Recent Advances in Biomedical Engineering

Recent Advances in Biomedical Engineering

Edited by
Dr Ganesh R Naik

In-Tech
intechweb.org

Published by In-Teh

In-Teh
Olajnica 19/2, 32000 Vukovar, Croatia

www.intechweb.org
Additional copies can be obtained from:
publication@intechweb.org

First published October 2009
Printed in India

Technical Editor: Zeljko Debeljuh

Recent Advances in Biomedical Engineering,
Edited by Dr Ganesh R Naik

p. cm.
ISBN 978-953-307-004-9

Preface

Background and Motivation

The field of biomedical engineering has expanded markedly in the past ten years. This growth is supported by advances in biological science, which have created new opportunities for development of tools for diagnosis and therapy for human disease. The discipline focuses both on development of new biomaterials, analytical methodologies and on the application of concepts drawn from engineering, computing, mathematics, chemical and physical sciences to advance biomedical knowledge while improving the effectiveness and delivery of clinical medicine.

Biomedical engineering now encompasses a range of fields of specialization including bioinstrumentation, bioimaging, biomechanics, biomaterials, and biomolecular engineering. Biomedical engineering covers recent advances in the growing field of biomedical technology, instrumentation, and administration. Contributions focus on theoretical and practical problems associated with the development of medical technology; the introduction of new engineering methods into public health; hospitals and patient care; the improvement of diagnosis and therapy; and biomedical information storage and retrieval.

Much of the work in biomedical engineering consists of research and development, spanning a broad array of subfields. Prominent biomedical engineering applications include the development of biocompatible prostheses, various diagnostic and therapeutic medical devices ranging from clinical equipment to micro-implants, common imaging equipment such as MRIs and EEGs, biotechnologies such as regenerative tissue growth, and pharmaceutical drugs and biopharmaceuticals.

Processing of biomedical signals, until a few years ago, was mainly directed toward filtering for removal of noise and power line interference; spectral analysis to understand the frequency characteristics of signals; and modeling for feature representation and parameterization. Recent trends have been towards quantitative or objective analysis of physiological systems and phenomena via signal analysis. The field of biomedical signal analysis has advanced to the stage of practical application of signal processing and pattern analysis techniques for efficient and improved noninvasive diagnosis, online monitoring of critically ill patients, and rehabilitation and sensory aids for the handicapped. Techniques developed by engineers are gaining wider acceptance by practicing clinicians, and the role of engineering in diagnosis and treatment is gaining much deserved respect.

The major strength in the application of computers in biomedical signal analysis lies in the potential use of signal processing and modeling techniques for quantitative or objective

analysis. Analysis of signals by human observers is almost always accompanied by perceptual limitations, interpersonal variations, errors caused by fatigue, errors caused by the very low rate of incidence of a certain sign of abnormality, environmental distractions, and so on. The interpretation of a signal by an expert bears the weight of the experience and expertise of the analyst; however, such analysis is almost always subjective. Computer analysis of biomedical signals, if performed with the appropriate logic, has the potential to add objective strength to the interpretation of the expert. It thus becomes possible to improve the diagnostic confidence or accuracy of even an expert with many years of experience.

Developing an algorithm for biomedical signal analysis, however, is not an easy task; quite often, it might not even be a straightforward process. The engineer or computer analyst is often bewildered by the variability of features in biomedical signals and systems, which is far higher than that encountered in physical systems or observations. Benign diseases often mimic the features of malignant diseases; malignancies may exhibit a characteristic pattern, which, however, is not always guaranteed to appear. Handling all of the possibilities and degrees of freedom in a biomedical system is a major challenge in most applications. Techniques proven to work well with a certain system or set of signals may not work in another seemingly similar situation. This book intends to provide an insight into the above mentioned applications.

Intended Readership

The book is directed at engineering students in their final year of undergraduate studies or in their graduate studies. Most undergraduate students majoring in biomedical engineering are faced with a decision, early in their program of study, regarding the field in which they would like to specialize. Each chosen specialty has a specific set of course requirements and is supplemented by wise selection of elective and supporting coursework. Also, many young students of biomedical engineering use independent research projects as a source of inspiration and preparation but have difficulty identifying research areas that are right for them. Therefore, a second goal of this book is to link knowledge of basic science and engineering to fields of specialization and current research.

Practicing engineers, computer scientists, information technologists, medical physicists, and data processing specialists working in diverse areas such as medical, bio signals, biomedical applications, and hospital information systems may find the book useful in their quest to learn advanced techniques for signal analysis. They could draw inspiration from other applications of signal processing or analysis, and satisfy their curiosity regarding computer applications in medicine and computer aided medical diagnosis.

The book is partly a textbook and partly a monograph. It is a textbook because it gives a detailed introduction to Bio medical engineering techniques and applications. It is simultaneously a monograph because it presents several new results and ideas and further developments and explanation of existing algorithms which are brought together and published in the book for the first time. Furthermore, the research results previously scattered in many scientific journals and conference papers worldwide, are methodically collected and presented in the book in a unified form. As a result of its twofold character the book is likely to be of interest to graduate and postgraduate students, engineers and scientists working in the field of biomedical engineering, communications, electronics, computer science, optimization, and neural networks. Furthermore, the book may also be of interest to researchers working in

different areas of science, because a number of results and concepts have been included which may be advantageous for their further research. One can read this book through sequentially but it is not necessary since each chapter is essentially self-contained, with as few cross references as possible. So, browsing is encouraged.

The editor would like to thank the authors, who have committed so much effort to the publication of this work.

Dr Ganesh R Naik
RMIT University,
Melbourne, Australia
ganesh.naik@rmit.edu.au

Contents

1

Micro Macro Neural Network to Recognize Slow Movement: EMG based Accurate and Quick Rollover Recognition

Takeshi Ando, Jun Okamoto and Masakatsu G. Fujie
Faculty of Science and Engineering, Waseda University
Japan

1. Introduction

The wearable robots to support many kinds of movements have been developed for the elder and disabled people all over the world (Hayashi et al., 2005, Furusho et al., 2007, Kawamura et al., 1997), because we are facing the elder dominated society. A surface ElectroMyoGram (EMG) signal, which is measured a little before the start of the movement, is expected as the trigger signal of movement support.

We have been also developing an EMG controlled intelligent trunk corset, shown in Fig. 1, to support rollover movement, since it is one of the most important activities of daily living (ADL). Especially, the rollover movement of bone cancer metastasis patients is focused as the target movement. The bone cancer metastasis patients feel sever pain when they conduct the rollover movement. The core of the intelligent trunk corset system is a pneumatic rubber muscle that is operated by the EMG signals from the trunk muscle. As shown in Fig. 2, in our study, we first analyzed the EMG signal (Ando et al., 2007) that is used as the input signal for the intelligent corset to recognize a rollover movement. Second, we proposed an original neural network algorithm to recognize the rollover quickly and with high accuracy (Ando et al., 2008a). Finally, we developed the mechanisms of the intelligent corset to assist rollover movement using the pneumatic rubber actuator (Ando et al., 2008b).

In this chapter, the proposed original neural network, called the Micro-Macro Neural Network (MMNN), is introduced. In addition, the methodology to determine the optimal structure of the MMNN to recognize the rollover movement is established. This paper is organized as follows; Section 2 summarizes the related neural network to recognize some movements based on the EMG signal. Section 3 discusses the traditional neural network known as Time Delay Neural Network (TDNN) and MMNN structures, Section 4 establish the methodology to determine the optimal structure of MMNN, and the rollover recognition result using the optimal MMNN is compared with that using traditional TDNN. Section 5 presents a summary and future work.

2. Related neural network to recognize movement using EMG signal

Since the recognition of rollover is based on noisy and complex EMG signals, a highly robust system that is unaffected by the possible misalignment of electrodes, individual differences, or surrounding electrical conditions is necessary to recognize EMG signals accurately. A Neural Network (NN) is one of the learning machines that use EMG signals to recognize movement (Kuribayashi et al., 1992, Fukuda et al., 1999, Wang et al., 2002, Kiguchi et al., 2003, Hou et al., 2004, Zecca et al., 2002). NN is capable of nonlinear mapping, generalization, and adaptive learning. There are generally two kinds of NN that recognize a time series-signal. One is the Time Delay Neural Network (TDNN) (Waibel, 1989), in which a delay is introduced in the network and past data (the data collected before the current measurement point) is set as the input signal of the network. The other is the Recurrent Neural Network (RNN) (Kelly et al., 1990, Tsuji et al., 1999), which uses feedback from the output signal of the output layer as the input signal of the input layer. To avoid needless time-stretch properties and to reduce calculation amounts and costs, we selected TDNN as the base neural network for the work reported here.

Many researchers have used TDNN to recognize movements from EMG signals. For example, Hincapie et al. (Hincapie et al., 2004) estimated the movement of the affected side of a patient by using EMG data of the unaffected side in their development of a prosthetic upper limb. Hirakawa et al. (Hirakawa et al., 1989) and Farry et al. (Farry et al., 1996) recognized movement using frequency domain information of the EMG signal. Huang et al. (Huang et al., 1999) proposed the feature vector, composed of an integrated EMG, Zero Crossing and variance, to recognize eight-finger movement. Finally, Nishikawa et al. (Nishizawa et al., 1999) recognized ten kinds of movements using a Gabor-transformed EMG signal.

Fig. 1. Intelligent trunk corset to support rollover movement

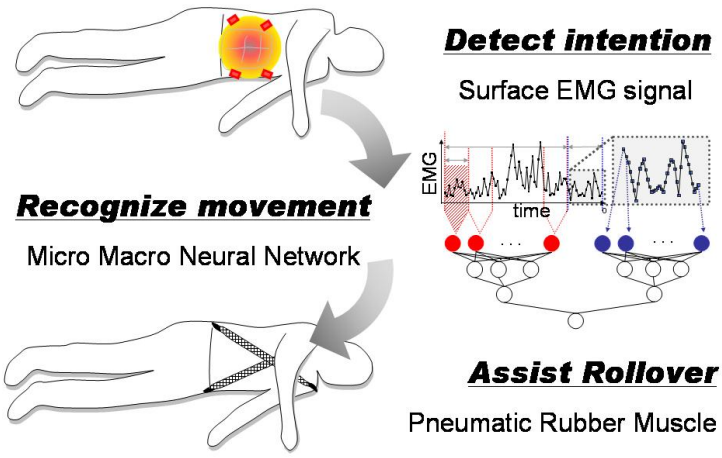


Fig. 2. Concept of the intelligent corset using EMG signal, original neural network and pneumatic actuator.

However, all of these related research efforts share two common problems, which are slow response time and incorrect recognition of the movement. Consequently, we previously proposed the original algorithm called the Micro Macro Neural Network (MMNN), composed of the Micro Part, which detects a rapid change in the strength of the EMG signal, and the Macro Part, which detects the tendency of the EMG signal toward a continuing increase or continuing decrease, to improve the response time and accurate recognition of the rollover movement based on the EMG signal as input. However, the methodology to design or optimize the structure of the MMNN is not established, because there are many parameters to determine the structure of the MMNN.

3. Micro - Macro Neural Network (MMNN)

3.1 Traditional Time Delay Neural Network

For the learning machine in this research, we selected the three-layer feed-forward type of Time Delay Neural Network (TDNN) as the structure of the network and the back propagation (BP) method with a momentum term as the leaning algorithm, which is a standard neural network to recognize time-series signals. In addition, we selected the sequential adjustment method to modify the weight and threshold of each unit. The relations between each pair of units in the TDNN are shown in (1), (2), and (3).

$$net_i^m = \sum_{j=1}^{n_{m-1}} \omega_{ij}^m x_j^{m-1} + \theta_i^m \tag{1}$$

$$x_i^m = f(net_i^m) \tag{2}$$

$$f(net) = 1/(1+\exp(-u_0 net)) \tag{3}$$

where m = 2 and 3, i = 1,…,n_m, n_m is the number of the m^{th} layer unit, $\omega^m{}_{ij}$ is the weight between the $(m\text{-}1)^{th}$ layer's i^{th} unit and the m^{th} layer's j^{th} unit, $x^m{}_i$ i is the output of the m^{th} layer's i^{th} unit, $\theta^m{}_i$ is the threshold in the m^{th} layer's i^{th} unit, and u_0 is the constant to decide the gradient of the sigmoid function.

In this study, the number of input layer units was typically 75, and, that is, the input of the input layer was EMG signals, *semg (t-i)* (i=0,1,...,74). In other words, the time it took the TDNN system to recognize the rollover movement from the inputted EMG data was 0.075 (msec) (Zecca et al., 2002).

3.2 Concept of Micro-Macro Neural Network

Using TDNN, previous researchers focused on upper limb movement, which is a relatively fast movement. Since the movement takes only a short time, less time-series EMG data is inputted into the system. The advantage of this short data length is that there are fewer calculations to be done and, therefore, less cost; the disadvantage is that less input data means more false recognitions.

We focused on the rollover movement, which is a relatively slow movement. Since the movement takes a relatively long time, it is possible to have more time-series EMG data inputted into the system.

We checked the impact of past time-series EMG data using TDNN on the recognition result. The structure of TDNN was as follows: the number of input layer units was 1700, the number of hidden layer units was 850, and the number of output layer units was 1. We determined the number of input units as 1700 based on our EMG experiment (Ando et al., 2007), which showed that the shortest time spent on rollover was 1.7 (sec) without taking into account the time for any previous rollover movement.

To check the importance of each unit in TDNN, the contribution rate of the weight of each input unit was calculated by (4).

$$R_{contribution}(i) = \frac{\sum_{j=1}^{N/2} \left| \omega_{ij}^{m} \right|}{\sum_{i=1}^{N} \sum_{j=1}^{N/2} \left| \omega_{ij}^{m} \right|} \times 100 \tag{4}$$

where $R_{contribution}(i)$ is the contribution rate of the weight of input unit i, whose data is the EMG data of i (msec) before the current measured point, N = 1700, m = 2.

As a result, it was found that the weights of units in the range of -1 to -10 (msec) were higher than those of the other units in TDNN (See Fig. 3). It is natural that the EMG data nearest to the time of measurement has a large impact on the recognition result. However, it is worth noting that the contribution rate of the inputted EMG data before -10 (msec) is almost constant. Even though the importance of data from 10-75 (msec) before is the same as that of data from 76-1700 (msec) before, the latter data was not used to recognize the rollover movement in the traditional TDNN (See Section 3.1). Therefore, in the traditional TDNN,

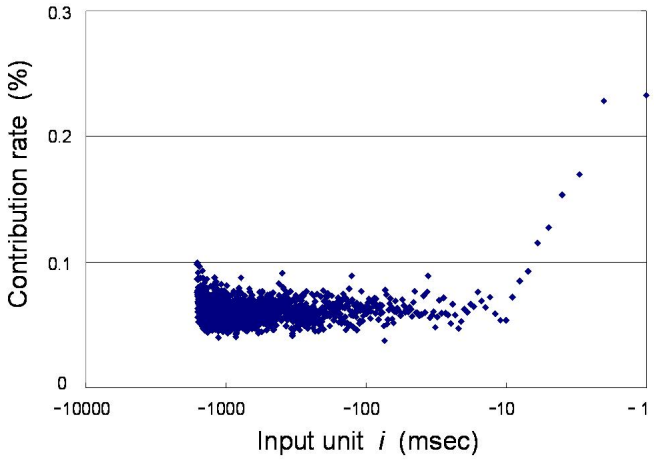


Fig. 3. Contribution rate as a function of input unit

whose input unit number was 75 (msec), a later response and a higher incidence of false recognition were evident.
When long past time-series EMG data is used in TDNN, the advantage of this long data length is that more input data means faster response and less false recognition. The disadvantage is the large amount of calculations and its cost.
In the proposed Micro Macro Neural Network (MMNN), some of the long past time-series data was compressed. Therefore, the amount and cost of the calculations do not increase.
The basic concept of the Micro Macro Neural Network (MMNN) is to use the long past time-series EMG data to discriminate the movement accurately and quickly without increasing the calculation cost by compressing some of the long past data.

3.3 Structure of the Micro-Macro Neural Network

Basically, we upgraded the traditional TDNN to MMNN (Fig. 4). The most important feature of MMNN is that it can handle an increased amount of input data to the neural network without increasing the number of calculations. Traditional TDNN is defined in our network as the Micro Part. The input data, ${}_{micro}x_n^1$ in the Micro Part is defined as following;

$$ {}_{micro}x_n^1 = semg\ (t - n + 1) \tag{5}$$

where n = 1,2,....,N_{micro}, and N_{micro} is the number of input unit in Micro part

As can be seen in Fig. 5, the data for $-T_{micro} < t < 0$ is the Micro Part, and the data for $-(T_{macro} + T_{micro}) < t < -T_{micro}$ is the Macro Part. In addition, the input data, ${}_{macro}x_n^1$ in the Macro Part is divided into several T_{ARV} (msec), and the average rectified value (ARV) of the EMG signal among the T_{ARV} values, calculated by (6), is defined as the input value of the Macro Part.

$$ {}_{macro}x_n^1 = \frac{\sum_{i=t-(n-1)T_{ARV}}^{t-nT_{ARV}} |semg(i)|}{T_{ARV}} \tag{6} $$

where n = 1,2,....,N_{macro}
Therefore, the number of input units of the Macro Part is expressed by the following equation:

$$ N_{macro} = T_{macro} / T_{ARV} \tag{7} $$

where N_{macro} is the number of input units of the Macro Part.
The relations between each pair of units in both the Macro Part and the Micro Part are shown in (1), (2), and (3) above.
The output data of the Micro part and Macro part is defined as the input data of the Integrated Layer. In the Integrated layer, the output signal is calculated using also (1), (2) and (3).

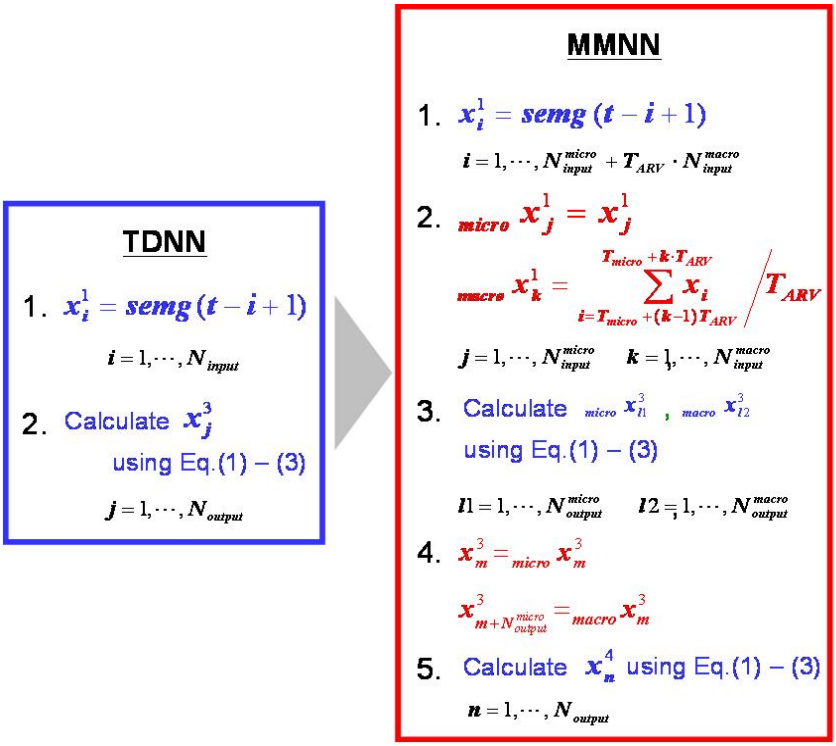


Fig. 4. Development of MMNN algorithm from TDNN algorithm

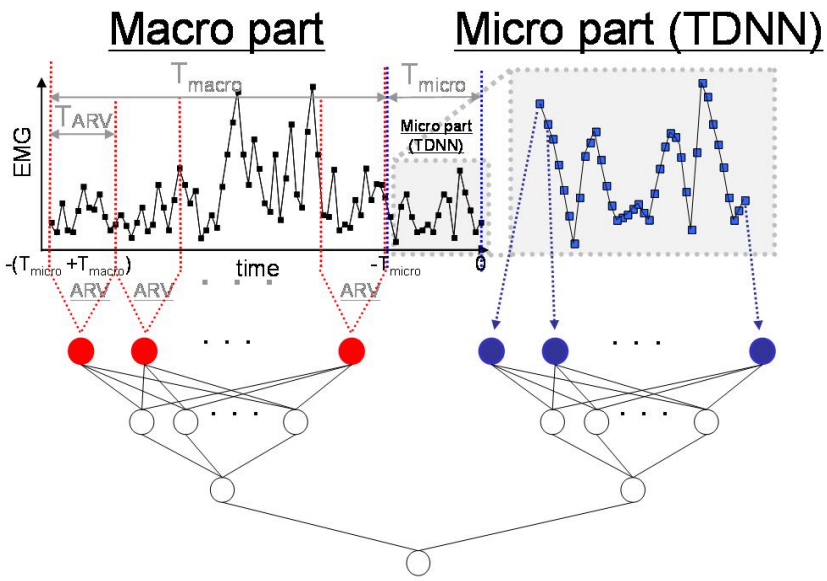


Fig. 5. Micro Macro neural network. Note that MMNN is divided into the Micro Part and the Macro Part. The Micro Part is TDNN using the data for T_{micro} as the input signal. The input data of the Macro Part uses the data for T_{macro}, which is the ARV of the EMG signal among all T_{ARV} values.

4. Optimal structure of proposed MMNN and rollover movement recognition

4.1 Objective

The structure of the MMNN is complex, because many parameters determine the structure of the MMNN. In this section, based on the contribution rate shown in Fig. 3 and an experiment about rollover recognition using MMNN, the optimal parameters in MMNN are determined.

4.2 Methodology of rollover recognition experiment

We defined the rollover movement as a continuous movement involving a deliberate change of posture from a supine position to a lateral or prone position. In this research, rollover movements were performed thirty times in advance by each of three young, healthy male subjects. EMG signals obtained from the internal oblique (IO) muscle were selected as the input signals based on our previous study (Ando et al., 2007). The EMG signals were sampled at a rate of 1000 (Hz), rectified with a second-order, low-pass filter with a cut-off frequency of 20 (Hz), and normalized by the 100% maximal voluntary contraction (MVC) method (Zaman et al., 2005, Kumar et al., 1989), which shows the ratio of muscle activity in the MVC of the IO muscle to the measured EMG signal (Helen et al., 2002).

As the learning data for every rollover type, 20% of the data (18 out of 90 rollovers – 30 for each of the three subjects) was randomly selected (Kuribayashi et al., 1992, Fukuda et al., 1999). The other 80% of the data was used as test data. Because the numbers of learning and

test data were small, the *k*-fold cross validation estimation (k = 5) was used to prevent degradation of the accuracy based on the selection of learning data.
The time required to recognize the rollover was measured using TDNN. Furthermore, by synchronizing the EMG data with the data of a 3D motion-capture system, VICON612 (sampling frequency; 100 (Hz) and measurement accuracy; 1 (mm)) , the start of rollover movement was recognized.

4.3 Evaluation index

The recognition results of the test data were evaluated according to the response by the indexes presented below.
(1) The response time, $t_{response}$, is the time from the start of the rollover movement to the recognition of the rollover movement by the neural network .

$$t_{response} = t_{recognition} - t_{movement} \tag{8}$$

where $t_{recognition}$ is the time when the rollover is recognized, and $t_{movement}$ is the time when the rollover starts.

(2) Movement recognition rate before starting movement, P_{start}

$$P_{start} = N_{before} / N_{total} \tag{9}$$

where P_{start} is the ratio of N_{before}, the number of times rollover was recognized before the movement started to N_{total}, the total number of rollover movements.

(3) Number of false recognition rate, N_{false}
N_{false} is the number of times when false recognition occurred, that is, the times that NN recognized a rollover movement even though no rollover was actually conducted.

4.4 Structure of TDNN and recognition result

As stated above, for the learning machine in this research, we selected the three-layer feed-forward type NN and the back propagation method with momentum term, which is a standard neural network for recognizing time-series signals. The number of input layer units was 75. The unit numbers of the hidden layer and the output layer were 38 and 1, respectively.
As shown in Fig. 6 (b), when TDNN was used, the recognition results were as follows: $t_{response}$ was -25 (S.D. 59) (msec), P_{start} was 38% (138 out of 360 trials), and N_{false} was 151 out of 360 trials.

4.5 Optimal structure of MMNN and recognition result

The structure of MMNN was resolved based on many parameters.
First, in the Micro Part, which is the traditional TDNN, the number of input layer units was fixed at 10 (T_{micro} = 10 (msec) in Fig. 4), because the contribution rates in -1 ~ -10 (msec) are higher than those at other input times, as shown in Fig. 2. The number of hidden layer units

was fixed at 5, and the number of output layer units was fixed at 1. The number of hidden layer units was determined based on the "rule of thumb" as follow;

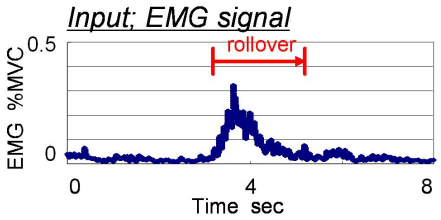
$$N_{hidden} = (N_{input} + N_{output}) / 2 \tag{10}$$

where N_{hidden}, N_{input}, and N_{output} are the numbers of hidden layer, input layer and output layer units.

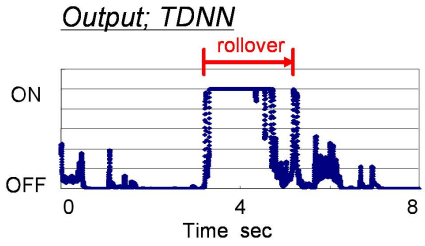
Second, the optimal structure of the Macro Part was determined as follows. The value of T_{ARV} was changed from 5 to 100 (T_{ARV} = 5, 10, 15,…., 100), and the value of N_{macro}, the number of input layer units, was changed from 5 to 70 (N_{macro} = 5, 10,15,…., 70). Additionally, according to the rule of thumb, the number of hidden layer units was set at N_{macro} /2 (if N_{macro} was even) or (N_{macro} +1)/2 (if N_{macro} was odd). Based on our EMG experiment (Ando et al., 2007), which showed that the shortest time spent on rollover was 1.7 (msec), we applied (11) when we calculated the response time for each rollover movement using MMNN, without taking into account the time for any previous rollover movement.

$$\begin{aligned} T_{ARV} \times N_{macro} &\le 1700 - N_{micro} \\ &= 1700 - 10 = 1690 \end{aligned} \tag{11}$$

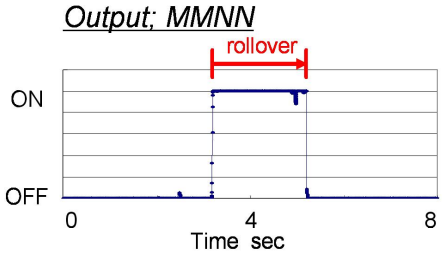
We obtained the best results for response with changing values of T_{ARV} and N_{macro} when T_{ARV} = 40 (sec) and N_{macro} = 40. With these conditions, the average $t_{response}$ for MMNN was -65 (S.D. 55) (msec). The average $t_{response}$ for TDNN was -25 (S.D. 59) (msec). Negative values mean the rollover was recognized before the movement started. Therefore, the recognition time of MMNN was 40 (S.D. 49) (msec) faster that that of TDNN.
Furthermore, as shown in Table 1, the P_{start} was 86% (310 out of 360 times), and N_{false} was only 50 in 360 trials.
Figure 6 shows an example of MMNN (T_{ARV} = 40 (msec), N_{macro} = 40). When the results of TDNN in Fig. 6(b) and the MMNN in Fig. 6(c) are compared, the following observations are clear: TDNN registers a false recognition four times, and, most importantly, the response speed in recognizing rollover is faster, steadier, and more accurate when MMNN is used than when TDNN is used.



(a) Input signal to NN



(b) Output signal from the TDNN



(c) Output signal from the MMNN

Fig. 6. Comparison between recognition of rollover by TDNN and MMNN. Note that TDNN fails to recognize the rollover at 0-2 (msec) and 5-7 (msec); however, it does recognize the rollover after the movement starts. In contrast, MMNN recognizes the rollover before the movement starts. EMG signal data is included for reference.

Neural Network	P_{start} %	N_{false}
TDNN	38	51/360
MMNN	86	150/360

Table 1. P_{start} and N_{false} of TDNN and MMNN

4.6 Discussion

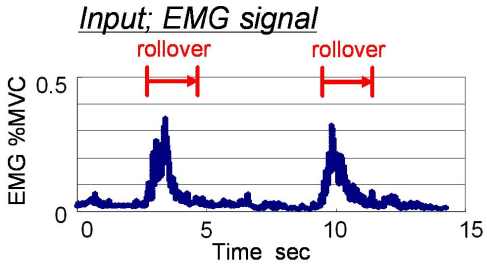
In Section 4.5, the effectiveness of MMNN in recognizing the rollover is shown in comparison with the effectiveness of TDNN.

In this section, the output signal of not only optimal MMNN but also the output of the Micro part and Macro part in the recognition of the rollover movement is discussed to show the characteristics of MMNN. In other words, first, the number of input layer units in TDNN was 10, and the input of the input layer was defined to show the characteristics of the Micro part as EMG signals, *semg (t-i)* (*i*=0,1,..., 9). Second, the number of input layer units in TDNN

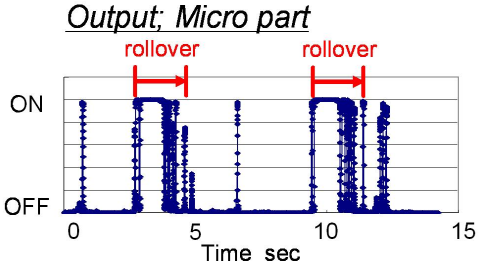
was 40, and the input of the input layer was defined to show the characteristics of the Macro part as the average reflected values among 40 (msec).
Table 2 shows the result of recognition time and the number of the false recognition using the TDNN (Input: 75 (msec) raw EMG signal), MMNN (Input: 10 (msec) raw EMG and 40 ARV EMG among 40 (msec)), only Micro Part (Input: 10 (msec) raw EMG) in MMNN and only Macro Part (Input: 40 ARV EMG among 40 (msec)) in MMNN. The response times using only the Micro part and the Macro part were $t_{response}$ = -50 (S.D. 26) (msec) and $t_{response}$ = 1 (S.D. 55) (msec). The number of false recognitions using only the Micro part and only the Macro part was N_{false} = 210 (in 360 times) and N_{false} = 56 (in 360 times).
When the input data was short past time-series data, the response time was short and the stability of the recognition was low. However, when the input data was the ARV of 40 (msec), the response time became longer and the stability increased.
The response time $t_{respons}$ did not show a significant difference ($p < 0.05$) between the optimized MMNN and TDNN using 10 (msec) time-series data as did the input data, that is, when using the only Micro part in MMNN. In addition, the number of false recognitions was almost the same when the number in the optimized MMNN was compared with that in TDNN using the ARV of 40 (msec), that is, when using the only Macro part in MMNN. Therefore, the advantages of quick response in the Micro part (See Fig. 7 (b)), and the stable recognition of the Macro part (See Fig. 7 (c)), are combined in the developed optimal MMNN. As a result, the MMNN is an NN that features quick response and little false recognition (See Fig. 7 (d)).

Neural Network	$t_{response}$ *msec*	N_{false}
TDNN (Input: 75 (msec) raw EMG)	-25 (S.D. 59)	150/360
MMNN (Input: 10 (msec) raw EMG and 40 ARV EMG among 40 (msec))	-65 (S.D. 55)	51/360
Only Micro part in MMNN (Input: 10 (msec) raw EMG)	-50 (S.D. 26)	210/360
Only Macro part in MMNN (Input: 40 ARV EMG among 40 (msec))	1 (S.D. 55)	56/360

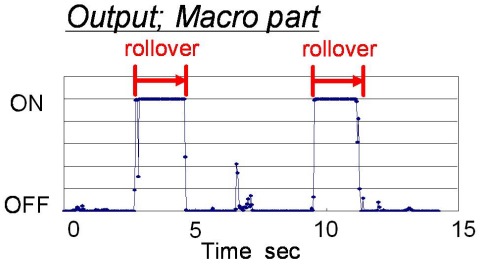
Table 2. Features of TDNN, MMNN, and Micro and Macro parts of MMNN



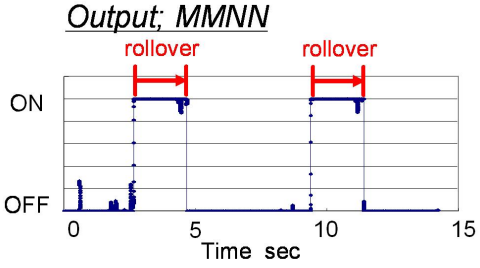
(a) Input signal to the NN



(b) Output from the Micro part in MMNN



(c) Output from the Macro part in MMNN



(d) Output from the MMNN

Fig. 7. Comparison with the output of TDNN, Micro part in MMNN, Macro Part in MMNN and MMNN. Note that the EMG signal is included for reference as (a). (b) is the output of Micro part and shows the quick and unstable response. (c) is the output of Macro part and shows the slow and stable response and (d) is output of MMNN and show quick and stable response.

5. Summary and future work

We have been studying patients with cancer bone metastasis who have a very short time to live. Specifically, we have developed the EMG controlled intelligent corset to support the rollover movement.

In this paper, we described an original neural network that we developed, called the Micro Macro Neural Network (MMNN), for the purpose of recognizing and responding to the rollover movement based on inputted EMG signals.

First, the structure of the MMNN was optimized with N_{micro} = 10 in the Micro part and N_{macro} = 40 and T_{ARV} = 40 in the Macro part, and then the response and accuracy of the MMNN were analyzed. After that, the response and accuracy of the optimized MMNN in recognizing the rollover movement were compared with those of the traditional TDNN. Test results showed that recognition in MMNN was 40 (S.D. 49) (msec), which is quicker than the recognition in TDNN. Additionally, the number of false recognitions in MMNN was only one third of those in TDNN. Hence, we can verify that our MMNN is effective and useful in recognizing rollover based on inputted EMG signals, which are noisy and vary considerably from individual to individual. In addition, by comparing the recognition results of only the Micro part and only the Macro part, we found that the advantages of quick response in the Micro part and stable recognition in the Macro part are features of MMNN.

