NON-ITERATIVE MISSING SAMPLES RECOVERY OF ECG SIGNALS BY LMMSE ESTIMATION FOR AN AUTOREGRESSIVE CYCLOSTATIONARY MODEL

Amir Weiss and Arie Yeredor

School of Electrical Engineering
Tel-Aviv University
P.O. Box 39040, Tel-Aviv 69978, Israel
{amirwei2@mail, arie@eng}.tau.ac.il

ABSTRACT

Electrocardiography (ECG) measured using wearable wireless sensors is already commonly used for several years, as one of the products of the emerging Telemedicine field, which is one the main branches in eHealth applications. In this work we address the problem of missing samples recovery of such ECG (digital) signals, resulting from temporally-local communication dropouts. We propose a new model for the ECG signal based on its conspicuous quasi-periodical characteristics in short time intervals, along with a compatible estimation procedure tailored to the proposed model. We extend the autoregressive (AR) model, previously proposed by Prieto-Guerrero *et al.*, to a cyclostationary AR model, and our proposed estimation scheme incorporates a first phase of model parameters estimation, followed by a Linear Minimum Mean Squared Error (LMMSE) estimation phase of the missing samples. We demonstrate significant improvement compared to the AR method in simulation experiments using real ECG data.

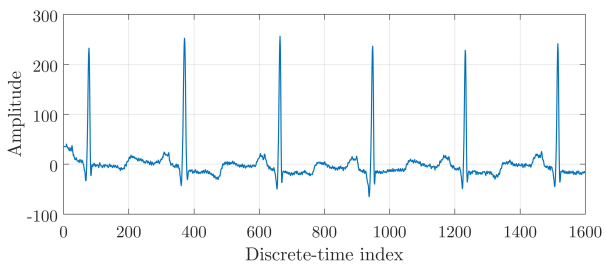
Index Terms— ECG, missing samples recovery, cyclostationary, LMMSE estimation.

1. INTRODUCTION

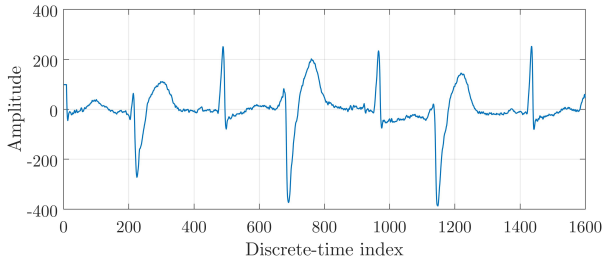
Missing samples recovery is a common well-known problem which was addressed in various contexts such as audio restoration [1, 2], image completion [3, 4], data recovery in electricity distribution systems [5], biomedical applications [6, 7] and more. In the context of biomedical applications, the interest in *eHealth* applications has increased significantly in recent years [8–10], and in particular also in a field termed *Telemedicine*, which is the use of telecommunication and information technology to provide clinical health care from a distance. Telemedicine can be beneficial to patients in remote regions or in case of emergencies, where reliable real-time monitoring is of great interest. One form of telemedicine is performing an Electrocardiography (ECG) with wearable wireless sensors using some wireless communication link, e.g., WiFi. The ECG signal is sampled and transmitted to a near by computing unit, from which it is sent to the desired destination (e.g., to an emergency medical center) via some other communication link (e.g., cellular or satellite). During the transmission process of the digital ECG signal, dropouts in communication links might occur, and in some case the Transmission Control Protocol (TCP) will allow packet loss for the sake of lower latency. In such cases, where the TCP is not modified especially for the ECG transmission, considering that an accurate diagnostic by a physician depends very much on a clear readable ECG signal, possible losses in transmission may become crucial. To this end, missing samples recovery of the ECG signal is clearly of interest and was previously addressed by Prieto-Guerrero *et al.* in [11] and later in [12], where a reconstruction method for ECG signals was proposed, based on autoregressive (AR) modeling. In this paper, we propose an enhanced model for the ECG signal and accordingly a different estimation procedure of the missing samples. The model relies on the observed ECG signal quasi-periodical characteristics in short time intervals, which can be easily seen in Fig. 1, presenting two different typical ECG signals as representative examples. Therefore, a suitable model would encompass both its random and quasi-periodical characteristics. One plausible option for such a model is an AR *cyclostationary* random process, which will preserve the advantage of correlations to near by samples and will also enjoy the benefit of the periodical statistical properties. In this paper we consider such a model along with a compatible non-iterative efficient estimation procedure, and demonstrate its superiority over the previously suggested AR method.

