MISSING DATA PROBLEM IN RANDOM ELECTROCARDIOGRAM SIGNAL PROCESSING

GAN THIAM YEE

A dissertation submitted in partial fulfillment of the
requirements for the award of the degree of
Master of Sciences (Mathematics)

Faculty of Sciences
Universiti Teknologi Malaysia

JANUARY 2014

Dedicated to

My supervisor, Dr. Ismail
and
All my friends…

ACKNOWLEDGEMENTS

First of all, I would like to express my greatest gratitude to Dr. Ismail Mohamad as my supervisor of this Dissertation who always being patient and guiding me in this study. Throughout my study, his guidance, suggestions, inspirations and caring are the greatest contribution for me to keep on. Once again, I am appreciating to all his efforts on me.

I am also grateful to the support from my friends especially the moral support from Jerry and the gang, who motivated me and built my momentum to finish the study. Besides, a special thanks to Liew, the one who provide useful and relevant resources. Lastly, to all many other people who did not listed in this limited space, thank you to you all.

ABSTRACT

Basically, signals are the entities that convey information and biomedical signals are the signals that carry information about the physiological process of organisms. Electrocardiogram (ECG) signal or known as heart signal is the signal that contains information about electrical activities in the heart. Since physiological signal are generated at low values and devices advancements are not sufficient to detect these small values perfectly, these signal tends to be missing from the record. As the noise interferes the signal at the same time, raw signal is practically unreliable to be interpreted directly. Hence, the random signal processing is required to obtain the signal as precise as possible. In this study, the missing probabilities of signal missingness were set to 0.1 at high values and 0.3 at low values. The noise to be reduced is Gaussian noise with zero mean and standard deviation 0.01 mV. A few methods have been applied to estimate the missing signal, including single mean imputation, empirical conditional mean imputation and Holt-Winters exponential smoothing. For noise filtering, the approach used is the Finite Impulse Response (FIR) Wiener filter. The study finds that the empirical conditional mean imputation is the best method among the three to estimate missing signal due to its accuracy, adequacy and simplicity. However, it appears that the FIR Wiener filter does not compatible with the estimation from empirical conditional mean imputation and does not further improve the signal quality by removing noise in general.

ABSTRAK

Pada asasnya, isyarat merupakan entiti yang menyampaikan maklumat dan isyarat bioperubatan adalah isyarat yang membawa maklumat tentang proses fisiologi organism. Isyarat elektrokardiogram (ECG) atau dikenali sebagai isyarat jantung adalah isyarat yang mengandungi maklumat mengenai aktiviti elektrik di dalam jantung. Oleh kerana isyarat fisiologi dihasilkan pada nilai yang rendah dan kemajuan peranti tidak mencukupi untuk mengesan nilai-nilai yang kecil ini dengan sempurna, isyarat ini sering terlepas dari pada rekod. Apabila gangguan isyarat belaku pada masa yang sama, isyarat mentah boleh dikatakan tidak boleh dipercayai untuk ditafsirkan secara langsung. Oleh itu, pemprosesan isyarat rawak diperlukan untuk mendapatkan isyarat setepat mungkin. Dalm kajian ini, kabarangkalian kehilangan isyarat telah ditetapkan kepada 0.1 pada nilai yang tinggi dan 0.3 pada nilai yang rendah. Hingar yang akan dikurangkan adalah hingar Gaussian dengan sifar min and sisihan piawai 0.01 mV. Beberapa kaedah telah digunakan untuk menganggar isyarat yang telah hilang, termasuklah imputasi min tunggal, imputasi min bersyarat empirikal dan pelicinan eksponen Holt-Winters. Untuk hingar penapisan, pendekatakan yang digunakan adalah Penapis Wiener sambutan dedenyut terhingga (FIR). Kajian ini mendapati imputasi min bersyarat empirikal adalah kaedah yang terbaik antara tiga untuk mengganggar isyarat yang hilang kerana ketepatan, kecukupan and keringkasanya. Walaubagaimanapun, ia kelihatan seperti penapis Wiener FIR tidak serasi dengan anggaran dari empirikal imputasi min dan tidak dapat meningkatkan kualiti isyarat dengan mengurangkan hingar secara amnya.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS AND SYMBOLS

