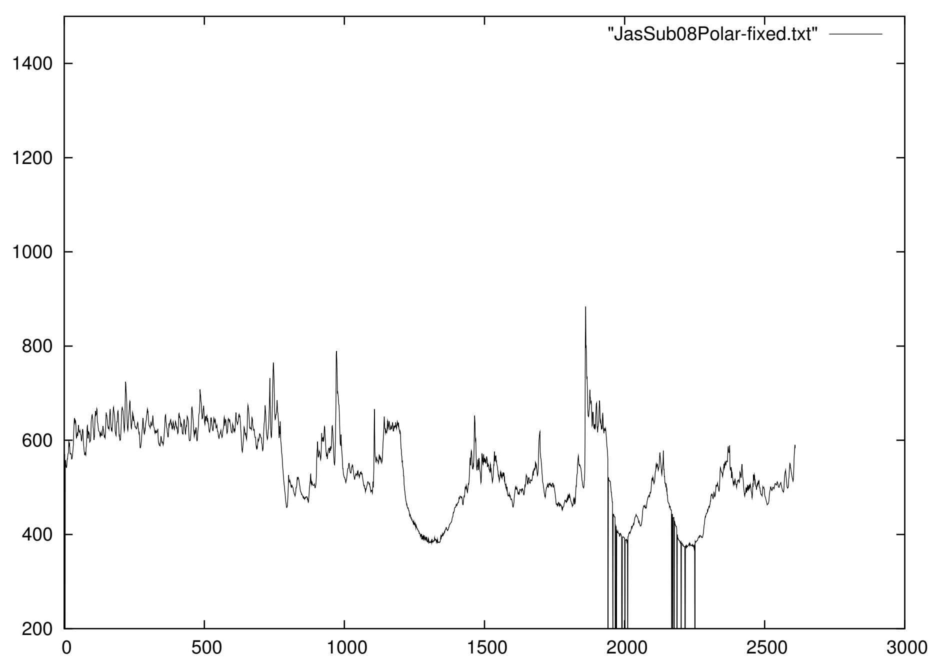


Figure 3.15: Subject 8 Polar corrected by Human 1

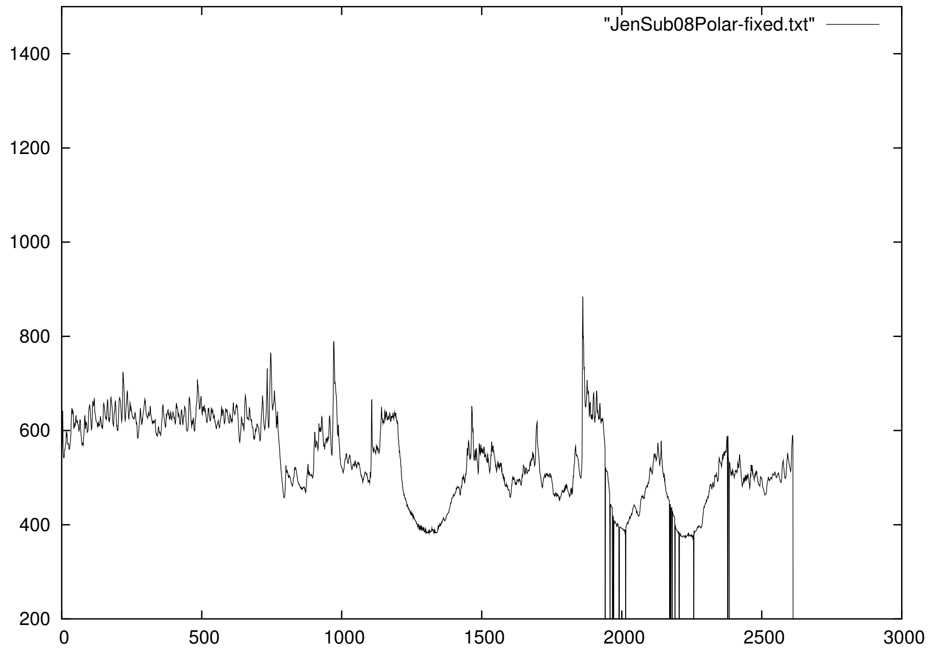


Figure 3.16: Subject 8 Polar corrected by Human 2

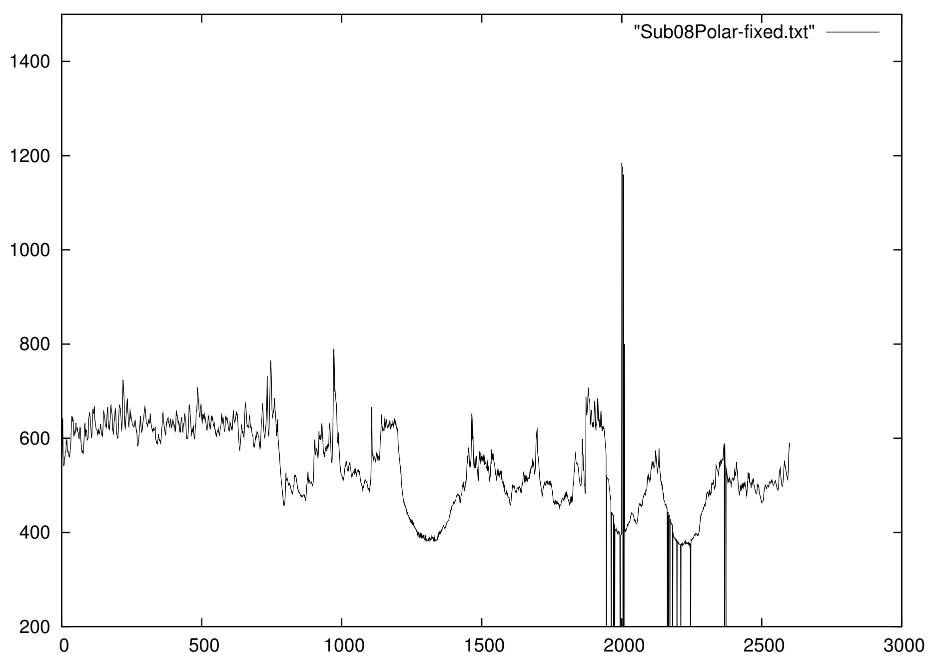


Figure 3.17: Subject 8 Polar corrected by Computer

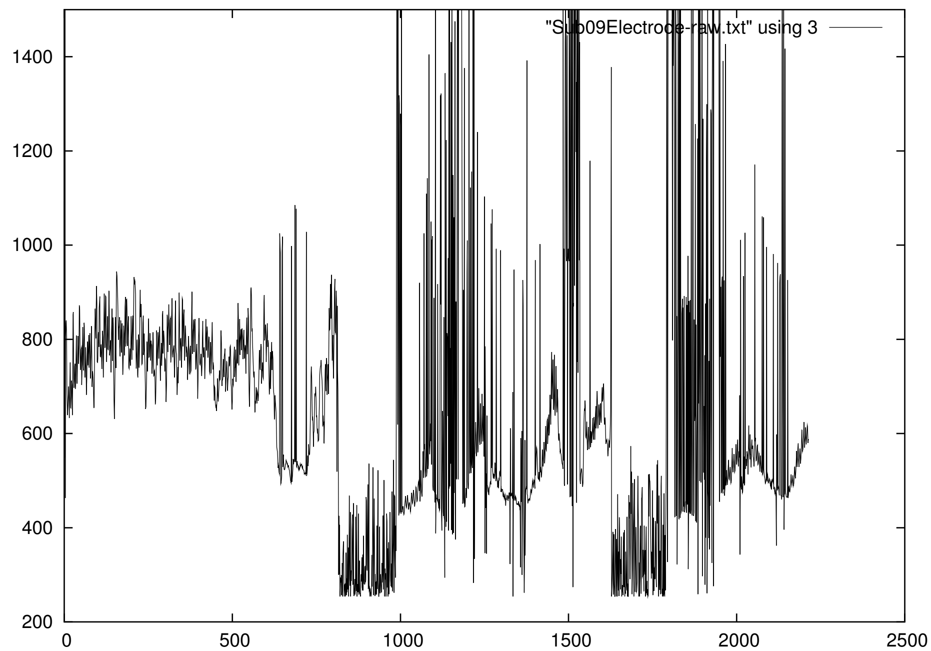


Figure 3.18: Subject 9 Electrode uncorrected

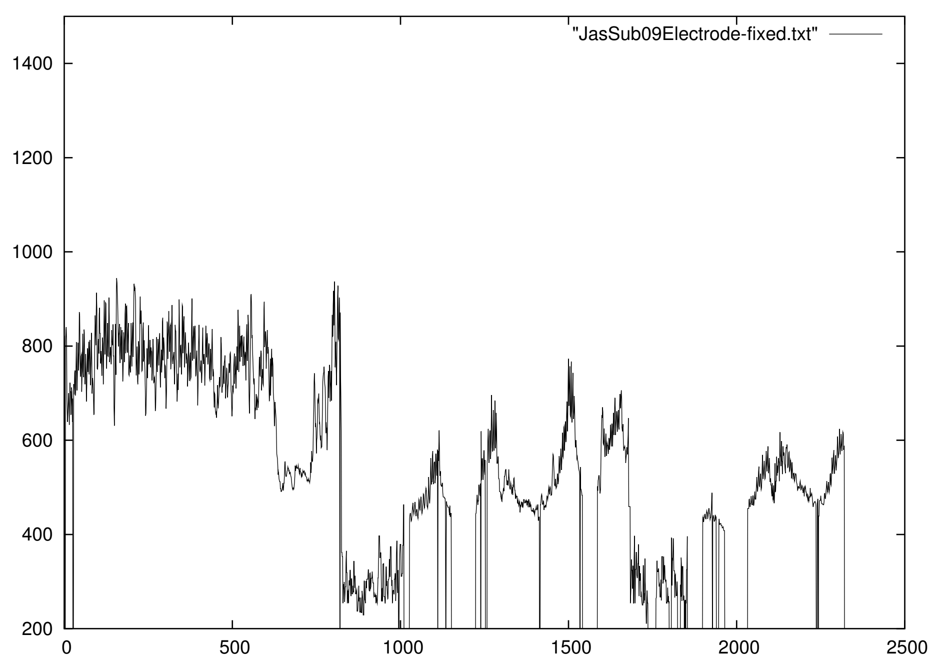


Figure 3.19: Subject 9 Electrode corrected by Human 1