The rest of this paper is structured as follows. In Section II we propose a model for the ECG signal and formulate the problem of missing samples recovery in this context. In Section III we present the proposed estimation procedure for the recovery of the missing samples. Comparative simulation results on real ECG data are presented in Section IV, and Sec-

The first author would like to thank the Yitzhak and Chaya Weinstein Research Institute for Signal Processing for a fellowship.



(a)



(b)

Fig. 1: Examples of two ECG signal taken from the MIT-BIH Arrhythmia Database. (a) ECG 100 (b) ECG 200.

tion V is devoted to conclusions.

2. THE PROPOSED MODEL AND PROBLEM FORMULATION

Due to the quasi-periodical characteristics of the ECG signal, we propose to model it as an AR process generated by a cyclostationary (uncorrelated) driving-noise. More formally, denote $x[n]$ as the discrete-time signal which represents the ECG signal samples. Thus, $x[n]$ is considered as an AR process of (known) order p satisfying

$$x[n] = -\sum_{k=1}^{p} a_k \cdot x[n-k] + v[n], \tag{1}$$

with

$$v[n] = \sqrt{d[n]} \cdot w[n], \tag{2}$$

where $w[n]$ is a zero-mean unit variance white noise, $\{a_k\}_{k=1}^{p}$ are the unknown deterministic AR parameters and $d[n]$ is some unknown deterministic periodic (positive) function, i.e., $\forall n : d[n] = d[n+T] > 0, T \in \mathbb{Z} \setminus \{1\}$. Note that in the proposed model, the cyclic period T is considered an unknown deterministic parameter as well. Clearly, $v[n]$ is a cyclostationary random process, which serves as the driving-noise of the process $x[n]$.

The measured ECG signal is represented by the vector of samples $\boldsymbol{x} \in \mathbb{R}^{N\times 1}$, where its n-th element is $(\boldsymbol{x})_n = x[n]$. We assume that $K \in \mathbb{Z}$ consecutive samples are missing in a known position. In addition, we also assume $K \ll T$. To this end, denote $\boldsymbol{x}^{\mathrm{T}} \triangleq \left[\boldsymbol{y}_{\text{pre}}^{\mathrm{T}}\ \boldsymbol{\theta}^{\mathrm{T}}\ \boldsymbol{y}_{\text{post}}^{\mathrm{T}}\right] \in \mathbb{R}^{1\times N}$, where $\boldsymbol{\theta} \in \mathbb{R}^{K\times 1}$ is the vector of missing samples and $\boldsymbol{y}_{\text{pre}} \in \mathbb{R}^{N_1\times 1}, \boldsymbol{y}_{\text{post}} \in \mathbb{R}^{N_2\times 1}$ ($N_1 + N_2 = N - K$) are the vectors of available measurements before and after the missing samples interval, respectively (where $(\cdot)^{\mathrm{T}}$ denotes the transpose). Thus, our goal is to estimate the vector of unknown samples $\boldsymbol{\theta}$ from the vector of measurements $\boldsymbol{y}^{\mathrm{T}} \triangleq \left[\boldsymbol{y}_{\text{pre}}^{\mathrm{T}}\ \boldsymbol{y}_{\text{pre}}^{\mathrm{T}}\right] \in \mathbb{R}^{1\times N-K}$. In the following section we propose a non-iterative estimation procedure by which the necessary deterministic model parameters are estimated first, and based on these estimated parameters, the missing samples are then estimated according to the Linear Minimum Mean Square Error (LMMSE) criterion.