In the future, we will incorporate MMNN into our rollover support system that uses pneumatic rubber muscles, and then we will test the effectiveness of the total system in clinical tests with cancer patients in terminal care.

6. Acknowledgement

This work was supported in part by the Global Center of Excellence Program, "Global Robot Academia", Waseda University, Tokyo, Japan; and Grant-in-Aid for Scientific Research (A) (20240058) and Grant-in-Aid for Scientific Research (20700415), Japan. In addition, this work was advised by Dr. Takahashi (M.D.), Shizuoka Cancer Center.

7. References

Ando Takeshi; Okamoto Jun & Masakatsu G. Fujie (2007). The development of roll-over support system with EMG control for bone metastasis patients, *Proceedings of 2007 IEEE International Conference on Robotics and Automation*, pp. 1244 -1249, Roma, Italy, April 2007.

Ando Takeshi;Okamoto Jun & Masakatsu G. Fujie (2008a). The development of roll-over support system with EMG control for bone metastasis patients, *Proceedings of 30th Annual International IEEE EMBS Conference*, 5228-5233, Vancouver, Canada, August 2008.

Ando Takeshi; Okamoto Jun & Masakatsu G. Fujie (2008b). "Intelligent corset to support rollover of cancer bone metastasis patients", *Proceedings of the 2008 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp.723-728, Nice, France, September 2008.

K. A. Farry; I. D. Walker & R. G. Baraniuk (1996). Myoelectric Teleoperarion of a Complex Robotic Hand, *IEEE Transactions on Robotics and Automation*, Vol.12, No.5, pp.775-787, 1996, 1042-296X.

Fukuda, O; Tsuji, T.; Shigeyoshi, H. & Kaneko, M. (1999). "An EMG controlled human supporting robot using neural network", *Proceedings of 1999 IEEE/RSJ International Conference on Intelligent Robots and Systems*, vol.3, pp.1586 - 1591, 0-7803-5184-3, Kyongju, South Korea, Octorber 1999.

Junji Furusho; Chengqiu Li; Shouji Morimoto; Miwa Tokuda; Takehito Kikuchi & Yasunori Hashimoto (2007). Development of Shear-type MR Brakes and their Application to Ankle-Foot Orthoses, *Proceedings of 2007 IEEE/ICME International Conference on Complex Medical Engineering*, 1283-1287, Beijing, China, June 2007.

Hayashi, T.; Kawamoto, H.; Sankai, Y.(2005). Control method of robot suit HAL working as operator's muscle using biological and dynamical information, *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems*, 3063 – 3068, 0-7803-8912-3, Edmonton, Canada, August 2005.

Juan Gabriel Hincapie; Dimitra Blana; Edward Chadwick & Robert F Kirsch (2004). Adaptive Neural Network Controller for an Upper Extremity Neuroprosthesis, *Proceedings of the 26th Annual International Conference of the IEEE EMBS*, pp.4133-4136, San Francisco, USA, September 2004.

A. Hirakawa; K. Shimohara; Y. Tokunaga (1989). EMG pattern analysis and classification by neural network", *Proc. of IEEE International Conference on Syst., Man and Cybern.*, pp.1113-1115, Cambridge, USA, November 1989.

Helen J. Hislop & Jacqueline Montgomery (2002). *Daniels and Worthingham's Muscle Testing: Techniques of Manual Examination*, W B Saunders Co; 7th, 9780721692999, pp.51-55, 2002.

Yanfeng Hou; Zurada, J.M. & Karwowski W. (2004). Prediction of EMG signals of trunk muscles in manual lifting using a neural network model, *Proceedings of 2004 IEEE International Joint Conference on Neural Networks*, Vol. 3, pp.1935 – 1940, Budapest, Hungary, July 2004.

H. P. Huang & C. Y. Chen (1999). Development of a Myoelectric Discrimination System for a Multi-Degree Prosthetic Hand, *Proceedings of the 1999 IEEE International Conference on Robotics and Automation*, pp. 2392-2397, Detroit, U.S.A., May 1999.

Kawamura S.; Hayakawa Y.; Tamai M. & Shimizu T. (1997). A design of motion-support robots for human arms using hexahedron rubber actuators., *Proc. of the 1997 IEEE/RSJ International Conference on Intelligent Robots and Systems*, Vol. 3, 1520 – 1526, Grenoble, France, Octorber 1997.

M. F. Kelly; P. A. Parker & R. N. Scott (1990). The Application of Neural Networks to Myoelectric Signal Analysis, A Preliminary Study, *IEEE Transctions on Biomedical Engineering*, Vol. 37, No. 3, pp.221-230, 1990, 0018-9294.

K. Kiguchi; K. Iwami; M. Yasuda; K. Watanabe & T. Fukuda (2003). An exoskeletal robot for human shoulder joint motion assist, *IEEE/ASME Transactions on Mechatronics*, 8(1), 2003, pp.125-135, 1083-4435.

Kumar S.; Chaffin D. & Redfern M. (1989). EMG of trunk muscles in isometric and isokinetic MVC, *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, Images of the Twenty-First Century, vol.3, pp.1029-1030, Seattle, USA, November 1989.

Kuribayashi, K.; K. Okimura & T. Taniguchi (1992). A discrimination system using neural network for EMG-controlled prostheses, *Proceedings of IEEE International Workshop on Robot and Human Communication*, pp. 63-68, Yokohama, Japan, 1992

D. Nishikawa; W. Yu; H. Yokoi & Y. Kakazu (1999). EMG Prosthetic Hand Controller Discriminating Ten Motions using Real-time Learning Method, *Proceedings of the 1999 IEEE/RSJ International Conference on Intelligent Robotics and Systems*, pp.1592-1597, Kyongju, South Korea, Octorber 1999.

T. Tsuji; O. Fukuda; H. Ichinobe & M. Kaneko (1999). A log- linearixed Gaussian mixture network and its application to EEG pattern classification, *IEEE Trans. Systems, Man and Cybernetics – Part C: Application and Reviews,* Vol.29, No. 1, pp.60-72, 1094-6977, 1999.

A. Waibel; T. Hanazawa; G. Hinton; K. Shikano & K. Lang (1989). Phoneme Recognition Using Time-Delay Neural Network, *IEEE Transaction on Acoustic, Speech, and Signal processing*, Vol. 37, No.3, 1989, pp.328 - 339.

Lin Wang & Buchanan, T.S. (2002). Prediction of joint moments using a neural network model of muscle activations from EMG signals, *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 10(1) , 2002, pp.30-37, 1534-4320.

M. Zecca; Silvestro Micera; M. C. Carrozza & P. Dario (2002). Control of Multifunctional Prosthetic Hands by Processing Electromyographic Signal, *Critical Rev. in BIO. Eng.*, 30(4-6), 2002, pp. 459-485, 0278-940X.

Abdullah Al zaman; Ahad, M.A.; Ferdjallah, M. & Wertsch, J.J. (2005). A new approach for muscle fatigue analysis in young adults at different MVC levels, *48th Midwest Symposium on Circuits and Systems*, vol.1, pp.499 – 502, 0-7803-9197-7, August 2005.

2

Compression of Surface Electromyographic Signals Using Two-Dimensional Techniques

Marcus V. C. Costa[1], João L. A. Carvalho[1], Pedro A. Berger[2], Adson F. da Rocha[1] and Francisco A. O. Nascimento[1]
[1]*Department of Electrical Engineering*
[2]*Department of Computer Science*
University of Brasília
Brazil

1. Introduction

Surface electromyographic (S-EMG) signals provide non-invasive assessment of muscle function and structure (Merletti & Parker, 2004). The transmission and/or storage of S-EMG signals may be an issue, since the data volume of such signals is potentially large, depending on factors such as sampling rate, quantization precision, number of channels, and experiment duration. Although several techniques have been proposed for the compression of biomedical signals such as the electrocardiogram (ECG) and the electroencephalogram (EEG) (Naït-Ali & Cavaro-Ménard, 2008), few techniques are available for the compression of S-EMG signals.

The use of image encoding standards may provide several benefits to the compression of S-EMG signals. Such standards are well-established and widely-used, and fast and reliable implementations of these algorithms are readily-available in several operational systems, software applications, and portable systems. These aspects could be positive in the implementation of S-EMG data banks, as well as in telemedicine applications.

This chapter discusses the use of two-dimensional data encoders for compression of S-EMG signals measured during isometric contractions. The S-EMG data is first arranged into a $N \times M$ matrix, composed of M signal segments of length N. Then, a preprocessing step is used to increase the two-dimensional correlation of this matrix. This is achieved by rearranging the columns of the matrix such that the correlation between adjacent columns is maximized. Finally, arithmetic encoding is used to compress the column-order list, and an off-the-shelf image compression algorithm is used for reducing the data size.

The proposed approach is evaluated on eighteen S-EMG recordings measured on the *biceps brachii* muscle of four healthy male volunteers during isometric exercise. We graphically show the increase in two-dimensional correlation provided by the proposed preprocessing stage, and quantitatively demonstrate the improvement in compression efficiency achieved when such stage is used. Using the proposed approach, we show that off-the-shelf image encoders — namely the JPEG2000 and the H.264/AVC-intra algorithms — may be used for efficient compression of S-EMG signals. Finally, we quantitatively compare the performance

of these two algorithms with the S-EMG encoders proposed by Norris et al. (2001) and Berger et al. (2006).

2. Electromyographic Signals

Electromyography assesses muscle function by studying electrical signals generated by the musculature. Electromyographic signals (EMG) are relevant to the comprehension of human musculature and for the diagnosis of several neuromuscular diseases (Basmajian & De Luca, 1985; Merletti & Parker, 2004). However, the storage and transmission of such signals in telemedicine applications are still challenging, as the amount of data to transmit and store increases with sampling rate, sampling precision, number of channels, number of individuals, and other factors. Signal compression techniques may be useful in reducing the data size of electromyographic signals.

2.1 Motor unit and action potentials

In order to understand the nature of the EMG signal, it is important to first understand muscle physiology and the way muscles produce bioelectrical signals. There are three types of muscles in the human body: cardiac muscle (specialized heart tissue, with peculiar characteristics); skeletal muscle (also known as voluntary muscle, because it is controlled consciously); and smooth muscle (involuntary muscle, controlled unconsciously). The latter covers the surface of internal organs and is responsible for functions such as compressing the esophagic channel to complete the deglutition, or controlling the blood flow to several tissues.

The object of study of electromyography is skeletal muscle, which is connected directly or indirectly (by tendons) to the bones. Skeletal muscles work in antagonic pairs: while one of the muscles contracts, the other, which is responsible for the opposite movement of the joint, relaxes, producing different types of movements.

Skeletal muscle is composed of multiple bundles of muscle fibers. Muscle fibers are long and cylindrical multinucleated cells. In normal human skeletal muscle, muscle fibers do not contract individually. Instead, they contract in small groups called motor units. A motor unit is composed by a motor neuron, neuromuscular junctions, and the muscle fibers innervated by this neuron. The motor unit is the smallest functional unit of striated muscle (Basmajian & De Luca, 1985).

The impulse that originates at the motor neuron, propagates along the spinal neuraxon, and reaches the muscle fiber, is called motor action potential, and is responsible for starting the muscle contraction process. When this impulse reaches the muscle fibers, it generates a muscle action potential. The wave created at the neuromuscular junction due to the excitation of the group of fibers that compose a motor unit is called motor unit action potential (MUAP). The MUAP propagates from the innervation area to the tendinous insertion, as well as in the opposite direction (Basmajian & De Luca, 1985).

2.2 Surface vs. intramuscular electromyography

Electromyography is based on extracellular assessment of the events described above. Intramuscular and surface electromyographic techniques are supplementary to each other, and both are relevant instruments for physiological investigation. Intramuscular

electromyography is an invasive technique which uses needles or microelectronic devices placed directly in contact with the muscle. This technique is more appropriate and widely accepted for clinical applications, although it causes pain and discomfort to the patient.
Surface electromyography, also known as non-invasive electromyography, uses metal electrodes (typically Ag/AgCl) placed over the skin. S-EMG has wide application in areas such as biofeedback, prosthesis control, ergonomics, occupational medicine, sports medicine, and movement analysis. S-EMG allows painless access to neuromuscular functions, thus providing additional versatility. The measured S-EMG signal is a record of the sum of several MUAPs. The MUAPs are activated asynchronously, thus the measured signal is stochastic and highly complex (Figure 1). The extraction of clinical relevant patterns from S-EMG signals is a difficult problem. Consequently, the knowledge about such signal is not as broad as in the field of electrocardiography, for example (Merletti & Parker, 2004).

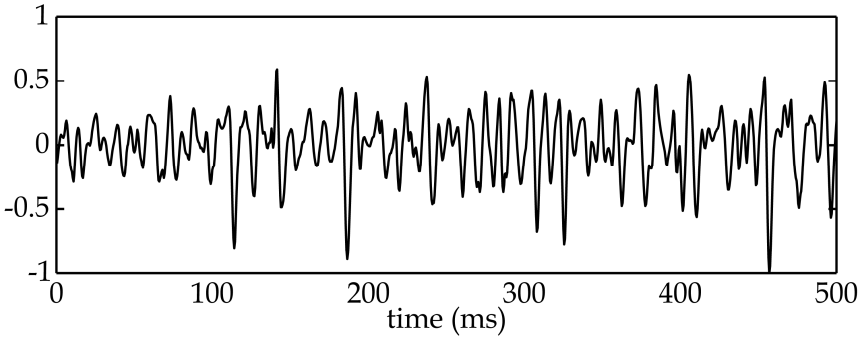


Fig. 1. S-EMG signal measured on the *biceps brachii* during an isometric contraction.

3. Signal Compression Techniques

Digitization of biomedical signals has been used for many different applications, such as ambulatory monitoring, phone line transmission, database storage, and several other uses in health and biomedical engineering. These applications have helped in diagnostics, patient care, and remote treatment. One example is the digital transmission of ECG signals, from the patient's home or ambulance to the hospital. This has been proven useful in cardiac diagnosis.
Biomedical signals need to be digitally stored or transmitted using a large number of samples per second, and with a large number of bits per sample, in order to ensure waveform fidelity, required for visual inspection. The use of signal compression techniques is fundamental for cost reduction and technical feasibility of storage and transmission of biomedical signals.
The purpose of any signal compression technique is the reduction of the amount of bits used to represent a signal. This must be accomplished while preserving the morphological characteristics of the waveform. In theory, signal compression is the process where the redundant information contained in the signal is detected and eliminated. Shannon (1948) defined redundancy as one minus "the ratio of the entropy of a source to the maximum value it could have while still restricted to the same symbols".

Signal compression has been widely studied in the past decades (Gersho & Gray, 1992; Jayant & Noll, 1984; Sayood, 2005; Salomon, 2006). Signal compression techniques are commonly classified into two categories: lossless and lossy compression. Lossless compression means that the decoded signal is identical to the original signal. In lossy compression, a controlled amount of distortion is allowed. Lossy signal compression techniques generally achieve higher compression rates than lossless techniques.

3.1 Lossless compression

Lossless signal compression techniques are less efficient with respect to compression rate than lossy compression. Lossless compression may be used in combination with lossy compression techniques, especially in cases where the maximum allowed distortion has been reached, and additional compression is needed. Among several lossless compression techniques, we highlight run-length encoding (Jayant & Noll, 1984), Huffman encoding (Huffman, 1952), arithmetic encoding (Witten et al., 1987), and differential encoding (Jayant & Noll, 1984).

3.1.1 Run-length encoding

Data files frequently present sequentially repeated characters, i.e., a character run. For example, text files use several spaces to separate sentences and paragraphs. Digital signals may contain the same value, or the same character representing that value, sequentially repeated many times in its data file. This indicates that the signal is not changing.

Figure 2 shows an example of run-length encoding (Jayant & Noll, 1984) of a data set that contains runs of zeros. Each time the encoder finds a zero in the entry data, two values are written in the output data. The first of these values is a zero indicating that run-length codification started. The second value is the amount of zeros in the sequence. If the run of zeros in the input data set is in average larger than two, then run-length encoding will achieve data compression.

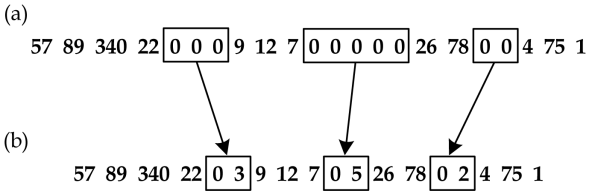


Fig. 2. Example of run-length encoding: (a) original signal; (b) encoded signal.

3.1.2 Huffman encoding

In Huffman encoding (Huffman, 1952), the data are represented as a set of variable-length binary words. The length depends on the frequency of occurrence of the symbols used for representing each signal value. Characters that are used often are represented with fewer bits, and seldom-used characters are represented with more bits.

Figure 3 shows an example of how a Huffman code is generated, given a symbol X, and its characteristic probability of occurrence, $p(X)$. The character codes are generated by combining the bits of a tree with ramifications, adding their probabilities, and then restarting the process until only one character remains. This process generates a tree with

ramifications linked to bits 0 and 1. The codes for each character are taken in the inverse path of these ramifications. Note that initial character arrangement is not relevant. In this example, we chose to encode the upper ramifications with bit 0 and the lower ones with bit 1. However, the opposite representation could also have been used. Any decision criteria may be used in ramifications with equal probabilities.
Huffman encoding has the disadvantage of assigning an integer number of bits to each symbol. This is a suboptimal strategy, because the optimal number of bits per symbol depends on the information content, and is generally a rational number.

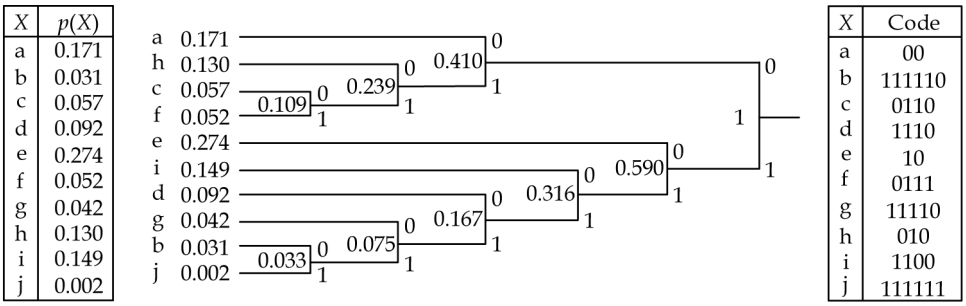


Fig. 3. Huffman code construction example.

3.1.3 Arithmetic encoding

Arithmetic encoding (Witten et al., 1987) is based on Huffman encoding concepts, but is more sophisticated, and achieves compression rates closer to the theoretical limits. Character sequences (symbols) are represented by individual codes, depending on their probability of occurrence, or a probability model.
At initialization, the [0,1) interval is divided into subintervals, with length proportional to its associated symbol probability. Every time a symbol appears in the message, its corresponding subinterval is divided into subintervals proportional to the symbol probabilities. When the end of the message is reached, the algorithm chooses a floating point value within the interval associated with the last encoded symbol. The binary representation of this value represents the message. This principle is illustrated in Figure 4, which shows the step-by-step encoding process for the message "BACADEA", with probability model: $p(A)=3/7$, $p(B)=1/7$, $p(C)=1/7$, $p(D)=1/7$ and $p(E)=1/7$.

3.1.4 Differential encoding

Differential pulse code modulation (DPCM) (Jayant & Noll, 1984) refers to signal compression techniques that represent the digital signal as the sequence of differences between successive samples. Figure 5 shows an example of how this is performed. The first sample in the encoded signal is equal to the first sample of the original signal. Subsequent encoded samples are equal to the difference between the current sample and the previous sample of the original signal. Using this technique, the encoded signal has a smaller amplitude dynamic range than the original signal. Therefore, fewer bits are needed to store or transmit the encoded signal. DPCM is used in combination with Huffman encoders in several biomedical signal compression algorithms.

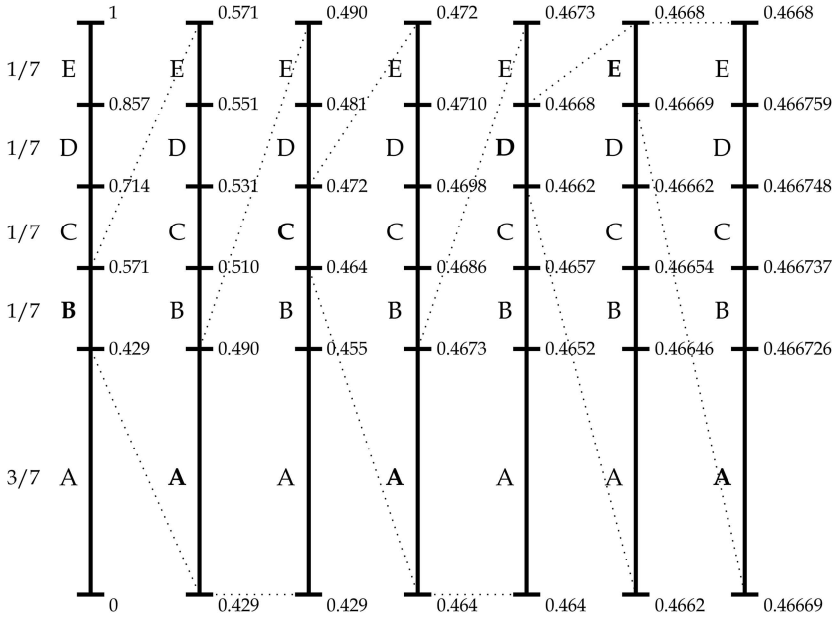


Fig. 4. Step-by-step arithmetic encoding process for the message "BACADEA", with probability model: $p(A)=3/7$, $p(B)=1/7$, $p(C)=1/7$, $p(D)=1/7$ and $p(E)=1/7$.

(a)

44 73 78 60 63 29 46 42 52 31 ...

(b)

44 29 5 -18 3 -34 17 -4 10 -21 ...

Fig. 5. Differential encoding example: (a) original signal; (b) encoded signal.

3.2 Lossy compression

There are two main categories of lossy compression techniques used with biomedical signals: direct methods and transform compression.

3.2.1 Direct methods

Direct methods encode signals in time domain. These algorithms depend on the morphology of the input signal. In most cases, these methods are complex, and provide lower compression efficiency than transform encoders (discussed in the next section).

Direct compression methods use sophisticated processes and "intelligent" signal decimation and interpolation. In other words, these methods extract K "significant" samples of the original length-N signal, $x(n)$, i.e.:

$$\left(n, x(n)\right),\ n = 0,...,N-1 \rightarrow \left(n_k, x(n_k)\right),\ k = 0,...,K-1\ , \tag{1}$$

where $K<N$. The process of selecting significant samples is based on the characteristics of the signal, and on a tolerance interval criterion for the reconstruction error.
The reconstruction of values between significant samples is performed by interpolation, using the generic expression (Sörnmo & Laguna, 2006):

$$\tilde{x}(n) = \begin{cases} x(n) & n = n_0,...,n_{K-1}; \\ f_{n_0,n_1}(n), & n = n_0+1,...,n_1-1; \\ \vdots & \vdots \\ f_{n_{K-2},n_{K-1}}(n), & n = n_{K-2}+1,...,n_{K-1}-1. \end{cases} \tag{2}$$

The interpolation function, $f_{n_{k-1},n_k}(n)$, is generally a first-order polynomial, and connects pairs of consecutive significant samples. Only K samples are stored, instead of N, resulting in a reduction of the amount of stored or transmitted data.

3.2.2 Transform encoding

Among the several methods for signal compression, techniques based on transforms achieve the best performance in terms of compression gain and waveform fidelity. For a data vector x, we can define an orthogonal transform as a linear operation given by a linear transformation T, such that:

$$y = Tx\ , \tag{3}$$

where y represents a vector of transformed coefficients, and T satisfies the orthogonality condition:

$$T^t = T^{-1}\ . \tag{4}$$

Transform compression is based on a simple premise: when the signal is processed by a transform, the signal energy (information) that was distributed along all time-domain samples can be efficiently represented with a small number of transform coefficients.
This is illustrated in Figure 6, where a two-dimensional signal is shown along with its corresponding coefficients in transform domain. In this example, we use the discrete cosine transform (DCT). The DCT is used in the most widely used standard for image compression, the Joint Picture Expert Group (JPEG) format.
Currently, the most widely used transform for encoding biomedical signals is the discrete wavelet transform (Daubechies, 1988; Mallat, 1989; Vetterli & Kovačević, 1995; Strang & Nguyen, 1996). In this transform, a signal containing N samples is filtered using a pair of filters that decompose the signal into low (L) and a high (H) frequency bands. Each band is undersampled by a factor of two; that is, each band contains $N/2$ samples. With the appropriate filter design, this process is reversible. This procedure can be extended to two-dimensional signals, such as images. In Figure 7, we show an example of wavelet decomposition for a gray-scale image with 256×256 pixels. Similarly to what was observed

using the DCT, many of the coefficients in the high-frequency subbands have amplitudes close to zero (dark pixels), and it is possible to compress the image by discarding them.

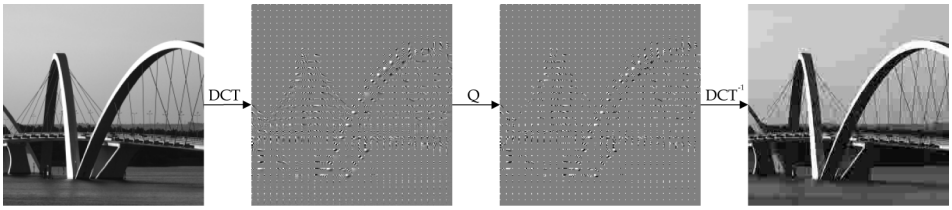


Fig. 6. Example of DCT-based image compression, applied in blocks of 8×8 pixels (Q denotes quantization).

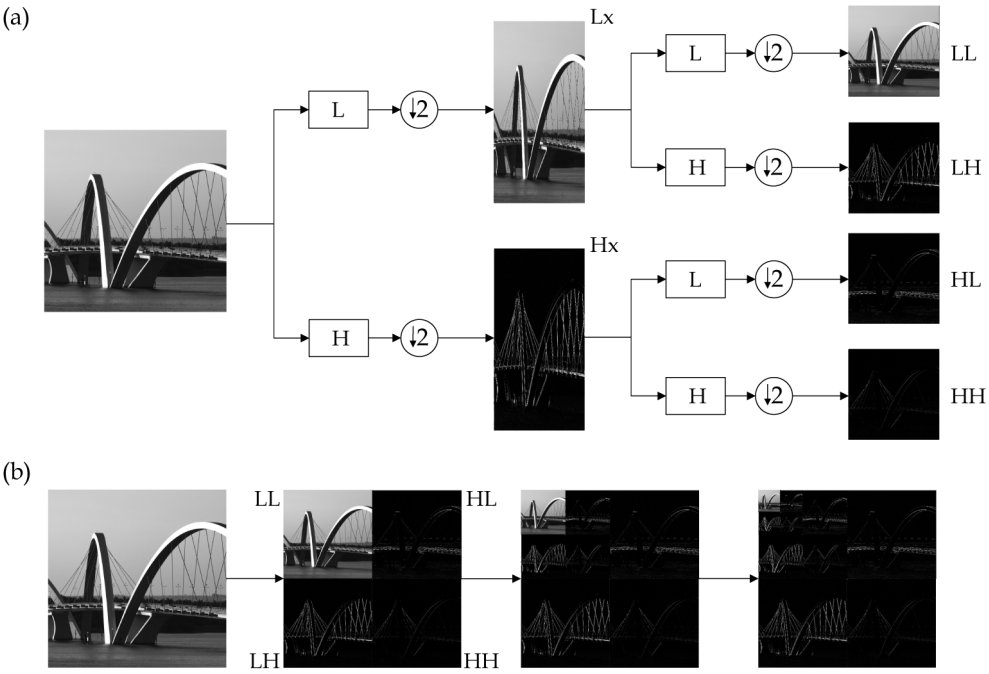


Fig. 7. Example of a 2D wavelet decomposition: (a) step-by-step one-level decomposition; (b) three-level decomposition.

Wavelet transform compression is evolving with respect to the way in which the coefficients are encoded. When a coefficient in a low-frequency subband is nonzero, there is a high probability that, at positions that correspond to high frequencies, the coefficients are also nonzero. Thus, the nonzero coefficients can be represented in a tree, starting at a low frequency root. Figure 8 illustrates this concept. Each coefficient in the LL band of level 1 has a corresponding coefficient in the other bands. The position of each coefficient in level 1 is mapped into four daughter-positions in each subband of level 2. An efficient way of encoding the coefficients that are nonzero is to encode each tree of coefficients, beginning

with the root decomposition level. The coefficients at the lower levels are encoded, and followed by their children coefficients in the higher level, until a null coefficient is found. The next coefficients of the tree have large probability of also being null, and are replaced by a code that identifies a tree of zeros (zerotree code). This method is called embedded zerotree wavelet (EZW) (Shapiro, 1993). A similar approach, commonly used in wavelet coefficient encoding, is the set partitioning in hierarchical trees (SPIHT) method (Said & Pearlman, 1996).

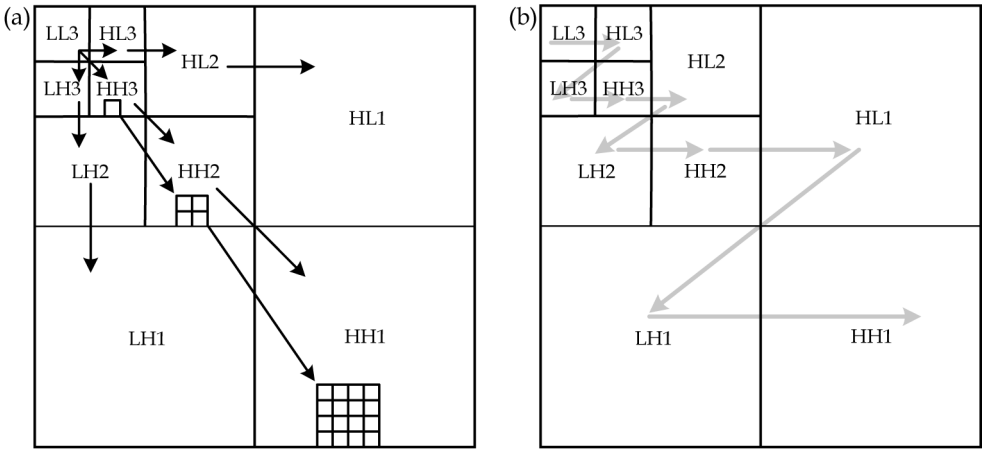


Fig. 8. EZW method: (a) zerotree data structure; (b) coefficients scanning order.

Modern image and video encoding algorithms are based on transform-domain approaches. For example, the JPEG2000 algorithm (Taubman, 2000; Taubman & Marcellin, 2002; Acharya & Tsai, 2004) is a state-of-the-art image encoding standard which uses embedded block coding with optimal truncation (EBCOT) (Taubman, 2000) on the subband samples of the discrete wavelet transform of the image. The H.264/AVC encoder (Richardson, 2003; Wiegand et al., 2003; Sullivan et al., 2004) is the latest standard for video compression, and uses a low-complexity integer discrete cosine transform.

4. Compression of S-EMG Signals

Different approaches have been proposed for compression of electromyographic signals. Norris & Lovely (1995) investigated the compression of electromyographic signals using adaptive differential pulse code modulation (ADPCM). Guerrero & Mailhes (1997) compared different compression methods based on linear prediction and orthogonal transforms. They showed that methods based on the wavelet transform outperform other compression methods. The use of the EZW algorithm has also been proposed for compression of electromyographic signals (Wellig et al., 1998; Norris et al., 2001). More recently, Berger et al. (2006) proposed an algorithm for compression of electromyographic signals using the wavelet transform and a scheme for dynamic bit allocation of the coefficients using a Kohonen layer. In a recent work, Brechet et al. (2007) adopted the

discrete wavelet packet transform decomposition with optimization of the mother wavelet and of the basis of wavelet packets, followed by an EZW-like encoder. The use of speech encoding methods has also been reported in the literature (Guerrero & Mailhes, 1997; Carotti et al., 2006). Carotti et al. (2006) proposed a scheme for compression of simulated and real EMG signals with algebraic code excited linear prediction (ACELP) and the results were evaluated with several spectral and statistics measurements. Filho et al. (2008a) adopted a multiscale multidimensional parser algorithm. Paiva et al. (2008) proposed adaptive EMG compression using optimized wavelet filters. Two of these methods are discussed in detail below.

Norris et al. (2001) proposed a S-EMG compression algorithm based on one-dimensional EZW, and compared its performance with that of standard wavelet compression methods, using both isometric and isotonic signals. They showed that the EZW based approach performed consistently better than standard wavelet approaches.

The algorithm proposed by Berger et al. (2006) is a transform domain encoder that uses a dynamic bit allocation scheme that adapts to the local signal statistics in wavelet domain. The S-EMG signal is segmented in blocks of 2048 samples, and the wavelet transform is applied to each block. The transform coefficients go through a normalization stage, which leads to a large number of null coefficients. The bit allocation scheme, which is implemented using a Kohonen neural network, assigns a specific amount of bits to quantize the coefficients in each frequency sub-band interval (Figure 9). The Kohonen layer represents a dictionary of 64 possible bit allocation vectors. Thus, six bits are used to identify the bit allocation scheme used in each transform block. The transform coefficients in each block are quantized using the associated bit allocation vector, and the quantized coefficients, Kohonen dictionary index, and normalization factors are then Huffman encoded and stored/transmitted. In the decoder, the quantized coefficients are recovered using the overhead information (dictionary index and normalization factor), and the decoded signal segments are obtained using the inverse wavelet transform.