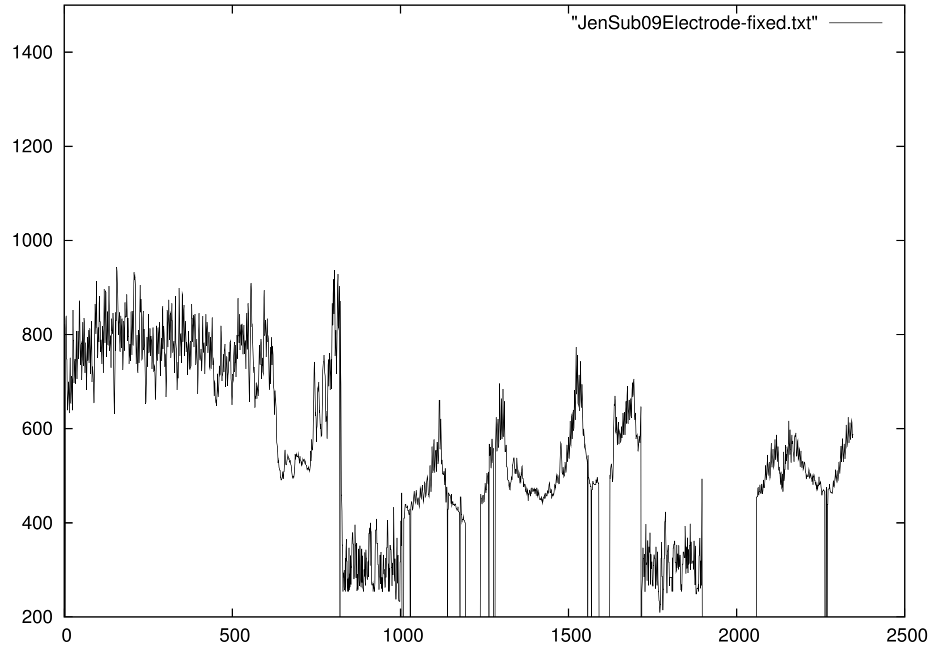


Figure 3.20: Subject 9 Electrode corrected by Human 2

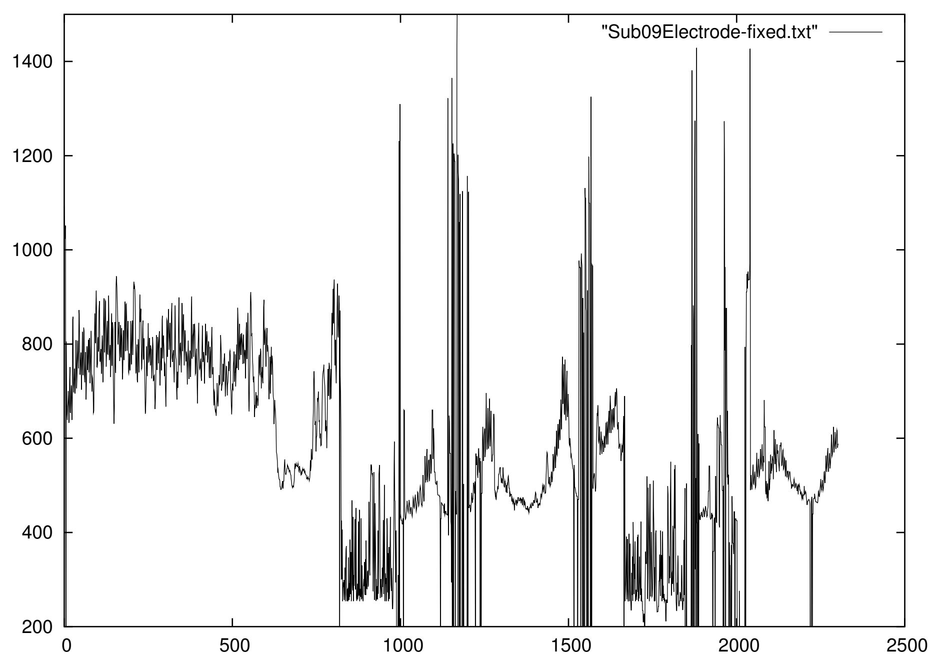


Figure 3.21: Subject 9 Electrode corrected by Computer

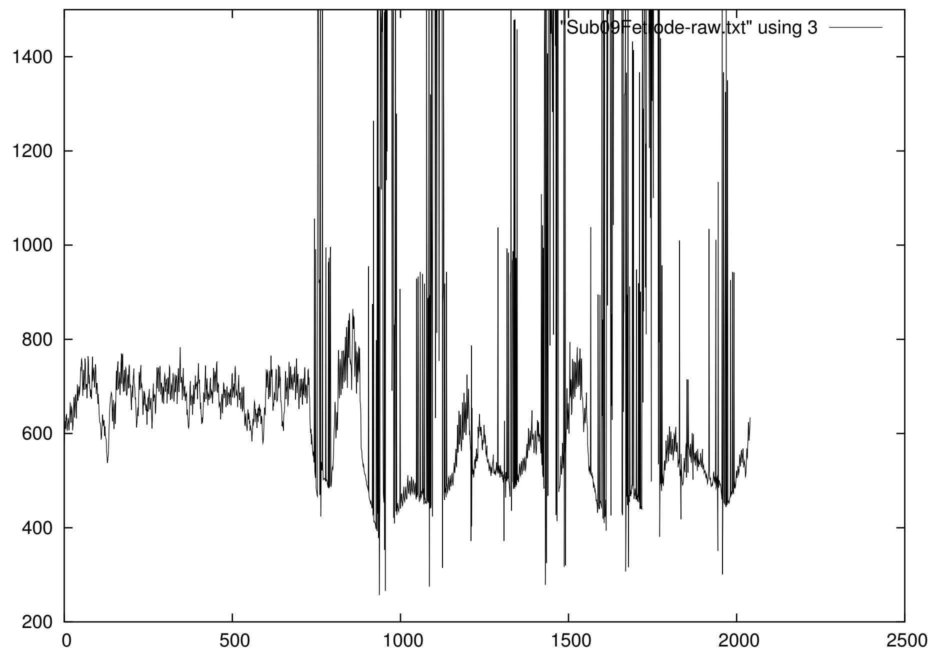


Figure 3.22: Subject 9 Fetrode uncorrected

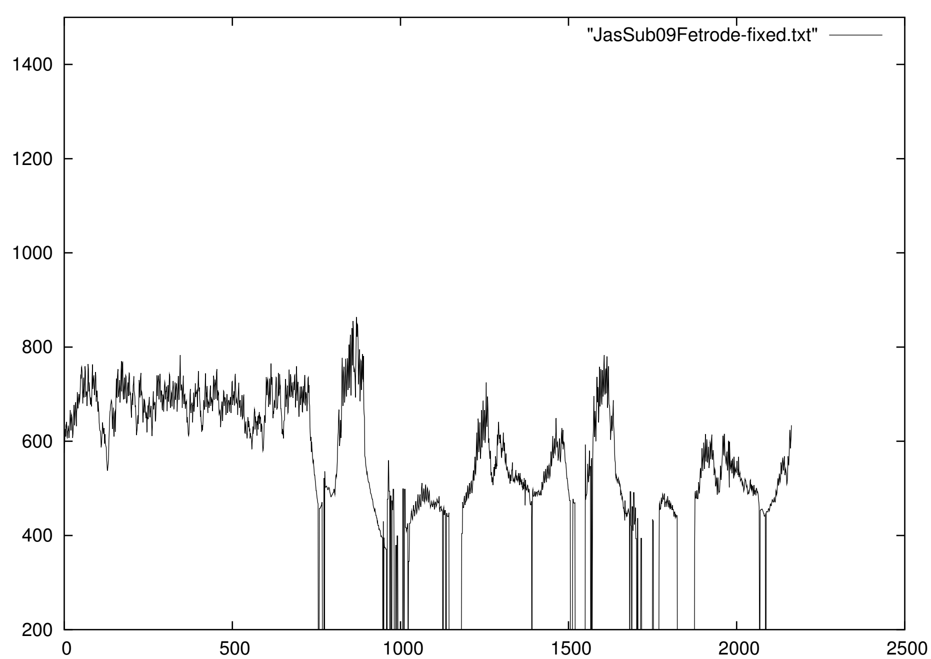


Figure 3.23: Subject 9 Fetrode corrected by Human 1

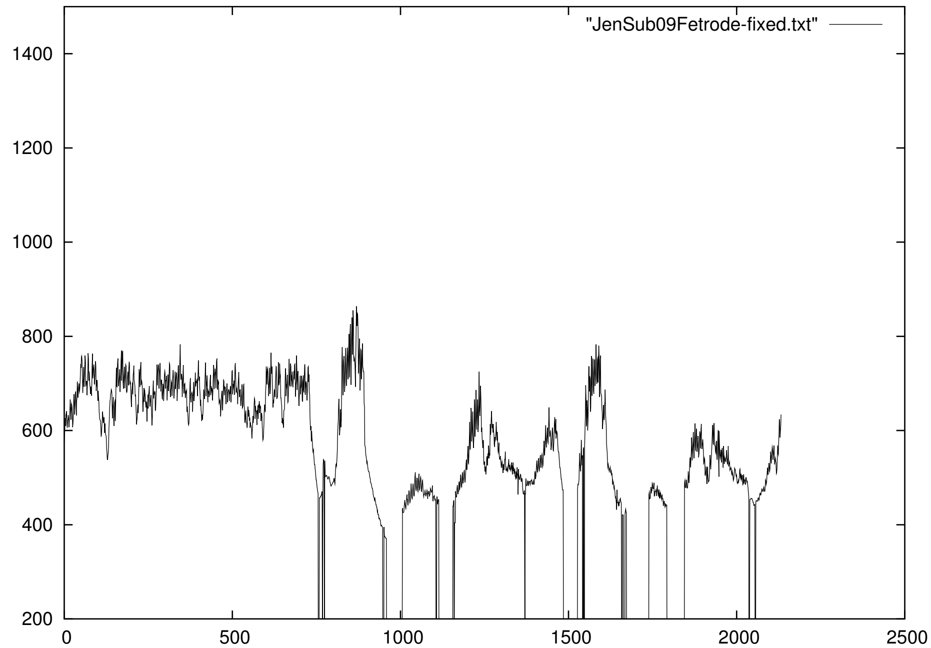