3. THE PROPOSED METHOD

Given the vector of measurements $\boldsymbol{y}$, estimate the vector of missing samples $\boldsymbol{\theta}$ according to the following procedure:

1. Estimate the first $p+1$ autocorrelation lags of $x[n]$ from the available samples $\boldsymbol{y}$ using ordinary correlation estimates;

2. Using the estimates from step 1, estimate the AR parameters (of order p) by solving the Yule-Walker equations (e.g., [13]);

3. Using the estimated AR parameters from step 2, filter the vector $\left[\boldsymbol{y}_{\text{pre}}^{\mathrm{T}}\ \boldsymbol{0}^{\mathrm{T}}\ \boldsymbol{y}_{\text{post}}^{\mathrm{T}}\right] \in \mathbb{R}^{1\times N}$ (where $\boldsymbol{0} \in \mathbb{R}^{K\times 1}$ is the all-zeros vector) with the corresponding Finite Impulse Response (FIR) inverse AR filter to obtain the estimated driving-noise vector, denoted by $\hat{\boldsymbol{v}}$;

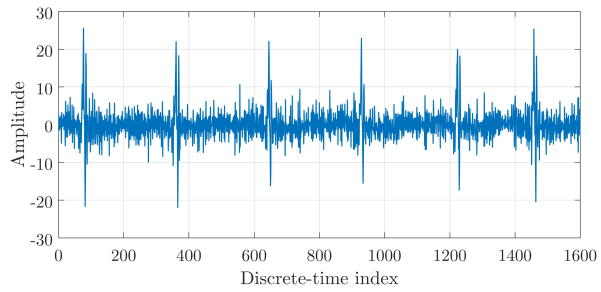
4. Using the Discrete Fourier Transform (DFT) of the squared driving-noise sequence $\hat{\boldsymbol{v}}^2 \triangleq \hat{\boldsymbol{v}} \circ \hat{\boldsymbol{v}}$, where $\circ$ denotes the Hadamard (element-wise) product, estimate the cyclic period T by the reciprocal of the (non-DC) highest peak-location multiplied by the DFT length (and rounded to an integer); denote this estimate as $\hat{T}$;

5. Based on $\hat{T}$, estimate $d[n]$, the time-varying variance of the driving-noise $v[n]$, by averaging all the full periods (only) of length $\hat{T}$ of $\hat{\boldsymbol{v}}^2$ and then concatenating the estimated period as necessary to obtain the total length; denote this sequence as $\hat{d}[n]$;

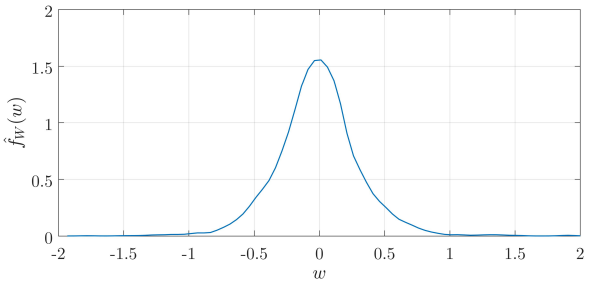
6. Using the results of step 2 and 5, construct the estimated covariance matrix of $\boldsymbol{x}$: notice that according to our model (neglecting edge-effects)

$$\boldsymbol{\Gamma}\boldsymbol{x} = \boldsymbol{D}^{\frac{1}{2}}\boldsymbol{w} \rightarrow \boldsymbol{C}_{\boldsymbol{xx}} \triangleq E\left[\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}\right] = \boldsymbol{\Gamma}^{-1}\boldsymbol{D}\boldsymbol{\Gamma}^{-\mathrm{T}}, \tag{3}$$

where $\boldsymbol{D}^{\frac{1}{2}} \in \mathbb{R}^{N\times N}$ is a diagonal matrix whose diagonal elements are $\left(\boldsymbol{D}^{\frac{1}{2}}\right)_{nn} = \sqrt{d[n]}$, $\boldsymbol{\Gamma} \in \mathbb{R}^{N\times N}$ is a (Toeplitz) matrix representation of the FIR (inverse