5. Two-Dimensional S-EMG Compression

One-dimensional signals may be compressed using image encoding algorithms, simply by rearranging the signal samples into a two-dimensional matrix, which may be processed as an image. This approach has been previously demonstrated in the field of electrocardiography. For example, the JPEG2000 image encoding algorithm has been efficiently applied to ECG signal compression (Bilgin et al., 2003; Chou et al., 2006). Similarly, the H.264/AVC-intra encoder has also been used (Filho et al., 2008b). Two-dimensional encoding methods based on the discrete wavelet transform, followed by SPIHT coefficient encoding, have also been proposed (Lu et al., 2000; Pooyan et al., 2004; Miaou & Chao, 2005; Moazami-Goudarzi et al., 2005; Rezazadeh et al., 2005; Tai et al., 2005; Sharifahmadian et al., 2006; Sahraeian & Fatemizadeh, 2007). This section investigates the use of image encoding algorithms for compression of S-EMG signals.

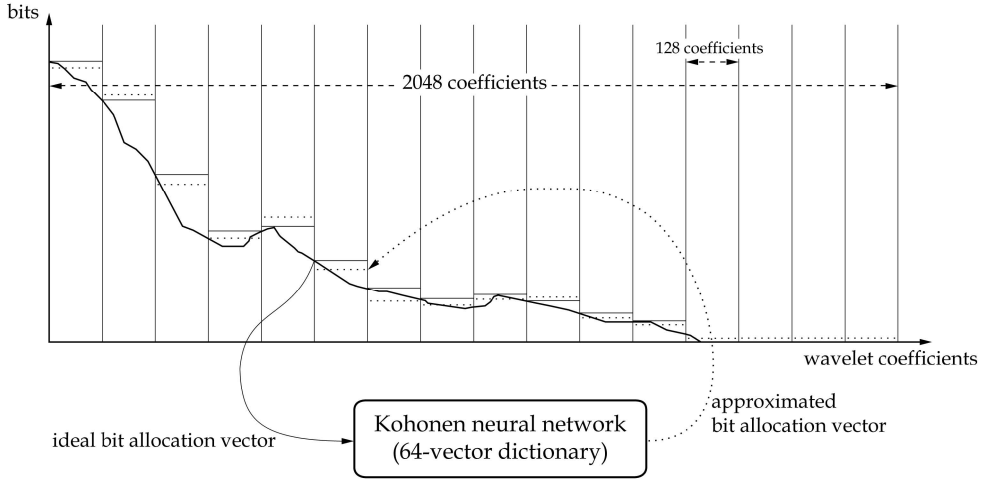


Fig. 9. Illustration of the quantization algorithm used in the encoder proposed by Berger et al. (2006). Each block of 2048 wavelet coefficients is divided into 16 sub-bands with 128 spectral lines. The ideal bit allocation vector is constructed using the number of bits necessary to quantize the largest coefficient in each sub-band. This is used as input of a Kohonen neural network, which outputs the approximated bit allocation vector.

5.1 Two-dimensional S-EMG modeling

S-EMG signals of isometric contractions are usually modeled as stochastic processes. The typical autocorrelation function associated with S-EMG signals indicates a stationary or, at least, locally stationary process.

In two-dimensional compression of S-EMG signals, the one-dimensional vector is segmented in M windows of length N, and arranged into a $N \times M$ matrix (Figure 10). This leads us to a two-dimensional real random field, $x_s[n,m]$, where: $n = 0, 1, \ldots, N-1$; $m = 0, 1, \ldots, M-1$; $s \in \Omega$; and $\Omega = \{1,2,\ldots\}$ is the sample space (Figure 11). The 2D matrix associated with each possible realization corresponds to a different two-dimensionally rearranged S-EMG signal.

The two-dimensionally arranged data can be seen as a homogeneous random field, because it is assumed that one-dimensional S-EMG signal is a stationary stochastic process. A random field is called homogeneous if its expected value of $x[n,m]$, μ_x, is independent of position $\vec{v}$, i.e., $\mu_x(\vec{v}) = \mu_x$. For rearranged S-EMG data, $\mu_x = 0$.

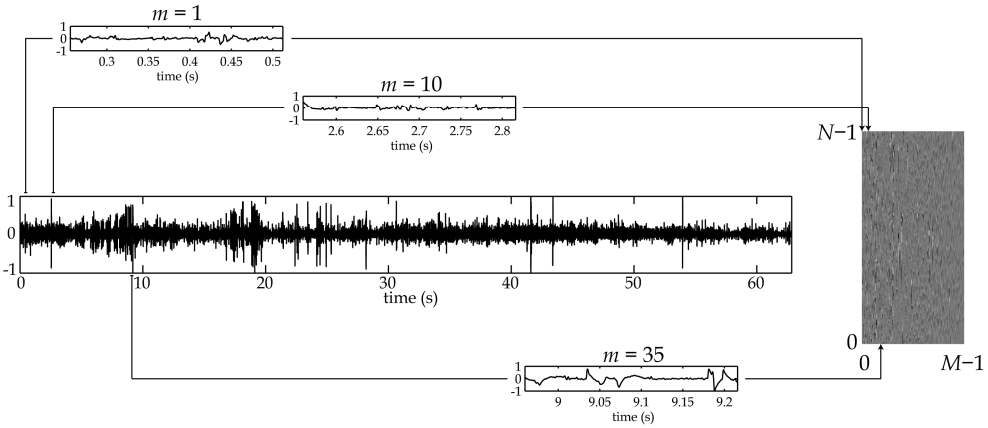


Fig. 10. One-dimensional S-EMG signal rearranged into a two-dimensional matrix.

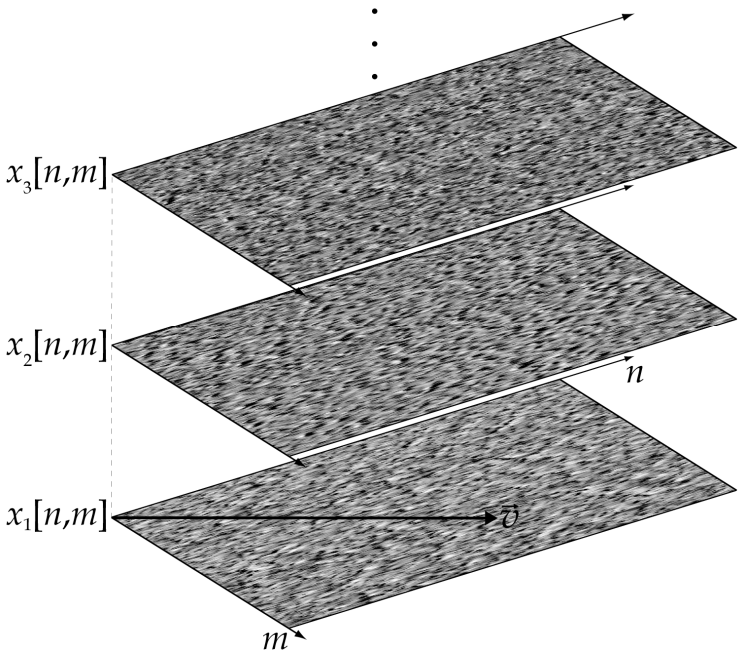


Fig. 11. Two-dimensional real random field, x_s [n,m], showing multiple realizations of two-dimensionally arranged S-EMG signals.

Homogeneous random fields present translation invariant autocorrelation functions, i.e.:

$$R_x(\vec{v}_1,\vec{v}_2) = R_x\{x(\vec{v}_1)x(\vec{v}_2)\} \equiv R_x(\vec{v}_1-\vec{v}_2) \equiv R_x(\vec{v}_2-\vec{v}_1)\ . \tag{5}$$

If we denote the position vectors $\vec{v}_1$ and $\vec{v}_2$ by their respective pair of discrete coordinates *m*, *n* and *j*, *i*, respectively, then the autocorrelation function can be expressed as

$$R_x(n,m,i,j) = R_x[n-i,m-j] = R_x[i-n,j-m] \,. \tag{6}$$

If we use the discrete variables *k* and *r* to denote the coordinate differences *n*−*i* and *m*−*j*, respectively, the above equation can be rewritten as

$$R_x(n,m,i,j) = R_x[k,r] = R_x[-k,-r] \,. \tag{7}$$

In general, the autocorrelation function of a random field is a function of four variables. However, the autocorrelation function of a homogeneous random field (e.g., S-EMG data) is a function of only two variables, *k* and *r*:

$$R_x[k,r] = E\{x[n,m]x[n+k,m+r]\} = \sum_{n=0}^{N-1}\sum_{m=0}^{M-1} x[n,m]x[n+k,m+r] \,. \tag{8}$$

The autocovariance function, $C_x[k,r]$, is defined as

$$C_x[k,r] = E\{x[n,m]x[n+k,m+r] - \mu^2\} = \sum_{n=0}^{N-1}\sum_{m=0}^{M-1}\left(x[n,m]x[n+k,m+r] - \mu^2\right) \,. \tag{9}$$

Under the assumption that a class of rearranged S-EMG data forms a homogeneous random field, the autocorrelation function, $R_x[k,r]$, may be assumed to be of the form

$$R_x[k,r] = \left(R_x[0,0] - \mu^2\right)e^{-\alpha|k|-\beta|r|} + \mu^2 \,, \tag{10}$$

where *a* and β are positive constants (Rosenfeld & Kak, 1982), and where, by definition,

$$R_x[0,0] = E\left\{\left(x[n,m]\right)^2\right\} = \sum_{n=0}^{N-1}\sum_{m=0}^{M-1} x[n,m]^2 \,, \text{ and} \tag{11}$$

$$\mu = E\{x[n,m]\} = \frac{1}{NM}\sum_{n=0}^{N-1}\sum_{m=0}^{M-1} x[n,m] = 0 \,. \tag{12}$$

For rearranged S-EMG data, $\mu = 0$, and the autocorrelation function is reduced to:

$$R_x[k,r] = R_x[0,0]e^{-\alpha|k|-\beta|r|} = C_x[k,r] \,. \tag{13}$$

Constants *a* and β can be distinct, due to the nature of rearranged S-EMG data. This means that the autocorrelation function can be used to model two-dimensional data with different degrees of correlation in the horizontal and vertical directions, by specifying the values of *a* and β. In our method, one direction corresponds to linear time data sampling, with strong

correlation, and the other corresponds to window step, and leads to weak correlation. The correlation along the window step direction may be increased using column reordering based on inter-column correlation, as discussed in the next section.
Figure 12a presents the theoretical autocorrelation function, calculated using equation (13), with a=0.215 and β=0.95. Figure 12b presents the autocorrelation function associated with the S-EMG shown in Figure 10, after column reordering. These results demonstrate that two-dimensionally arranged S-EMG data presents two-directional correlation and two-dimensional redundancy. Therefore, this type of data may be compressed using image compression techniques. In the next section, we present a technique for maximizing two-dimensional S-EMG correlation and thus improving compression efficiency.

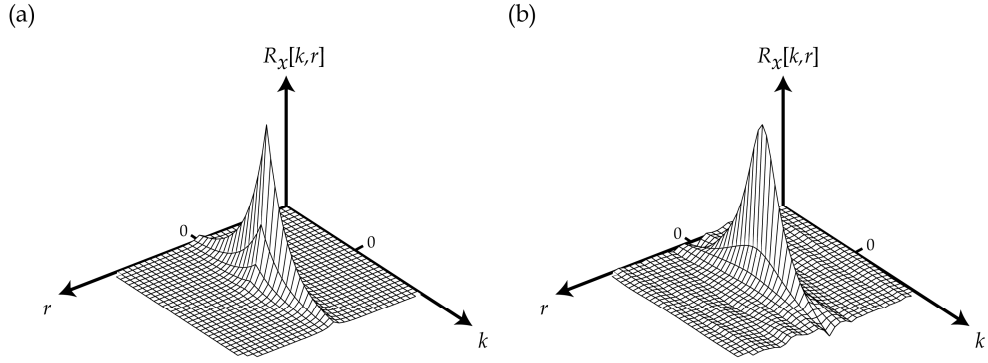


Fig. 12. Autocorrelation functions: (a) computed from the theoretical model, using a=0.215 and β=0.95; (b) computed from the data shown in Figure 10.

5.2 Correlation sorting

Adjacent samples of S-EMG signals are typically moderately temporally-correlated. When the S-EMG signal is arranged into a 2D matrix, this feature is preserved along the vertical dimension (columns). However, such correlation is generally lost along the horizontal dimension (rows). In order to increase 2D-compression efficiency, we attempt to increase the correlation between adjacent columns, by rearranging the columns based on their cross-correlation coefficients.
The matrix of column cross-correlation coefficients (R) is computed from the covariance matrix C, as follows:

$$R(u,w)=\frac{C(u,w)}{\sqrt{C(u,u)\cdot C(w,w)}}\ . \tag{14}$$

Then, the pair of columns that present the highest cross-correlation coefficient is placed as the first two columns of a new matrix. The column that presents the highest cross-correlation with the second column of the new matrix is placed as the third column of the new matrix, and so forth. A list of column positions in annotated. This procedure is similar

to that used by Filho et al. (2008b) for reordering segments of ECG signals, but the similarity metric used in that study was the mean squared error. Figure 13 illustrates the result of applying the proposed column-correlation sorting scheme to a S-EMG signal arranged in 2D representation.

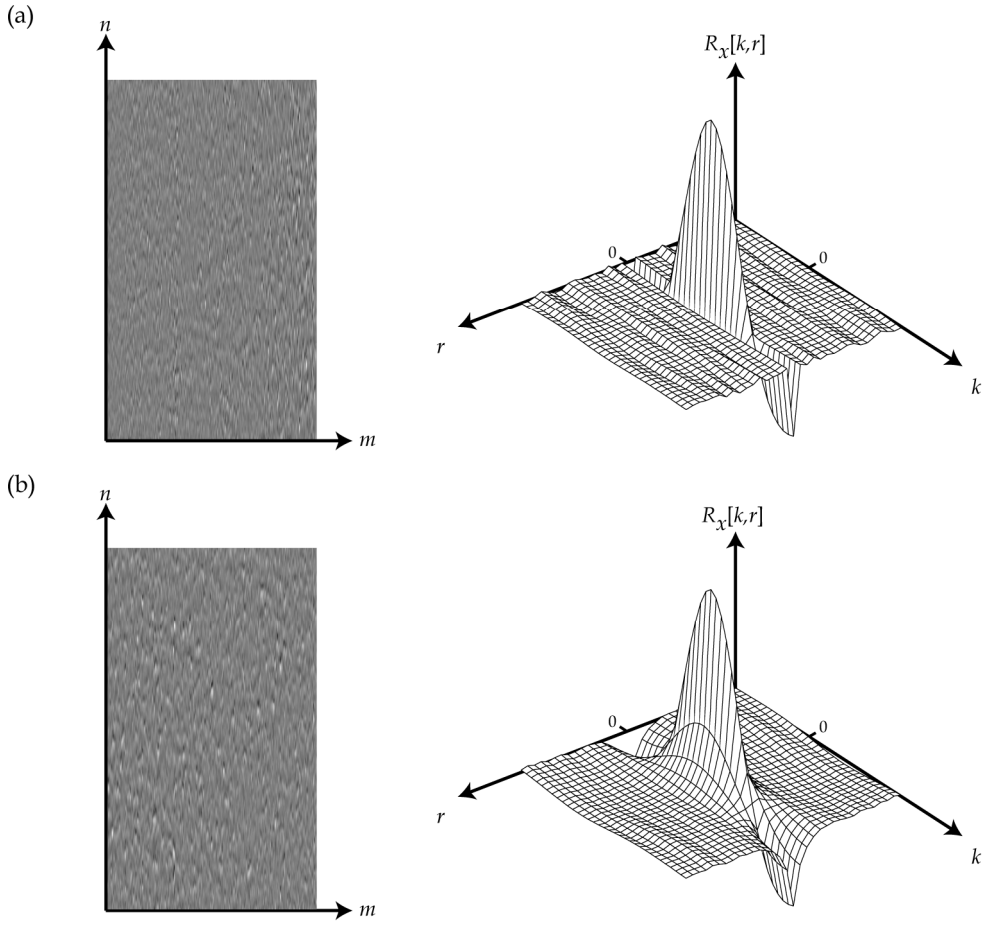


Fig. 13. Two-dimensionally arranged S-EMG signal (left) and associated autocorrelation function (right): (a) without correlation sorting; (b) with correlation sorting.

5.3 Image compression techniques applied to 2D-arranged S-EMG

Figure 14 shows a block diagram of the proposed encoding scheme. The method consists in segmenting each S-EMG signal into 512-sample windows, and then arranging these segments as different columns of a two-dimensional matrix, which can then be compressed using 2D algorithms. In this work, we investigated the use of two off-the-shelf image encoders: the JPEG2000 algorithm, and the H.264/AVC encoder.

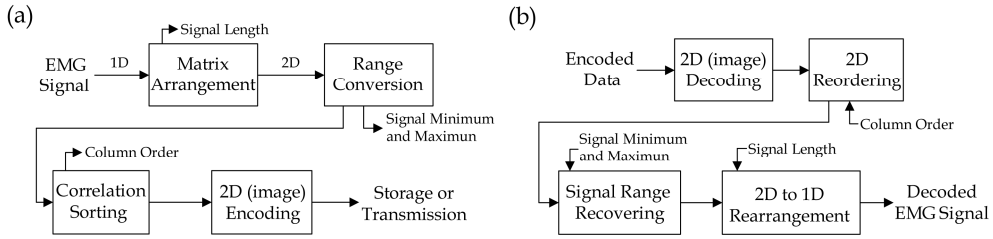


Fig. 14. Block diagram of the proposed compression algorithm: (a) encoder; (b) decoder.

The number of columns in the 2D matrix is defined by the number of 512-sample segments. The last (incomplete) segment is zero-padded. The matrix is scaled to the 8-bit range (0 to 255). The columns are rearranged, based on their cross-correlation coefficients. The matrix is encoded using one of the above-mentioned image encoders. The list of original column positions is arithmetically encoded. Scaling parameters (maximum and minimum amplitudes) and number of samples are also stored (uncompressed).
The encoded matrix is recovered using the appropriate image decoder, and the S-EMG signal is reconstructed by scaling the signal back to its original dynamic range and then rearranging the matrix columns back into a one-dimensional vector.

5.4 Experimental methods

A commercial electromyograph (Delsys, Bagnoli-2, Boston, USA) was used for signal acquisition. This equipment uses active electrodes with a pre-amplification of 10 V/V and a pass-band of 20–450 Hz. The signals were amplified with a total gain of 1000 V/V, and sampled at 2 kHz using a 12-bit data acquisition system (National Instruments, PCI 6024E, Austin, TX, USA). LabView (National Instruments, Austin, TX, USA) was used for signal acquisition, and Matlab 6.5 (The MathWorks, Inc., Natick, MA, USA) was used for signal processing.
Isometric contraction EMG signals were obtained from 4 male healthy volunteers with 28.3 ± 9.5 years of age, 1.75 ± 0.04 m height, and 70.5 ± 6.6 kg weight. Signals were measured on the *biceps brachii* muscle. In the beginning of the protocol, the maximum voluntary contraction (MVC) was determined for each subject. The signals were collected during 60% MVC contraction, with an angle of 90° between the arm and the forearm, and with the subject standing. The protocol was repeated 5 times for each volunteer, with a 48-hour interval between experiments. One of the volunteers was absent during two of the sessions. Therefore, a total of 18 EMG signals were acquired.
The JPEG2000 algorithm was evaluated with compression rates ranging from 0.03125 to 8 bits per pixel. The H.264/AVC encoder was used in intraframe (still image) mode, with DCT quantization parameter values ranging from 51 to 1.
The compression quality was evaluated by comparing the reconstructed signal with the original signal. The performance of the compression algorithm was measured by two quantitative criteria: the compression factor (CF) and the square root of the percentage root mean difference (PRD). These two criteria are widely used for evaluating the compression of S-EMG signals. The compression factor is defined as

$$\mathrm{CF}(\%) = \frac{Os - Cs}{Os} \cdot 100 \,, \tag{15}$$

where Os is the number of bits required for storing the original data, and Cs is the number of bits required for storing the compressed data (including overhead information). The PRD is defined as

$$\mathrm{PRD}(\%) = \sqrt{\frac{\sum_{n=0}^{N-1} \left(x[n] - \tilde{x}[n]\right)^2}{\sum_{n=0}^{N-1} x^2[n]}} \cdot 100 \,, \tag{16}$$

where x is the original signal, $\tilde{x}$ is the reconstructed signal, and N is the number of samples in the signal.

5.5 Results

Figure 15 shows the mean PRD (as a function of CF) measured on the set of 18 isometric S-EMG signals, using the JPEG2000 and H.264/AVC-intra compression algorithms, after correlation-based column-reordering. The quality decreases (PRD increases) when the compression factor is increased. With the JPEG2000 algorithm, compression factors higher than 88% causes significant deterioration of the decoded signal. With the H.264/AVC-intra algorithm, the results show significant degradation for compression factors higher than 85%. Figure 16 illustrates the compression quality for a S-EMG signal measured during isometric muscular activity. The central 2500 samples of the original, reconstructed, and error signals are shown. In this example, correlation sorting (*c.s.*) was used, with 75% compression factor. The PRD was measured to be 2.81% and 4.65% for the JPEG2000 and H.264/AVC-intra approaches, respectively. The noise pattern observed for both approaches seems visually uncorrelated with the signal.

Table 1 shows mean PRD values measured using different compression algorithms, for isometric contraction signals. The JPEG2000-based method provided slightly better reconstruction quality (lower PRD) than the EZW-based algorithm by Norris et al. (2001) for compression factors values ≤85%. However, this difference was not statistically significant. Compared with the method by Berger et al. (2006), JPEG2000 showed moderately inferior overall performance. This is especially true for 90% compression, in which its performance is comparable to that achieved by Berger et al. The H.264/AVC-based method showed low overall performance. The signal acquisition protocols used by Norris et al. (2001) and Berger et al. (2006) were similar to the one used in this work: 12-bit resolution, 2 kHz sampling rate, S-EMG isometric contractions measured on the *biceps brachii* muscle. However, some details of the acquisition protocols were not discussed in the work by Norris et al., (e.g., the distance between electrodes). The signals used in that work may present characteristics that are relevantly different from the those of the signals used in this work.

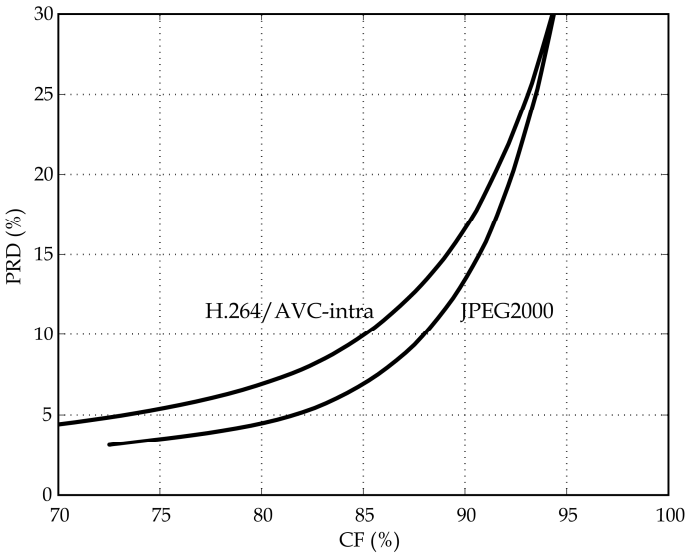


Fig. 15. Compression performance comparison (CF vs. PRD) between the JPEG2000 and H.264/AVC-intra image encoders, using the correlation sorting preprocessing step.

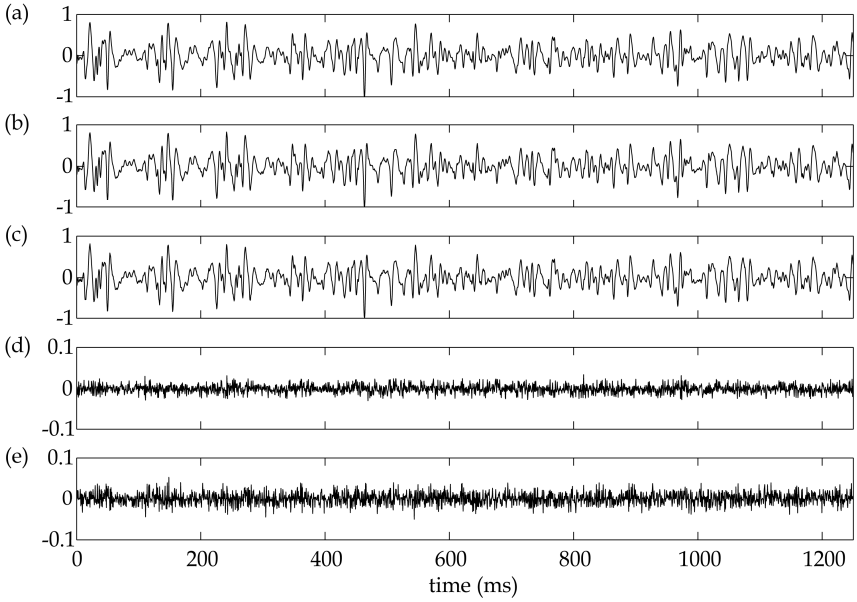


Fig. 16. Representative results for a 1250-ms segment of a S-EMG signal. (CF=75%): (a) uncompressed; (b) *c.s.* + JPEG2000; (c) *c.s.* + H.264/AVC-intra; (d) JPEG2000 reconstruction error; (e) H.264/AVC-intra reconstruction error. Reconstruction errors are magnified by 10-fold.

Compression Factor	75%	80%	85%	90%
Norris et al.	3.8	5	7.8	13
Berger et al.	2.5	3.3	6.5	13
JPEG2000	3.58	4.60	7.05	13.63
c.s. + JPEG2000	3.50	4.48	6.92	13.44
H.264/AVC-intra	5.51	7.03	10.01	16.68
c.s. + H.264/AVC-intra	5.37	6.90	9.93	16.62

Table 1. Mean PRD (in %) for isometric contraction signals.

The improvement in compression performance achieved using the proposed preprocessing stage (correlation-based column reordering) was not significant (Table 1). Column reordering increases inter-column correlation and improves compression efficiency. However the addition of overhead information increases the overall data size, resulting in similar PRD values. Better results may be achieved in the context of isotonic contractions, in which data redundancy is more significantly increased by the proposed approach.

6. Conclusions

This chapter presented a method for compression of surface electromyographic signals using off-the-shelf image compression algorithms. Two widely used image encoders were evaluated: JPEG2000 and H.264/AVC-intra. We showed that two-dimensionally arranged electromyographic signals may be modeled as random fields with well-determined autocorrelation function properties. A preprocessing step was proposed for increasing inter-column correlation and improving 2D compression efficiency.

The proposed scheme was evaluated on surface electromyographic signals measured during isometric contractions. We showed that commonly available algorithms can be effectively used for compression of electromyographic signals, with a performance that is comparable or better than that of other S-EMG compression algorithms proposed in the literature. We also showed that correlation sorting preprocessing may potentially improve the performance of the proposed method.

The JPEG2000 and H.264/AVC-intra image encoding standards are well-established and widely-used, and fast and reliable implementations of these algorithms are readily-available in several operational systems, software applications, and portable systems. These are important aspects to be considered when selecting a compression scheme for specific biomedical applications, and represent promising features of the proposed approach.

7. References

Acharya, T. & Tsai, P. S. (2004). *JPEG2000 Standard for Image Compression: Concepts, Algorithms and VLSI Architectures*. John Wiley & Sons, ISBN 9780471484226, Hoboken, NJ, USA.

Basmajian, J. V. & De Luca, C. J. (1985). *Muscles Alive: Their Functions Revealed by Electromyography*. Williams & Wilkins, ISBN 9780683004144, Baltimore, USA.

Berger, P. A.; Nascimento, F. A. O.; do Carmo, J. C. & da Rocha, A. F. (2006). Compression of EMG Signals with Wavelet Transform and Artificial Neural Networks, *Physiological Measurement*, Vol. 27, No. 6, pp. 457–465, ISSN 1361-6597.

Bilgin, A.; Marcellin, M. W. & Altbach, M. I. (2003). Compression of Electrocardiogram Signals using JPEG2000. *IEEE Transactions on Consumer Electronics*. Vol. 49, No. 4, pp. 833–840, ISSN 0098-3063.

Brechet, L.; Lucas, M.-F.; Doncarli, C. & Farina, D. (2007). Compression of biomedical signals with mother wavelet optimization and best-basis wavelet packet selection. *IEEE Transactions on Biomedical Engineering*, Vol. 54, No. 12, pp. 2186–2192, ISSN 0018-9294.

Carotti, E. S. G.; De Martin, J. C.; Merletti, R. & Farina, D. (2006). Compression of surface EMG signals with algebraic code excited linear prediction. *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 1148–1151, ISBN 142440469X, Tolouse, France, May 2006.

Chou, H-H.; Chen, Y-J.; Shiau, Y-C. & Kuo, T-S. (2006). An effective and efficient compression algorithm for ECG signals with irregular periods. *IEEE Transactions on Biomedical Engineering*, Vol. 53, No. 6, pp. 1198–1205, ISSN 0018-9294.

Daubechies, I. (1988). Orthogonal bases of compactly supported wavelets. *Communications on Pure and Applied Mathematics.* Vol. 41, No. 7, pp. 909–996, ISSN 0010-3640.

Filho, E. B. L.; da Silva, E. A. B. & de Carvalho, M. B. (2008a). On EMG signal compression with recurrent patterns. *IEEE Transactions on Biomedical Engineering*, Vol. 55, No. 7, pp. 1920–1923, ISSN 0018-9294.

Filho, E. B. L.; Rodrigues, N. M. M.; da Silva, E. A. B.; de Faria, S. M. M.; da Silva, V. M. M. & de Carvalho, M. B. (2008b). ECG signal compression based on DC equalization and complexity sorting. *IEEE Transactions on Biomedical Engineering*, Vol. 55, No. 7, pp. 1923–1926, ISSN 0018-9294.

Gersho, A. & Gray, R. (1992). *Vector quantization and signal compression*. Kluwer Academic Publishers, ISBN 0792391810, Norwell, MA, USA.

Guerrero, A. P. & Mailhes, C. (1997). On the choice of an electromyogram data compression method. *Proceedings of the 19th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 1558–1561, ISBN 0780342623, Chicago, IL, USA, Oct. 30-Nov. 2 1997.

Huffman, D. A. (1952). A method for the construction of minimum-redundancy codes. *Proceedings of the Institute of Radio Engineers*, Vol. 40, No. 9, pp. 1098–1101.

Jayant, N. S. & Noll, P. (1984). *Digital coding of waveforms – principles and application to speech and video*. Prentice Hall, Inc., ISBN 9780132119139, Englewood Cliffs, NJ, USA.

Lu, Z.; Kim, Y. D. & Pearlman, A. W. (2000). Wavelet compression of ECG signals by the set partitioning in hierarchical trees algorithm. *IEEE Transactions on Biomedical Engineering*. Vol. 47, No. 7, pp. 849–856, ISSN 0018-9294.

Mallat, S. G. (1989). A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence.* Vol. 11, No. 7, pp. 674–693, ISSN 0162-8828.

Merletti, R. & Parker, P. (2004). *Electromyography: Engineering and Noninvasive Applications*, John Wiley & Sons – IEEE Press, ISBN 9780471675808, Hoboken, NJ, USA.

Miaou, S. & Chao, S. (2005). Wavelet-based lossy-to-lossless ECG compresion in a unified vector quantization framework. *IEEE Transactions on Biomedical Engineering.* Vol. 52, No. 3, pp. 539–543, ISSN 0018-9294.

Moazami-Goudarzi, M.; Moradi, M. H. & Abbasabadi, S. (2005). High performance method for electrocardiogram compression using two dimensional multiwavelet transform. *IEEE 7th Workshop on Multimedia Signal Processing*, pp. 1–5, ISBN 0780392884, Shanghai, Oct. 30-Nov. 2 2005.

Naït-Ali, A. & Cavaro-Ménard, C. (2008). *Compression of Biomedical Images and Signals,* ISTE - John Wiley & Sons, ISBN 9781848210288, London, UK - Hoboken, NJ, USA.

Norris, J. A.; Englehart, K. & Lovely, D. (2001). Steady-state and dynamic myoelectric signal compression using embedded zero-tree wavelets, *Proceedings of 23rd Annual International Conference of the IEEE Engineering in Medicine Biology Society*, pp. 1879–1882, ISBN 0780372115, Istanbul, Turkey, Oct. 2001.

Norris, J. F. & Lovely, D. F. (1995). Real-time compression of myoelectric data utilizing adaptive differential pulse code modulation. *Medical and Biological Engineering and Computing*, Vol. 33, No. 5, pp. 629–635, ISSN 0140-0118.

Paiva, J. P. L. M.; Kelencz, C. A.; Paiva, H. M.; Galvão, R. K. H. & Magini, M. (2008). Adaptive wavelet EMG compression based on local optimization of filter banks, *Physiological Measurement*, Vol. 29, No. 7, pp. 843–856, ISSN 1361-6597.

Pooyan, M.; Moazami-Goudarzi, M. & Saboori, I. (2004). Wavelet compression of ECG signals using SPIHT algorithm. *IEEE International Journal of Signal Processing*, Vol. 1, No. 4, pp. 219–225, ISSN 2070-397X.

Rezazadeh, I. M.; Moradi, M. H. & Nasrabadi, A. M. (2005). Implementing of SPIHT and sub-band energy compression (SEC) method on two-dimensional ECG compression: a novel approach. *27th Annual International Conference of the Engineering in Medicine and Biology Society,* pp. 3763–3766. ISBN 0780387414, Shanghai, 17-18 Jan. 2006.

Richardson I. E. G. (2003). *H.264 and MPEG-4 Video Compression: Video Coding for Next-generation Multimedia*. Wiley, ISBN 9780470848371, UK.

Rosenfeld, A. & Kak, A. C. (1982). *Digital picture Processing (Volume 1): 2nd Ed.* Academic Press. Inc, ISBN 0125973012, San Diego, CA, USA.