Figure 3.24: Subject 9 Fetrode corrected by Human 2

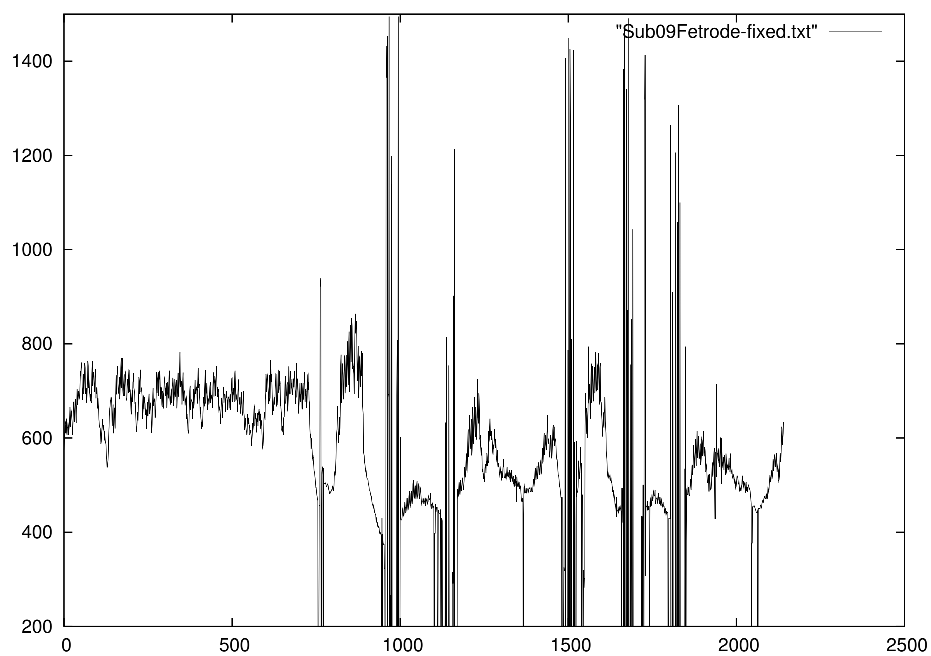


Figure 3.25: Subject 9 Fetrode corrected by Computer

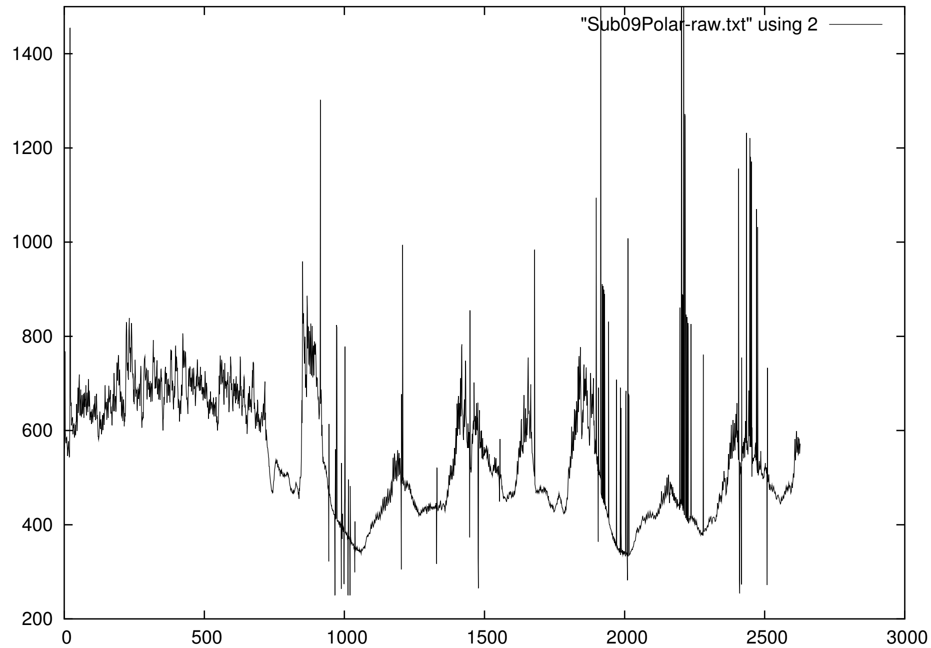


Figure 3.26: Subject 9 Polar uncorrected

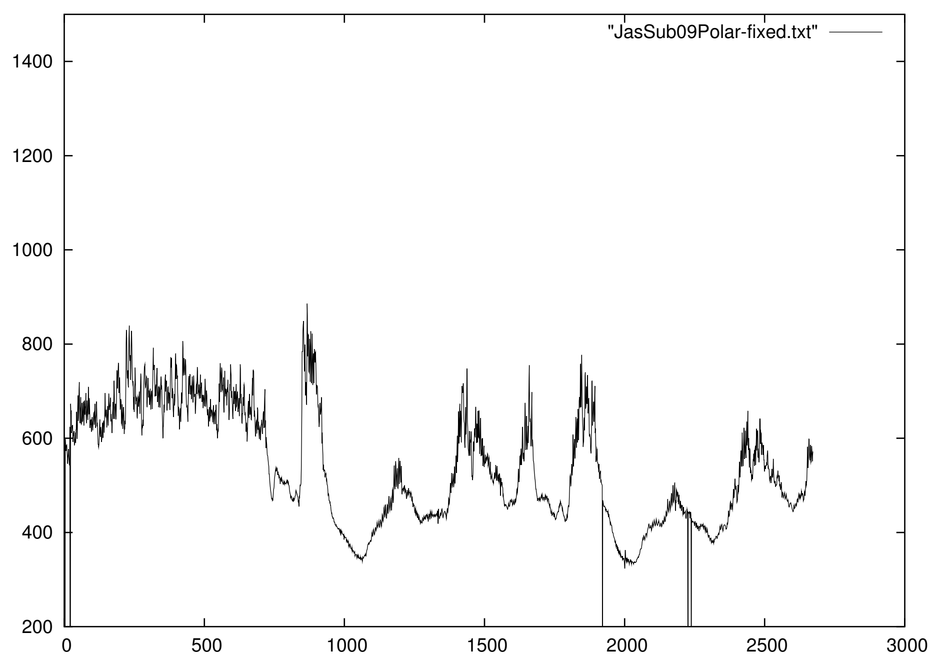


Figure 3.27: Subject 9 Polar corrected by Human 1

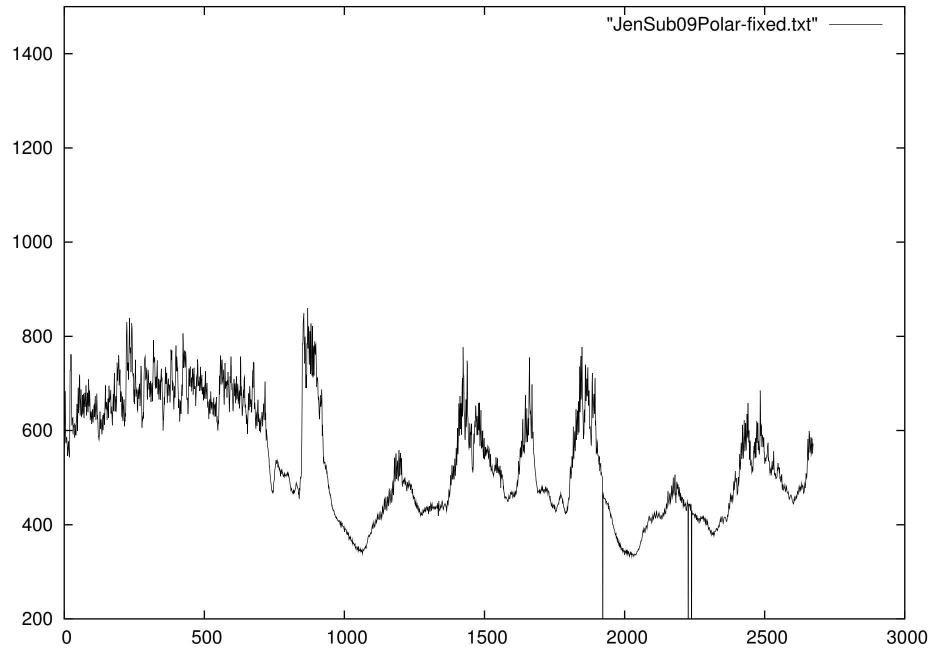


Figure 3.28: Subject 9 Polar corrected by Human 2

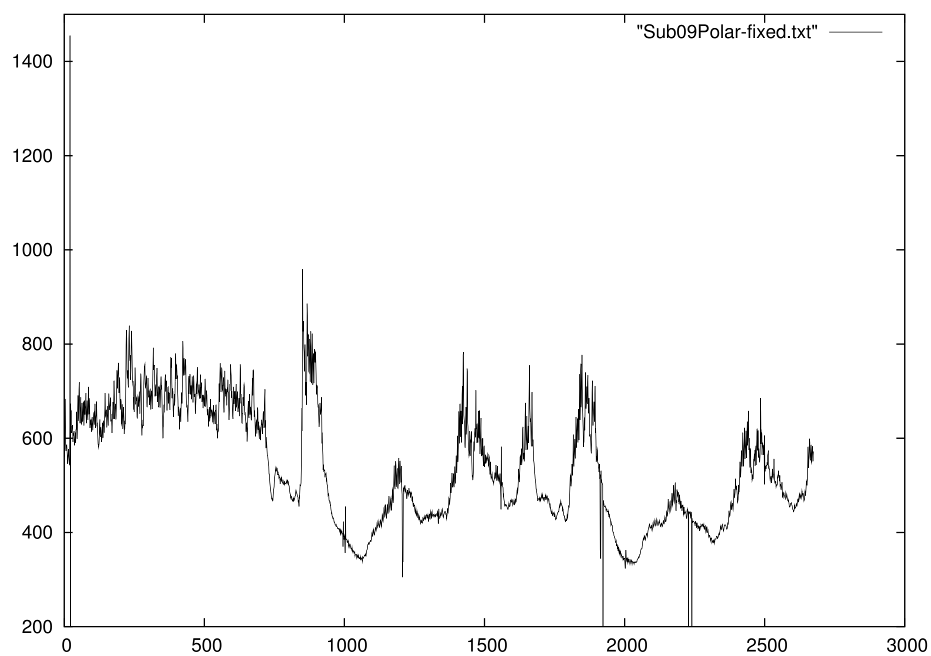


Figure 3.29: Subject 9 Polar corrected by Computer

Chapter 4

Conclusions

This thesis presents a method for the automatic correction of IBI data. This automated method was evaluated against a pair of human graders using the train and test paradigm. The current results indicate a 97.5% agreement in areas that humans deem to be correctable. This is probably not a high enough rate of agreement for use in sensitive clinical studies, but it may be useful in situations where the errors introduced by invalid data are more problematic than any issues that may be introduced by a poor agreement rate.

4.1 Contribution

This study adds to the body of literature on the automated correction of heart interbeat intervals. In particular, this is the only study the author is aware of that contains a detailed evaluation of the practices of human graders in correcting IBI data gathered under conditions that may interfere with the acquisition of good data. This alone provides a significant foundation for future studies. The attempt to correct heavily corrupted data with an algorithm designed to be applied while data are gathered is also, to the author's knowledge, unique. This has the potential to be a tremendous boon to researchers who wish to perform analysis of IBI data in situations where an offline correction process is not feasible.