AV	-	Atrioventricular
A^2	-	Anderson-Darling Statistics
$a[t]$	-	Amplitude Indicator Function
bpm	-	Beat per minute
$b(t)$	-	Trend Component of Time Series
ECG	-	Electrocardiogram
f	-	Frequency
FIR	-	Finite Impulse Response
$F(t)$	-	Estimated Signal from Missing Signal
$\vec{\boldsymbol{h}}$	-	Basis Vector of Filter Function
$\boldsymbol{H}$	-	Matrix of filter function
H_0	-	Null Hypothesis
H_1	-	Alternative Hypothesis
Hz	-	Hertz
$i[t]$	-	Missingness Indicator Function
$L(t)$	-	Level Component of Time Series
m_0	-	Number of Low Value Observations
m_1	-	Number of High Value Observations
MAE	-	Mean Absolute Error
MAR	-	Missing at Random
MCAR	-	Missing Completely at Random
mV	-	Milli-volt
$N[t]$	-	Generated noise
NMAR	-	Not Missing at Random

p_1	-	Probability of Missingness for Strong Signal
p_2	-	Probability of Missingness for Weak Signal
RMSE	-	Root Mean Square Error
$\boldsymbol{R}_{xy}$	-	Matrix of cross-correlation for $x[t]$ and $y[t]$
$\boldsymbol{R}_y$	-	Matrix of autocorrelation for $y[t]$
s	-	Seasonal length
SA	-	Sinoatrial
$S(t)$	-	Seasonal Component of Time Series
t	-	Time index
U_1	-	Uniform Random Variable
U_2	-	Uniform Random Variable
$u[t]$	-	Sequence of Uniform Random Numbers
V	-	Root Mean Square Voltage
$x[t]$	-	Sequence of Observed Signal
$y[t]$	-	Sequence of Real Signal
$\hat{y}[t]$	-	Estimation of Real Signal
Z_1	-	Standard Gaussian Random Variable
Z_2	-	Standard Gaussian Random Variable
β_0	-	Regression Parameter of Intercept
β_1	-	Regression Parameter of Gradient
$\delta[t]$	-	Probability of Missingness
ε	-	Sum of Squared Error
μV	-	Micro-volt
π_0	-	Proportion of Observable Signal
π_1	-	Proportion of Missing Signal

CHAPTER 1

INTRODUCTION

1.1 Background of the Study

Based on the definition of Lathi (2009), Meade and Dillon (1991), a signal can be viewed as a set of data or information in the form of measurable quantity. Most of the time, signals are the functions of dependent on time, such as biomedical signal, speech signal and econometrics signal. However, the definition of signal is not only restricted in this, signals are not necessary be the functions of time always. They can be function of space, like electrical charge distributed over a body or images over a two-dimensional surface. In general, biomedical signals carry information in several forms which reflect the nature and activities of physiological processes (Rangayyan, 2002). They can be hormones and neurotransmitters as biochemical form, potential and current as electrical form or pressure and temperature as physical form.

On the other hand, Lathi (2009) has defined that systems are the entities that may modify signals or extract additional information from the signals. Systems can be hardwares which made up of physical components. Electrical, mechanical and hydraulic systems are some common examples. Also, a system can be in term of software; an algorithm that computes the output signal with the given input signal. Signals are naturally come along with noises, the undesired components or disturbances. In biomedical signal, the noise can be physiological interferences, such as muscle contraction interference which usually are due to the body movement or external factors like electromagnetic signal from power cable and electrical stimuli. In reality, these factors are most likely unavoidable.

The blood circulatory system is a system of blood flow for humans and animals. It consists of three major parts, the heart, the blood and the blood vessels in organism. Humans are made of up many tiny cells, which every single cell need oxygen and nutrients to survive and work. Other than that, waste products from cellular activities like carbon dioxide will be transported away from the cells, then from body by blood. The main function of blood circulatory system is to provide continuous blood flow in the body and to ensure the blood reach each cell in the body (Houghton, 2007).