Sahraeian, S. M. E. & Fatemizadeh, E. (2007). Wavelet-based 2-D ECG data compression method using SPIHT and VQ coding. *The International Conference on "Computer as a Tool", EUROCON, 2007*, pp. 133–137. ISBN 9781424408139, Warsaw, 9-12 Sept. 2007.

Said, A. & Pearlman, W. A. (1996). A new, fast, and efficient image codec based on set partitioning in hierarchical trees. *IEEE Transactions on Circuits and Systems for Video Technology*, Vol. 6, No. 3, pp. 243–250. ISSN 1051-8215.

Salomon, D. (2006). *Data Compression: The Complete Reference, 4th ed.*, Springer, ISBN 9781846286025, London, UK.

Sayood, K. (2005). *Introduction to Data Compression, 3rd ed.*, Morgan Kaufmann Publishers, ISBN 9780126208627, San Francisco, CA, USA.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal,* Vol. 27, July and October, pp. 379-423 and 623-656.

Shapiro, J. M. (1993). Embedded image coding using zerotrees of wavelet coefficients. *IEEE Transactions on Signal Processing,* , Vol. 41, No. 12, pp. 3445–3462, ISSN 1053-587X.

Sharifahmadian, E. (2006). Wavelet compression of multichannel ECG data by enhanced set partitioning in hierarchical trees algorithm. *28th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 5238–5243, New York City, USA, July 30-Aug. 3 2006.

Sörnmo, L. & Laguna, P. (2006). Electrocardiogram (ECG) signal processing. In: *Wiley Encyclopedia of Biomedical Engineering (Vol. 2)*, Metin Akay (Ed.), pp. 1298–1313, John Wiley & Sons, ISBN 9780471249672.

Strang, G. & Nguyen, T. (1996). *Wavelets and Filter Banks*. Wellesley-Cambridge Press, ISBN 0961408871, Wellesley, MA, USA.

Sullivan, G.; Topiwala, P. & Luthra, A. (2004). The H.264/AVC advanced video coding standard: Overview and introduction to the fidelity range extensions. *Proceedings of SPIE Conference on Applications of Digital Image Processing XXVII, Special Session on Advances in the New Emerging Standard: H.264/AVC*, Vol. 5558 (2), pp. 454–474. ISBN 0819454966, Denver, CO, USA, 2-6 August 2004.

Tai, S-C.; Sun C-C. & Yan, W-C. (2005). A 2-D ECG compression method based on wavelet transform and modified SPIHT. *IEEE Transactions on Biomedical Engineering,* Vol. 52, No. 6, pp. 999–1008, ISSN 0018-9294.

Taubman, D. S. & Marcellin, M. W. (2002), *JPEG2000: Image Compression Fundamentals, Standards and Practice*, Springer, ISBN 079237519X, Boston, USA.

Taubman, D. S. (2000). High performance scalable image compression with EBCOT. *IEEE Transactions on Image Processing*, Vol. 9, No. 7, pp. 1158–1170, ISSN 1057-7149.

Vetterli, M. & Kovačević, J. (1995) *Wavelets and Subband Coding*, Prentice-Hall, ISBN 0130970808, Englewood Cliffs, NJ, USA.

Wellig, P.; Zhenlan, C.; Semling, M. & Moschytz G. S. (1998). Electromyogram data compression using single-tree and modified zero-tree wavelet encoding. *Proceedings of the 20th Annual International Conference of the IEEE Engineering in Medicine and Biology Societ,* ISBN 0780351649, pp. 1303–1306. Hong Kong, China, Oct. 29-Nov. 1 1998.

Wiegand, T.; Sullivan, G. J.; Bjontegaard, G. & Luthra, A. (2003). Overview of the H.264/AVC video coding standard. *IEEE Transactions on Circuits and Systems for Video Technology*, Vol. 13, No. 7, pp. 560–576, ISSN 1051-8215.

Witten, I.; Neal, R. & Cleary, J. (1987). Arithmetic coding for data compression. *Communications of the ACM*, Vol. 30, No. 6, pp. 520–540, ISSN 0001-0782.

3

A New Method for Quantitative Evaluation of Neurological Disorders based on EMG signals

Jongho Lee[1], Yasuhiro Kagamihara[2] and Shinji Kakei[1]
[1]*Behavioral Physiology, Tokyo Metropolitan Institute for Neuroscience*
[2]*Tokyo Metropolitan Neurological Hospital*
Japan

1. Introduction

In this chapter, we propose a novel method to make a quantitative evaluation of neurological disorders based on EMG signals from multiple muslces. So far, some researchers tried to evaluate arm movements in various conditions (Nakanishi et al., 1992; Nakanishi et al., 2000; Sanguineti et al., 2003). They captured some features of movement disorders in patients with neurological diseases such as Parkinson's disease or cerebellar atrophy. However, the scope of these analyses was limited to movement kinematics. The problem here is that the movement kinematics, in general, cannot specify its causal muscle activities (i.e. motor commands) due to the well-known redundancy of the musculo-skeletal system. Thus, in order to understand central mechanisms for generation of pathological movements, it is essential to capture causal anomaly of the motor commands directly, rather than to observe the resultant movement indirectly (Manto, 1996; Brown et al., 1997). It should be also emphasized that the new method must be simple and noninvasive for wider clinical application.

To address these issues, we developed a novel method to identify causal muscle activities for movement disorders of the wrist joint. In order to determine causal relationship between muscle activities and movement disorders, we approximated the relationship between the wrist joint torque calculated from the movement kinematics and the four EMG signals using a dynamics model of the wrist joint (see Section 3.2). Consequently, we found that the correlation between the wrist joint torque and the EMG signals were surprisingly high for cerebellar patients as well as for normal controls (see Section 3.3). These results demonstrated a causal relationship between the activities of the selected muscles and the movement kinematics. In fact, we confirmed the effectiveness of our method, identifying the causal abnormality of muscle activities for the cerebellar ataxia in detail (see Section 3.4).

Finally, we further extended our analysis to calculate parameters that characterize pathological patterns of the muscle activities (see Section 4.1). We will conclude this chapter by discussing the application and clinical value of these parameters (see Section 4.2).

2. Materials and Methods

2.1 Experimental apparatus

In order to make a quantitative evaluation of neurological disorders, we developed a system for quantitative evaluation of motor command using wrist movements (Lee et al, 2007). Specifically, we intended to analyze the causal relationship between movement disorders and abnormal muscle activities. In addition, the system was also designed to be non-invasive and used handily at the bedside.

An outline of the system is shown in Figure 1. It consists of four components, a wrist joint manipulandum, a notebook computer, a small Universal Serial Bus (USB) analog-to-digital (A/D) converter interface and a multi-channel amplifier for surface electromyogram (EMG) signal. Movement of the wrist joint is measured with 2 position sensors of the manipulandum at 2 kHz sampling rate, and the wrist position is linked to the position of the cursor on the computer display. In other words, the manipulandum worked as a mouse for the wrist joint. Consequently, we can analyze the relationship between movement disorders and muscle activities, while subjects perform various wrist movement tasks using the manipulandum.

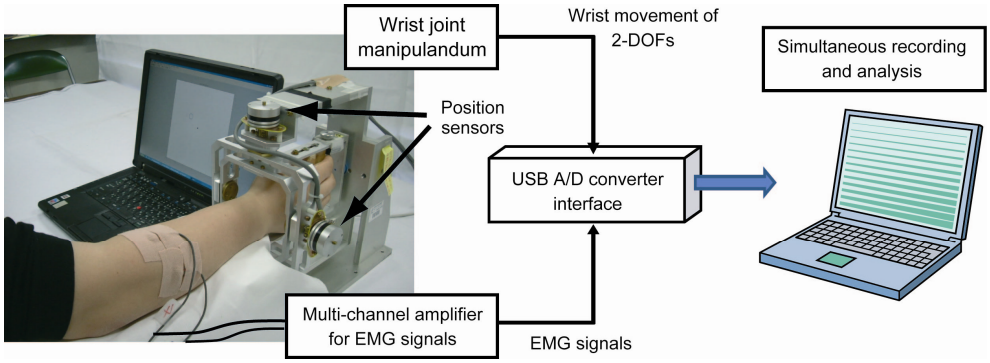


Fig. 1. Outline of the quantitative evaluation system for motor function using wrist movement

2.2 Experimental task

Subjects sat on a chair and grasped the manipulandum with his/her right hand. The forearm was comfortably supported by an armrest. As the experimental task, we asked subjects to perform step-tracking wrist movements (Figure 2A) and pursuit wrist movements (Figure 2B).

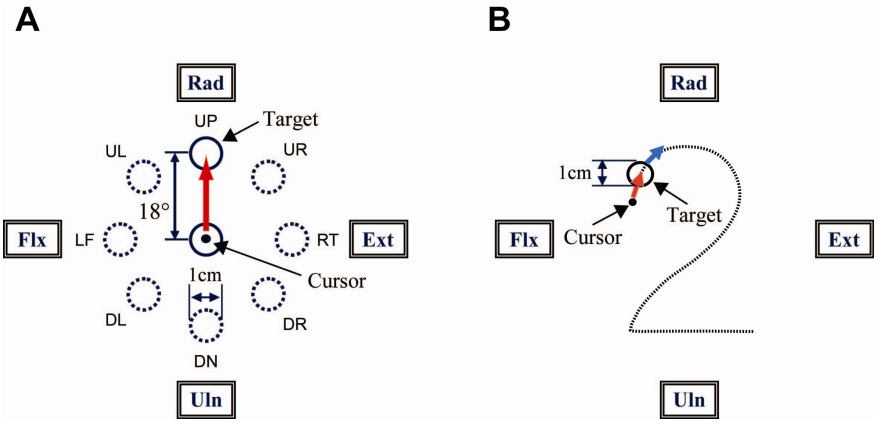


Fig. 2. Experimental tasks : step-tracking wrist movement (A) and pursuit wrist movement (B). To make these wrist movement tasks, the subject holds the forearm in the neutral position, midway between full pronation and full supination.

1. Step-tracking wrist movement
When a circular target, whose diameter was 1 cm, was displayed at the center of the monitor, the subject was required to move the cursor into the target. When a new target was shown at a place equivalent to 18 degrees of the wrist joint movement, the subjects had to move the cursor immediately to the new target as rapidly and accurately as possible. The subject performed the step-tracking wrist movement for the target of 8 directions (UP, UR, RT, DR, DN, DL, LF, UL). For this task, eight patients clinically diagnosed as cerebellar disorders and eight normal controls participated as subjects. Each subject performed this task 3 times.

2. Pursuit wrist movement
When a circular target, whose diameter was 1 cm, was displayed at the upper left of the monitor (X=-10° , Y=8°), the subject was required to move and hold the cursor into the target. After 3 seconds, the target moves by making the path of the figure 2 at the constant speed (mean velocity = 4.97deg/sec). At that time, the subjects had to enter the cursor into the moving target continuously. For this task, eight patients clinically diagnosed as cerebellar disorders, four patients clinically diagnosed as Parkinson's disease and eight normal controls participated as the subjects. Each subject performed this task 5 times.

During the task, four channels of EMG signals and two degree of freedom wrist movements were sampled and recorded at 2 kHz.

2.3 Recording muscle activities

We recorded surface EMG signals from four wrist prime movers: extensor carpi radialis (ECR), extensor carpi ulnaris (ECU), flexor carpi ulnaris (FCU) and flexor carpi radialis (FCR). The EMG signals were recorded with Ag-AgCl electrodes, amplified and sampled at 2 kHz. Typical locations of the surface electrodes for these four muscles are shown in Figure 3A. The position of each electrode was adjusted for each subject to maximize EMG signals of each muscle for a specific movement. In a few healthy control volunteers, we confirmed

effectiveness of the adjustment with high correlation between the surface EMG signals and the corresponding EMG signals recorded with needle electrodes from the same muscles identified with evoked-twitches.

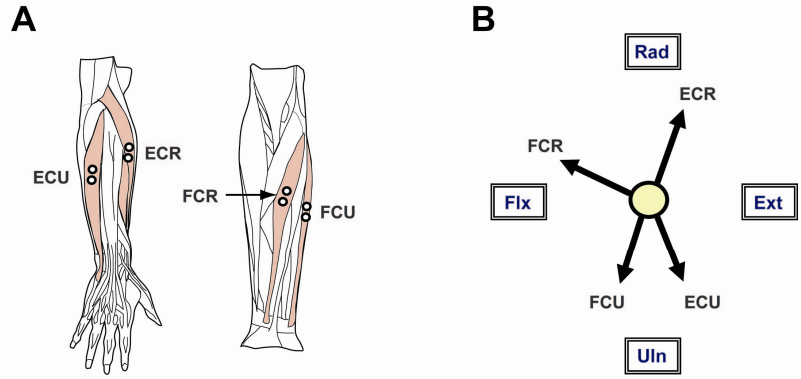


Fig. 3. Muscles related to the wrist joint (A) and pulling direction of each muscle (B). (A) The four wrist prime movers whose activities were recorded: extensor carpi radialis (ECR), extensor carpi ulnaris (ECU), flexor carpi ulnaris (FCU) and flexor carpi radialis (FCR). We did not distinguish extensor carpi radialis longus (ECRL) and extensor carpi radialis brevis (ECRB), because they have quite similar actions on the wrist and their activities are indistinguishable with surface electrodes. (B) The arrow indicates the pulling direction of each muscle. Muscle pulling directions for ECR, ECU, FCU, FCR were 18.4, 159.5, 198.3, and 304.5° clockwise from UP target.

There are two reasons why we chose these four muscles. First, the mechanical actions of these muscles are evenly distributed to cover the wrist movement for any direction (Figure 3B). Second, it is easy to record their activities with surface electrodes (Figure 3A). This is an essential clinical benefit to record muscle activities without pain, sparing use of invasive needle or wire electrodes. It should be also noted that use of no more than four surface electrodes contributes greatly to minimize time needed to set up recording.

2.4 Normalization of EMG Signals

It is well known that EMG signals are closely correlated with activities of α-motoneurons, which represent the final motor commands from the CNS. These motor commands generate muscle contraction, which results in muscle tension. It is established that a second order, low-pass filter is sufficient for estimating muscle tension from the raw EMG signal (Mannard & Stein, 1973). However, although the low-pass filtered EMG signal is proportional to muscle tension, the proportional constant varies due to variability of skin resistance or relative position of the electrode on a muscle for each recording. Therefore, for a quantitative analysis, it is necessary to normalize the EMG signals. For this purpose, we asked each subject to generate isometric wrist joint torque for the PD of each muscle. Namely, for each muscle, we set the amplitude of the EMG signals for 0.8 Nm of isometric wrist joint torque as 1. Then, the normalized EMG signals were digitally rectified and then filtered with a low-pass filter of a second order.

In this study, we used a Butterworth low-pass filter of a second order with cut-off frequency of 4Hz. Most critically, we considered the filtered EMG signals as muscle tensions, and used them to estimate the wrist joint torque (Mannard & Stein, 1973; Koike & Kawato, 1995). In this study, we called the filtered EMG signals as muscle tension shortly.

3. Identification of causal muscle activities for movement disorders

In this section, we will describe the results of identification for the step-tracking movement of the wrist for various directions. Specifically, we identified causal abnormality of muscle activities for movement disorders of cerebellar patients, confirming effectiveness of our method for analysis of movement disorders at the level of the motor command.

3.1 Movement disorders and causal anomaly in the muscle activities

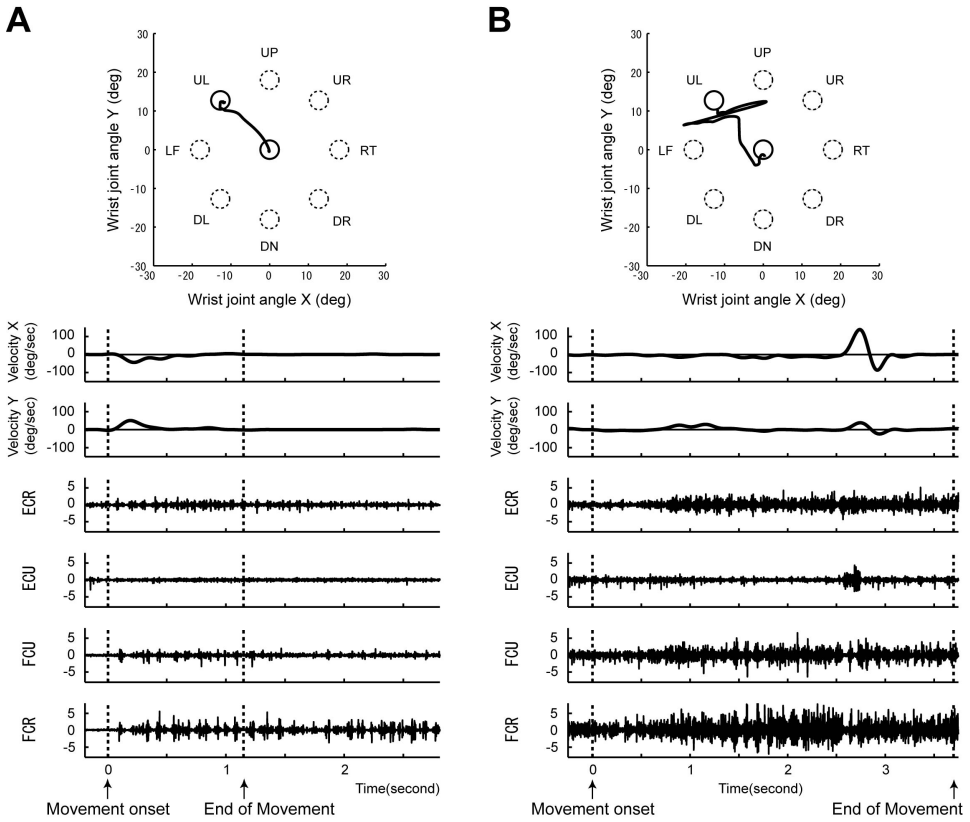


Fig. 4. Wrist joint kinematics and EMG signals for UL target. (A) An example of step-tracking movement for UL target in a normal control. The inset demonstrates a trajectory of the wrist joint. The top two traces show X-axis and Y-axis components of the angular velocity. The bottom four traces show EMG signals of ECR, ECU, FCU, FCR. (B) A corresponding example recorded from a cerebellar patient.

Figure 4 shows a trajectory, velocity profiles and EMG signals for a movement toward UL target. As shown in Figure 4A, the trajectory of a normal control was almost straight, and the angular velocity of a wrist joint showed a typical bell-shape profile for both X- and Y-components. In terms of EMG signals, FCR whose pulling direction (see Figure 3B) was directed to the UL target was active from the movement onset to the end. In addition, the activity of FCR lasted while a wrist was maintained in the UL target. ECR and FCU that had minor contribution for the direction also demonstrated moderate and probably cooperative activities. In contrast, ECU whose pulling direction was directed opposite to UL target, was inactive throughout the movement. Overall, the muscle activities and the mechanical actions of the four muscles can explain the movement quite reasonably in the normal control. The same was true for the cerebellar patients. Even the complex trajectory can be explained with the muscle activities as follows. As shown in Figure 4B, the initial downward movement was lead by inadvertent dominance of activities of FCU. Then simultaneous recruitment of FCR and ECR lifted the wrist upward. However, as the activities of FCR exceeded that of ECR, the wrist was pulled leftward. But a sudden burst of ECU and simultaneous shut-down of FCR and FCU ignited a diddling of the wrist.

3.2 A dynamics model of the wrist joint

In order to determine causal muscle activities for movement disorders quantitatively, we approximated the relationship between the wrist joint torque calculated from the movement kinematics and the four EMG signals using a dynamics model of the wrist joint.

The equations of the wrist joint torque calculated from the wrist joint kinematics (angle, angular velocity, angular acceleration) can be decomposed into the X-axis component and Y-axis component as follows.

$$M\ddot{\theta}_x(t) + B\dot{\theta}_x(t) + K\theta_x(t) = f_x(t) \quad (1)$$

$$M\ddot{\theta}_y(t) + B\dot{\theta}_y(t) + K\theta_y(t) + mgc\cos\theta_y(t) = f_y(t) \quad (2)$$

Where, $\theta_x(t)$ and $\theta_y(t)$ represent X-axis component and Y-axis component of the wrist joint angle. $\dot{\theta}_x(t)$, $\dot{\theta}_y(t)$, $\ddot{\theta}_x(t)$ and $\ddot{\theta}_y(t)$ indicate X-axis component and Y-axis component for angular velocity and angular acceleration of the wrist joint respectively. *M* is an inertial parameter and we calculated this parameter for each subject by measuring volume of the hand. *B* and *K* represent viscous coefficient and elastic coefficient. We set these coefficients as 0.03Nms/rad and 0.2Nm/rad for the step-tracking movement, based on the previous studies (Gielen & Houk, 1984; Haruno & Wolpert 2005). *m* and *c* are the mass and center of mass for the hand, and we calculated these parameters for each subject by measuring volume of the hand. *g* is acceleration of gravity (g=9.8m/s2). $f_x(t)$ and $f_y(t)$ denote X-axis component and Y-axis component of the wrist joint torque calculated from the wrist movement.

We assumed that the wrist joint torque were proportional to the linear sum of the four EMG signals. That is, considering the pulling direction of each muscle shown in Figure 3B, the relationship between the wrist joint torque and the muscle tension of four muscles are formalized as follows:

$$a_{1x}e_1(t) + a_{2x}e_2(t) - a_{3x}e_3(t) - a_{4x}e_4(t) = g_x(t) \quad (3)$$

$$a_{1y}e_1(t) - a_{2y}e_2(t) - a_{3y}e_3(t) + a_{4y}e_4(t) = g_y(t) \quad (4)$$

Where, $e_1(t)$, $e_2(t)$, e_3 (t), and $e_4(t)$ represent the muscle tension of ECR, ECU, FCU, and FCR, respectively. $g_x(t)$ and $g_y(t)$ represent X-axis component and Y-axis component of the wrist joint torque estimated from the four muscle tensions, respectively. a_{1x}~a_{4x} ($\geqq 0$) and a_{1y}~a_{4y} ($\geqq 0$) denote the parameters for the musculo-skeletal system of the wrist joint that convert the muscle tension into the wrist joint torque. It should be noted that the sign of each parameter works as a constraint to limit the pulling direction of each muscle.

In our previous study, we calculated the parameters a_{1x}~a_{4x} and a_{1y}~a_{4y} using the simple relationship between the wrist joint torque and the muscle tension for isometric contraction (Lee et al., 2007). However, there was no guarantee that these parameters obtained for an isometric condition were suitable to estimate dynamic wrist joint torques during movement. In fact, estimation of the dynamic wrist joint torques with these parameters was relatively poor for extreme movements, such as jerky movements of the cerebellar patients. Therefore, it is desirable to introduce alternative parameters obtained for movement conditions. In this study, we directly calculated these parameters from the relationship between the wrist joint torque and the muscle tension during movement, by optimizing a match between the wrist joint torque (equation (1) and (2)) and the linear sum of four muscle tensions (equation (3) and (4)) using the least squares method.

3.3 Performance of the model

Figure 5 shows an example of the match between the wrist joint torque calculated from the wrist movement (blue line) and the linear sum of the four muscle tensions (red line) for a normal control (A) and a cerebellar patient (B). As clearly seen in Figure 5 and Table 1, there were very high correlations between the wrist joint torque and the four muscle activities for both the cerebellar patients and the normal controls (R for normal controls = 0.81±0.08 (X-axis), 0.84±0.05 (Y-axis); R for cerebellar patients = 0.81±0.09 (X-axis), 0.81±0.05 (Y-axis)). The result strongly suggested that it is possible to identify causal anomaly of the muscle activities for each abnormal movement. Therefore, it should be possible to analyze central mechanisms for generation of pathological movements at the level of the motor command with high accuracy.

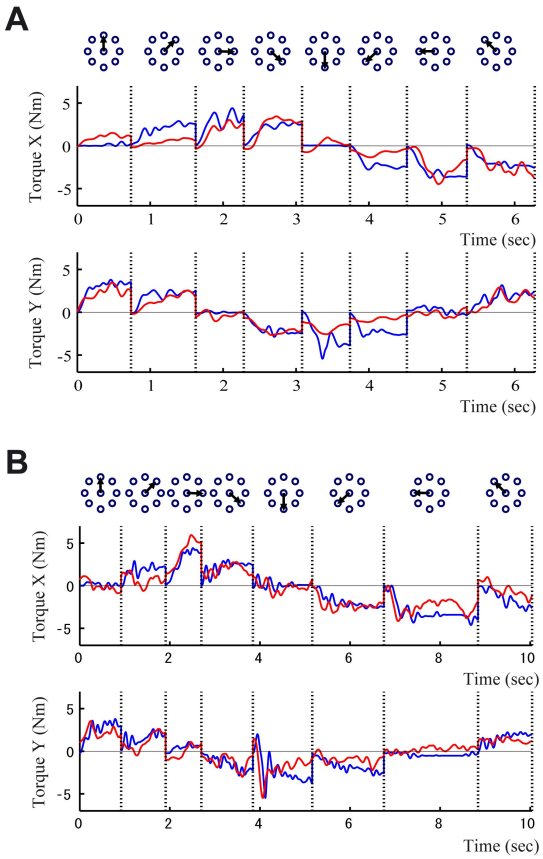


Fig. 5. Relationship between the wrist joint torque calculated from the wrist movement (blue line) and the linear sum of four muscle tensions (red line) for a normal control (A) and a cerebellar patient (B). Figures of top trace indicate the direction of wrist movement.

	Correlation R of normal control (n=8)	Correlation R of cerebellar patient (n=8)
Torque X	0.81±0.08	0.81±0.09
Torque Y	0.84±0.05	0.81±0.05

Table 1. Correlation between the wrist joint torque and the muscle activities.

3.4 Analysis of Causal Motor Commands for the Cerebellar Ataxia

In fact, we identified causal abnormality of muscle activities for cerebellar ataxia, confirming effectiveness of our method to analyze pathological movements at the level of the motor command.

Figure 6 demonstrates a typical example of one-to-one correlation between the muscle activities and the concomitant movement for the downward movement in Figure 5B. This figure summarizes relationship between the muscle activities (i.e.motor commands) and a

jerky wrist movement of a cerebellar patient for every 100msec. For instance, the initial movement (0msec) was away from the down target (i.e. upward) due to the excess activities of ECR that pulls the wrist upward. Then the wrist was redirected toward the down target due to the desirable predominance of the activities of FCU (100-300msec). However, 400msec after the onset, inadvertent activities of FCR pulled the wrist leftward, again, away from the target. In this way, it is possible with our system to determine the anomalous motor command for the cerebellar ataxia in further detail.

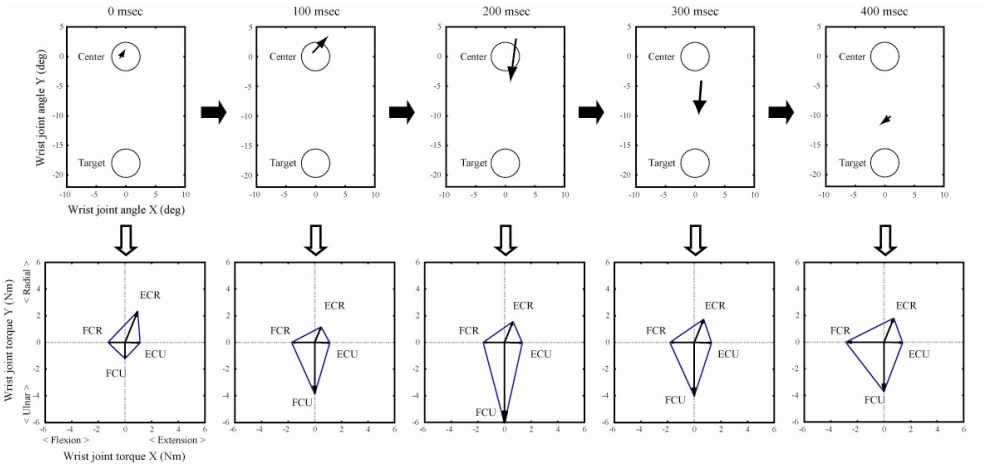


Fig. 6. Causal relationship between muscle activities and a jerky wrist movement of a cerebellar patient. Top panels show directions of the wrist movement for every 100msec. 0msec indicates the movement onset. Bottom panels show averaged activities of the four muscles for the corresponding time window. Muscle activities are represented as vectors.

4. Quantification of pathological patterns of muscle activities

Overall, it is possible to identify abnormal components of agonist selection for wrist movements by recording activities of as few as four (out of twenty-four) forearm muscles. We further extended our analysis to quantify the pathological patterns of muscle activities. In this section, we will describe two parameters that summarize variability and efficacy of muscle activities for the pursuit movement.

4.1 Parameters characterizing pathological patterns of muscle activities

To make a pursuit wrist movement, it is desirable to change muscle activities smoothly, because the target moves smoothly. On the other hand, it is also desirable to maximize contrast between activities of agonist and antagonist muscles to minimize energy consumption for a movement. As parameters characterizing the variability and the effectiveness of muscle activities, we defined "Variability of Total Contraction" (VTC) and "Directionality of Muscle Activity" (DMA) as follows. Indeed, we found these parameters were very different between control subjects and patients with neurological disorders, and therefore, were useful to quantify movement disorders.

4.1.1 Variability of Total Contraction (VTC)

VTC represents temporal variability of muscle activitites, as illustrated in Figure 6A. We first calculated amplitude of torque for each muscle using equation (5).

$$|\vec{T}_{Muscle}| = \sqrt{(a_x^{Muscle})^2 + (a_y^{Muscle})^2} \times e_{Muscle}(t) \tag{5}$$

Where, a_x^{Muscle} ($\geqq 0$) and a_y^{Muscle} ($\geqq 0$) denote the parameters for the musculo-skeletal system of the wrist joint, which convert muscle tension into the X-axis component and the Y-axis component of the wrist joint torque respectively. $e_{Muscle}(t)$ represents the muscle tension of each muscle.

$$VTC = \frac{\int \left(\sum_{Muscle=1}^{4} \left| \frac{d(|\vec{T}_{Muscle}|)}{dt} \right| \right) dt}{t} \tag{6}$$

Then, as described in equation (6), we calculated the instantaneous variability of the torque for the four muscles. Finally, the VTC was calculated by averaging the absolute value of the variation with movement duration *t* to normalize it for movement duration.

4.1.2 Directionality of Muscle Activity (DMA)

DMA was evaluted as the ratio of wrist joint torque to the total muscle torque as shown in Figure 6B and equation (8). We first calculated the wrist joint torque from four muscle activities as follows:

$$|\vec{\tau}_{EMG}| = \sqrt{(g_x(t))^2 + (g_y(t))^2} \tag{7}$$

Where, $g_x(t)$ and $g_y(t)$ represent X-axis component and Y-axis component of the wrist joint torque estimated from the four muscle tensions (see equations (3) and (4)).

$$DMA = \frac{\int \frac{|\vec{\tau}_{EMG}|}{\sum_{Muscle=1}^{4} |\vec{T}_{Muscle}|} dt}{t} \tag{8}$$

Then, as described in equation (8), we calculated the ratio of the wrist joint torque to the sum of the torque of the individual muscles, and finally, the DMA was calculated by averaging the ratio for movement duration *t* as a nomalization.

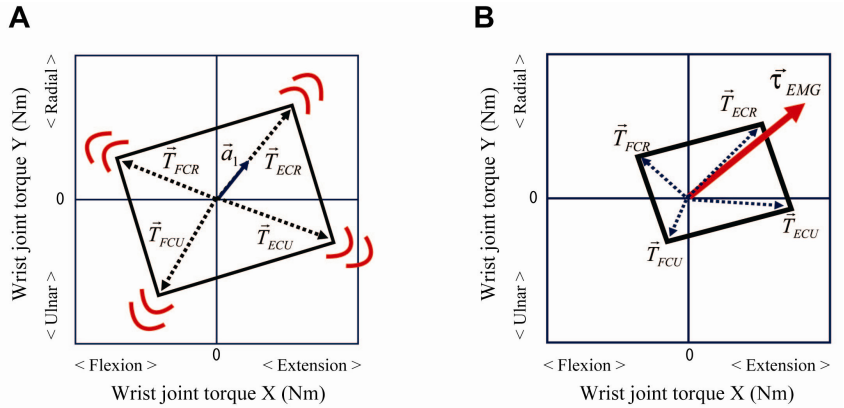


Fig. 6. Explanation of Variability of Total Contraction (VTC) (A) and Directionality of Muscle Activity (DMA) (B).

4.2 VTC and DMA in neurological disorders

In order to evaluate usefulness of VTC and DMA, we calculated these parametes for patients with cerebellar atrophy and patients with Parkinson's disease, as well as for normal control subjects. Figure 7 summarizes the results. The VTC indicates variability of muscle activities. Therefore, if there are a number of abrupt changes in the muscle activities, the VTC gets higher. For instance, in case of cerebellar patients, muscle activities keep fluctuating intensely due to the cerebellar ataxia. As a result, VTCs for the cerebellar patients tend to be higher than control subjects with much smoother muscle activities, as shown in Figure 7A. In contrast, VTCs for patients with Parkinson's disease tend to be smaller due to faint modulation of muscle activities (Figure 7A).

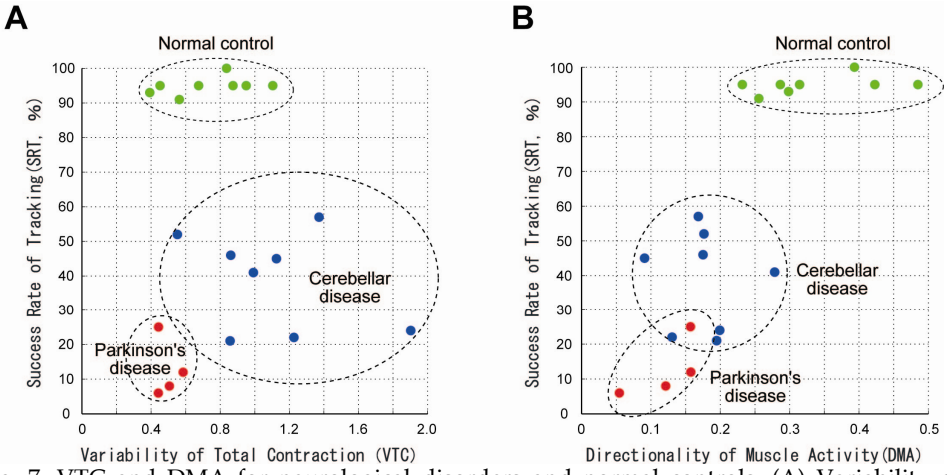


Fig. 7. VTC and DMA for neurological disorders and normal controls. (A) Variability of Total Contraction (VTC), (B) Directionality of Muscle Activity (DMA). SRT indicates the rate (%) of the cursor within the target for the pursuit movement.

