Estimation of Missing Data in Multi-channel Physiological Time-series by Average Substitution with Timing from a Reference Channel

Philip Langley, Susan King, Kun Wang, Dingchang Zheng, Roberto Giovannini, Marjan Bojarnejad, Alan Murray

Newcastle University and Freeman Hospital, Newcastle upon Tyne, UK

Abstract

The Computing in Cardiology Challenge 2010 was to develop a computer algorithm for reconstructing missing sections of physiological data.

For cardiac related signals our algorithm obtained beat timings from a reference timing channel. Missing beats were estimated from the average of non-missing beats. ECG derived respiration was used for missing respiratory data.

Score for event 1 was 59 and for event 2 was 72.

Good scores were achieved despite beat-to-beat reconstructions that lacked important physiological detail, serving to illustrate that it is essential to evaluate the clinical impact of reconstruction algorithms.

1. Introduction

Physiological data continuously acquired from the clinical environment is often corrupted by noise, artifact or interruption, so that sections of data are un-analysable. If sufficiently accurate estimations of these sections of data could be derived, a more complete analysis could then be achieved and that was the motivation for the Computing in Cardiology/PhysioNet 2010 Challenge [1].

Average beat substitution, which estimates missing beats from averages of available beats, might provide accurate estimates of missing data from recordings from patients who have stable cardiovascular characteristics. However, when applied to data from patients with subtle but important physiological beat-to-beat variations, for example T wave alternans, such variability would not be reconstructed.

We investigated whether crude estimates of missing time series data based on average beat substitution, and which lack basic physiological detail, could provide accurate reconstructions as quantified by the Challenge scores based on RMS differences and correlations [1].

2. Methods

The dataset comprised signals exhibiting variations relating to the cardiac cycle (eg ECG, blood pressure) and respiration. For the cardiac related signals the algorithm used average beat substitution as the estimate of missing beats, while for respiration, a surrogate respiratory signal was derived from the ECG as illustrated in figure 1.

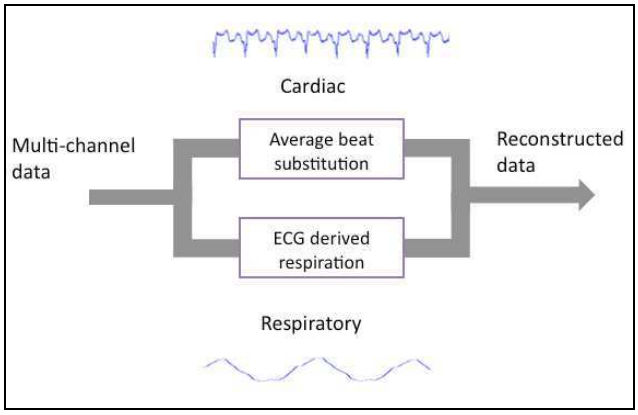

Figure 1. Block diagram illustrating different reconstruction schemes used for cardiac and respiratory data.

For cardiac signals, the timing of missing beats was estimated from the timing of beats in a reference timing channel. The reference timing channel was preferentially chosen to be an ECG channel if available. Beat locations were obtained using a QRS detection algorithm as illustrated in figure 2.

Average beat substitution reconstructs each missing beat from the average of available beats. However, it is necessary to consider RR interval variability when generating the beat average. An average beat generated from beats with significantly different RR intervals would be highly distorted. One solution is to use only beats that have the specific RR interval of the beat being reconstructed to calculate the average beat. However, this requires a search across all the beats to find those with RR intervals within a close tolerance of the specific RR

interval, and it is possible that such beats may not be present.

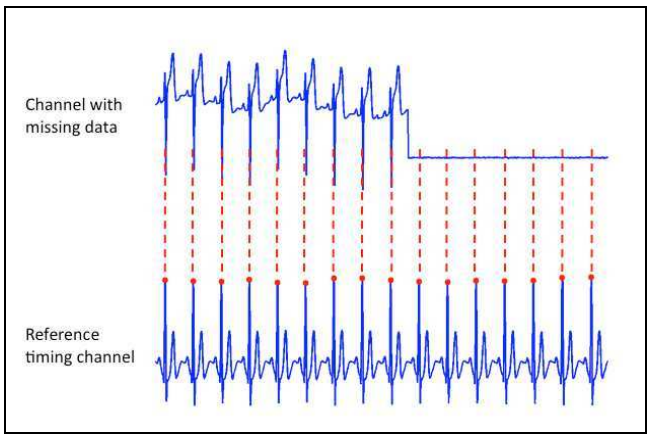

Figure 2. A 10 s strip of ECG with missing data (top trace) and a simultaneously recorded further ECG lead without missing data (bottom trace). The lead without missing data provides the reference timing channel from which the locations of missing beats are determined by QRS detection.