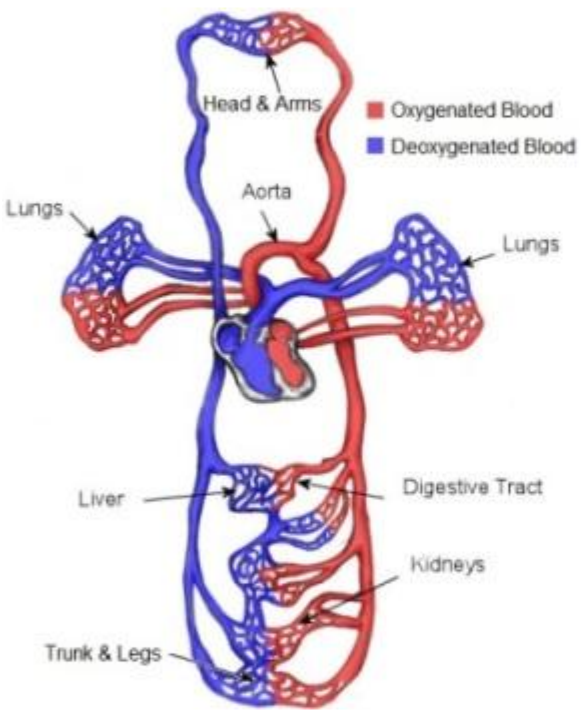

Figure 1.1 Diagram of Blood Circulatory System.

The heart is one of the important organs for most of the multicellular life forms. It contracts to pump the oxygenated blood along with the nutrients throughout the body, so that the basic requirements of cellular activities are fulfilled. Cardiac signal or electrocardiogram (ECG) signal generate as the heart contracts and relaxes, then record by electrocardiograph. It describes the electrical activities of the heart. Since every normal and complete cycle of ECG are coming from a heartbeat, the heart's activities can be observed from the ECG signal, and heart conditions such as diseases or abnormalities can be identified easily by analyze the shape of the ECG signal.

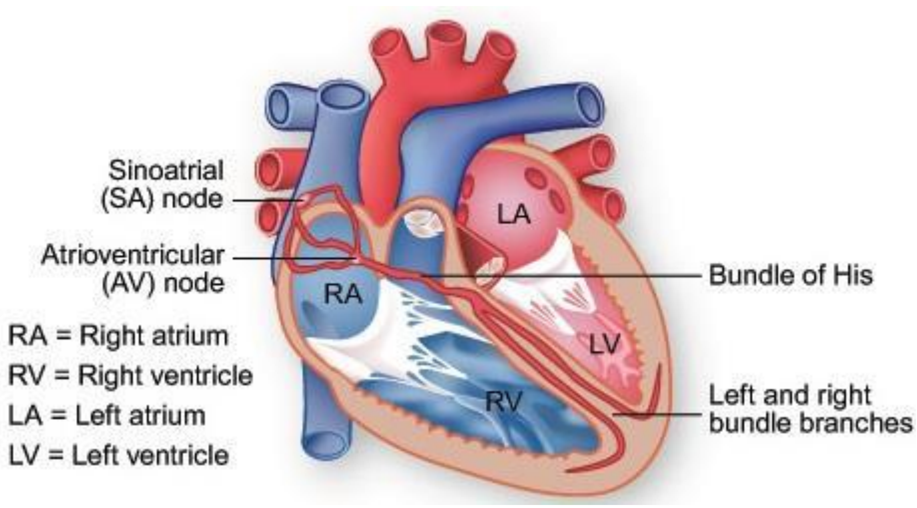

Figure 1.2 Simple illustration of the human heart structure.

Basically, human's heart divided to four major chambers, namely left atrium, right atrium, left ventricle and right ventricle. The atria collect blood from the other part of the body and pump the blood entered to the ventricles whereas the ventricles pump the blood away from heart. At the same time, the left parts of the heart contain the blood rich with oxygen which referred as oxygenated blood, while the right parts contain deoxygenated blood, the impure blood that has less oxygen amount and higher concentration in waste products (Katz, 2011).