The DMA represents directionality of muscle activities, and thereby indicating contrast between activities of agonist and the antagonist muscles. By definition, if agonists are activated selectively with complete suppression of antagonists, DMA gets highest. In contrast, DMA is low in case of co-contraction with comparable activities for agonists and antagonists. As a result, DMAs for cerebellar patients are usually very low due to significant co-contraction (see Figure 4B for example) as shown in Figure 7B. On the other hand, in case of patients with Parkinson's diseases, DMAs are also low due to poor modulation of agonist activities.

Overall, VTC or DMA captures characteristic patterns of the muscle activities for patients with cerebellar disorders and patients with Parkinson's disease. Moreover, it is possible to make more detailed characterization of pathological muscle activities by combining these parameters (Figure 8). If we use more useful parameters in combination with VTC and DMA, it will be possible to make more sophisticated evaluation of movement disorders in a high dimensional space of parameters that quantify patterns of muscle activities. Consequently, it could be possible to evaluate effects of newly developed treatments for neurological diseases in the parameter space.

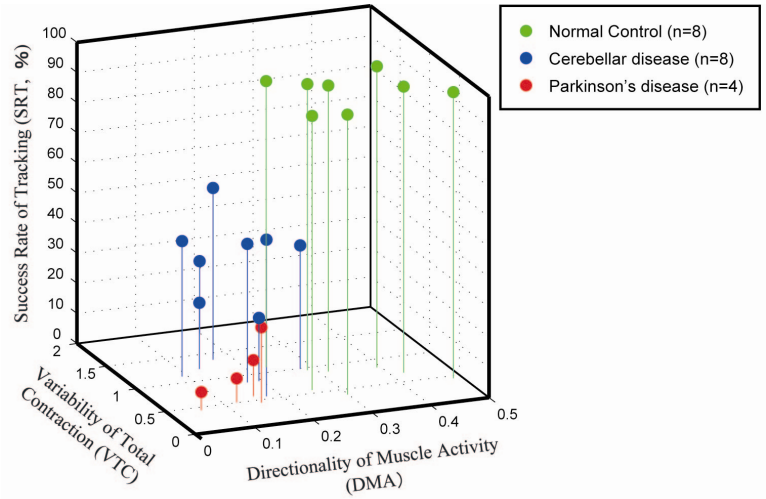


Fig. 8. Comprehensive assessment of muscle activities (i.e. motor commands) for neurological disorders and normal control. Green spheres, blue spheres and red spheres indicate normal controls, cerebellar patients and Parkinson's patients, respectively.

5. Discussion and conclusion

In this chapter, we proposed a new method to make a quantitative evaluation for movement disorders based on the EMG signals. In the following discussion, we will focus on three points: 1) Why it is essential to analyze muscle activities for evaluation of neurological disorders; 2) How effective our proposed method is. 3) Application of our proposed method.

Some researchers tried to make quantitative evaluation of the motor function for the arm movement (Nakanishi et al., 1992; Nakanishi et al., 2000; Sanguineti et al., 2003). For example, by analyzing the position, velocity and acceleration of arm during a circular movement on the digitizer, Nakanishi et al. evaluated the motor function of the arm in patients with neurological disorders including cerebellar deficits and Parkinson's disease. However, their analysis was limited to the movement kinematics. Unfortunately, the movement kinematics cannot specify its causal muscle activities due to the well-known redundancy of the musculo-skeletal system. In other words, completely different sets of muscle activities (causes) end up with the same kinematics (result). Thus, in order to understand central mechanisms for generation of pathological movements, it is essential to capture causal anomaly of the motor commands directly, rather than to observe the resultant movement indirectly (Manto, 1996; Brown et al., 1997). In addition, the movement kinematics provides no information about muscle tonus that is a crucial factor to diagnose movement disorders. Overall, it is essential to examine muscle activities to make more fundamental evaluation of neurological disorders.

In this study, we proposed a new method to identify causal muscle activities for movement disorders of the wrist joint. However, there are twenty-four muscles in the forearm that have significant effects on the wrist joint. If we had to record activities of all these muscles to reconstruct the movement kinematics, we would have to use a number of (i.e. twenty-four pairs of needle electrodes and it would take painful hours for just placing the electrodes. In this chapter, we proposed a new method to determine abnormal components of agonist selection for various wrist movements by recording activities of as few as four forearm muscles without pain. Consequently, with our proposed method, it is easy to analyze central mechanisms for generation of pathological movement. In fact, we confirmed the effectiveness of our proposed method, identifying the causal abnormality of muscle activities for the cerebellar ataxia with high accuracy.

So far, our method is limited to examine the wrist movement, rather than the whole arm. Nevertheless, the wrist joint is suitable to examine important motor functions of the arm. Basically, not only six wrist muscles but also eighteen finger muscles are relevant to control the two degrees of freedom of the wrist joint (Brand, 1985). This anatomical setup allows the wrist joint a uniquely wide variety of motor repertoires. For instance, the wrist joint plays an essential role in hand writing which requires the finest precision of all the motor repertoires. It should be emphasized that its role is not just a support for finger movements. On the other hand, the wrist is also capable to generate and/or transmit considerable torque as seen in the arm wrestling. Overall, our method is capable to examine wide range of natural or disordered movements by the wrist joint. However, in future, it is desirable to expand our method to analyze movements of any body part including the whole arm or gait.

Our proposed method is not limited to analysis of motor deficits. We will further apply this method to evaluation of rehabilitation or guidance of treatment for neurological diseases. As a first step, we examined parameters characterizing pathological patterns of muscle activities and demonstrated their usefulness to evaluate pathological muscle activities. These parameters, if combined appropriately, are useful to characterize complex patterns of muscle activities in a way easy to recognize visually. The high-dimensional parameter space

is also useful to evaluate effects of a medical treatment as a shift toward or away from the normal control in the parameter space. In other words, this system is potentially a navigation system for medical treatments based on the motor commands.

We are now preparing to use this system for evaluation and navigation of rehabilitation. We expect that an earliest sign of favorable or infavorable effects of rehabilitation emerges as subtle changes in muscle activities long before *visible* changes in movement kinematics. Our method may be also useful for evaluation of treatments currently available like the deep brain stimulation therapy or available in a near future, such as gene therapies whose targets are in the central nervous system and whose effects appear as, probably, slow renormalization of the motor commands.

6. Acknowledgement

We thank Dr. Yasuharu Koike for his invaluable advices on the approximation of the wrist joint model. We also thank Drs. Yoshiaki Tsunoda and Seaka Tomatsu for helpful discussions.

7. References

Brand, P.W. (1985). *Clinical mechanics of the hand*, Mosby, St. Louis

Brown, P.; Corcos, D.M. & Rothwell, J.C. (1997). Does parkinsonian action tremor contribute to muscle weakness in Parkinson's disease?, *Brain*, pp. 401-408

Gielen, G.L. & Houk, J.C. (1984). Nonlinear viscosity of human wrist, *Journal of Neurophysiology*, pp. 553-569

Haruno, M. & Wolpert, D.M. (2005). Optimal control of redundant muscles in step-tracking wrist movements, *J. Neurophysiol.*, pp.4244-4255

Koike, Y. & Kawato, M. (1995). Estimation of dynamic joint torques and trajectory formation from surface electromyography signals using a neural network model, *Biological Cybernetics*, pp. 291-300

Lee, J.; Kagamihara, Y. & Kakei, S. (2007). Development of a quantitative evaluation system for motor control using wrist movements—an analysis of movement disorders in patients with cerebellar diseases, *Rinsho Byori*, pp. 912-921

Mannard, A. & Stein, R. (1973). Determination of the frequency response of isometric soleus muscle in the cat using random nerve stimulation, *Journal of Physiology*, pp. 275-296

Manto, M. (1996). Pathoplysiology of Cerebellar dysmetria: The imbalance between the agonist and the antagonist electromyographic activities, *European Neurology*, pp. 333-337

Nakanishi, R.; Yamanaga, H.; Okumura, C.; Murayama, N. & Ideta, T. (1992). A quantitative analysis of ataxia in the upper limbs, *Clinical neurology*, pp. 251-258

Nakanishi, R.; Murayama, N.; Uwatoko, F.; Igasaki, T. & Yamanaga, H. (2000). Quantitative analysis of voluntary movements in the upper limbs of patients with Parkinson's disease, *Clinical Neurophysiology*, Vol. 18, pp. 37-45

Sanguineti, V.; Morassoa, P.G.; Barattob, L.; Brichettoc, G.; Mancardic, G.L. & Solaro, C. (2003). Cerebellar ataxia: Quantitative assessment and cybernetic interpretation, *Human Movement Science*, pp. 189-205

4

Source Separation and Identification issues in bio signals: A solution using Blind source separation

Ganesh R Naik and Dinesh K Kumar
School of Electrical and Computer Engineering, RMIT University
Melbourne, Australia

1. Introduction

The problem of source separation is an inductive inference problem. There is not enough information to deduce the solution, so one must use any available information to infer the most probable solution. The aim is to process these observations in such a way that the original source signals are extracted by the adaptive system. The problem of separating and estimating the original source waveforms from the sensor array, without knowing the transmission channel characteristics and the source can be briefly expressed as problems related to blind source separation (BSS). Independent component analysis (ICA) is one of the widely used BSS techniques for revealing hidden factors that underlie sets of random variables, measurements, or signals. ICA is essentially a method for extracting individual signals from mixtures of signals. Its power resides in the physical assumptions that the different physical processes generate unrelated signals. The simple and generic nature of this assumption ensures that ICA is being successfully applied in diverse range of research fields.

Source separation and identification can be used in a variety of signal processing applications, ranging from speech processing to medical image analysis. The separation of a superposition of multiple signals is accomplished by taking into account the structure of the mixing process and by making assumptions about the sources. When the information about the mixing process and sources is limited, the problem is called "blind". ICA is a technique suitable for blind source separation - to separate signals from different sources from the mixture. ICA is a method for finding underlying factors or components from multidimensional (multivariate) statistical data or signals (Hyvarinen et al., 2001; Hyvarinen and Oja, 2000).

ICA builds a generative model for the measured multivariate data, in which the data are assumed to be linear or nonlinear mixtures of some unknown hidden variables (sources); the mixing system is also unknown. In order to overcome the under determination of the algorithm, it is assumed that the hidden sources have the properties of non-Gaussianity and statistical independence. These sources are named Independent Components (ICs). ICA

algorithms have been considered to be information theory based unsupervised learning rules. Given a set of multidimensional observations, which are assumed to be linear mixtures of unknown independent sources through an unknown mixing source, an ICA algorithm performs a search of the unmixing matrix by which observations can be linearly translated to form independent output components. When regarding ICA, the basic framework for most researchers has been to assume that the mixing is instantaneous and linear, as in Infomax. ICA is often described as an extension to Principal Component Analysis (PCA) that uncorrelates the signals for higher order moments and produces a non-orthogonal basis. More complex models assume for example, noisy mixtures (Hansen, 2000; Mackay, 1996), nontrivial source distributions (Kaban, 2000; Sorenson, 2002), convolutive mixtures (Attias and Schreiner, 1998; Lee, 1997), time dependency, underdetermined sources (Hyvarinen et al., 1999; Lewicki and Sejnowski, 2000), mixture and classification of independent component (Kolenda, 2000; Lee et al., 1999). A general introduction and overview can be found in (Lee, 1998).

2. Challenges of source separation in Bio signal processing

In biomedical data processing, the aim is to extract clinically, biochemically or pharmaceutically relevant information (e.g metabolite concentrations in the brain) in terms of parameters out of low quality measurements in order to enable an improved medical diagnosis (Niedermeyer and Da Silva, 1999; Rajapakse et al., 2002). Typically, biomedical data are affected by large measurement errors, largely due to the noninvasive nature of the measurement process or the severe constraints to keep the input signal as low as possible for safety and bio-ethical reasons. Accurate and automated quantification of this information requires an ingenious combination of the following four issues:

- An adequate pre-treatment of the data,
- The design of an appropriate model and model validation,
- A fast and numerically robust model parameter quantification method and
- An extensive evaluation and performance study, using in-vivo and patient data, up to the embedding of the advanced tools into user friendly user interfaces to be used by clinicians

A great challenge in biomedical engineering is to non-invasively asses the physiological changes occurring in different internal organs of the human body. These variations can be modeled and measured often as biomedical source signals that indicate the function or malfunction of various physiological systems. To extract the relevant information for diagnosis and therapy, expert knowledge in medicine and engineering is also required.

Biomedical source signals are usually weak, geostationary signals and distorted by noise and interference. Moreover, they are usually mutually superimposed. Besides classical signal analysis tools (such as adaptive supervised filtering, parametric or non parametric spectral estimation, time frequency analysis, and higher order statistics), Intelligent Blind Signal Processing (IBSP) techniques can be used for pre-processing, noise and artefact reduction, enhancement, detection and estimation of biomedical signals by taking into account their spatio-temporal correlation and mutual statistical dependence.

Exemplary ICA applications in biomedical problems include the following:

- Fetal Electrocardiogram extraction, i.e removing/filtering maternal electrocardiogram signals and noise from fetal electrocardiogram signals (Niedermeyer and Da Silva, 1999; Rajapakse et al., 2002).
- Enhancement of low level Electrocardiogram components (Niedermeyer and Da Silva, 1999; Rajapakse et al., 2002)
- Separation of transplanted heart signals from residual original heart signals (Wisbeck et al., 1998)
- Separation of low level myoelectric muscle activities to identify various gestures (Calinon and Billard, 2005; Kato et al., 2006; Naik et al., 2006, 2007)

One successful and promising application domain of blind signal processing includes those biomedical signals acquired using multi-electrode devices: Electrocardiography (ECG) (Niedermeyer and Da Silva, 1999; Rajapakse et al., 2002; Scherg and Von Cramon, 1985; Wisbeck et al., 1998), Electroencephalography (EEG)(Niedermeyer and Da Silva, 1999; Rajapakse et al., 2002; Vig´ario et al., 2000;Wisbeck et al., 1998), Magnetoencephalography (MEG) (H¨am¨al¨ainen et al., 1993; Mosher et al., 1992; Parra et al., 2004; Petersen et al., 2000; Tang and Pearlmutter, 2003; Vig´ario et al., 2000) and Surface Electromyography (sEMG). sEMG is an indicator of muscle activity and related to body movement and posture. It has major applications in biosignal processing; next section explains sEMG and its applications.

3. BSS and Surface Electromyography

Surface EMG is the electrical recording of the spatial and temporal integration of the Motor Unit Action Potential (MUAP) originating from different motor units. It can be recorded non-invasively and used for dynamic measurement of muscular function. It is typically the only in vivo functional examination of muscle activity used in the clinical environment. The signal contains the information that is related to the anatomy and physiology of the muscle. In clinical application, the signal is used for the diagnosis of neuro-muscular disease or disorder. Another application of sEMG is for device control application where the signal is used for controlling devices such as prosthetic devices, robots, and human-machine interface. sEMG is a quick and easy process that facilitates sampling of a large number of MUAPs (Basmajian and Deluca, 1985; Enderle et al., 2005). In sEMG recordings multiple sensors are used to record some physiological phenomena. Often these sensors are located close to each other, so that they simultaneously record signals that are highly correlated with each other. Therefore, the sensors not only record the muscle activity transmitted by volume conduction from a few dynamic muscles but also from artificial signals, such as noise independent of muscle activities, that overlap with actual muscle activity which may be present in all sensors. Extraction of the useful information from such kind of sEMG becomes more difficult for low level of contraction mainly due to the low signal-to-noise ratio. At low level of contraction, sEMG activity is hardly discernible from the background activity. Therefore to correctly identify the number of individual muscles (sources) sEMG needs to be decomposed. There is little or no prior information of the muscle activity, and the signals have temporal and spectral overlap, making the problem suitable for BSS (James and Hesse, 2005;Jung et al., 2000). ICA is a statistical technique for obtaining independent sources, s from their linear mixtures, x when neither the original sources nor the actual

mixing matrix, A are unknown. This is achieved by exploiting higher order signal statistics and optimization techniques.

For independent component analysis we assume that the observed signals x consists of n underlying sources $s=(s_1,s_2,\ldots,s_n)$, that are unknown, but mutually statistically independent and that these sources were mixed by an unknown (linear) mixing process A

$$x = As(t) \tag{1}$$

with $x=(x_1,x_2,\ldots,x_n)$, $m>n$ where each component s_i has zero mean. The crucial assumption is statistical independence of these source components, which can be expressed mathematically by the joint probability density function as

$$p(s_1,s_2,\ldots,s_n)=\prod_{i=1}^{n} p_i(s_i) \tag{2}$$

Given these assumptions it is possible to separate the recorded data x *through* the linear transformation

$$u(t)=Wx(t) \tag{3}$$

into independent components by applying statistical independence on the output u of this un mixing process and recover the original sources from the observed mixtures. Here both the mixing matrix A and the sources s are unknown, therefore these techniques are called *blind source separation* (Hyvarinen et al., 2001; Hyvarinen and Oja, 2000). The block diagram approach of ICA for source separation is shown in figure 1.

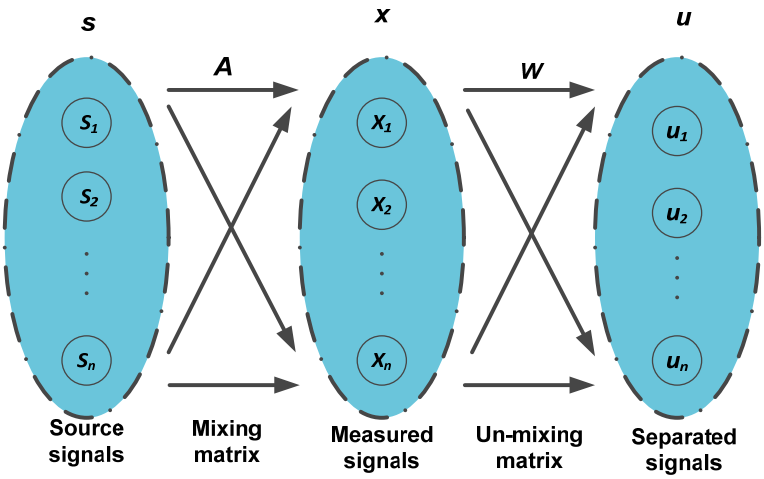


Fig. 1. Flowchart of the Independent Component Analysis (ICA). Here $s(t)$ are the sources. $x(t)$ are the mixtures, A is mixing matrix and W is un-mixing matrix.

As mentioned above, the signals that can be separated need to be non-Gaussian and independent. For the purpose of applying ICA to sEMG recordings, signals can be considered as independent and non-Gaussian and the mixing matrix can be considered to be stationary and linear. Hence this paper analyses the conditions using a stationary mixing matrix.

3.1 Source separation of sEMG

MUAP separation is a new biomedical application of ICA. In previous applications of ICA to sEMG, researchers have treated the sEMG activity from entire muscles as ICs. Each muscle contains up to 100 individual motor units and the sEMG activity from an entire muscle is the superposition of the activity from each motor unit within the muscle. It has been shown that it is possible to apply ICA to isolate sEMG signals from individual muscles (Azzerboni et al., 2002; Mckeown et al., 2002). Treating sEMG activity from entire muscles as ICs is useful in some applications, especially when studying muscle activity in performing movements. For example, ICA has been used to determine the exact sequence of muscle contractions in swallowing by McKeown et al. (Mckeown et al., 2002) in order to diagnose dysphagia (disorder of swallowing). The focus on treating sEMG activity from entire muscles as ICs arises from a desire to analyse human movement. The most important application of sEMG is as a clinical tool for neuromuscular disease diagnosis. In clinical applications physicians seek to analyse individual motor units. BSS techniques such as ICA is proposed as a novel approach for isolating individual MUAPs from sEMG interference patterns by treating individual motor units as independent sources. This is relevant to clinical sEMG as motor unit crosstalk can make it difficult to study individual MUAPs (Kimura, 2001).

During the sEMG recordings of the digitas muscles to identify the hand gestures for human computer interface, the cross talk due to the different muscles can result in unreliable recordings. The simplest and most commonly used method to improve the quality of the recording is rejection (Barlow, 1979). This is done by discarding a section of the recording that has artefact exceeding a threshold. This method is simple, but causes a significant loss of data and its reliability is questionable since it is predominantly based on visual examination. There is little safeguard that prevents the removal of some small but important features of the signal. It is also very dependent on the technician making it less dependable, and very expensive.

The other commonly used techniques to improve the quality of bio signals recordings include spectral filtering, gating and cross-correlation subtraction (Bartolo et al., 1996). Spectral filtering is often not useful due to the overlap of the frequency spectrum of the desired signals and the artefact component. On the other hand, gating and subtraction may introduce discontinuity in the reconstructed signal. In the recent past, techniques such as time domain (Hillyard and Galambos, 1970; Verleger et al., 1982), and frequency domain regression (Whitton et al., 1978; Woestenburg et al., 1983), have been attempted. However, simple regression in time domain can over-compensate the artefacts (Peters, 1967; Weerts and Lang, 1973). The regression techniques depend on the availability of a good regressing channel - a separate channel to record the corresponding artefact as a reference. This is often not possible when recording sEMG. Therefore, better artefact removal techniques are necessary to overcome the disadvantages of the previous methods. One property of the

sEMG is that the signal originating from one muscle can generally be considered to be independent of other bioelectric signals such as ECG, EOG, and signals from neighbouring muscles. This opens an opportunity of the use of ICA for this application.

A number of researchers have reported the use of ICA for separating the desired sEMG from the artefacts and from sEMG from other muscles. While details differ, the basic technique is that different channels of sEMG recordings are the input of ICA algorithm. The outputs of ICA are the ICs and the estimated unmixing matrix W. He et al. (He et al., 2006) have used ICA to remove ECG artefact from sEMG data. A variation of the same has been attempted by the Djuwari et al. (Djuwari et al., 2003), for removing ECG artefact from sEMG of the lumbar muscles. They attempted to overcome the limitation of the number of signals to be equal to the number of recordings and remove the ambiguity of the order. Their work utilized ICA in two sequential steps. In the first step, ICA with multichannel sEMG recordings that was corrupted with ECG artefact as the input gave one pure ECG signal in one of its row. In the next step, vector z found by concatenating the row of the output matrix $u = Wx$ contained the ECG artefact and each single row of x in turn was used as its input. The output of this step is a matrix $y = Bz$ that contains ECG artefact in row and the 'cleaned' sEMG of corresponding channel in its other row. While in both cases, the visual inspection suggested the successful removal of the artefact, and statistical analysis seem to suggest an improvement compared to other techniques, because of the unknown properties of the signal, the quality of the signal before and after could not be compared in a better way. Similar work is also reported by Yong et al. (Hu et al., 2007) where ICA has been employed to filter the sEMG of the lumbar muscles. Azzerboni et al. (Azzerboni et al., 2004) demonstrated the artefacts removal in sEMG using ICA and Discrete Wavelet Transform (DWT). ICA has also been used by Nakamura et al. (Nakamura et al., 2004), to decompose the sEMG recordings in terms of the MUAPs. In their paper, they have acknowledged the drawbacks and the necessary conditions required for the success of the ICA, but have not demonstrated the suitability of their experimental data for ICA application. The earlier work done by the researchers have mainly focused on sEMG source separation and identification. However further source separation issues need to be investigated.

3.2 Validity of the basic ICA model for sEMG applications

The application of ICA to the study of sEMG and other bio signals assumes that several conditions are verified, at least approximately: the existence of statistically independent source signals, their instantaneous linear mixing at the sensors, and the stationarity of the mixing and the ICs. The independence criterion considers solely the statistical relations between the amplitude distributions of the signals involved, and not the morphology or physiology of neural structures. Thus, its validity depends on the experimental situation, and cannot be considered in general. There are however, two other practical issues that must be considered:

1. Firstly, to ensure that the mixing matrix is constant the sources must be fixed in space (this was an implied assumption as only the case of a constant mixing matrix was considered). This is satisfied by sEMG as motor units are in fixed physical locations within a muscle, and in this sense applying ICA to sEMG is much simpler than in other biomedical signal processing applications such as EEG or fMRI in which the sources can move (Jung et al., 2001).

2. Secondly, in order to use ICA it is essential to assume that signal propagation time is negligible. Signals from Gaussian sources cannot be separated from their mixtures using ICA (Mckeown et al., 1999) because Gaussianity is a measure of independence. Mathematical manipulation demonstrates that all matrices will transform this kind of mixtures to another Gaussian data. However, a small deviation of density function from Gaussian may make it suitable as it will provide some possible maximization points on the ICA optimization landscape, making Gaussianity based cost function suitable for iteration. If one of the sources has density far from Gaussian, ICA will easily detect this source because it will have a higher measure of non Gaussianity and the maximum point on the optimization landscape will be higher. If more than one of the independent sources has non Gaussian distribution, those with higher magnitude will have the highest maximum point in the optimization landscape.

Given a few signals with distinctive density and significant magnitude difference, the densities of their linear combinations will tend to follow the ones with higher amplitude. Since ICA uses density estimation of a signal, the components with dominant density will be found easily. The fundamental principle of ICA is to determine the unmixing matrix and use that to separate the mixture into the ICs. The ICs are computed from the linear combination of the recorded data. The success of ICA to separate the independent components from the mixture depends on the properties of the recordings. However there are few issues involved in ICA for sEMG applications. Three main problems that need to be addressed:

- issue related to identifying dependency and independency nature of the
- sources
- order of the separated signals and
- normalisation of the estimated ICs

This research proposes the imposition of sEMG conditions on ICA to overcome these limitations, resulting in semi-blind ICA. In order to validate the above mentioned theory two types of sEMG (bio signals) were analysed. First one is to identification of various complex gestures based on decomposition of myo electric signal and the second one is to identification of different vowel utterances based on facial sEMG signals. The experimental methodology, results and discussion related to above mentioned experiments are explained next.

4. Methodology

Experiments were conducted to evaluate the performance of the hand gesture recognition and facial muscle activity using surface EMG. Experiments were performed to determine the reliability of the use of facial sEMG to identify the unspoken vowel of an individual. The study focused on inter-experimental variations, to determine whether the person repeated the same set of muscle activation strategies for the same speech patterns. This was done with the aim of determining the reliability of the use of facial sEMG for identifying unspoken vowels, and for human computer interface. It was also done to establish whether normal people speak with the same muscle activation strategy.

4.1 Hand gesture sEMG and Facial sEMG recording procedure

For the hand gesture experiments five subjects whose ages ranging from 21 to 32 years (four males and one female) were chosen. The experiments were conducted on two different days on all five subjects. For the data acquisition a proprietary surface EMG acquisition system by Delsys (Boston, MA, USA) was used. Four electrode channels were placed over four different muscles as indicated in the table 1 and figure 2. A reference electrode was placed at Epicondylus Medialis.

Channel	Muscle	Function
1	Brachioradialis	Flexion of forearm
2	Flexor Carpi radialis (FCR)	Abduction and flexion of wrist
3	Flexor Carpi Ulnaris (FCU)	Adduction and flexion of wrist
4	Flexor digitorum superficialis (FDS)	Finger flexion while avoiding wrist flexion

Table 1. Placement of electrodes over the skin of the forearm

Before placing the electrodes subject's skin was prepared by lightly abrading with skin exfoliate to remove dead skin that helps in reducing the skin impedance to less than 60 Kilo Ohm. Skin was also cleaned with 70% v/v alcohol swab to remove any oil or dust on the skin surface. The experiments were repeated on two different days. Subject was asked to keep the forearm resting on the table with elbow at an angle of 90 degree in a comfortable position. Three hand actions were performed and repeated 12 times at each instance. Each time raw signal sampled at 1024 samples/second was recorded. A suitable resting time was given between each experiment. There was no external load. The gesture used for the experiments are listed below and details have been provided in table 1:

- Wrist flexion (without flexing the fingers).
- Finger flexion (ring finger and the middle finger together without any wrist flexion).
- Finger and wrist flexion together but normal along centre line

The hand actions and gestures represented low level of muscle activity. The hand actions were selected based on small variations between the muscle activities of the different digitas muscles situated in the forearm. The recordings were separated using ICA to separate activity originating from different muscles and used to classify against the hand actions. Experiments were conducted on the single subject on two different days to test the inter day variations. A male subject is participated in the experiment. The experiment used 4 channel EMG configurations as per the recommended recording guidelines (Fridlund and Cacioppo, 1986). A four channel, portable, continuous recording MEGAWIN equipment (from MEGA Electronics, Finland) was used for this purpose. Raw signal sampled at 2000 samples/ second was recorded. Prior to the recording, the male participant was requested to shave his facial hair. The target sites were cleaned with alcohol wet swabs. Ag/AgCl electrodes (AMBU Blue sensors from MEDICOTEST, Denmark) were mounted on appropriate locations close to the selected facial muscles: the right side *Zygomaticus Major, Masseter & Mentalis* and left side *Depressor anguli oris*. The inter electrode distance was kept constant at 1cm for all the channels and the experiments.

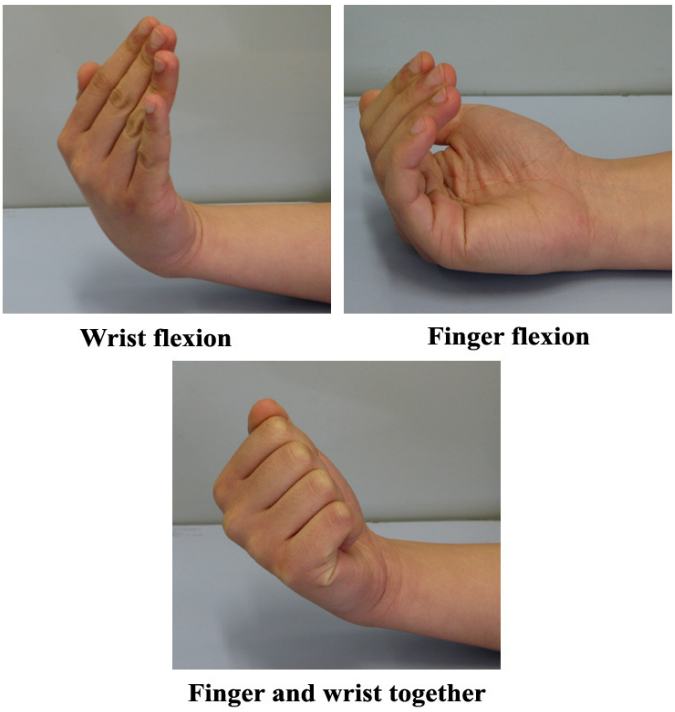


Fig. 2. Three hand gestures during the hand gesture experiment

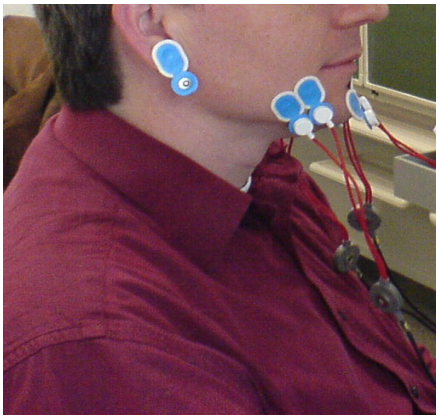


Fig. 3. Facial vowel utterance during the experiment

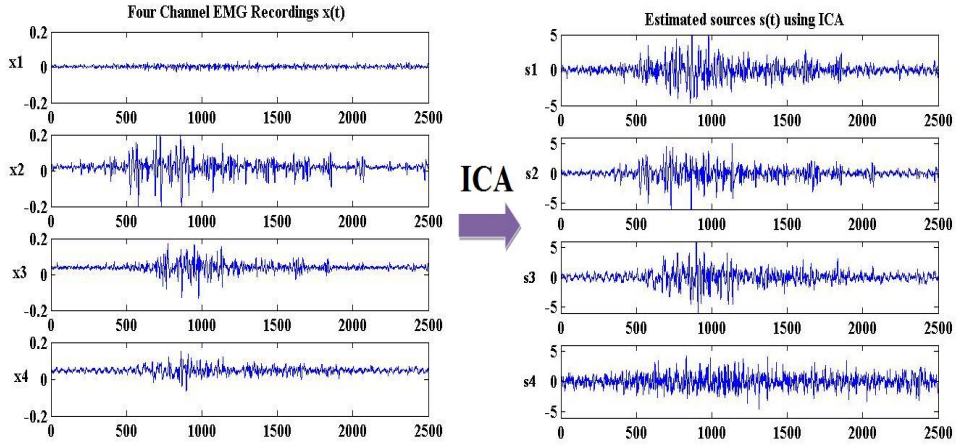


Fig. 4. Estimated four channel source signals $s(t)$ from a four channel Hand sEMG recording $x(t)$ for one of the hand gesture actions using fast ICA algorithm

Controlled experiments were conducted where the subject was asked to speak 5 English vowels (/a/, /e/, /i, /o, /u/). Each vowel was spoken separately such that there was a clear start and ends of the utterance. During this utterance, facial sEMG from the muscles was recorded. sEMG from four channels were recorded simultaneously. The recordings were visually observed, and the recordings with any artifacts -typically due to loose electrodes or movement, were discarded. The experiment was repeated for ten times. A suitable resting time was given between each experiment.

4.2 Data Analysis

The aim of these experiments were to test the use of ICA along with known properties of the muscles for separation of sEMG signals for the purpose of identifying hand gestures and to test the use of ICA on the facial sEMG signals for identifying speakers. Similar data analysis was performed to test the reliability of the ICA on facial sEMG and hand gesture sEMG.