4.2 Future work

One improvement that could be made to this experiment is in the area of human grading. The human graders used in this study had only minimal training before they began correcting the IBI data. Before seeing the results of the human scoring, we did not anticipate how differently two humans would correct IBI data, nor did we anticipate the effect this would have on our training method. The low agreement rate between the humans appears to have had a negative impact on training the method to properly identify errors, since the computer's agreement rate tracks the human agreement rate in Figure 3.5. This graph allows us to make the hypothesis that an increased rate of agreement between human graders would increase the computer's agreement with the humans in a similar fashion. One way to improve the human grading would be to give the human graders a better understanding of the physiological principles underlying the correction of IBI data before they began the correction process. Understanding the physiological causes of errors allows the human graders to have a metric outside of the data to gauge the appropriateness of a particular correction. If a suitable physiological explanation for a particular error cannot be determined, any applied correction should be viewed as suspect.

The training used in the current version of the program was primarily focused on error detection. Additional training methods should be examined in order to increase the frequency of appropriate classification. There are a number of potential improvements to the algorithm that could make classification training easier. One such improvement would be to fire each of the rules for every flagged IBI instead of terminating the search after the first acceptable correction. The best rule could be chosen based on which one fit the neighboring data best. This improvement would allow the possibility of different weights on the rules based on their frequency of application by human graders.

While the current method has some measures in place to evaluate IBIs within the context of their neighbors, each IBI is still evaluated individually. The ability to evaluate an entire neighborhood of IBIs is an important improvement that should be examined in further detail. This is particularly necessary when marking IBI values as uncorrectable. The frequency of errors plays a large role in determining the recoverability of a particular section of IBI data, and this information should be used appropriately. Identifying areas of uncorrectable data are dependent on more than a simple frequency count, however, and further study is required to determine the best approach.

An evaluation of the method's impact on analyses that may be dependent on the output would be an appropriate extention to the current work. IBI data are being examined and corrected with the intent of introducing them to other processes for further calculation. It is important to know the effect of the corrections not just on the data itself, but on the output of these additional stages. If the errors that the computer makes have little impact on further stages, the current implementation may be acceptable for some applications.

In conclusion, this thesis provides a foundation for further exploration of the automated correction of heart interbeat interval data. The current method may not be sufficient for the rigorous standards of clinical studies, but as the field of research on real time analysis of IBI data grows it may prove to be a useful component of more complex systems.

Bibliography

[1] L.A.N. Amaral, A.L. Goldberger, P.C. Ivanov, and H. E. Stanley, ”Modeling heart rate variabilitiy by stochastic feedback”, *Computer Physics Communications*, vol. 121-122, pp. 126-128, 1999.

[2] A.K. Barros and N. Ohnishi, ”Heart instantaneous frequency (HIF): An alternative approach to extract heart rate variability” *IEEE Transactions on Biomedical Engineering* vol. 48, pp. 850-855, 2001.

[3] G.G. Berntson, J.T Bigger, Jr., D.L. Eckberg, P. Grossman, P.G. Kaufmann, M. Malik, H. Nagaraja, S. Porges, J.P. Saul, P.H. Stone, and M.W. Van Der Molen, ”Heart rate variability: origins, methods, and interpretive caveats”, *Psychophysiology*, vol. 34, pp. 623-648, 1997.

[4] G.G. Berntson, K.S. Quigley, J.F. Jang, and S.T. Boysen, ”An approach to artifact identification: application to heart period data”, *Psychophysiology*, vol. 27, no. 5, pp. 586-598, 1990.

[5] P.F. Binkley, E. Nunziata, G.J. Haas, S.D. Nelson, and R.J. Cody, ”Parasympathetic withdrawal is an integral component of autonomic imbalance in congestive heart failure: demonstration in human subjects and verification in a paced canine model of ventricular failure”, *Journal of the American College of Cardiology*, vol. 18, no. 2, pp. 464-472, 1991.

[6] R.M. Carney, R.D. Saunders, K.E. Freedland, P. Stein, M.W. Rich, A.S. Jaffe, ”Association of depression with reduced heart rate variability in coronary artery disease”, *The American Journal of Cardiology*, vol. 76, no.8, pp. 562-564, 1995.

[7] J.T. Catalano, *Guide to ECG Analysis*. Lippincott Williams & Wilkins, 2002.

[8] M. Cheung, ”Detection of and recovery from errors in cardiac interbeat intervals”, *Psychophysiology*, vol. 18, no. 3, pp. 341-346, 1981.

[9] M.B. Conover, *Understanding Electrocardiography*. Mosby, 1996.

[10] T.G. Farrell, Y. Bashir, T. Cripps, M. Malik, J. Poloniecki, E.D. Bennett, D.E. Ward, and A. J. Camm, ”Risk stratification for arrhythmic events in postinfarction patients based on heart rate variability, ambulatory electrocardiographic variables and the signal-averaged electrocardiogram”, *Journal of the American College of Cardiology*, vol. 18, no. 3, pp. 687-697, 1991.

[11] B.H. Friedman and J.F. Thayer ”Autonomic balance revisited: panic anxiety and heart rate variability”, *Journal of Psychosomatic Research*, vol. 44, no. 1, pp. 133-151, 1998.

[12] M.A. García-González, R. Pallàs-Areny, ”A novel robust index to assess beat-to-beat variability in heart rate time-series analysis”, *IEEE Transactions on Biomedical Engineering*, vol. 48, no. 6., pp. 617-621, 2001.