In most cases, the ventricles have larger and thicker muscle wall compared to the atria and the left ventricle is around three times thicker than the left ventricle. The thicker muscle wall can contribute to higher pressure exertion. It is because atria just have to pump the blood to ventricles but the ventricles need to pump the blood to other parts of body through the blood vessels. Besides, there are valves prevent the blood flowing backward. The atrioventricular valves prevent blood flow form ventricles to atria and the semilunar valves ensure the blood flow away from ventricles to the vessels.

Statistical data analyses are important as it is widely applied in various fields. However, Watanabe and Yamaguchi (2004) mentioned that data collection methods are not always ideal. Sometimes, it makes the data collection rate lower than the initially expected one. Sörnmo and Laguna (2005) proposed, for some reasons, it appears that missing data problem do occur in signal processing as well. Thus, methods to deal with these missing values problems have been developed. Traditionally, the methods to overcome the problem are simply ignoring the existence of the missing data or substitute the missing data with the mean of the collected data. Nowadays, there are several algorithms to estimate the parameters or the values of the process for various forms of data incompleteness.

1.2 Problem Statement

Human bodies are made up by several organs and some of them are sources of biomedical signals. For example, endocrine glands release signals in chemical form called hormones, which are actually a few types of proteins. The heart is also one of the common examples, it generates the biomedical signal in the form of electrical potential. However, heart is not the only source of bioelectric potential, other organs such as brain, muscles, stomach and so on.

Physiological systems are dynamic, that is, they are interacting with each other, in various ways like feedback and collateral effects. Since organs other than the heart can release electrical signal, signals from some other organs act as noises that corrupt the ECG signals generated by the heart while the heart signal is the only interested signal. In addition, external interferences such as power cables of the ECG device also contribute in reducing the quality of the desired ECG signals.

Commonly, the magnitudes of biomedical signals are generated at a very low level, millivolt or even microvolt at their sources. Of course, ECG is one of the low level magnitude signals. Sensitive transducers and devices are required to record such signals. When the instruments are not capable to detect all the signals especially the low amplitude signals, missing signals will occur. As missing data are also possible in signal analysis, the raw signals obtained can be unreliable.

When both factors above are considered, it is obvious that biomedical researchers always have the problem in separating the noise and signal and retrieving the missing signals to obtain an accurate ECG signals for the purpose of biomedical diagnosis and treatment (Catalano, 2002). Without a good approach, analysis of an inaccurate signal might lead to a wrong conclusion. It must not be tolerated as this is a matter of life and death, a little inappropriate judgment or action might cause death.

In the fields of statistics, there are many choices of statistical software for the data analysis. In recent years, the developments of these software tools are rapid and have provided conveniences in analysis and process of data for anyone as long as they own a personal computer. Yet, these tools are only designed to handle complete data. Even though there are some statistical packages provide missing data processing but they can only conduct simple processing. As an illustration, the statistical software usually use listwise deletion, exclude the subjects with missing data or mean imputation, substitute the missing values with the sample mean to deal with incomplete data.

Based on the problem statements mentioned earlier, the following research questions can be formulated:-

1. What is the optimal method to deal with corrupted ECG signal with missing signals?
2. How to restore the ECG signals that are corrupted by noises and missing signals?

1.3 Objectives

From the problem statements above, the following are the objectives of the study:-

1. To identify an optimal algorithm to deal with missing ECG signal problem.
2. To evaluate the performance of imputations and exponential smoothing that deal with missing data problem in signal.
3. To estimate the de-noised ECG with missing data.

1.4 Significance of the Study

Through this study, the results are helpful in the development of biomedicine field, or more precisely, the biomedical signal analysis. A filter can be defined as a noise removal tool and its major function is to obtain the signal as true as possible. Analyzing signal filter is important because a good filter can remove a high proportion of noise that lies within the signal and give an accurate signal output without weakening the quality of the signal itself.

While dealing with noise, the study also tackles the missing signal problem. When the number of subject increases, the missing data is more probably to occur. The traditional methods like ignoring the missing data and mean imputation are not always a suitable approach to conduct analysis of signal with missingness as missing signals are not always occurred at random. Simply ignoring or guessing the signal values, might lead to a incorrect results and hence, a biased justification (Watanabe and Yamaguchi, 2004).