For hand gesture actions each experiment was repeated 12 times and each experiment lasted approximately 2.5 seconds. The sampling rate was 1024 samples per second. There were four channel (recordings) electrodes and four active muscles associated with the hand gesture, forming a square 4×4 mixing matrix. For facial muscle experiments, there were approximately 5000 samples of the data for each utterance of vowels (a/e/i/o/u). 10 set of these recording were considered. Since there were four channel recordings electrodes and four active muscles associated with each utterance of vowel, this formed 4×4 mixing matrix

For both experimental datasets, the sEMG recordings were separated using fast ICA matlab algorithm which has been developed and proposed by the team at the Helsinki University of Technology (Hyvarinen and Oja 1997). The mixing matrix A was computed for the first set of data only. The independent sources of motor unit action potentials that mix to make the EMG recordings were computed using the following.

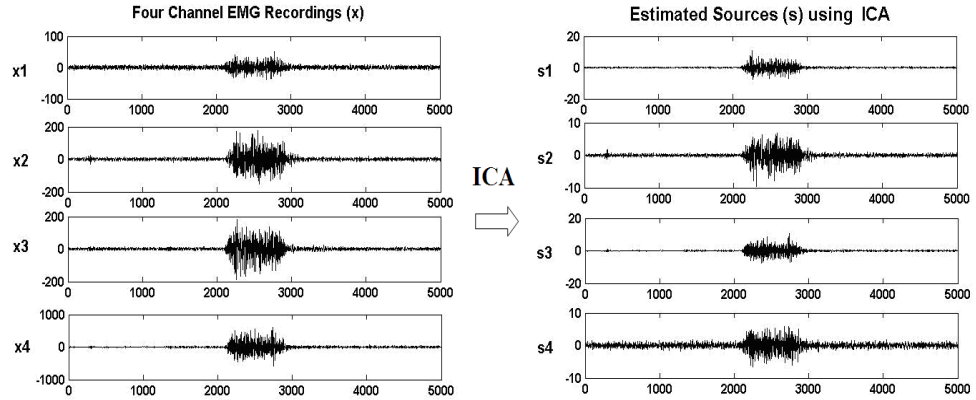


Fig. 5. Estimated four channel source signals $s(t)$ from a four channel Facial sEMG recording $x(t)$ for one of the hand gesture actions using fast ICA algorithm

$$s(t)=Wx(t) \tag{4}$$

where, W is the inverse of the mixing matrix A. This process was repeated for each of the three hand gesture experiments. Four sources were estimated for each experiment. Samples of four channels of muscle activity for hand gesture and facial muscle activity, after source separation using Fast ICA are shown in figures 4 and 5. After separating the four sources *sa*, *sb, sc* and *sd,* each of these was segmented to sample length. Root Mean Squares (RMS) was computed for each separated sources using the following.

$$S_{rms} = \sqrt{\frac{1}{N}\sum_{i=1}^{n} s_i^2} \tag{5}$$

where s is the source and N is the number of samples. This results in one number representing the muscle activity for each channel for each hand action and muscle activity for facial muscle. Our analysis demonstrates that this is a simple yet very efficient measure of the muscle activity when the muscle activity for each of the muscles has been separated from the sEMG recordings.

RMS value of muscle activity of each source represents the muscle activity of that muscle and is indicative of the force of contraction generated by each muscle. Taking a ratio of these activities gives a relative combination of the activity from each of these muscles responsible for the muscle activity. A constant mixing matrix A and set of weight matrix for neural networks was used for each subject making the system configured for each individual.

The above process was repeated for all three different hand actions 12 times and for each of the participants. The process had been repeated for the facial muscle sEMG for the five vowels (a/e/i/o/u). The outcome of this was 10 set of examples, each example pertaining to speech of five vowels. These results were used further for neural network analysis.

4.3 Neural Network analysis

As a first step, the networks were trained using the randomly chosen training data. Performances were also monitored during the training phase in order to prevent overtraining of the network. The similar ANN architecture was used to test the reliability of Hand gesture sEMG and facial sEMG. The ANN consisted of two hidden layers with a total of 20 nodes. Sigmoid function was the threshold function and the type of training algorithm for the ANN was gradient descent and adaptive learning with momentum with a learning rate of 0.05 to reduce chances of local minima.

The systems were tested using data that was not the training data. During testing, the ANN with weight matrix generated during training was used to classify RMS of the muscle activity separated using un-mixing matrix generated during training. The ability of the network to correctly classify the inputs against known data's was used to determine the efficacy of the technique.

For hand gesture actions 12 sets of examples were used to train a back-propagation neural network. The inputs to the network were the 4 RMS values for each gesture and the output of the network were the three gestures. A back propagation neural network was then trained with the RMS values as the inputs and the gesture numbers as the targets. This network was then tested for the test data. For facial sEMG we used 10 sets with 4 inputs and 3 outputs by taking different combinations of vowels (a/i/u), (i/o/u), (a/o/u), (e/i/u) etc. The inputs to the network were the 4 RMS values for each vowel utterance and the output of the network were the three vowels. Similar to the hand gesture analysis, a back propagation neural network was trained with the RMS values as the inputs and the vowel utterance numbers as the targets. This network was then tested for the test data.

5. Results and observations

The aim of this research was to test the reliability and to determine the efficacy of the semi-blind ICA technique to decompose sEMG into muscle activity from individual muscles and classify the activity from these muscles to identify the hand gestures and speakers. The ability of the system to accurately classify the decomposed sEMG against the known hand gestures has been tabulated in table 2. For comparative purposes and to evaluate the ability of the system, the classification of different vowels has been tabulated in table 3.

5.1 Hand gesture Identification using decomposed sEMG

This is the result of classification of RMS of the decomposed sEMG using ICA generated un-mixing matrix from the training data and classified using a neural network trained with the help of the training data. The accuracy was computed based on the percentage of correct classified data points to the total number of data points. These results indicate an over all classification accuracy of 100% for all the experiments. The results demonstrate that this technique can be used for the classification of different hand gesture actions when the muscle activity is low. The results also indicate that the system is resilient to differences in subjects and inter-day variations.

Number of participants	Wrist flexion		Finger Flexion		Finger flexion and wrist flexion	
	Day one	Day two	Day one	Day two	Day one	Day two
Subject 1	100%	100%	100%	100%	100%	100%
Subject 2	100%	100%	100%	100%	100%	100%
Subject 3	100%	100%	100%	100%	100%	100%
Subject 4	100%	100%	100%	100%	100%	100%
Subject 5	100%	100%	100%	100%	100%	100%

Table 2. Experimental results for Hand Gesture Identification using muscle activity separated from sEMG using ICA

5.2 Vowel classification using decomposed sEMG

The result of the experiment demonstrates the performance of the subject for different days in classifying the RMS values of the 3 vowels.

	Correctly Classified Vowels			Correctly Classified Vowels	
	Day 1	Day 2		Day 1	Day 2
/a/	(60%)	(60%)	/e/	(60%)	(60%)
/o/	(55%)	(65%)	/i/	(55%)	(65%)
/u/	(65%)	(60%)	/u/	(65%)	(60%)

Table 3. Experimental results for vowel classification using muscle activity separated from facial sEMG using ICA

The result of the use of these RMS values to train the ANN using data from individual subjects showed easy convergence. The results of testing the ANN to correctly classify the test data based on the weight matrix generated using the training data is tabulated in table 3 for two different set of vowels. The accuracy was computed based on the percentage of correct classified data points to the total number of data points. The results indicate an overall average accuracy of about 60%.

5.3 Comparative evaluation of hand sEMG with facial sEMG applications

Independent Component Analysis with back propagation neural network was successfully classified the hand gesture surface EMG signals. To measure the efficiency of ICA for source separation, similar analysis was performed on facial sEMG signals. In order to measure the quality of the separation of hand gesture muscle activities in comparison to facial muscle activities, we used the mixing matrix analysis. The surface EMG signals (wide-band source signals) are a linear decomposition of several narrow-band sub components: $s(t) = s_1(t) + s_2(t) + s_3(t) + \ldots + s_n(t)$ where $s_1(t)$, $s_2(t),\ldots, s_n(t)$ are 2500 samples in length each, which are obtained from the recorded signals $x_1(t)$, $x_2(t),\ldots, x_n(t)$ by using ICA. Such decomposition can be modeled in the time, frequency or time frequency domains using any suitable linear transform. We obtain a set of un-mixing or separating matrices: $W_1, W_2, \ldots, W_n$ where W_1 is the un-mixing matrix for sEMG sensor data $x_1(t)$ and W_n is the

un-mixing matrix for sEMG sensor data $x_n(t)$. If the specific sub-components of interest are mutually independent for at least two sub-bands, or more generally two subsets of multi-band, say for the sub band "*p*" and sub band "*q*", then the global matrix

$$G_{pq} = W_p \times W_q^{-1} = P \tag{6}$$

will be a sparse generalized permutation matrix *P* with special structure and only one non-zero (or strongly dominating) element in each row and each column (Cichocki and Amari, 2003). This follows from the simple mathematical observation that in such case both matrices W_p and W_q represents pseudo-inverses (or true inverse in the case of square matrix) of the same true mixing matrix *A* (ignoring non-essential and unavoidable arbitrary scaling and permutation of the columns) and by making an assumption that sources for two multi-frequency sub-bands are independent (Cichocki and Amari, 2003). The above assumption is applied for different hand gestures, and some convincing results were derived, which demonstrate that ICA is clearly able to isolate the four independent sources from hand muscle sEMG recordings. The results of two un-mixing matrices which are obtained from one of the hand gesture are given below, which satisfies the equation (6):

$$G = W_1 * W_2^{-1} = \begin{bmatrix} 0.0800 & \mathbf{-1.0094} & 0.0271 & 0.0927 \\ 0.0670 & -0.0046 & 0.0307 & \mathbf{-1.2610} \\ 0.0143 & 0.0295 & \mathbf{0.8062} & 0.0273 \\ \mathbf{2.1595} & 0.3787 & -0.0729 & 0.0686 \end{bmatrix}$$

Determinant (G) = 2.2588

In this example the dominant values in each row (ICA does have order and sign ambiguity, hence only absolute values will be taken into consideration) demonstrate that ICA is able to isolate the four sources (*s1, s2, s3* and *s4)* from four sEMG recordings (*x1, x2, x3* and *x4)* successfully. To justify this hypothesis, the determinant of the matrix *G* was computed. From the mathematical point of view, *n* vectors in R_n are linearly dependent if and only if the determinant of the matrix formed by the vectors is zero (Meyer, 2000). In each instance, results which are higher than one were obtained. These results clearly justified that ICA is able to isolate four independent sources from the four channel hand muscle recordings.

Similar analyses were performed on facial muscles: Four sources (*s1, s2, s3* and *s4)* were decomposed from four recordings (*x1, x2, x3* and *x4)* using fastICA algorithm. In order to check the quality of the source separation, the global matrices for each narrow-band components was computed. The following results show one of the examples of facial sEMG signals, which also satisfy the equation (6).

$$G = W_1 * W_2^{-1} = \begin{bmatrix} 0.0485 & \mathbf{-1.1738} & 0.0891 & \mathbf{-1.1105} \\ \mathbf{-0.8019} & \mathbf{1.0171} & \mathbf{0.7873} & 0.1669 \\ \mathbf{-0.8377} & 0.0142 & \mathbf{1.1837} & \mathbf{-1.0169} \\ \mathbf{-1.4905} & 0.0192 & \mathbf{-1.3557} & 0.4750 \end{bmatrix}$$

Determinant (G) = 0.0013 (Which is very close to Zero)

By inspecting the above matrix we are certain that the values are dependent (sources are dependent), cause in each row there are more than one dominant value. To clarify this we computed the determinant of the global matrix *G* and the result are very close to zero which from matrix theory explains that the sources are dependent (Meyer, 2000).

The above analysis demonstrates the importance of mixing matrix analysis for source separation and identification of surface EMG signals. For the results it is evident that the above analysis could be used as a pre-requisite tool to measure the reliability of SEMG-based systems, especially those classifying recorded such bio-signals.

6. Discussion

The results demonstrated the applications and limitations of ICA for Hand gesture sEMG and facial sEMG. Similar data analysis on both hand gesture sEMG and Facial sEMG has helped to verify the reliability of ICA.

6.1 Applications

In this chapter, a new system to classify small level of muscle activity to identify hand gesture using a combination of independent component analysis (ICA), known anatomy and neural network configured for the individual has been proposed. It has been tested with 5 volunteer participants and the experiments were repeated on different days. The results indicate the ability of the system to perfectly recognise the hand gesture even though the muscle activity was very low and there were number of muscles simultaneously active for each of the gesture.

There are number of researchers who have reported attempts to identify hand and body gestures from sEMG recordings but with low reliability. This may be attributed to low signal to noise ratio and large cross-talk between different simultaneously active muscles. ICA is a recently developed signal processing and source separation tool and has been employed to separate the muscle activity and remove artefacts to overcome this difficulty. While ICA has been extremely useful for audio based source separation, its application for sEMG is questionable due to the random order of the separated signals and magnitude normalisation. This paper reports research that overcomes this shortcoming by using prior knowledge of the anatomy of muscles along with blind source separation. Using a combination of the model and ICA approaches with a neural network configured for the individual overcomes the order and magnitude ambiguity. The results indicate that the classification of the muscle activity estimated from sEMG using ICA gave 100% accuracy. These results indicate that muscle activity separated from sEMG recordings using ICA is a good measure of the subtle muscle activity that results in the hand gestures.

6.2 Limitations

The results on facial sEMG analysis demonstrated that, the proposed method provides interesting result for inter experimental variations in facial muscle activity during different vowel utterance. The accuracy of recognition is poor when the system is used for testing the training network for all subjects. This shows large variations between subjects (inter-subject variation) because of different style and speed of speaking. This method has only been tested for limited vowels. This is because the muscle contraction during the utterance of vowels is relatively stationary while during consonants there are greater temporal variations.

The results demonstrate that for such a system to succeed, the system needs to be improved. Some of the possible improvements that the authors suggest will include improved electrodes, site preparation, electrode location, and signal segmentation. This current method also has to be enhanced for large set of data with many subjects in future. The authors would like to use this method for checking the inter day and inter experimental variations of facial muscle activity for speech recognition in near future to test the reliability of ICA for facial SEMG

7. Conclusions

BSS technique has been considered for decomposing sEMG to obtain the individual muscle activities. This paper has proposed the applications and limitations of ICA on hand gesture actions and vowel utterance.

A semi blind source separation using the prior knowledge of the biological model of sEMG had been used to test the reliability of the system. The technique is based on separating the muscle activity from sEMG recordings, saving the estimated mixing matrix, training the neural network based classifier for the gestures based on the separated muscle activity, and subsequently using the combination of the mixing matrix and network weights to classify the sEMG recordings in near real-time.

The results on hand gesture identification indicate that the system is able to perfectly (100% accuracy) identify the set of selected complex hand gestures for each of the subjects. These gestures represent a complex set of muscle activation and can be extrapolated for a larger number of gestures. Nevertheless, it is important to test the technique for more actions and gestures, and for a large group of people.

The results on vowel classification using facial sEMG indicate that while there is a similarity between the muscle activities, there are inter-experimental variations. There are two possible reasons; (i) people use different muscles even when they make the same sound and (ii) cross talk due to different muscles makes the signal quality difficult to classify Normalisation of the data reduced the variation of magnitude of facial SEMG between different experiments. The work indicates that people use same set of muscles for same utterances, but there is a variation in muscle activities. It can be used a preliminary analysis

for using Facial SEMG based speech recognition in applications in Human Computer Interface (HCI).

8. References

Attias, H. & Schreiner, C. E. (1998). Blind source separation and deconvolution: the dynamic component analysis algorithm, *Neural Comput*. Vol. 10, No. 6, 1373–1424.

Azzerboni, B. Carpentieri, M. La Foresta, F. & Morabito, F. C. (2004), Neural-ica and wavelet transform for artifacts removal in surface emg, *Proceedings of IEEE International Joint Conference*', pp. 3223–3228, 2004.

Azzerboni, B. Finocchio, G. Ipsale, M. La Foresta, F. Mckeown, M. J. & Morabito, F. C. (2002). Spatio-temporal analysis of surface electromyography signals by independent component and time-scale analysis, *in Proceedings of 24th Annual Conference and the Annual Fall Meeting of the Biomedical Engineering Society EMBS/BMES Conference*, pp. 112–113, 2002.

Barlow, J. S. (1979). Computerized clinical electroencephalography in perspective. *IEEE Transactions on Biomedical Engineering*, Vol. 26, No. 7, 2004, pp. 377–391.

Bartolo, A. Roberts, C. Dzwonczyk, R. R. & Goldman, E. (1996). Analysis of diaphragm emg signals: comparison of gating vs. subtraction for removal of ecg contamination', *J Appl Physiol*., Vol. 80, No. 6, 1996, pp. 1898–1902.

Basmajian & Deluca, C. (1985). *Muscles Alive: Their Functions Revealed by Electromyography*, 5th edn, Williams & Wilkins, Baltimore, USA.

Bell, A. J. & Sejnowski, T. J. (1995). An information-maximization approach to blind separation and blind deconvolution. *Neural Computations,* Vol. 7, No. 6, 1995, pp. 1129–1159.

Calinon, S. & Billard, A. (2005). Recognition and reproduction of gestures using a probabilistic framework combining pca, ica and hmm, *in Proceedings of the 22nd international conference on Machine learning*, pp. 105–112, 2005.

Djuwari, D. Kumar, D. Raghupati, S. & Polus, B. (2003). Multi-step independent component analysis for removing cardiac artefacts from back semg signals, *in 'ANZIIS'*, pp. 35–40, 2003.

Enderle, J. Blanchard, S. M. & Bronzino, J. eds (2005). *Introduction to Biomedical Engineering*, Second Edition, Academic Press, 2005.

Fridlund, A. J. & Cacioppo, J. T. (1986). Guidelines for human electromyographic research. *Psychophysiology,*Vol. 23, No. 5, 1996, pp. 567–589.

H¨am¨al¨ainen, M. Hari, R. Ilmoniemi, R. J. Knuutila, J. & Lounasmaa, O. V. (1993). Magnetoencephalography; theory, instrumentation, and applications to noninvasive studies of the working human brain, *Reviews of Modern Physics*, Vol. 65, No. 2, 1993, pp. 413 - 420.

Hansen (2000), Blind separation of noisy image mixtures. *Springer-Verlag*, 2000, pp. 159–179.

He, T. Clifford, G. & Tarassenko, L. (2006). Application of independent component analysis in removing artefacts from the electrocardiogram, *Neural Computing and Applications*, Vol. 15, No. 2, 2006, pp. 105–116.

Hillyard, S. A. & Galambos, R. (1970). Eye movement artefact in the cnv. *Electroencephalography and Clinical Neurophysiology*, Vol. 28, No. 2, 1970, pp. 173–182.

Hu, Y. Mak, J. Liu, H. & Luk, K. D. K. (2007). Ecg cancellation for surface electromyography measurement using independent component analysis, *in IEEE International Symposium on'Circuits and Systems*, pp. 3235–3238, 2007.

Hyvarinen, A. Cristescu, R. & Oja, E. (1999). A fast algorithm for estimating overcomplete ica bases for image windows, *in International Joint Conference on Neural Networks*, pp. 894–899, 1999.

Hyvarinen, A. Karhunen, J. & Oja, E. (2001). *Independent Component Analysis*, Wiley-Interscience, New York.

Hyvarinen, A. & Oja, E. (1997). A fast fixed-point algorithm for independent component analysis, *Neural Computation*, Vol. 9, No. 7, 1997, pp. 1483–1492.

Hyvarinen, A. & Oja, E. (2000). Independent component analysis: algorithms and applications, *Neural Network*, Vol. 13, No. 4, 2000, pp. 411–430.

James, C. J. & Hesse, C. W. (2005). Independent component analysis for biomedical signals, *Physiological Measurement,* Vol. 26, No. 1, R15+.

Jung, T. P. Makeig, S. Humphries, C. Lee, T. W. McKeown, M. J. Iragui, V. & Sejnowski, T. J. (2001). Removing electroencephalographic artifacts by blind source separation. *Psychophysiology,* Vol. 37, No. 2, 2001, pp. 163–178.

Jung, T. P. Makeig, S. Lee, T. W. Mckeown, M. J., Brown, G., Bell, A. J. & Sejnowski, T. J. (2000). Independent component analysis of biomedical signals, *In Proceeding of Internatioal Workshop on Independent Component Analysis and Signal Separation*' Vol. 20, pp. 633–644.

Kaban (2000), Clustering of text documents by skewness maximization, pp. 435–440.

Kato, M. Chen, Y.-W. & Xu, G. (2006). Articulated hand tracking by pca-ica approach, *in Proceedings of the 7th International Conference on Automatic Face and Gesture Recognition,* pp. 329–334, 2006.

Kimura, J. (2001). *Electrodiagnosis in Diseases of Nerve and Muscle: Principles and Practice*, 3rd edition, Oxford University Press.

Kolenda (2000). Independent components in text, Advances *in Independent Component Analysis*, Springer-Verlag, pp. 229–250.

Lapatki, B. G. Stegeman, D. F. & Jonas, I. E. (2003). A surface emg electrode for the simultaneous observation of multiple facial muscles, *Journal of Neuroscience Methods*, Vol. 123, No. 2, 2003, pp. 117–128.

Lee, T. W. (1998). *Independent component analysis: theory and applications*, Kluwer Academic Publishers.

Lee, T. W. Lewicki, M. S. & Sejnowski, T. J. (1999). Unsupervised classification with non-gaussian mixture models using ica, *in Proceedings of the 1998 conference on Advances in neural information processing systems*, MIT Press, Cambridge, MA, USA, pp. 508–514, 1999.

Lewicki, M. S. & Sejnowski, T. J. (2000). Learning overcomplete representations, *Neural Computations*, Vol. 12, No. 2, pp. 337–365, 2006.

Mackay, D. J. C. (1996). Maximum likelihood and covariant algorithms for independent component analysis, *Technical report*, University of Cambridge, London.

Manabe, H. Hiraiwa, A. & Sugimura, T. (2003). Unvoiced speech recognition using emg - mime speech recognition, *in proceedings of CHI 03 extended abstracts on Human factors in computing systems*, ACM, New York, NY, USA, 2003, pp. 794–795.

Mckeown, M. J. Makeig, S. Brown, G. G. Jung, T.-P. Kindermann, S. S. Bell,A. J. & Sejnowski, T. J. (1999). Analysis of fmri data by blind separation into independent spatial components, *Human Brain Mapping,* Vol. 6, No. 3, 1999, pp. 160–188.

Mckeown, M. J. Torpey, D. C. & Gehm, W. C. (2002). Non-invasive monitoring of functionally distinct muscle activation during swallowing, *Clinical Neurophysiology,* Vol. 113, No. 3, 2002, pp. 354–366.

Mosher, J. C. Lewis, P. S. & Leahy, R.M. (1992). Multiple dipole modeling and localization from spatio-temporal meg data, *IEEE Transactions on Biomedical Engineering*, Vol. 39, No. 6, 1992, pp. 541–557.

Naik, G. R. Kumar, D. K. Singh, V. P. & Palaniswami, M. (2006). Hand gestures for hci using ica of emg, *in Proceedings of the HCSNet workshop on Use of vision in human-computer interaction*, Australian Computer Society, Inc., pp. 67–72, 2006.

Naik, G. R. Kumar, D. K. Weghorn, H. & Palaniswami, M. (2007). Subtle hand gesture identification for hci using temporal decorrelation source separation bss of surface emg, *in 9th Biennial Conference of the Australian Pattern Recognition Society on 'Digital Image Computing Techniques and Applications*, pp. 30–37, 2007.

Nakamura, H. Yoshida, M. Kotani, M. Akazawa, K. & Moritani, T. (2004). The application of independent component analysis to the multi-channel surface electromyographic signals for separation of motor unit action potential trains: part i-measuring techniques, *Journal of electromyography and kinesiology : official journal of the International Society of Electrophysiological Kinesiology,* Vol. 14, No. 4, 2004, pp. 423–432.

Niedermeyer, E. & Da Silva, F. L. (1999). *Electroencephalography: Basic Principles, Clinical Applications, and Related Fields*, Lippincott Williams and Wilkins; 4th edition .

Parra, J. Kalitzin, S. N. & Lopes (2004). Magnetoencephalography: an investigational tool or a routine clinical technique?, *Epilepsy & Behavior,* Vol. 5, No. 3, 2004, pp. 277–285.

Parsons (1986), Voice and speech processing., Mcgraw-Hill.

Peters, J. (1967). Surface electrical fields generated by eye movement and eye blink potentials over the scalp, *Journal of EEG Technology*, Vol. 7, 1967, pp. 1129–1159.

Petersen, K. Hansen, L. K. Kolenda, T. & Rostrup, E. (2000).On the independent components of functional neuroimages, *in processing of Third International Conference on Independent Component Analysis and Blind Source Separation*, pp. 615–620, 2000.

Rajapakse, J. C. Cichocki, A. & Sanchez (2002). Independent component analysis and beyond in brain imaging: Eeg, meg, fmri, and pet, *in Proceedings of the 9th International Conference on Neural Information Processing*, pp. 404–412, 2002.

Scherg, M. & Von Cramon, D. (1985). Two bilateral sources of the late aep as identified by a spatio-temporal dipole model, *Electroencephalogr Clin Neuro-physiol*., Vol. 62, No. 1,1985, pp. 32–44.

Sorenson (2002). Mean field approaches to independent component analysis. *Neural Computation*, Vol. 14, 2002, pp. 889–918.

Tang, A. C. & Pearlmutter, B. A. (2003). *Independent components of magnetoencephalography: localization*', 2003, pp. 129–162.

Verleger, R. Gasser, T. & Mocks, J. (1982). Correction of eog artefacts in event related potentials of the eeg: aspects of reliability and validity. *psychophysiology*, Vol. 19, No. 2, 1982,pp. 472–480.

Vig´ario, R. S¨arel¨a, J. Jousm¨aki, V. H¨am¨al¨ainen, M. & Oja, E. (2000). Independent component approach to the analysis of eeg and meg recordings, *IEEE transactions on bio-medical engineering*, Vol 47, No. 5, 2002, pp. 589–593.

Weerts, T. C. & Lang, P. J. (1973). The effects of eye fixation and stimulus and response location on the contingent negative variation (cnv), *Biological psychology*, Vol. 1,No. 1, 1973, pp. 1–19.

Whitton, J. L. Lue, F. & Moldofsky, H. (1978). A spectral method for removing eye movement artifacts from the eeg, *Electroencephalography and clinical neurophysiology,* Vol. 44, No. 6, 1978, pp. 735–741.

Wisbeck, J. Barros, A. & Ojeda, R. (1998). Application of ica in the separation of breathing artifacts in ecg signals.

Woestenburg, J. C. Verbaten, M. N. & Slangen, J. L. (1983).The removal of the eye-movement artifact from the eeg by regression analysis in the frequency domain, *Biological psychology*, Vol. 16, No. 1, 193, pp. 127–147.

5

Sources of bias in synchronization measures and how to minimize their effects on the estimation of synchronicity: Application to the uterine electromyogram

Terrien Jérémy[1], Marque Catherine[2], Germain Guy[3] and Karlsson Brynjar[1,4]
[1]*Reykjavik University*
Iceland
[2]*Compiègne University of technology*
France
[3]*CRC MIRCen, CEA-INSERM*
France
[4]*University of Iceland*
Iceland

1. Introduction

Preterm labor (PL) is one of the most important public health problems in Europe and other developed countries as it represents nearly 7% of all births. It is the main cause of morbidity and mortality of newborns. Early detection of a PL is important for its prevention and for that purpose good markers of preterm labor are needed. One of the most promising biophysical markers of PL is the analysis of the electrical activity of the uterus. Uterine electromyogram, the so called electrohysterogram (EHG), has been proven to be representative of uterine contractility. It is well known that the uterine contractility depends on the excitability of uterine cells but also on the propagation of electrical activity to the whole uterus. The different algorithms proposed in the literature for PL detection use only the information related to local excitability. Despite encouraging results, these algorithms are not reliable enough for clinical use. The basic hypothesis of this work is that we could increase PL detection efficiency by taking into account the propagation information of the uterus extracted from EHG processing. In order to quantify this information, we naturally applied the different synchronization methods previously used in the literature for the analysis of other biomedical signals (i.e. EEG).

The investigation of the coupling between biological signals is a commonly used methodology for the analysis of biological functions, especially in neurophysiology. To assess this coupling or synchronization, different measures have been proposed. Each measure assumes one type of synchronization, i.e. amplitude, phase… Most of these measures make some statistical assumptions about the signals of interest. When signals do

not respect these assumptions, they give rise to a bias in the measure, which may in the worst case, lead to a misleading conclusion about the system under investigation. The main sources of bias are the noise corrupting the signal, a linear component in a nonlinear synchronization and non stationarity. In this chapter we will present the methods that we developed to minimize their effects, by evaluating them on synthetic as well as on real uterine electromyogram signals. We will finally show that the bias free synchronization measures that we propose can be used to predict the active phase of labor in monkey, where the original synchronization measure does not provide any useful information. In this chapter we illustrate our methodological developments using the nonlinear correlation coefficient as an example of a synchronization measure in which the methods can be used to correct for bias.

2. Uterine electromyography

The recording of the electrical activity of the uterus during contraction, the uterine electromyography, has been proposed as a non invasive way to monitor uterine contractility. This signal, the so called Electrohysterogram (EHG), is representative of the electrical activity occurring inside the myometrium, the uterine muscle. The EHG is a strongly non stationary signal mainly composed of two frequency components called FWL (Fast Wave Low) and FWH (Fast Wave High). The characteristics of the EHG are influenced by the hormonal changes occurring along gestation. The usefulness of the EHG for preterm labor prediction has been explored as it is supposed to be representative of the uterus contractile function.

2.1 Preterm labor prediction by use of external EHG

Gestation is known to be a two-step process consisting of a preparatory phase followed by active labor (Garfield & al., 2001). During the preparatory phase, the uterine contractility evolves from an inactive to a vigorously contractile state. This is associated to an increased myometrial excitability, as well as to an increased propagation of the electrical activity to the whole uterus (Devedeux & al., 1993; Garfield & Maner, 2007).

Most studies have focused on the analysis of the excitability of the uterus using two to four electrodes. It is generally supposed that the increase in excitability is mainly observable through an increase in the frequency of FWH (Buhimschi & al., 1997; Maner & Garfield, 2007). Some authors, like (Buhimschi & al., 1997), also used the energy of the EHG as potential parameter for the prediction of preterm labor. This parameter is however highly dependent on experimental conditions like the inter-electrode impedance. A relatively recent paper used the whole frequency content, i.e. FWL + FWH, of the EHG for PL prediction (Leman & al., 1999). This study, based on the characterization of the time-frequency representation of the EHG, demonstrated that a fairly accurate prediction can be made as soon as 20 weeks of gestation in human pregnancies.

In spite of very exciting results, this method is not currently used in routine practice due to the discrepancy between the different published studies, a strong variability of the results obtained and thus a not sufficient detection ratio for clinical use. Increasingly, teams working in this field tried to increase the prediction ratio by taking into account the propagation phenomenon in addition to the excitability (Euliano & al., 2009; Garfield & Maner, 2007). A uterus working as a whole is a necessary condition to obtain efficient

contractions capable of dilating the cervix and expulsing the baby. The study of the propagation of the electrical activity of the uterus has been performed in two different ways. The first approach consists, like for skeletal muscle, in observing and characterizing the propagation of the electrical waves (Karlsson & al., 2007; Euliano & al., 2009). The second one consists in studying the synchronization of the electrical activity at different locations of the uterus during the same contraction by using synchronization measures (Ramon & al., 2005; Terrien & al., 2008b). The work presented in this chapter derived from this second approach.

2.2 Possible origins of synchronization of the uterus at term

The excitability is mainly controlled at a cellular level by a modification of ion exchange mechanisms. Propagation is mainly influenced by the cell-to-cell electrical coupling (intercellular space, GAP junctions). More precisely, the propagation is a multi-scale phenomenon. At a cellular level, it mainly takes place through GAP junctions (Garfield & Hayashi, 1981; Garfield & Maner, 2007). At a higher scale, there is preferential propagation pathways called bundles which represent group of connected cells organized as packet (Young, 1997; Young & Hession, 1999). The organization of the muscle fibers might also play an important role in propagation phenomenon and characteristic. Contrary to skeletal muscle, the fibers of uterus are arranged according to three different orientations. The role of the nerves present in the uterus is still debated but may be responsible of a long distance synchronization of the organ (Devedeux & al., 1993).

The recent studies focusing on the propagation characterization used multi electrode grids position on the woman abdomen in order to picture the contractile state of the uterus along the contraction periods. The most common approach uses the intercorrelation function in order to detect a potential propagation delay between the activities of two distant channels. It has been shown that there is nearly no linear correlation between the raw electrical signals (Duchêne & al., 1990; Devedeux & al., 1993) so all these studies used the envelope (≈ instantaneous energy) of the signals to compute propagation delays. Only recently, two studies have used synchronization parameters on the EHG in order to analyze the propagation/synchronization phenomenon involved (Ramon & al., 2005; Terrien & al., 2008b).

3. Synchronization measures

If we are interested in understanding or characterizing a particular system univariate signal processing tools may be sufficient. The system of interest is however rarely isolated and is probably influenced by other systems of its surrounding. The detection and comprehension of these possible interactions, or couplings, is challenging but of particular interest in many fields as mechanics, physics or medicine. As a biomedical example, we might be interested in the coupling of different cerebral structures during a cognitive task or an epilepsy crisis. To analyze this coupling univariate tools are no longer sufficient and we would need multivariate or at least bivariate analysis tools. These tools have to be able to detect the presence or not of a coupling between two systems but also to indicate the strength and the direction of the coupling (Figure 1). A coupling measure or a synchronization measure has so to be defined.

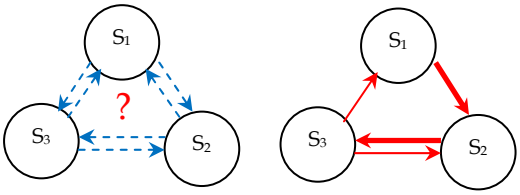


Fig. 1. Schema of synchronization analysis between 3 systems. These methods are able to detect the presence or absence, the strength and the direction of the couplings defining a coupling pattern.