[13] D.R. Gold, A. Litonjua, J. Schwartz, E. Lovett, A. Larson, B. Nearing, G. Allen, M. Verrier, R. Cherry, and R. Verrier, ”Ambient pollution and heart rate variability”, *Circulation*, vol. 101, p. 1267-1273, 2000.

[14] P. Grossman, K.H.L. Janssen, and D. Vaitl, *Cardiorespiratory and Cardiosomatic Psychophysiology*. New York and London: Plenum Press, 1983.

[15] P. Grossman, J. van Beek, and C. Wientjes ”A comparison of three quantification methods for estimation of respiratory sinus arrhythmia”, *Psychophsiology*, vol. 27, no. 6, pp. 702-714, 1990.

[16] J. Hayano, Y. Sakakibara, A. Yamada, M. Yamada, S. Muka, T. Fujinami, K. Yokoyama, Y. Watanabe, and K. Takata, ”Accuracy of assessment of cardiac vagal tone by heart rate variability in normal subjects”, *The American Journal of Cardiology*, vo. 67, pp. 199-204, 1991.

[17] A. Hoover and E. Muth, ”A real-time index of vagal activity”, *International Journal of Human-Computer Interaction*, vol. 17, no. 2, pp. 197-209, 2004.

[18] J.R. Jennings, W.K. Berg, J.S. Hutheson, P. Obrist, S. Porges, and G. Turpin, ”Publication guidelines for heart rate studies in man”, *Psychophysiology*, vol. 18, no. 3, pp. 226-231, 1981.

[19] P.G. Katona and F. Jih, ”Respiratory sinus arrhythmia: noninvasive measure of parasympathetic cardiac control”, *Journal of Applied Physiology*, vol. 39, no. 5, pp.801-805, 1975.

[20] R.E. Kleiger, J.P. Miller, J.T Bigger, Jr., and A.J. Moss, ”Decreased heart rate variability and its association with increased mortality after acute myocardial infarction” *The American Journal of Cardiology*, vol. 59, no. 4, pp. 256-262, 1987.

[21] M.T. La Rovere, J.T. Bigger, Jr., F.I. Marcus, A. Mortara, and P.J. Schwartz, ”Baroreflex sensitivity and heart-rate variability in prediction of total cardiac mortality after yocardia infarction. ATRAMI (Autonomic Tone and Reflexes After Myocardial Infarction) Investigators.” *Lancet*, vol. 351, pp. 478-484, 1998.

[22] M. Malik, T. Cripps, T. Farrell, and A.J. Camm ”Prognostic value of heart rate variability after myocardial infarction. A comparison of different data processing methods.” *Medical and Biological Engineering and Computing*, vol. 27, pp. 603-611, 1989.

[23] A.E. Mezzacappa, D. Kinklon, F. Earls, and J.P. Saul, ”The utility of spectral analytic techniques in the study of the autonomic regulation of beat-to-beat heart rate variability”, *International Journal of Methods in Psyciatric Research*, vol. 4, pp. 29-44, 1994.

[24] O. Odemuyiwa, M. Malik, T. Farrell, Y. Bashir, J. Poloniecki, and J. Camm, ”Comparison of the predictive characteristics of heart rate varibility index and left ventricular ejection fraction for al-cause mortality, arrhythmic events and suddend death after acute myocardial infarction”, *The American Journal of Cardiology*, vol. 68, no. 5, pp. 434-439, 1991.

[25] P. Ponikowski, S.D. Anker, T.P. Chua, R. Szelemej, M. Piepoli, S. Adamopoulos, K. Webb-Peploe, D. Harrington, W. Banasiak, K. Wrabec, and A.J. Coats, ”Depressed heart rate variability as an independent predictor of death in chronic congestive heart failure secondary to ischemic or idiopathic dilated cardiomyopathy”, *The American Journal of Cardiology*, vol. 79, no. 12, pp. 1645-1650, 1997.

[26] C.A. Pope, 3rd, R.I. Verrier, E.G. Lovett, A.C. Larson, M.E. Raizenne, R.E. Kanner, J. Schwartz, G.M. Villegas, D.R. Gold, and D.W. Dockery, ”Heart rate variability associated with particulate air pollution”, *American Heart Journal*, vol. 138, no. 5, pp. 804-807, 1999.

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

[28] J.P. Saul, ”Beat-to-beat variations of heart rate reflect modulation of cardiac atuonomic outlflow”, *Physiology*, vol. 5, pp. 32-37, 1990.

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

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

[31] P.E. Tikkanen, ”Nonlinear wavelet and wavelet packet denoising of electrocardiogram signal”, *Biological Cybernetics*, vol. 80, pp. 259-267, 1999.

[32] P.E. Tikkanen, L.C. Sellin, H.O. Kinnunen, and H.V. Huikuri, ”Using simulated noise to define optimal GT intervals for computer analysis of ambulatory ECG”, *Medical Engineering & Physics*, vol. 21, pp. 15-25, 1999.

[33] H. Tsuji, F.J. Venditti, E.S. Manders, J.C. Evans, M.G. Larson, C.L. Feldman, and D. Levy, ”Reduced heart rate variability and mortality risk in an elderly cohort”, *Circulation*, vol. 90, pp. 878-883, 1994.

[34] H. Tsuji, M.G. Larson, F.J. Venditti, Jr., E.S. Manders, J.C. Evans, C.L. Feldman, and D. Levy, ”Impact of reduced heart rate variability on risk for cardiac events”, *Circulation*, vol. 94, pp. 2850-2855, 1996.