Our approach was to generate the average beat sample by sample taking into account the location of each sample within the current beat. So, if the sample being reconstructed was in the first 2/3rd of the beat interval, say N_s samples from the start of the beat ($N_s < 2xN_{rr}/3$, where N_{rr} is the number of sample points in the beat interval), the amplitude was the average of the signal amplitudes at points N_s samples from the start of all available beats. Similarly, if the sample was in the last third of the beat interval, N_d samples from the end of the beat interval ($N_d <= N_{rr}/3$), the amplitude was the average of the signal amplitudes at all points N_d samples from the end of all available beats. This method has the advantage that all available beats can be used to generate the average beat, but has the disadvantage of a discontinuity at the sample point $2xN_{rr}/3$. This discontinuity can be very significant when the signal is changing rapidly at that sample point, so we chose the point to lie within the quiescent TP interval for many of the ECG signals, ie $2xN_{rr}/3$, rather than the more intuitive $N_{rr}/2$ which in many cases coincided with the rapidly descending T wave. Figure 3 illustrates the calculation of the signal amplitude for a sample point lying within the last third of the beat interval.

To reconstruct the respiratory signal, an ECG derived respiration algorithm was used [2]. Briefly, principal component analysis was applied to the matrix of the collection of ECG beats occurring over the period of the missing respiratory signal. The eigenvector of the first principal component provided the surrogate of the respiratory signal.

A final processing step for both cardiac and respiratory reconstructed signals was to compensate for offsets in the means of reconstructed and actual signals. The mean of the reconstructed signal was made equal to the estimate of the mean of the actual signal. The estimate of the mean of the actual signal was from the mean of the last 10 s before signal loss.

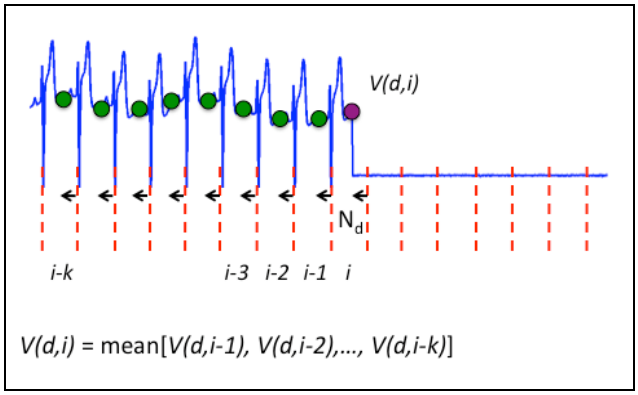

Figure 3. Calculation of signal amplitude (V) for a sample in the last third of the first missing beat interval (purple dot), illustrating that the amplitude is the average of the signal N_d samples from the end in all available beats (green dots). i represents the beat number, k the number of available beats and d the offset of the sample from the end of the beat interval.

3. Results

Table 1 shows the results for events 1 and 2 when applying the algorithm to the three Challenge data sets, with 100 reconstructions in each.

Table 1. Scores for each event and data set.

Data set	Event 1	Event 2
A	54.3	66.5
B	67.8	78.0
C	59.2	71.8

Data set A was the training set from which the missing data was known and data sets B and C were test sets from which the missing data was unknown [1]. Figure 4 presents an analysis of our results in terms of the type of missing data for data set A.

4. Discussion

We have developed an algorithm to estimate missing data by using beat averaging for cardiac parameters and ECG derived respiration. The algorithm achieved the best scores for the cardiac related signals. The simple algorithm was able to achieve very high scores for some recordings despite lacking physiological detail. Average

substitution places identical beats at each beat location so the reconstructions lack important beat-to-beat variability associated with the real data. Figure 5 presents some examples showing actual and reconstructed data from 10 recordings from data set A. The 10 recordings were chosen to illustrate reconstructions over the full range of scores achieved by the algorithm.

A major limitation for the respiratory reconstructions was that ECG derived respiration did not always accurately determine the phase of respiration. Further development of the algorithm to match the phase of the reconstructed respiratory signal to the estimated phase of the actual signal could address this limitation.

In conclusion, crude estimates of missing data from beat averages gave good results in terms of the Challenge scores. However, it would be essential to validate any algorithms against the clinical parameters of interest.

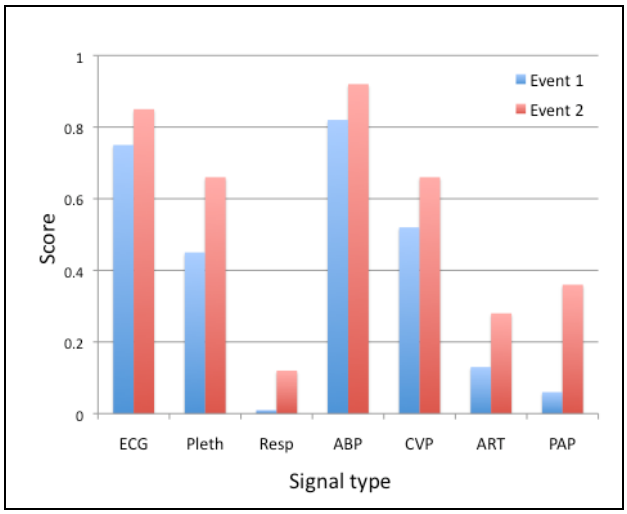

Figure 4. Mean scores achieved for different missing signal types from data set A.