(a)



(b)

Fig. 2: (a) The estimated driving-noise $\hat{\boldsymbol{v}} = \widehat{\boldsymbol{\Gamma}}\boldsymbol{x}$, obtained by filtering ECG 100 using the FIR inverse AR filter, obtained from its estimated AR parameters. (b) The estimated marginal PDF $\hat{f}_W(w)$ of $\hat{\boldsymbol{w}} = \widehat{\boldsymbol{D}}^{-\frac{1}{2}}\hat{\boldsymbol{v}}$ obtained from ECG 100.

AR) filter which was used in step 3, $\boldsymbol{w} \in \mathbb{R}^{N\times 1}$ is the vector representation of the white noise $w[n]$ (with $(\boldsymbol{w})_n = w[n]$), $\boldsymbol{C}_{\boldsymbol{xx}}$ is the covariance matrix of $\boldsymbol{x}$, and $(\cdot)^{-\mathrm{T}}$ denotes the inverse of the transpose. Thus,

$$\widehat{\boldsymbol{C}}_{\boldsymbol{xx}} \triangleq \widehat{\boldsymbol{\Gamma}}^{-1}\widehat{\boldsymbol{D}}\widehat{\boldsymbol{\Gamma}}^{-\mathrm{T}}, \tag{4}$$

where $\widehat{\boldsymbol{C}}_{\boldsymbol{xx}}, \widehat{\boldsymbol{\Gamma}}$ and $\widehat{\boldsymbol{D}}$ are the estimates of $\boldsymbol{C}_{\boldsymbol{xx}}, \boldsymbol{\Gamma}$ and $\boldsymbol{D}$ (based on $\boldsymbol{y}$), respectively.

7. Using the (symmetric) permutation matrix $\boldsymbol{\Pi} \in \mathbb{R}^{N\times N}$, such that $\boldsymbol{z}^{\mathrm{T}} \triangleq \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\Pi}^{\mathrm{T}} = \begin{bmatrix}\boldsymbol{\theta}^{\mathrm{T}} & \boldsymbol{y}^{\mathrm{T}}\end{bmatrix}$, construct the LMMSE estimate of $\boldsymbol{\theta}$ from $\boldsymbol{y}$ based on $\widehat{\boldsymbol{C}}_{\boldsymbol{xx}}$ (e.g., [14])

$$\hat{\boldsymbol{\theta}} = \widehat{\boldsymbol{C}}_{\boldsymbol{\theta y}}\widehat{\boldsymbol{C}}_{\boldsymbol{yy}}^{-1}\boldsymbol{y} = -\widehat{\boldsymbol{\Lambda}}_{\boldsymbol{\theta\theta}}^{-1}\widehat{\boldsymbol{\Lambda}}_{\boldsymbol{\theta y}}\boldsymbol{y}, \tag{5}$$

where

$$\widehat{\boldsymbol{C}}_{\boldsymbol{zz}}^{-1} \triangleq \boldsymbol{\Pi}\widehat{\boldsymbol{C}}_{\boldsymbol{xx}}^{-1}\boldsymbol{\Pi}^{\mathrm{T}} \triangleq \begin{bmatrix}\widehat{\boldsymbol{C}}_{\boldsymbol{\theta\theta}} & \widehat{\boldsymbol{C}}_{\boldsymbol{\theta y}} \\ \widehat{\boldsymbol{C}}_{\boldsymbol{y\theta}} & \widehat{\boldsymbol{C}}_{\boldsymbol{yy}}\end{bmatrix}^{-1} \triangleq \begin{bmatrix}\widehat{\boldsymbol{\Lambda}}_{\boldsymbol{\theta\theta}} & \widehat{\boldsymbol{\Lambda}}_{\boldsymbol{\theta y}} \\ \widehat{\boldsymbol{\Lambda}}_{\boldsymbol{y\theta}} & \widehat{\boldsymbol{\Lambda}}_{\boldsymbol{yy}}\end{bmatrix}. \tag{6}$$