Since ECG is the graphical recorder of the electrical activities of heart over the period of time, it is important that the ECG have a good filter to reduce the noises and a suitable algorithm to deal with missing signals. An accurate cardiac signal is important to reveal many information of the heart, such as its rhythm, conduction abnormalities or enlargement of the chambers because the heart is an important organ for humans and all animals. Information of the heart, especially the heart of patients with heart diseases is significant for diagnosis, therapy and treatment control. Hence, the ECG should be able to extract correct information about one’s heart.

The study also can make contribution in the developments of statistic fields. People nowadays can easily access to statistical methods via the statistical tools, handle data with massive volume, carry out complex data analysis with ease, and process data in short instant. Watanabe and Yamaguchi (2004) stated, most of the advance algorithms that handle missing data are still not abundantly available even the swift enrichments in statistical packages. Usually, completeness of data is the requirement of these software tools. Well-understanding in a missing data processing algorithm can be a great help to include the algorithm in software currently available. Thus, findings of the study are important in the evolution of statistical software.

1.5 Scopes of the Study

The study mainly focuses on reducing noise from biomedical signal with missing data. In this study, the biomedical signal that was studied is the ECG signal. Apart from that, the study also covered the application of mean imputation and multiple imputations on missing ECG signal estimation. The ECG data was also modelled and estimated as seasonal time series, by Holt-Winters Exponential Smoothing. Then, comparisons were made on the estimation results and estimation performance.

The ECG data used in the study are adopted from the online database, http://physionet.org/physiobank/database/#ecg. The ECG data used was categorized in the MIT-BIH ECG Compression Test Database (cdb) section with data code 11442_01 in the online database. The ECG data taken from the database are complete, that is, there are no any missing data occurred and it is assumed has no noise, in simpler word, it is clean. The ECG last for 20.476 seconds and consists of 5120 reading values in total and measured in millivolt (mV). Then, the frequency of the ECG can be determined in Hertz (Hz) and it is obvious that the ECG sampled at the frequency of 250 Hz. Yet, the ECG signal data are not fully utilized, a sample of 2000 readings was drawn from 0:05.188 second to 0:13.184 second for the purpose of study. According to Sörnmo and Laguna (2005), the low values of the signals have higher tendency to missing. So, it can be concluded that, most of the ECG signals is probably missing as most of the sample is low value.

Since the study requires data modification, the ECG signal are the sampled with noises. The simulation of noises was made to corrupt the clean sample. A set of normally distributed random numbers were to be generated for each reading in the sample. The random numbers were taken as the Gaussian random noises for the available ECG data. This sample of noise was generated from the online random numbers generator, RANDOM.ORG, http://www.random.org/gaussian-distributions/. The series of noise, N was generated so that it is normally distributed with mean 0.0

and standard deviation 0.01. These random numbers were generated in 2000 terms at six significant figures. In the views of the biomedical signals, frequency of the noise in the study is 250 Hz, same as the frequency of the ECG signal sample. Other than that, the classification of missing data that was studied is the not missing at random (NMAR) case. Since the missingness mechanism is not random, the missingness mechanism will be generated. For simulation purposes, a set of four significant figures uniform distributed random numbers will be generated from RANDOM.ORG as well. The uniform random numbers generated as the missing signal threshold. The missing signal probabilities are set to $p_1 = 0.3$ and $p_2 = 0.1$ for low amplitude and high amplitude signals respectively. Then, the data from the sample will be eliminated according to the values of ECG signal and missingness indicator. The partial eliminated data set was assumed to be the incomplete signal with the missingness mechanism of not missing at random.

REFERENCES

Baken, R. J. and Olikoff, R. F (2000). *Clinical Measurement of Speech and Voice*. (2nd ed.). San Diego: Singular Publishing Group.

Bourne, R. (2010). *Fundamentals of Digital Imaging in Medicine*. London: Springer-Verlag.

Bowman, K. P. (2006). *An Introduction to Programming with IDL: Interactive Data Language*. London: Elsevier Inc.