There are a numerous synchronization measures in the literature. The interested reader can find a review of the different synchronization measures and their applications for EEG analysis in (Pereda & al., 2005). Each of them makes a particular hypothesis on the nature of the coupling. As simple examples, it can be an amplitude modulation or a frequency modulation of the output of one system in response to the output of another one. These measures can be roughly classified according to the approach that they are based on (Table 1).

Approach	Synchronization measure
Correlation	Linear correlation coefficient
	Coherence
	Nonlinear correlation coefficient
Phase synchronization	Phase entropy
	Mean phase coherence
Generalized synchronization	Similarity indexes
	Synchronization likelihood

Table 1. Different approaches and associated synchronization measures.

To this non exhaustive list of measures, we could add two other particular classes of methods. The methods presented Table 1 are bivariate methods. In the case of more than two systems possibly coupled to each other, these methods might give an erroneous coupling pattern. Therefore multivariate synchronization methods have been introduced recently (Baccala & Sameshima 2001a, 2001b; Kus & al., 2004). The main associated synchronization measures are the partial coherence and the partial directed coherence. The last class of method is the event synchronization. One example of derived synchronization measure is the Q measure (Quian Quiroga & al., 2002).
In this work we will treat in more detail the nonlinear correlation coefficient in the context of a practical approach. In our context of treating bias in synchronization measures, we chose this particular measure since in previous study the linear correlation coefficient was not able to highlight any linear relationship between the activity of different part of the uterus during contractions. The methods of correcting for bias presented in this work however allowed us to use this measure to show the real underlying relation in the signals. We however want to stress that the methods presented here can be used with any other synchronization measures.

3.1 Linear correlation coefficient

The linear correlation coefficient represents the adjustment quality of a relationship between two time series x and y, by a linear curve. It is simply defined by:

$$r^2 = \frac{\mathrm{cov}^2(x,y)}{\mathrm{var}(x).\mathrm{var}(y)} \tag{1}$$

where cov and var stand for covariance and variance respectively.
This model assumes a linear relationship between the observations x and y. In many applications this assumption is false. More recently, a nonlinear correlation coefficient has been proposed in order to be able to model a possible nonlinear relationship (Pijn & al., 1990).

3.2 Nonlinear correlation coefficient

The nonlinear correlation coefficient (H^2) is a non parametric nonlinear regression coefficient of the relationship between two time series x and y. In practice, to calculate the nonlinear correlation coefficient, a scatter plot of y versus x is studied. The values of x are subdivided into bins; for each bin, the x value of the midpoint (p_i) and the average value of y (q_i) are calculated. The curve of regression is approximated by connecting the resulting points (p_i, q_i) by segments of straight lines; this methodology is illustrated figure 2. The nonlinear correlation coefficient H^2 is then defined as:

$$H^2_{y/x} = \frac{\sum_{k=1}^{N} y(k)^2 - \sum_{k=1}^{N} (y(k) - f(x(k)))^2}{\sum_{k=1}^{N} y(k)^2} \tag{2}$$

where $f(x)$ is the linear piecewise approximation of the nonlinear regression curve. This parameter is bounded by construction between [0, 1]. The measure H^2 is asymmetric, because $H^2_{y/x}$ may be different to $H^2_{x/y}$ and can thus gives information about the direction of coupling between the observations. If the relation between x and y is linear $H^2_{y/x} = H^2_{x/y}$ and is close to r^2. In the case of a nonlinear relationship, $H^2_{y/x} \neq H^2_{x/y}$ and the difference ΔH^2 indicates the degree of asymmetry. H^2 can be maximized to estimate a time delay τ between both channels for each direction of coupling. Both types of information have been used to define a measure of the direction of coupling and successfully applied to EEG by (Wendling & al., 2001).

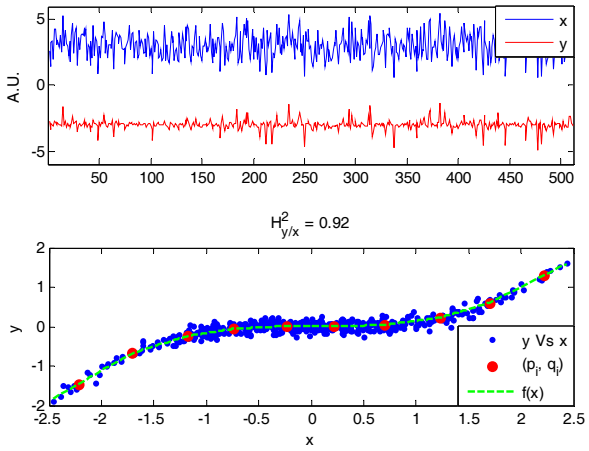


Fig. 2. Original data $x = N(0, 1)$ and $y = (x/2)^3 + N(0, 0.1)$ (upper panel) and construction of the piecewise linear approximation of the nonlinear relationship between x and y in order to compute the parameter H^2 (lower panel). For comparison, the linear correlation coefficient r^2 is only 0.64.

This method is non parametric is the sense that it does not assume a parametric model of the underlying relationship. The number of bins needs however to be defined in a practical application. Our experience shows that this parameter is not crucial regarding the performances of the method. It has to be set anyway in accordance to the nonlinear function that might exist between the input time series. Similarly to what is expressed by the Shannon theorem, the sampling rate of the nonlinear function must be sufficient to model properly the nonlinear relationship. The limit case of 2 bins might give a value close or equal to the linear correlation coefficient. The hypothetic result that we might obtain with a very high number of bins highly depends on the relationship between the time series. It may tend to an over estimation due to an over fitting of the relationship corrupted by noise. We so suggest evaluating the effect of this parameter on the estimation of the relationship derived from a supposed model of the relationship or clean experimental data.

4. Effect of noise in synchronization measure

4.1 Denoising methods

Noise corrupting the signals is the most common source of bias. It is present in nearly all real life measurements in varying quantities. The noise can come from the environment of the electrodes and the acquisition system, e.g. powerline noise, electronic noise, or from other biological systems not under investigation like ECG, muscle EMG ... To reduce the influence of this noise on the synchronization measure, one may use digital filters to increase the signal to noise ratio (SNR) expressed in decibel (dB). We have to differentiate linear filters like classical Butterworth filters, and nonlinear filters like wavelet filters. Nonlinear filters are filters that can make the distinction between the signal of interest and the part of the noise present in the same frequency band in order to remove it. With linear filter it is not the case and we have to set the cutting frequency according to the bandwidth

of the signal of interest. This kind of filter cannot remove the noise present in the signal bandwidth without distorting the signal itself.
In synchronization analysis, only linear filters have been used in the literature to our knowledge. However, linear filters are known to dephase the filtered signal. In order to avoid this distortion, phase preserving filters are used instead. Practically, this is realized by filtering two times the noisy signal, one time in the forward direction and the second time in the reverse direction to cancel out the phase distortion.

4.2 Example

To model and illustrate the effect of noise on synchronization measures, we used two coupled chaotic Rössler oscillators. This model has been widely used in synchronization analysis due to is well known behavior. The model is defined by:

$$\begin{aligned}
\dot{x}_1 &= -\omega_1(t)\, y_1 - z_1 \\
\dot{y}_1 &= \omega_1(t)\, x_1 + 0.15\, y_1 \\
\dot{z}_1 &= 0.2 + z_1(x_1 - 10) \\
\dot{x}_2 &= -\omega_2(t)\, y_2 - z_2 + C(t)(x_2 - x_1) \\
\dot{y}_2 &= \omega_2(t)\, x_2 + 0.15\, y_2 \\
\dot{z}_2 &= 0.2 + z_2(x_2 - 10)
\end{aligned} \tag{3}$$

The function $C(t)$ allows us to control the coupling strength between the two oscillators. The system was integrated by using an explicit Runge-Kutta method of order 4 with a time step Δt = 0.0078. For this experiment we used the following Rössler system configuration: ω_1 = 0.55, ω_2 = 0.45 and C = 0.4. On the original time series we added some Gaussian white noise in order to obtain the following SNR = {30; 20; 15; 10; 5; 0} dB. The synchronization analysis was then realized on the filtered version of the noisy signals using a 4th order phase preserving Butterworth filter. The results of this experiment are presented figure 3.
As we can see, the measured coupling drops dramatically for SNR below 20 dB. The filtering procedure is able to keep the measured coupling close to the reference down to 10 dB. For more noise, the measured coupling deviated significantly from the real value due to the non negligible amount of noise inside the bandwidth of the signals. The results obtained with a simple linear filter are surprisingly good. It can be explained by the very narrow bandwidth of the Rössler signals. The amount of noise present in the bandwidth of the signals is very small as compared to the total amount of noise added in the whole frequency band. In this situation, the use of nonlinear filter might be interesting. A study of the possible influences of the nonlinear filtering methods on the synchronization measures has to be done first and might be interesting for the community using synchronization measures.

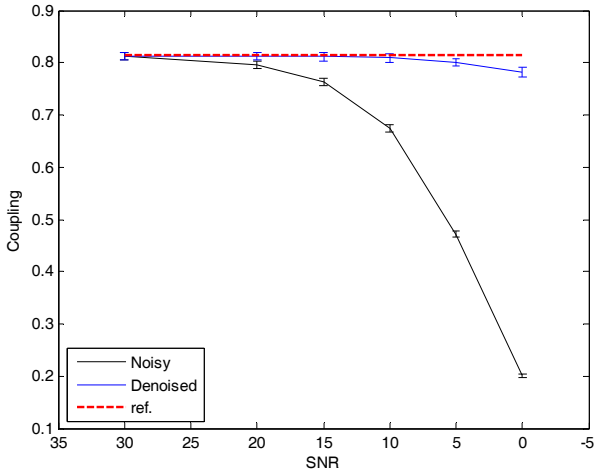


Fig. 3. Evolution of the coupling as a function of the imposed SNR before (Noisy) and after denoising (Denoised). The reference synchronization value is plotted by a horizontal dashed line.

The main physiological noises corrupting external EHG are the maternal skeletal EMG and ECG. The noise and the EHG present overlapping spectra. A specific nonlinear filter has been developed for denoising properly these EHG (Leman & Marque, 2000). Internal EHG, like the signals used here, are less corrupted and allow the use of classical phase preserving linear filters. An analysis of the possible effects of this type of denoising will have to be done for an application of synchronization analysis of external EHG, as it is performed on pregnant women.

5. Nonlinearity testing with surrogate measure profile

To test a particular hypothesis on a time series, surrogate data are usually used. They are built directly from the initial time series in order to fulfill the conditions of a particular null hypothesis. One common hypothesis is the nonlinearity of the original time series. The procedure involves the analysis of the statistics of the surrogates as compared to the statistic found with the original data in order to define its z-score. The z-score assumes that the surrogate measure profile presents a Gaussian distribution. If this is not the case the test might be erroneous.

We propose to use a surrogate corrected value instead of the z-score of a particular statistic. We also derive a statistical test based on the fitting of the surrogate measure profile distribution. We demonstrate the proposed method on the nonlinear correlation coefficient (H^2) as the initial statistic. The performance of the corrected statistic was evaluated on both synthetic and real EHG signals.

5.1 Surrogate data

Surrogate data are time series which are generated in order to keep particular statistical characteristics of an original time series while destroying all others. They have been used to test for nonlinearity (Schreiber & Schmitz, 2000) or nonstationarity (Borgnat & Flandrin, 2009) of time series for instance. The classical approaches to constructing such time series are phase randomization in the Fourier domain and simulated annealing (Schreiber & Schmitz, 2000). Depending on the method used to construct the surrogates, a particular null hypothesis is assumed. The simulated annealing approach is very powerful since nearly any null hypothesis might be chosen according to the definition of an associated cost function. As a first step, we chose the Fourier based approach.

The Fourier based approach consists mainly in computing the Fourier transform, F, of the original time series $x(t)$.

$$X(f) = F\{x(t)\} = A(f)\, e^{i\,\Phi(f)} \tag{4}$$

where A(f) is the amplitude and Φ(f) the phase. The surrogate time series is obtained by rotating the phase Φ at each frequency f by an independent random variable φ taking values in the range [0, 2π) and going back to the temporal domain by inverse Fourier transform F^{-1}, that is:

$$\tilde{x}(t) = F^{-1}\left\{\tilde{X}(f)\right\} = F^{-1}\left\{A(f)\, e^{i\left[\Phi(f)+\varphi(f)\right]}\right\} \tag{5}$$

By construction, the surrogate has the same power spectrum and autocorrelation function as the original time series but not the same amplitude distribution. This basic construction method has been refined to assume different null hypothesis. We used the iterative amplitude adjusted Fourier transform method to produce the surrogates (Schreiber & Schmitz, 2000). Basically, this iterative algorithm starts with an initial random shuffle of the original time series. Then, two distinct steps will be repeated until a stopping criterion is met, i.e. mean absolute error between the original and surrogate amplitude spectrum. The first step consists in a spectral adaptation of the surrogate spectrum and the second step in an amplitude adaptation of the surrogate. At convergence, the surrogate has the same spectrum and amplitude distribution of the original time series, but all nonlinear structures present in the original time series are destroyed.

5.2 Use of surrogate measure profile

On each surrogate j we can compute a measure $\Theta_0(j)$. All values of $\Theta_0(j)$ form what we call a surrogate measure profile Θ_0. Surrogate measure profiles Θ_0 are usually used in order to give a statistical significance to a measure Θ_1 against a given null hypothesis H_0. The classical approach assumes that Θ_0 is normally distributed and uses the z-score. The empirical mean $\langle\Theta_0\rangle$ and standard deviation $\sigma(\Theta_0)$ of Θ_0 are calculated. The z-score of the observed value Θ_1 is then:

$$z = \left| \frac{\Theta_1 - \langle\Theta_0\rangle}{\sigma(\Theta_0)} \right| \tag{6}$$

The hypothesis test is usually considered as significant at a significance level $p < 0.05$ when $z \geq 1.96$. The z-score has been also directly used to measure the nonlinearity of a univariate or a multivariate system (Prichard & Theiler, 1994).

In practice, the normality assumption should be checked before using the z statistic. For that purpose, the Kolmogorov-Smirnov or Lilliefors test might be used. The Kolmogorov-Smirnov test uses a predefined normal distribution of the null hypothesis, i.e. known mean

and variance. The Lilliefors test is on the contrary based on a mean and variance of the distribution derived directly from the data.

5.3 Percentile corrected statistic and associated hypothesis test

The distributions of Θ_0 might be non Gaussian as attested by a Lilliefors test for example. In that case, the use of z-score statistics may be erroneous or at least meaningless. We propose to use instead a measure corrected according to the statistics of the surrogates. This measure, Θ_{cx}, is defined as:

$$\Theta_{cx} = \Theta_1 - P_x(\Theta_0) \tag{7}$$

where $P_x(y)$ stands for the x^{th} percentile of the data y.

The study of the statistical distribution of Θ_0 allows us to define a statistical test even when dealing with non Gaussian distributions. In practice, we have noticed that the distribution of Θ_0 follows approximately a Gamma law $\Gamma(a, \beta)$ when the distribution is not Gaussian. A distribution model can be fitted directly on the surrogate data by maximum likelihood estimation. This model allows us to easily define a statistical threshold for a given probability p, over which the observed value Θ_1 is considered as significant. The inverse of the Gamma cumulative distribution function, parameterized by the fitted a and β, gives the threshold knowing the chosen probability p.

In the context of using the nonlinear correlation coefficient H^2, we called the corrected measure Θ_{cx}, H^2_{cx} or surrogate corrected nonlinear correlation coefficient. This statistic is bounded between [-1, 1] where the sign roughly indicates a non significant test if the percentile x and the probability p coincide. According to the characteristics of the generated surrogate data in this study, the parameter H^2_{cx} represents the part of the original H^2 value unexplained by the linearity presents in the original time series.

From a practical point of view, the only parameter that has to be tuned is the number of surrogates used to construct the surrogate measure profile. This number must be large enough for a good estimation of the density function. It varies largely from one signal to another. The counterparts of choosing a very high number of surrogates is the time of computation especially with long original time series. After empirical evaluation of this parameter, we found that 10000 surrogates was a good compromise for our signals.

5.4 Results on synthetic signals

For this experiment we used the following Rössler system configuration: $\omega_1 = 0.55$ and $\omega_2 = 0.45$. The sampling rate was 256 Hz.

An instance of the coupled Rössler systems, with $C = 0.5$, is presented figure 4 as well as the corresponding surrogates measure profile. We can clearly see that the original synchronization value $H^2_{y/x}$ is above the imposed coupling value C. The relatively high values of the measure obtained with the surrogates suggest that a non negligible amount of the observed synchronization value is due to a linear component between the systems.

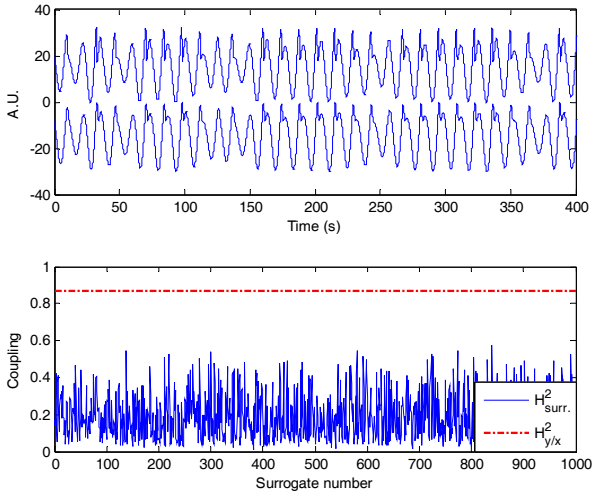


Fig. 4. Example of the output of the model for C = 0.5 (top panel) and surrogates measure profile (bottom panel).

The distribution of the surrogate profile is depicted figure 5. We can easily see that the distribution is highly non Gaussian and is best fitted by a Gamma law. A statistical test based on the z-value might thus be erroneous. The non Gaussianity was attested by a Lilliefors test applied on the experimental data. The 90 percentile derived from the fitted law was 0.38. The measured coupling, 0.87 as observed figure 4, is above the 90 percentile and thus attests of significant test. The proposed corrected measure, H^2_{cx}, is in this case 0.49 which is closer to the imposed coupling value C = 0.5 than the original measure.

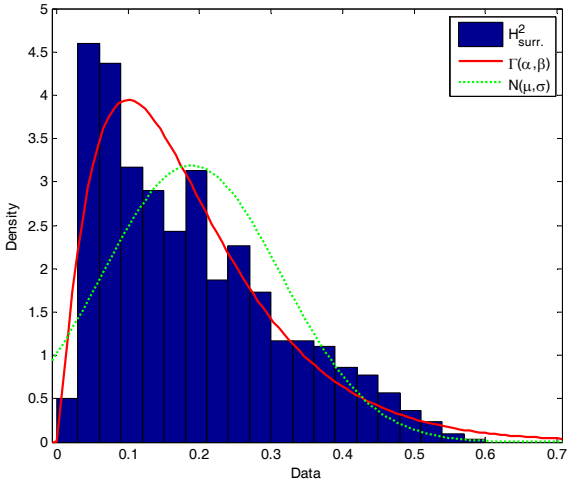


Fig. 5. Distribution of the surrogate values (Θ_0), Gamma law model (Γ(α,β), continuous line) and normal law model (N(μ,σ), dotted line).

The original synchronization values were always above the imposed coupling (Figure 6). For moderate couplings, below 0.5, the proposed correction gives nearly identical values as the imposed coupling. From a coupling of 0.5, the proposed correction underestimates the coupling strength between the systems. More importantly, we can notice that the difference between the original and the corrected values is nearly constant. It indicates that the nature of the relationship between the Rössler systems is identical whatever the imposed coupling strength. This might explain the underestimation of the corrected synchronization due to a "saturation" of the original synchronization at values near 1.

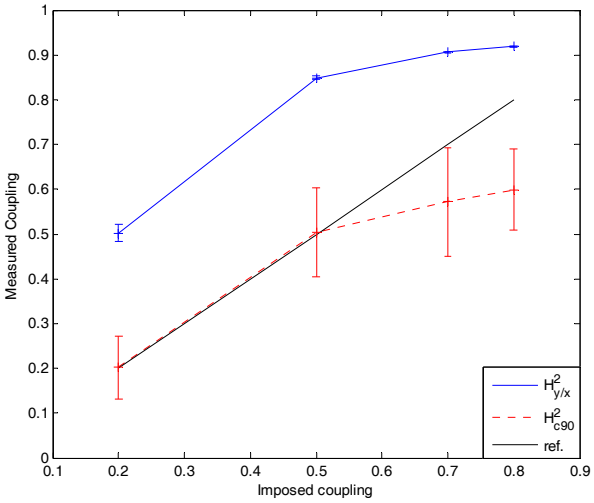


Fig. 6. Original and corrected H^2 estimations (μ±σ) for different imposed coupling values.

5.5 Results on real EHG signals

Uterine EMG was recorded on a monkey during labor. Two bipolar channels were sutured on the uterus approximately 7 cm apart. The two EMG channels were digitalized simultaneously at 50 Hz. A detailed description of the experimental setup can be found in (Terrien & al., 2008a). The EMG signals were then segmented manually to extract segments containing uterine contractions. The different segments were then band-pass filtered (1 - 4.7 Hz) to extract FWH according to (Devedeux & al., 1983) by a 4th order phase preserving Butterworth filter. We also showed that a time delay of EHG bursts highlight the synchrony between the signals (Terrien & al., 2008b). The time delay between bursts that we chose corresponds to the delay needed to maximize the cross-correlation function.

When applied to uterine EMG, we noticed very different behavior of H^2 in pregnancy and labor contractions as depicted figure 7. In this example, even if the two contractions present nearly the same original synchronization values (0.13 and 0.15), their surrogate measure profiles are very different. For the labor contraction, the synchronization measures obtained on surrogates are very low when compared to the original value contrary to the pregnancy EMG, where some surrogates present synchronization measure above the original one. This

may indicate a strong relationship between the nonlinear components of the EMG burst during labor which seems to be absent or less important during pregnancy. We consider these differences to be useful in differentiating labor and pregnancy contractions. Concerning the statistical test all labor contractions presented a significant test. For pregnancy contractions, the majority but not all the contractions did not test as significant.

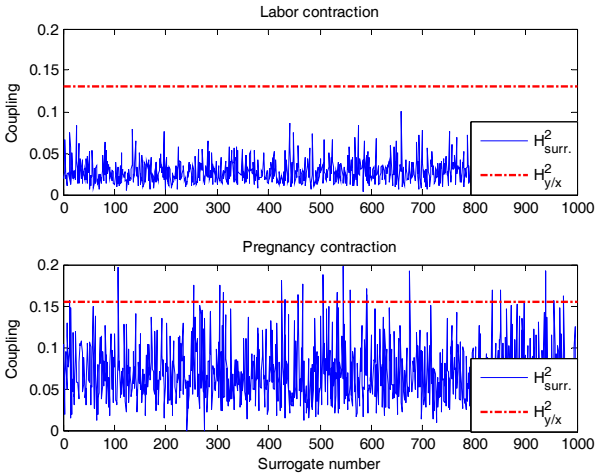


Fig. 7. Example of surrogates measure profile obtained with a labor contraction (top panel) and a pregnancy contraction (bottom panel).

5.6 Discussion

Surrogates are constructed to fulfill all characteristics of a null hypothesis that we want to evaluate on a time series. The statistical tests of the considered hypothesis use the z score which implicitly assumes the Gaussianity of the surrogate statistic distribution. In case of non Gaussian statistics, the usual test might fail or simply gives rise to erroneous conclusion. We proposed to use, instead of the z-score, a percentile corrected statistic. This corrected value is thought to be independent of the surrogate distribution. We derived a statistical test by simply fitting the surrogate distribution by a given distribution model and defining a statistical threshold. We demonstrated the satisfactory use of the proposed approach on synthetic and real signals as well. When applied to the nonlinear correlation coefficient, we showed that the statistics of the surrogate measure profile present a Gamma distribution, probably explained by the quadratic nature of the original statistic. For this particular statistic, the new value represents the part of the original value not explained by the linearity present in the original time series. The usefulness of this new measure has of course to be confirmed and tested on different type of data usually used in the field, EEG for example.

The use of this "new" synchronization measure on uterine EMG helped us to show two different behaviors of contractions. We think that this difference might help us in differentiating inefficient (pregnancy) and efficient (labor) contractions in the final aim of labor prediction. This difference in behavior might be explained by the increase in the nonlinearity of the EHG as labor approaches (Radhakrishnan & al., 2000). The surrogates

used in this study are also a stationarized version of the original time series. In the case of uterine EMG, we assumed that the EMG bursts were stationary and we imputed the difference between pregnancy and labor to a change in linearity only. Without testing this stationary assumption, we could not be sure about the origin of the observed differences, i.e. linearity or stationarity. The use of surrogates which preserve the non stationarity of the original time series might be helpful for that purpose (Schreiber & Schmitz, 2000). As a first way to answer this open question, we decided to study the influence of the non stationarity in synchronization analysis and to propose an approach able to take into account this information which is another source of bias of synchronization measure.

6. Dealing with non stationary signals

Most synchronization measures are only reliable in the analysis of long stationary time series. A stationary signal is a signal which has all statistical moments constant with time. This strong assumption might be relaxed since this property is impossible to verify. This relaxed condition is called "weak stationarity" of order n. A weak stationary signal of order n presents all moments up to n that do not vary with time. The stationarity of order 2 is often used (Blanco & al., 1995).
Many biological signals are however highly non stationary. Nevertheless, the coupling analysis of these non stationary signals is usually performed by using a sliding window in which the signals of interest are supposed to be stationary, or by directly using time dependant synchronization measures like time-frequency approach (Ansary-Asl & al., 2006). The most commonly used approach is the windowing method. The length of the window has to be set according to the characteristics of the signal of interest. A bad choice of this parameter might have dramatic effects on the obtained results. We propose to use instead a pre-processing step able to detect automatically the longer stationary segments of a signal of interest. This approach avoids making any trade off between the length of the segments and the stationary assumption.

6.1 The windowing approach

The windowing approach consists in computing the synchronization parameter in a window of finite length L, supposed to be the minimal stationary length of the signals of interest, and shifting the window by a time τ before computing another value. The time shift is often expressed as a percentage of overlapping between successive windows. The main problem of this method is the estimation of the minimal stationary length.

A tradeoff between the length of the analysis window and the stationary assumption has to be made. The length of the window also limits the accuracy of the time detection of abrupt changes that can reflect biological mechanisms in the underlying systems. As it can be seen figure 8, an increase in the length of the analysis window reduces the variance of the estimation but at a same time smoothes the boundary of the transition times, located in this example roughly at 204 and 460 s. The length of the window is thus an important parameter which has to be set according to a prior knowledge of the minimal length of the stationary parts of the signals or by trial and error.

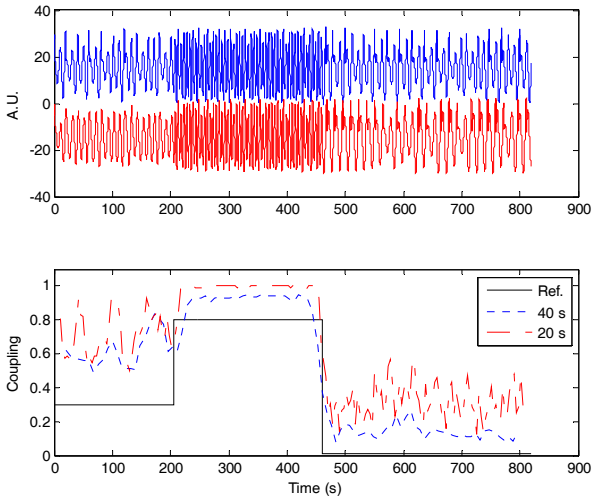


Fig. 8. Example of the output of the Rössler system (top panel) and the corresponding synchronization analysis using H^2 (bottom panel) obtained by the windowing approach for a window length of 40 s or 20 s. The coupling function $C(t)$ is presented as a continuous line (Ref.).

6.2 Piecewise stationary pre-segmentation (PSP) approach

Piecewise stationary segmentation algorithms are designed to detect all local stationary partitions composing a signal of interest. They are different from event segmentation algorithms which are designed to detect events of interest, stationary or not, inside a signal. They are mostly based on the analysis of the local statistical properties of the signal. In the context of synchronization analysis, we used advantageously one of these algorithms in order to detect the longer stationary parts inside the signals of interest before applying the traditional synchronization measure. Its results in a succession of windows of automatically locally adapted length. We call this pre-processing step: Piecewise stationary pre-segmentation or PSP.

The PSP algorithm has been proven to be useful as pre-treatment of synchronization analysis (Terrien & al., 2008b). In the case of different stationarity changes in the two channels, the univariate PSP (uPSP) method, described in (Terrien & al., 2008b), might fail to detect these changes. It is explained by the nature of this algorithm which only uses one of both channels for the segmentation. In order to be able to deal properly with this situation, we slightly modified the uPSP algorithm. Instead of using the auto spectrum of only one channel, the stationarity changes are detected using the cross spectrum, thus taking into account the statistical changes in both channels at the same time. We called this method bivariate PSP or bPSP for short. This algorithm, also based on the algorithm developed by Carré and Fernandez (Carré & Fernandez, 1998), can be described briefly as follows:

1. Decompose the signal x and y into successive dyadic partitions up to chosen decomposition level L+1
2. Compute and denoise, by undecimated wavelet transform, the log cross spectrum of each partition
3. Compute a binary tree of spectral distances between adjacent partitions
4. Search for the tree which minimizes the sum of the spectral distances by a modified version of the best basis algorithm of Coifman Wickerhauser
5. Apply the post processing steps described in (Carré & Fernandez, 1998) to deal properly with non dyadic partitions

The post processing steps consist mainly in applying the step 1 to 4 on each non terminal node with one level of decomposition and using the original best basis algorithm.

Our modified version of the best basis algorithm differs from the original one only by the node selection rule. This modification was necessary to differentiate the increase in the spectral distances due to the bias of the estimator, or due to signal symmetry around the considered cutting point. Each node $n_{i,j}$ has a cost, corresponding to the spectral distance, $c_{i,j}$. The classical decision rule concerning the selection of a father node $c_{i,j}$ is:

`if` ($c_{i,j} \leq c_{i+1,2j} + c_{i+1,2j+1}$) `then`
    Mark the node as a part of the best basis
`else`
    $c_{i,j} = c_{i+1,2j} + c_{i+1,2j+1}$
`endif`

The empirical modification of the selection rule is simply $\alpha.c_{i,j} \leq c_{i+1,2j} + c_{i+1,2j+1}$ with $a > 2$. We chose $a = 2.5$.

We showed that the use of the bPSP method avoids an arbitrary choice of the channel to which the stationary segmentation is based on and takes in to account the non stationarity of both signals present.

In a practical point of view, the parameters used in this method are mainly the number of decomposition levels in the segmentation procedure and in the wavelet denoising. These parameters are independent. The first one controls the minimal stationary length that the algorithm can detect. It must be roughly adapted to the signal of interest. A too high number of levels might increase the spectral estimation error, and lead to bad segmentation, due to an increased bias of the cross periodogram. The second parameter controlling the denoising of the spectra might lead to over smoothing of the estimated spectra and thus miss some important features in the different local stationary zones.

6.3 Results on synthetic signals

The configuration of the Rössler system used in this study is summarized table 2. The sampling rate used was 10 Hz.

Time t (s)	$\omega_1(t)$	$\omega_2(t)$	C(t)
0 - 204.8	0.65	0.55	0.3
204.8 - 307.2	1.2	0.55	0.01
307.2 - 460.8	1.2	1.1	0.8
460.8 - 563.3	0.65	1.1	0.01
563.3 - 819.1	0.65	0.5	0.5

Table 2. Parameters of the coupled Rössler system.

Figure 9 presents one example of the synthetic signals used and the corresponding synchronization analyses with H^2. The results obtained by the windowing approach show a synchronization pattern that approximately follows the coupling function. Important differences can be found during periods of low coupling between the two signals. Increasing the length of the window allowed us to significantly decrease the amplitude of the variations of the parameter but at the same time the boundaries of the different coupling periods become smoother. The bPSP approach shows marked transitions between the different coupling periods with relatively constant parameter values. More importantly, the algorithm is able to detect the change points situated at 307.2 and 563.3 s. This time instant corresponds to changes occurring in the second signals. The previous algorithm, presented in (Terrien & al., 2008b), would not have detected this transition, when using the top signal as reference, and not the transitions at 205 and 470 seconds when using the lower signal as reference. The differences between the coupling function $C(t)$ and the estimates are due to the intrinsic bias of H^2 as already highlighted in figure 6 of the paragraph 5.

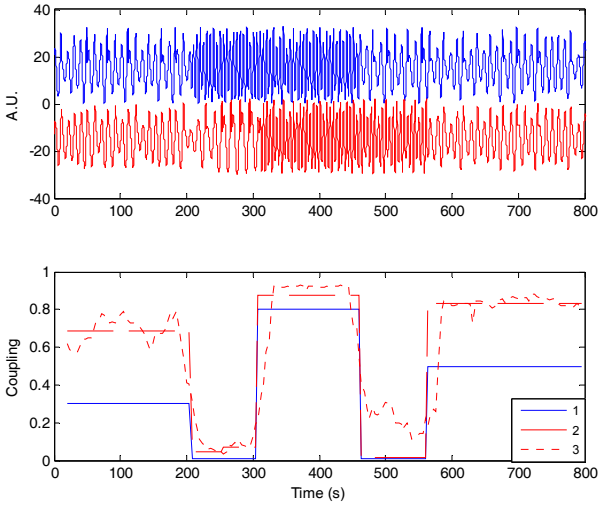


Fig. 9. Example of the output of the Rössler system (top panel) and the corresponding synchronization analysis using H^2 (bottom panel) obtained by the bPSP (2) and the windowing approach for a window length of 40 s (3). The coupling function C(t) is presented as a continuous line (1).

We might be interested in the robustness of a particular method or algorithm in order to apprehend its behavior in the presence of noise. This step is important since most biological signals are very noisy. The main parameters used in robustness analysis are the bias and the variance of the estimator.