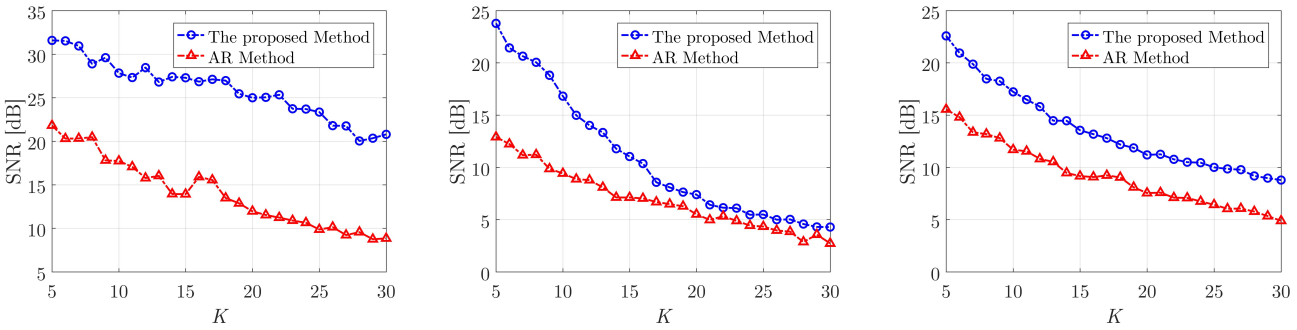
Before we proceed to demonstrate the proposed method's performance compared to another ECG missing samples recovery method, we would like to note the following:

- We assume that the given signal (with missing samples) contains at least a few cyclic periods, i.e., $T \ll N$. This is assumed in order to obtain a "good" estimate of the cyclic period.
- Although our method is only based on second-order statistics, it may be reasonable to assume that the driving-noise, and therefore also the ECG signal, are Gaussian distributed. This assumption was examined on real data and appears to be rather plausible. Fig. 2a shows an example of an ECG signal (ECG 100) which was filtered by the FIR inverse AR filter, obtained from its estimated AR parameters. It is readily seen that the filtered signal $\hat{\boldsymbol{v}}$ exhibits quasi-periodical characteristics. In addition, it can be easily seen from Fig. 2b, presenting the estimated Probability Density Punction (PDF) of the signal $\hat{\boldsymbol{w}} = \boldsymbol{D}^{-\frac{1}{2}}\hat{\boldsymbol{v}}$, that at least its marginal distribution is approximately Gaussian. Although the marginal Gaussianity does not necessarily imply full Gaussianity of the entire process, if the ECG signals are indeed approximately Gaussian then our LMMSE is approximately the MMSE.
- In step 7, ideally, we would take all the available measurements to estimate the missing samples. However, since we need to compute the estimated covariance matrix $\widehat{\boldsymbol{C}}_{\boldsymbol{xx}}$ by matrix multiplications, this would become computationally expensive for "very long" vectors. Thus, we use all the samples in $\boldsymbol{y}$ for the model parameters estimation (which is relatively cheap in the sense of computational load), and we use only the M "closest" samples to $\boldsymbol{\theta}$ for the estimation of the missing samples, where $M \in \mathbb{Z}$ is a (tunable) predefined parameter of the algorithm. Of course, if computational resources are abundant, all entries of $\boldsymbol{y}$ may be used.
- In order to mitigate initial conditions edge effects of the filtering process in step 3, we set $\hat{v}[n] = E[v[n]] = 0, \forall n \in \{1,\ldots,p+1\}\cup\{N_1+1,\ldots,N_1+K+p+1\}$.
- Notice that computing the inverse of $\widehat{\boldsymbol{C}}_{\boldsymbol{yy}}$ directly is very costly when N is large (recall we assume $K \ll N$). However, we may compute the inverse of $\widehat{\boldsymbol{C}}_{\boldsymbol{xx}}$ efficiently, since from (3) we have