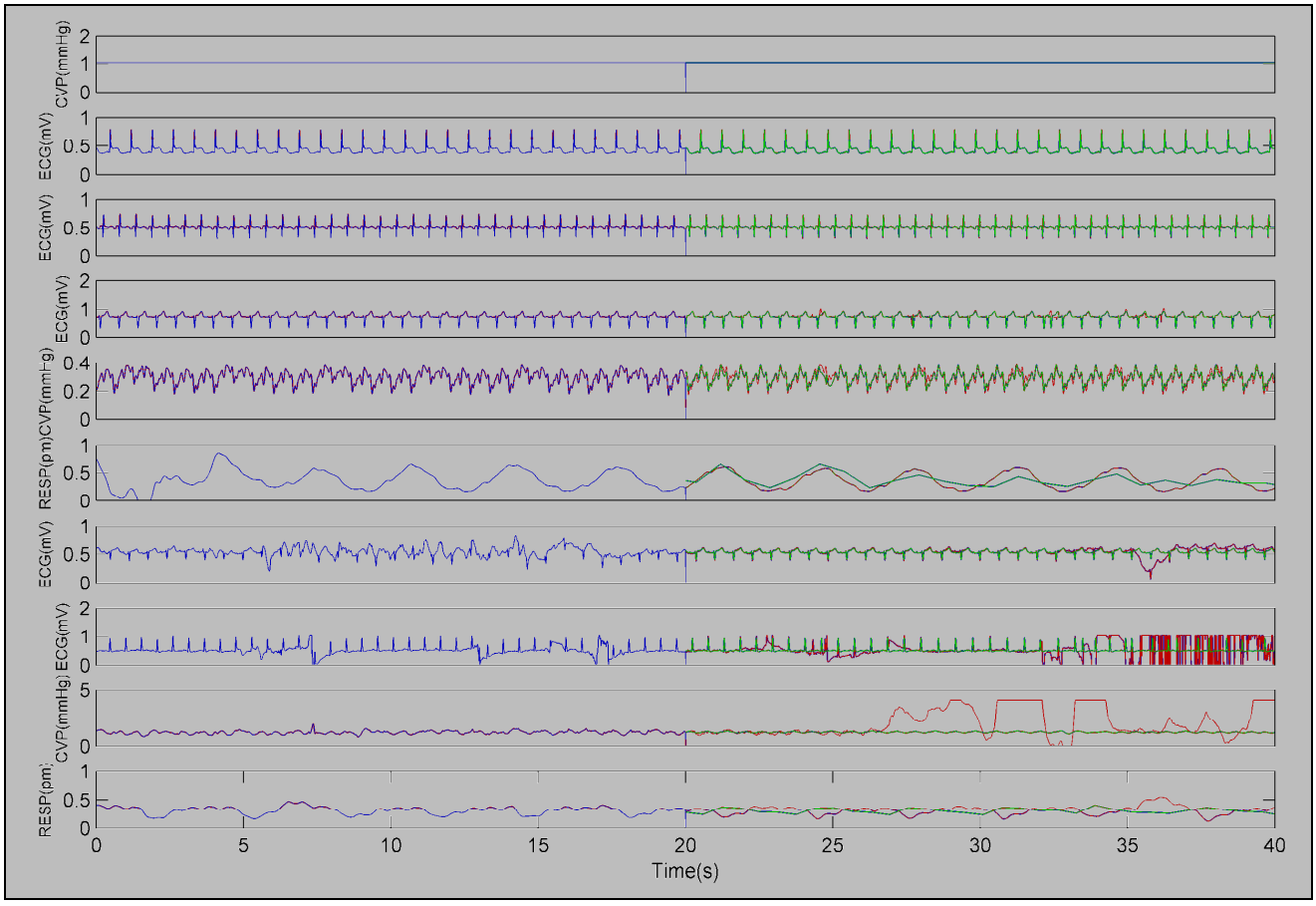

Figure 5. Examples of reconstructed missing data for a range of signal types. Blue trace is the signal immediately before data loss, green is the reconstructed signal and red is the actual signal. Where no red shows, the reconstructed signal overlaps the actual signal. Signals are arranged in score order with those at the top scoring highly and those at the bottom achieving low scores.

References

[1] Moody GB. The PhysioNet/Computing in cardiology challenge 2010: Mind the gap. Comput Cardiol 2010;37.

[2] Langley P, Bowers EJ, Murray A. Principal component analysis as a tool for analysing beat-to-beat changes in electrocardiogram features. Application to electrocardiogram derived respiration. IEEE TBME 2010;57:821-829.

Address for correspondence.

Philip Langley
Cardiovascular Physics and Engineering Research Group
Medical Physics Department
Freeman Hospital
Newcastle upon Tyne
NE7 7DN
UK
philip.langley@ncl.ac.uk