Catalano, J. T. (2002). *Guide to ECG Analysis*. (2nd ed.). Philadelphia: Lippincott Williams & Wilkins.

Chaudhuri, S., Pawar, T. D. and Duttagupta, S. (2009). *Ambulation Analysis in Wearable ECG*. New York: Springer.

Etten, W. C. V. (2005). *Introduction to Random Signals and Noise*. Chichester: John Wiley & Sons Ltd.

Gentle, J. E. (2003). *Statistics and Computing: Random Number Generation and Monte Carlo Methods*. New York: Springer.

Guerrero, A. P., Mailhes, C. and Castanié, F. (2007). Lost Sample Recovering of ECG Signals in e-Health Applications. *Proceedings of the 29th Annual International Conference of the IEEE EMBS* Cité Internationale. Lyon, France. 23-26 August, pp. 31-34

Houghton, G. (2007). *Blood: The Circulatory System*. New York: The Rosen Publishing Group, Inc.

Huff, J. (2006). *ECG Workout: Exercises in Arrhythmia Interpretation*. (5th ed.). Ambler: Lippincott Williams & Wilkins.

Katz, A. M. (2011). *Physiology of the Heart*. (5th ed.). Philadelphia: Lippincott Williams & Wilkins.

Lathi, B. P. (2009). *Principles of Linear System and Signal*. (2nd ed.). New Delhi: Oxford University Press.

Longford, N. T. (2005). *Missing Data and Small-Area Estimation: Modern Analytical Equipment for the Survey Statistician*. New York: Springer.

Madisetti, V. K. (Ed.) (2010). *The Digital Signal Processing Handbook: Digital Signal Processing Fundamentals*. (2nd ed.). Boca Raton: CRC Press.

Meade, M. L. and Dillon, C. R. (1991). *Signals and Systems: Models and Behaviour*. (2nd ed.). Dordrecht: Kluwer Academic Publishers.

Pierce, D. A. (1978). Seasonal Adjustment when Both Deterministic and Stochastic Seasonality are Present. In Zellner, A. (Ed.) *Seasonal Analysis of Economic Time Series*. (pp. 242-269). Washington, D. C.: Bureau of the Census.

Rangayyan, R. M. (2002). *Biomedical Signal Analysis: A Case-Study Approach*. Piscataway: IEEE Press.

Reddy, T. A. (2011). *Applied Data Analysis and Modeling for Energy Engineers and Scientists*. New York: Springer.

Ross, S. M. (2009). *Probability and Statistics for Engineers and Scientist*. (4th ed.). London: Elsevier Academic Press.

Rumsey, D (2006). *Probability for Dummies*. Indianapolis: Wiley Publishing, Inc.

Sherwood, L. (2013). *Human Physiology: From Cells to Systems*. (8th ed.). Belmont: Cengage Learning.

Smith, S. W. (2003). *Digital Signal Processing: A Practical Guide for Engineers and Scientists*. Burlington: Elsevier Science.

Sörnmo, L. and Laguna, P. (2005). *Bioelectrical Signal Processing in Cardiac and Neurological Applications*. San Diego: Elsevier Academic Press.

Telford, W. M., Geldart, L. P., Sheriff, R. E. (1990). *Applied Geophysics*. (2nd ed.). New York: Cambridge University Press.

Watanabe, M. and Yamaguchi, K. (Eds.) (2004). *The EM Algorithm and Related Statistical Models*. Basel: Marcel Dekker, Inc.

Warner, R. M. (1998). *Spectral Analysis of Time-Series Data*. New York: The Guilford Press.

Whitfield, J. (1995). *Electrical Craft Principle Volume 1*. (4th ed.). London: The Institution of Electrical Engineering.

Vaseghi, S. V. (2000). *Advanced Digital Signal Processing and Noise Reduction* (2nd ed.). Chichester: John Wiley & Sons Ltd.

Yates, R. D. and Goodman, D. J. (2005). *Probability and Stochastic Processes: A Friendly Introduction for Electrical and Computer Engineers* (2nd ed.). Danvers: John Wiley & Sons, Inc.