$$\boldsymbol{C}_{\boldsymbol{xx}}^{-1} = \boldsymbol{\Gamma}^{\mathrm{T}}\boldsymbol{D}^{-1}\boldsymbol{\Gamma}, \tag{7}$$

which means we only have to compute the reciprocal values of the sequence lying on the diagonal of $\boldsymbol{D}$, i.e., $(d[n])^{-1}$. Once we obtain $\widehat{\boldsymbol{C}}_{\boldsymbol{xx}}^{-1}$, $\widehat{\boldsymbol{\Lambda}}_{\boldsymbol{\theta\theta}}^{-1}$ is also computationally (relatively) cheap, since its dimension is $K \times K$. Thus, in practice, we use the right hand side formula of (5).

We now turn to assess the performance of the proposed method on real ECG data.



(a) ECG record 102 (b) ECG record 215 (c) ECG record 228

Fig. 3: Averaged local SNR of missing samples recovery versus K for three different ECG signal recordings (102, 215, 228) taken from the MIT-BIH Arrhythmia Database. It is clearly seen that the proposed method attains significantly higher SNR for each of the examined signals, and for every missing samples interval size. Results were obtained by averaging 1000 trials of randomly selected positions for the missing interval.

4. EXPERIMENTAL RESULTS

In this section we test and evaluate the performance of the proposed method on real ECG data. Test signals were taken from the MIT-BIH Arrhythmia Database [15], consisting of several ECG records for different study cases. We chose three signals from this database which present different characteristics (records: 102, 215 and 228), in segments of 10[sec] sampled at a rate of 360[Hz]. The AR order was set to $p = 50$ for two considerations. First, according to [11, 12], for $p > 50$ the model error variance does not decrease significantly anymore. In addition, since we compare our proposed method to the AR method, in order to create fair comparison conditions we choose the same value of p as chosen in [11]. For applying the AR method [11], 250 samples were taken for the forward and backward interpolation, and α, the cross-fading windows roll-off, was set to 2 as in [11, 12], again, for creating fair comparison conditions. Accordingly, we set $M = 250$ for the same reasoning.

The ECG test signals from the MIT-BIH Arrhythmia Database were zeroed in random positions so as to simulate lost samples, where the size of the lost sample interval was varied from 5 to 30 consecutive samples. All possible positions of the missing interval were allowed (with equal probability), except for first and last 250 samples, in order to enable the AR method both forward and backward interpolation for each missing segment. Note that this exception is unnecessary for the implementation of our proposed method.

We assess the performance by the local Signal to reconstruction error (Noise) Ratio (SNR) measure, defined as

$$\text{SNR} \triangleq 10\log_{10}\left(\frac{\sigma_\theta^2}{\sigma_{\theta-\hat{\theta}}^2}\right)\text{[dB]}, \tag{8}$$

where σ_θ^2 is the average power of the original signal in the missing interval and $\sigma_{\theta-\hat{\theta}}^2$ is the average power of the reconstruction error. Fig. 3 shows the average local SNR of missing samples recovery vs. the missing interval size K for the three ECG signal recordings (each of length $N = 3600$ samples, equivalent to 10[sec]). The average local SNR was obtained by averaging 1000 trails of randomly selected positions for the missing interval. As can be seen, the proposed method yields superior performance over the AR method for all tested signals and for every missing interval size which were examined. When K is small, the proposed method's gain over the AR method is significant and is more than 10[dB] in some cases for the examined signals. As K increases, the local SNR decreases in both methods and the gap is narrowed as well. Nevertheless, the gain is still significant and strictly positive for each of the tested signals and missing intervals sizes.

5. CONCLUSION

In the context of missing samples recovery of ECG signals, we proposed an autoregressive cyclostationary model, which was empirically shown to be more suitable due to its quasi-periodical characteristics. Additionally, we proposed an appropriate non-iterative estimation procedure according to this model, which is comprised of two phases - model parameters estimation, followed by the missing samples estimation based on the LMMSE criterion. In comparison to a previously suggested approach, our model and estimation procedure yield superior results and enhanced reconstructed ECG signals.

6. REFERENCES

[1] AJEM Janssen, R Veldhuis, and L Vries, "Adaptive interpolation of discrete-time signals that can be modeled as autoregressive processes," *IEEE Transactions*

on Acoustics, Speech, and Signal Processing, vol. 34, no. 2, pp. 317–330, 1986.

[2] JJK Ó Ruanaidh and WJ Fitzgerald, "Interpolation of missing samples for audio restoration," *Electronics Letters*, vol. 30, no. 8, pp. 622–623, 1994.

[3] Shwu-Huey Yen, Hsien-Yang Li, and Po-Yen Kuo, "Image completion using sample synthesis," in *Proc. of IEEE International Conference on Advanced Information Networking and Applications (AINA)*, 2017, pp. 328–335.

[4] Radhakrishna Achanta, Nikolaos Arvanitopoulos Darginis, and Sabine Süsstrunk, "Extreme image completion," in *Proc. of ICASSP*, 2017.

[5] Cristian Genes, Inaki Esnaola, Samir M Perlaza, Luis F Ochoa, and Daniel Coca, "Recovering missing data via matrix completion in electricity distribution systems," in *Proc. of IEEE International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, 2016, pp. 1–6.

[6] Diego Martín-Martínez, Pablo Casaseca-de-la Higuera, Marcos Martín-Fernández, Abbes Amira, Chunbo Luo, Christos Grecos, and Carlos Alberola-López, "A stochastic modelling framework for the reconstruction of cardiovascular signals," in *Proc. of the 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2014, pp. 676–679.

[7] Kyoung Jae Kim, Vibhor Agrawal, Ignacio Gaunaurd, Robert S Gailey, and Christopher L Bennett, "Missing sample recovery for wireless inertial sensor-based human movement acquisition," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 24, no. 11, pp. 1191–1198, 2016.

[8] Michelle Omoogun, Visham Ramsurrun, Shivanand Guness, Preetila Seeam, Xavier Bellekens, and Amar Seeam, "Critical patient ehealth monitoring system using wearable sensors," in *Proc. of the 1st International Conference on Next Generation Computing Applications (NextComp)*. IEEE, 2017, pp. 169–174.

[9] Andreas P Plageras, Christos Stergiou, George Kokkonis, Kostas E Psannis, Yutaka Ishibashi, Byung-Gyu Kim, and B Brij Gupta, "Efficient large-scale medical data (ehealth big data) analytics in internet of things," in *Proc. of IEEE 19th Conference on Business Informatics (CBI)*, 2017, vol. 2, pp. 21–27.

[10] Pei Huang, Borui Li, Linke Guo, Zhanpeng Jin, and Yu Chen, "A robust and reusable ecg-based authentication and data encryption scheme for ehealth systems," in *Proc. of IEEE Global Communications Conference (GLOBECOM)*, 2016, pp. 1–6.

[11] Alfonso Prieto-Guerrero, Corinne Mailhes, and Francis Castanié, "Lost sample recovering of ecg signals in e-health applications," in *Proc. of the 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBS)*, 2007, pp. 31–34.

[12] A Prieto-Guerrero, C Mailhes, and F Castanie, "Recovering electrocardiogram missing samples in wireless transmission," in *Proc. of Computers in Cardiology*. IEEE, 2009, pp. 845–848.

[13] Steven M Kay, *Fundamentals of statistical signal processing*, Prentice Hall PTR, 1993.

[14] Christopher M Bishop, *Pattern recognition and machine learning*, springer, 2006.

[15] A. L. Goldberger, L. A. N. Amaral, L. Glass, J. M. Hausdorff, P. Ch. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley, "PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals," *Circulation*, vol. 101, no. 23, pp. e215–e220, 2000 (June 13), Circulation Electronic Pages: http://circ.ahajournals.org/content/101/23/e215.full PMID:1085218; doi: 10.1161/01.CIR.101.23.e215.