We evaluated the robustness of the segmentation algorithm by Monte-Carlo simulations. For different noise (Gaussian white noise) levels, as express by the SNR, the bias and variance of the estimators were computed against the parameter values computed in the reference segments (segments that we would have obtained with a perfect segmentation).

This methodology allowed us to take into account the intrinsic bias of the synchronization measure.
The robustness analysis (bias and variance) for the parameters H^2 is presented figure 10. The stationary approach presents a lower bias than the windowing approach whatever the noise level. The variance obtained by the bPSP method is however greater. The analysis of the individual results showed that this high variance is mainly due to an over segmentation of each stationary zone.

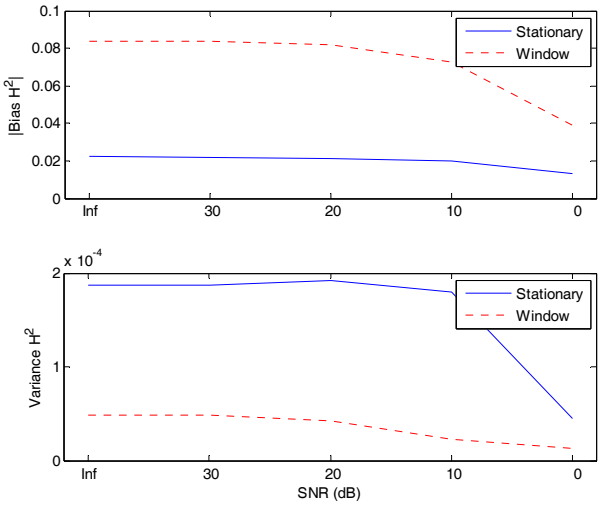


Fig. 10. Absolute value of the bias (top panel) and variance (bottom panel) obtained with the bPSP (continuous line) and the windowing approach (dotted line) for the parameter H^2.

6.4 Results on real EHG signals

The results of the segmentation of a broad band and narrow band contractile event recorded during labor are presented figure 11. We can clearly see, on both types of signals, that the algorithm is able to take into account changes occurring in both channels.

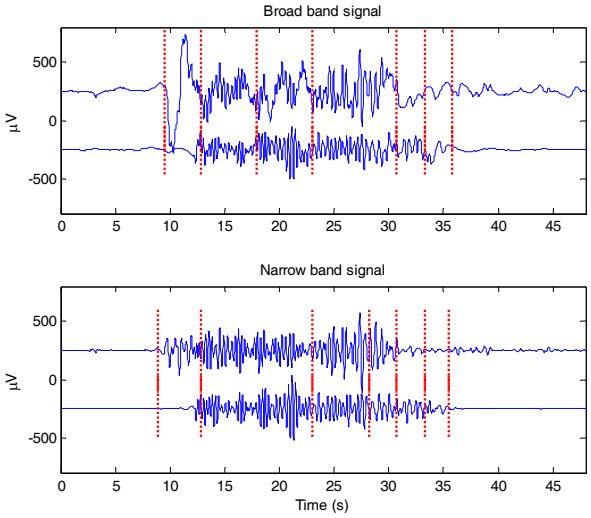


Fig. 11. Example of an electrical contraction burst recorded during the same contraction occurring in labor and their segmentation considering the broad band signals, raw signals (top panel), or a narrow band version of them, FWH filtered (bottom panel).

The results obtained with H^2 are presented figure 12. On the non shifted broad band signal the values of H^2 observed inside the contractile event are similar to those observed during the base lines (non EMG segments present before 10 s and after 35 s). The time shift of the signal of the second channel to compensate for the propagation delay of the contraction does not change the parameter pattern significantly. Only the first base line presents lower synchronization values as compared with no time shift. With the narrow band version of the signals, an increase in the parameter H^2 is clearly observable inside the contractile event when compared to the base line segments. The base lines still present relatively high values. The time shift of 0.18 s of the second signal causes a strong decrease in the base line values, while the values inside the contractile event increase or stay high. The effects of the time shift is less clear on the narrow band signal maybe due to the short delay between the two channels.

Looking at the results obtained with the windowing approach, no specific pattern can be observed in the same conditions. Moreover, the base lines present stronger or similar values of synchronicity than inside the contractions whatever the considered situation. Similar results were obtained for the other contractions tested.

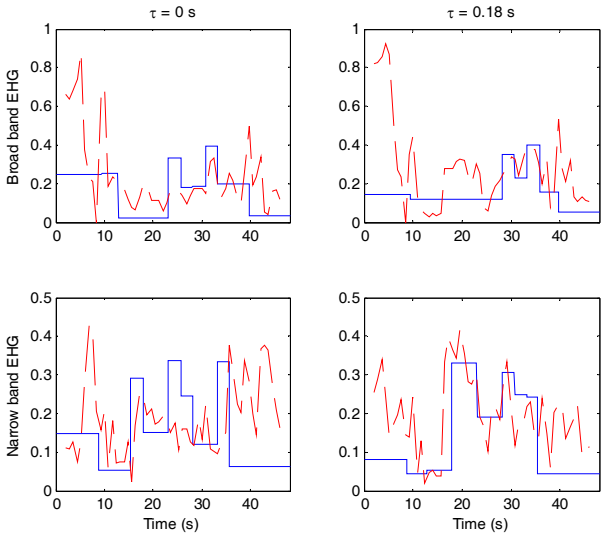


Fig. 12. H^2 profile obtained on the broad band (first line) or the narrow band (second line) signal of the contractile event presented figure 11 and for no time shift (first column) or a time shift of τ = 0.18 s (second column). The results obtained with the windowing approach in the same conditions are plotted with dashed lines.

6.5 Discussion

Most physiological signals are non stationary. Their characterization is therefore often difficult. In the context of synchronization analysis, the most commonly used approach is the windowing of the signals of interest before computing the different synchronization measures. The length of the analysis window is however an important parameter and controls the trade-off between stationary assumption of the signals in the window and the accuracy of the analysis. This parameter is often chosen as constant in time. We presented the advantages of using an automatic segmentation procedure of the signal that search for the longer locally adapted stationary parts in the context of synchronization analysis.

The bPSP approach shows reduced bias of the estimators when compared to the reference segmentation. The obtained variance is however higher than with the windowing approach due to over segmentation of the different stationary parts. A better control of the partition fusion procedure in the algorithm might allow us to reduce this over segmentation. The fusion procedure is based on a modification of the Coifman Wickerhauser algorithm, controlled by the parameter α. The adaptation of this parameter to the signal of interest might reduce this problem. The length of the minimal stationary zone is dependant of the levels of decomposition. Even if this parameter can be adapted to the signal of interest, a too high number of decomposition levels increases the bias of the spectral estimation due to the shortness of the analyzed data segments. This can introduce errors in the detection of the stationary parts of the signals. The use of other algorithms, which perform segmentation in a continuous time, can be a solution if precise time detection is needed. We have shown that

the PSP method can be applied to signals with different characteristics and that it gives satisfactory results when compared to the ones obtained with the windowing approach.
Specifically for real EHG, the numerous segments found on each analyzed burst confirm the high non stationarity of the EHG signal, even when band-pass filtered. This might indicate that the difference of surrogate measure profile characteristics between contractions is due to a difference in non stationarity rather than of nonlinearity, even if we cannot exclude definitively or totally this latter hypothesis. Both influences might coexist and be independent. More investigations are thus necessary in order to obtain a clear answer.
We have shown that the use of this method can clearly identify the proper treatment, filtering or time shift for example, needed to identify and highlight synchronization between different parts of the uterus during labor. It might be of use to monitor the evolution of the synchronization of the uterus from pregnancy to labor. This approach can also be used in the determination of the optimal synchronization measure for the uterine EMG.

7. Labor prediction

7.1 ROC curve analysis

In order to evaluate the possible use of the proposed parameters for the prediction of labor in monkey, we used the classical Receiver Operating Characteristic (ROC) curves. A ROC curve is a graphical tool permitting to evaluate a binary, i.e. two classes, classifier. A ROC curve is the curve corresponding to TPR (True Positive Rate or sensitivity) vs. FPR (False Positive Rate or 1 - Specificity) obtained for different parameter thresholds. ROC curves are classically compared by mean of the Area Under the Curve (AUC) and accuracy (ACC). The AUC was estimated by the trapezoidal integration method. We additionally used the Matthew's Correlation Coefficient (MCC) defined as:

$$MCC=\frac{TP*TN-FP*FN}{\sqrt{(TN+FN)(TN+FP)(TP+FN)(TP+FP)}} \tag{8}$$

where TP, TN, FP and FN stand respectively for True Positive, True Negative, False Positve and False Negative values.

7.2 Prediction using proposed corrected parameters

Under the hypothesis that the uterus synchronizes as labor approaches, we evaluated the potential of the nonlinear correlation coefficient as a predictor of labor on monkey. We evaluated the performance of the proposed parameters for labor prediction on a data set containing 35 pregnancy and 34 labor contractions. We compared the predictive capability of the nonlinear correlation coefficient (H^2) and of the 90 percentile corrected nonlinear correlation coefficient, H^2_{c90}. We have shown that the segmentation into piecewise stationary parts of EHG highlights synchronicity inside EHG burst, but gives rise to multiple values of H^2 within one burst (one for each stationary segment). For comparison purpose, we thus decided to also include the integral of the H^2 profile during the bursts, $\int H^2$, as a synchronization parameter for labor prediction.
The average results obtained on our data set with the parameter H^2_{c90} and H^2 are presented table 3. We can see that the original values of H^2 are very similar or slightly lower during

labor. This is in contradiction to what should be expected. Indeed, it is assumed that disorganized pregnancy contractions evolve into effectively synchronized labor contractions. The proposed H^2_{c90} synchronization value demonstrates a relatively important increase from pregnancy to labor, by a factor of nearly 10. The corrected measure is able to differentiate pregnancy and labor contractions and highlights the increase in synchronicity from pregnancy to the active phase of labor. The other synchronization value, $\int H^2$, shows a relative increase from pregnancy to labor by a factor of nearly 1.5 (table 3). The standard deviation obtained with this parameter is moreover relatively high in both situations. This corrected measure seems also to be able to differentiate pregnancy and labor contractions in spite of the high standard deviation obtained and a lower difference between pregnancy and labor.

Parameter	Pregnancy	Labor
H^2	0.1377±0.05	0.1111±0.02†
H^2_{c90}	0.0036±0.02	0.0310±0.02‡
$\int H^2$	3.0995±1.02	4.7104±1.12‡

Table 3. Mean synchronization measures along gestation (‡ and † indicate a significant difference at p = 0.01 and 0.05 respectively, Wilcoxon rank sum test).

The values obtained during pregnancy with H^2 were higher during pregnancy than during labor. This might indicate a decrease in the synchronization of the uterus in labor. Even if we have considered this option, the ROC curve obtained depicts a bad predictive parameter (Figure 13). Indeed, the first half of the curve lies on the diagonal, which is representative of a random decision. H^2_{c90} increases, as well as $\int H^2$, during labor. The ROC curves obtained are close together as depicted figure 13.

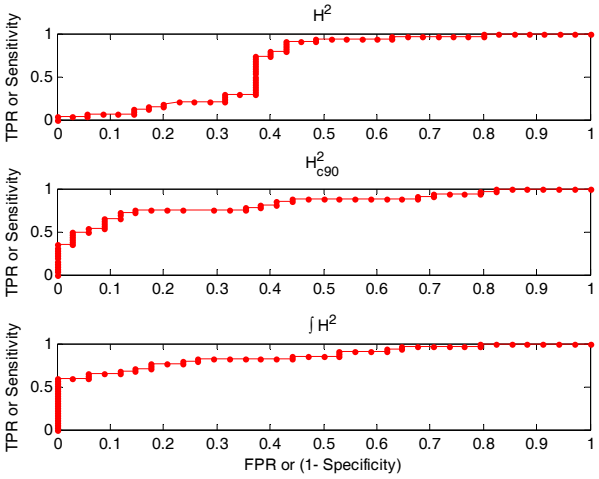


Fig. 13. ROC curve obtained with original, corrected H^2 measure with surrogates and integral of H^2 after piecewise stationary segmentation for the prediction of labor.

A summary of the ROC curves characteristics is presented table 4. The table indicates that both proposed parameters are better than the original parameter H^2. The performance values of the parameters H^2_{c90} and $\int H^2$ are very close. The main difference between the predictors concerns the first part of the curves, which might indicate a better separability between pregnancy and labor contractions by using the $\int H^2$ parameter. We can see for this parameter, the curve goes to higher TPR for low FPR as compared to the other parameter.

Parameter	AUC	ACC (%)	MCC
H^2	0.651	73.91	0.512
H^2_{c90}	0.827	80.30	0.607
$\int H^2$	0.863	79.71	0.596

Table 4. Comparison of ROC curves for labor prediction.

7.3 Discussion

By using a basic classifier, we have shown that the two new measures, H^2_{c90} and $\int H^2$, give better results than the original measure H^2 for the detection of labor on monkey. The performances of the two parameters are very close to each other. The use of $\int H^2$ seems appealing since the computation time required for this parameter is very low. The complementarity of these two approaches on EHG has to be evaluated in order to see if we can increase the prediction ratio. We used as a first step a linear classifier based on only one parameter. If we choose to combine several parameters, classifier like neural network or support vector machine classifiers might give better performances than classical linear ones using only one parameter at a time. To demonstrate this, two examples of support vector machine classifiers, using both H^2_{c90} and $\int H^2$, are illustrated Figure 14. They give better results than the previously obtained ones whatever the kernel used. The results obtained with a first order polynomial kernel are {ACC = 81.82; MCC = 0.639} and with a second order polynomial kernel {ACC = 87.88; MCC = 0.769}. This first attempt of combining both synchronization parameters, described here, shows that each parameter seems to carry, at least in part, different information about the relationship between both channels. This fact is mainly noticeable during pregnancy.

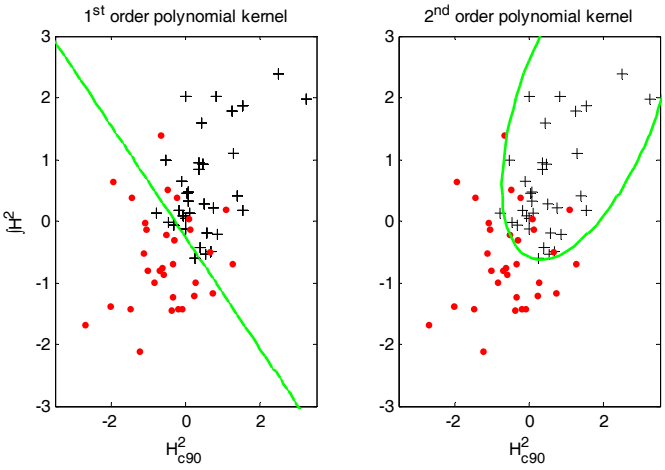


Fig. 14. Decision function (continuous line) obtained with a support vector machine parameterized by a 1st order polynomial (left) or a 2nd order polynomial kernel (right) trained on pregnancy (dot) and labor (cross) contractions dataset (normalized to have zero mean and unit variance on both components).

We have seen that taking into account nonlinearity and non stationarity in the EHG allows the prediction of labor with good accuracy. In the context of this work, we can only speculate as to what are the physiological origins of the observed changes, which occurred between pregnancy and labor. The increase in non-linearity, already observed (Radhakrishnan & al., 2000), as well as in non stationarity might be explained by different phenomena like:

- *The changes in the shape of the uterine cells action potential associated with the increase in the number of cell active at the same time*. The action potential during pregnancy presents a lower bursting frequency thus less complexity than during labor. The increase in the number of active cells may result in an increase in the interferences pattern between the action potential of cells that are not synchronized, increasing thus non stationarity.
- *The evolution of the propagation*. It is well admitted that the propagation of the electrical activity of the uterus increases as labor approaches. This might results in more synchronized active cells. It is probably the case at a local scale mainly due to the increase in the number of GAP junction close to delivery (Garfield & Hayashi, 1981; Blennerhassett & Garfield, 1991). The propagation is however also supported at a bigger scale by other structures like the bundles or the different orientation of the uterine muscle fibers. These complex structures that also evolve during gestation, might also modify the characteristics of the recorded signals.

We do believe that only a physiological mathematical model of the electrical activity of the uterus taking into account the specificity and complexity of the uterine contractility may answer clearly these open questions on the physiology of the uterine muscle.

The promising results obtained here are on signals from a single monkey and have to be tested on human EHG. This may have relevance in helping to solve a real public health problem, namely preterm labor.

8. Conclusion

Synchronization analysis methods are very powerful tools used to detect any coupling between two or more systems. All these methods suffer however from being very sensitive to the properties of the analyzed signals that may induce strong bias in the measure. The sources of bias might be simply experimental, like noise, or intrinsic to the signals of interest like non stationarity. In this chapter, we described the effects of these sources of bias on the synchronization analysis of synthetic as well as real EHG signals. We then proposed some general solutions rid the synchronization measures of bias. We finally defined two "new" synchronization measures that we used for analyzing the synchronization of the uterus during contractions.

The first method proposed consists in isolating the synchronization carried only by nonlinear components of a signal by using surrogate data analysis. This powerful method allowed us to differentiate to types of contractions: pregnancy and labor contractions. It permitted us then to propose a synchronization measure for labor prediction. The origins of this difference in terms of signal, as well as of physiological phenomenon, are still unknown but might be due partly to the increase in non stationarity of the EHG as labor approaches.

The second method uses a segmentation algorithm to segment stationary parts of the signals and thus to avoid a compromise between the analysis window length and stationarity assumption. In addition to the absence of this need of a trade-off between these two important factors, the proposed method, which is based on the non stationarity properties of both signals, thus improves a previously proposed method that uses only one of the signals for the segmentation. This general method can be used with any synchronization measures. It allowed us to highlight the synchrony noticeable within EHG bursts. We were able to define another synchronization measure, defined as the integral of the obtained synchronization profile.

We finally showed that the two synchronization measures proposed by correcting the original nonlinear correlation coefficient, H^2, are able to show an increase of the synchronization of the uterus during labor, when compared to pregnancy. Both parameters have nearly the same prediction performance with an accuracy of nearly 80%. The original measure, H^2, exhibited a decrease or at least no changes from pregnancy to labor, which would have lead to conclusion in contradiction with the physiology of the uterus. This is a typical example where a blind use of signal processing tools without taking into account possible sources of bias or limitation of the methods might lead to dramatic interpretation errors.

9. Acknowledgments

This work was supported by the Icelandic center of research, RANNIS, and the region Picardie-France in the project "Pôle périnatalité Enfance".

10. References

Ansari-Asl K., Senhadji L., Bellanger J.-J. & Wendling F. (2006), "Time-frequency characterization of interdependencies in nonstationary signals: application to epileptic EEG," *IEEE Trans Biomed Eng*, vol. 52, pp. 1218-1226.

Baccala L.A. & Sameshima K. (2001a), "Partial directed coherence: a new concept in neural structure determination," *Biol Cybern*, vol. 84, pp. 463-74.

Baccala L.A. & Sameshima K. (2001b), "Overcoming the limitations of correlation analysis for many simultaneously processed neural structures," *Prog Brain Res*, vol. 130, pp. 33-47.

Blanco S., Garcia H., Quiroga R.Q., L. Romanelli, and O. A. Rosso (1995), "Stationarity of the EEG Series," *IEEE Engineering in Medicine and Biology Magazine*, vol. 14, pp. 395-399.

Blennerhassett M.G. & Garfield R.E. (1991), "Effect of gap junction number and permeability on intercellular coupling in rat myometrium," Am J Physiol, vol. 261, pp. C1001-9.

Borgnat P. & Flandrin P. (2009), "Stationarization via surrogates," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2009, pp. P01001.

Buhimschi C., Boyle M.B. & R. E. Garfield (1997), "Electrical activity of the human uterus during pregnancy as recorded from the abdominal surface," *Obstet Gynecol*, vol. 90, pp. 102-11.

Carré P. & Fernandez C. (1998), "Research of stationary partitions in nonstationary processes by measurement of spectral distance with the help of nondyadic Malvar's decomposition," presented at IEEE-SP International Symposium on Time-Frequency and Time-Scale Analysis, Pittsburgh, PA, USA.

Devedeux D., Marque C., Mansour S., Germain G. & Duchene J. (1993), "Uterine electromyography: a critical review," *Am J Obstet Gynecol*, vol. 169, pp. 1636-53.

Duchêne J., Marque C. & S. Planque (1990), "Uterine EMG signal: Propagation analysis," presented at the Annual International Conference of the IEEE EMBS.

Euliano T.Y., Marossero D., Nguyen M.T., Euliano N.R., Principe J. & R. K. Edwards (2009), "Spatiotemporal electrohysterography patterns in normal and arrested labor," *Am J Obstet Gynecol*, vol. 200, pp. 54 e1-7.

Garfield R.E. & Hayashi R. H. (1981), "Appearance of gap junctions in the myometrium of women during labor," *Am J Obstet Gynecol*, vol. 140, pp. 254-60.

Garfield R.E., Maul H., Shi L., Maner W., Fittkow C., Olsen G. & Saade G.R. (2001), "Methods and devices for the management of term and preterm labor," *Ann N Y Acad Sci*, vol. 943, pp. 203-24.

Garfield R.E. & Maner W.L. (2007), "Physiology and electrical activity of uterine contractions," *Semin Cell Dev Biol*, vol. 18, pp. 289-95.

Karlsson B., Terrien J., Guðmundsson V., Steingrímsdóttir T. & Marque C. (2007), "Abdominal EHG on a 4 by 4 grid: mapping and presenting the propagation of uterine contractions," presented at 11th Mediterranean Conference on Medical and Biological Engineering and Computing, Ljubljana, Slovenia.

Kus R., Kaminski M. & Blinowska K.J. (2004), "Determination of EEG activity propagation: pair-wise versus multichannel estimate," *IEEE Trans Biomed Eng*, vol. 51, pp. 1501-10.

Leman H. & Marque C. (2000), "Rejection of the maternal electrocardiogram in the electrohysterogram signal," *IEEE Trans Biomed Eng*, vol. 47, pp. 1010-7.

Leman H., Marque C. & Gondry J. (1999), "Use of the electrohysterogram signal for characterization of contractions during pregnancy," *IEEE Trans Biomed Eng*, vol. 46, pp. 1222-9.

Maner W.L. & Garfield R.E. (2007), "Identification of human term and preterm labor using artificial neural networks on uterine electromyography data," *Ann Biomed Eng*, vol. 35, pp. 465-73.

Pereda E., Quiroga R. Q. & Bhattacharya J. (2005), "Nonlinear multivariate analysis of neurophysiological signals," *Progress in Neurobiology*, vol. 77, pp. 1-37.

Pijn J.P., Vijn P.C., Lopes da Silva F.H., Van Ende Boas W. & Blanes W. (1990), "Localization of epileptogenic foci using a new signal analytical approach," *Neurophysiol Clin*, vol. 20, pp. 1-11.

Prichard D. & Theiler J. (1994), "Generating surrogate data for time series with several simultaneously measured variables," *Physical Review Letters*, vol. 73, pp. 951-954.

Quian Quiroga R., Kreuz T. & Grassberger P. (2002), "Event synchronization: a simple and fast method to measure synchronicity and time delay patterns," *Phys Rev E Stat Nonlin Soft Matter Phys*, vol. 66, pp. 041904.

Radhakrishnan N., Wilson J. D., Lowery C., Murphy P. & Eswaran H. (2000), "Testing for nonlinearity of the contraction segments in uterine electromyography," *International Journal of Bifurcation and Chaos*, vol. 10, pp. 2785-2790.

Ramon C., Preissl H., Murphy P., Wilson J.D., Lowery C. & Eswaran H. (2005), "Synchronization analysis of the uterine magnetic activity during contractions," *Biomed Eng Online*, vol. 4, pp. 55.

Schreiber T. & Schmitz A. (2000), "Surrogate time series," *Physica D*, vol. 142, pp. 346-382.

Terrien J., Germain G. & Marque C. (2008a), "Ridge extraction from the time-frequency representation (TFR) of signals based on an image processing approach: Application to the analysis of uterine electromyogram AR TFR," *IEEE Trans Biomed Eng*, vol. 55, pp. 1496-1503.

Terrien J., Hassan M., Marque C. & Karlsson B. (2008b), "Use of piecewise stationary segmentation as a pre-treatment for synchronization measures," presented at 30th Annual International Conference of the IEEE EMBS, Vancouver, Canada.

Young R.C. & Hession R.O. (1999), "Three-dimensional structure of smooth muscle in term-pregnant human uterus," *Obstet Gynecol*, vol. 93, pp. 94-99.

Young R.C. (1997), "A computer model of uterine contractions based on action potential propagation and intercellular calcium waves," *Obstet Gynecol*, vol. 89, pp. 604-8.

Wendling F., Bartolomei F., Bellanger J.J. & Chauvel P. (2001), "Interpretation of interdependencies in epileptic signals using a macroscopic physiological model of the EEG," *Clin Neurophysiol*, vol. 112, pp. 1201-18.

6

Multichannel analysis of EEG signal applied to sleep stage classification

Zhovna Inna[1] and Shallom Ilan[1,2]
[1]*Ben Gurion University,* [2]*Audiocodes Company*
Israel

1. Introduction

The human brain is a complex organ with approximately 100 billion nerve cells (neurons) transmitting electrochemical signals. Regardless of what state we are in, whether asleep or awake, our brain produces brainwaves that can be observed and used for clinical and study applications. German psychiatrist named Hans Berger was the first to measure this electrical activity in humans in 1924, he called it electroencephalogram (EEG). It is a non-invasive method of measuring electrical activity of the brain by recording the brainwaves with electrodes placed on the scalp. Ever since his discovery, EEG has been used to diagnose many medical conditions, identifying the location of a suspected brain tumor, or a disease in the brain such as epilepsy and Parkinson's disease.
In this research the EEG method was employed for sleep disorders study. Most of us refer to sleep as a passive process; in fact the opposite is the truth, sleeping is an extremely active process. Sleep complexity is poorly understood during daily lives, our brain is more active during sleep than it is during the normal waking state. There is a distinct "architecture" of sleep, which includes five stages; four are defined as Non-Rapid Eye Movement and one as Rapid Eye Movement. These sleep stages patterns can be observed in EEG signal by change of waveforms, frequency and magnitude.
EEG is a widespread method for sleep disorders diagnostic and research. The challenge of EEG is the interpolation of recorded signals. This difficult and time consuming task is performed mainly manually by an EEG expert (technician or physiologist). In order to simplify this manual process, an automatic sleep stage detection and classification method should be analyzed. In this chapter a new method for automatic detection and classification of sleep stages using a multichannel signal analysis is proposed.

1.1 Previous Work

The idea of an automatic classification system for EEG signals in general and for sleep stages in particular, is not novel. There have been several researches utilizing various methods to achieve high results for automatic classification of EEG signals into sleep stages. One of the most common methods is the neural network and fuzzy rule method. Researches (Kerkeni et al., 2005), (Pinero et al., 2004), (Heiss et al., 2001), (Shimada & Shiina, 1999) and (Schaiboldet al., 2003) examined such methods along with some EEG signals featuring

extraction algorithm achieved acceptable results. Work (Kerkeni et al., 2005) and (Pinero et al., 2004) had accuracy of around 70%, (Shimada & Shiina, 1999) achieved 83% of classification using in addition multichannel information, (Schaiboldet al., 2003) reached 84.4% and (Heiss et al., 2001) succeeded in reaching 86.7% of accuracy. In (Gerla & Lhotska, 2006), the authors used Hidden Markov Model (HMM) for multichannel EEG signal analysis and principal component analysis (PCA) for dimension reduction; they accomplished only 70%-80% accuracy. Furthermore, the author of (Ghosh & Zhong, 2002) used the HMM method, however with an AR model for vector feature extraction, reaching nearly 80% of accuracy. In (Wanli & Luo, 2007) the authors used the conditional random field (CRF) method which is similar to the HMM method and attained merely 70% accuracy. Another analysis method is wavelet transform, used in (Qianli et al., 2005) and (Song et al., 2007), yielded no suitable result. In (Masaaki et al., 2002), for sleep stage classification, a waveform recognition method and decision–tree learning was used with hardly 70% accuracy. Clustering by k-mean is also a useable method, e.g. in (Agerwal & Gotman, 2001) the classification accuracy was 80.6%.
In this section we presented the recent researches in the sleep stage classification of EEG signals. Some of these researchers achieved quite good results, an accuracy of 80-86%. In spite of that it is still not good enough for clinical application, and more research needs to be done. In this research we try to achieve higher accuracy rate and we set as a wishful thinking target to cross the 90%.

2. Problem Definition

2.1 Physiological Background

2.1.1 ElectroEncephaloGraphy (EEG)

EEG is a non-invasive neurophysiologic measurement of an electrical activity produced by the brain. Usually, EEG measurement is performed during a physical task that stimulates the brain cells, such as blinking, talking, sleeping, etc. The measurement involves a set of electrodes placed on different areas of the outside surface of the scalp. Electrodes are sensors that sense the electrical activity of the drain through the scalp. The electrical activity is expressed as analog signals, which is being sampled and convert into a digital signal by an analog to digital converter. The digital data is collected and stored for further analysis.
The recorded EEG signals are characterized by frequency, patterns, and amplitude. Traditionally, EEG is defined by 5 frequency bands and 5 different wave forms (Zumsteg et al., 2004): Delta waves with frequency range of 0-4Hz / 0.5-3Hz, Theta waves with frequency range of 4-8Hz/3-7Hz, Alpha waves with frequency range between 8-11/12Hz, Beta waves with frequency range of 12/13-26Hz and Gamma waves with frequency of approximately 26-100Hz.
The expanded form of the EEG is the Video EEG. Video EEG consists of recording an electrical activity of the brain along with a simultaneous recording of audio and video of patient's environment. It can help physician to determine if there is a correlation between movement and abnormal brain activity. In this research, a video EEG has been used mostly for artifacts reduction.

2.1.2 Sleep Stages

Since the early 20th century, human sleep has been described as a succession of five recurring stages, (or sixth including awakening). Sleep stage transition is characterized by abrupt changes in frequencies and amplitudes of the EEG signal.

The first four stages are defined as a "Non-Rapid Eye Movement" (NREM) sleep and the fifth stage is defined as a "Rapid Eye Movement" (REM) sleep (Zumsteg et al., 2004), (Kelly, 1991), (Pace-Schot & Hobson,2002).

The NREM and REM sleep alternate in 90-110 minute cycles. A normal sleep pattern begins at about 80-90 minutes of NREM sleep, followed by an REM period for about 10-20 minutes. This NREM-REM cycle repeats about 4-6 times during the sleep.

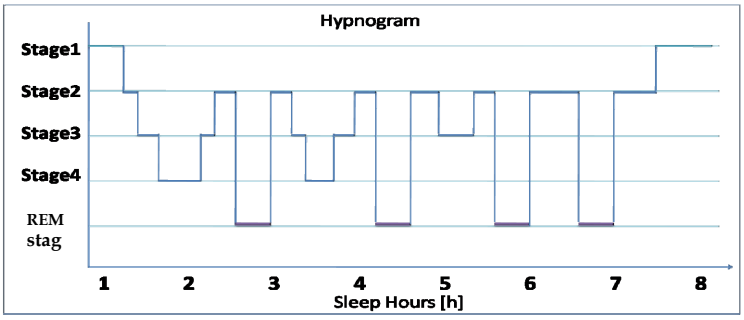


Fig. 1. Hypnogram - Typical sleep cycle.

The five stage cycle of sleep repeats itself throughout 7-8 hours during the sleep. Stage 1 starts by shutting the eyes, and cycles through stages 2, 3 and 4. From stage 4 the processes goes recursively back, when stage 1 is replaced by the REM sleep (Fig. 1). In the successive cycles of the night, the amount of stages 3 and 4 decreases, and the proportion of the cycles occupied by REM sleep tends to increase.

Wakefulness: At the wakefulness state the EEG pattern alternates between two main wave forms. One is the beta wave that has fast activity of 13-26 Hz and low voltage of 10-30 μV. The second wave form is the alpha wave that has higher voltage of 20-40 μV and slower activity of 8-12 Hz.

NREM sleep: The NREM sleep occurs for 75-80 % of total sleep time and it is characterized by low frequency and high voltage wave activity that correspond to increasing depths of sleep. According to the Academy of Sleep Medicine (ASM) the NREM sleep can be divided into four separate stages, stage 1 to stage 4.

Stage 1: The duration of stage 1 is about 5 to 10 minutes, it can be defined as a gateway state between the awake state and sleep state. This stage is characterized by relative low EEG voltage and slow movements of eye rolling. Alpha waves (8-13 Hz), seen in the awake state, disappear in the first stage and are replaced by theta waves (4-7 Hz).

Stage 2: Stage 2 takes approximately 45-55% of the total sleep. This stage is characterized by a lack of eye movements, sleep spindles, and K-complexes . Sleep spindles and K-complexes are two distinct brain wave forms appearing on the background of theta waves.

A “Sleep spindle” is a burst of brain activity visible on EEG, it consists of 11-15 Hz waves that occur for 0.5 to 1.5 seconds. A “K-complex” is a sudden, brief, high amplitude

waveform of EEG. It consists of a brief high-voltage peak, and lasts for longer than 0.5 seconds.

Stage 3: This stage refers to a deep sleep and happens for 35-45 minutes after falling asleep. Stage 3 takes approximately 12% of the NREM sleep. This stage is characterized by 20-40% of delta (slow) wave and high amplitude (>75 μV). Additionally, a "K-complex" and "Sleep spindle" can also appear at this stage.

Stage 4: Stage 4 is very similar to stage 3, in some cases both are regarded as one. Stage 4 refers to a very deep sleep. This stage presents around 13% of the NREM sleep and more than 50% of it is characterized by delta waves.

REM sleep: Most dreaming occurs during the REM sleep, therefore a burst of prominent rapid eye movement appears in the EEG at this stage. Adults spend about 20-25% of their sleep cycle in the REM sleep (approximately 10 out of 90 minutes of one cycle). The EEG in this period is aroused and it is very similar to stage 1, it exhibits a mixed frequency and low voltage with occasional bursts of "saw-tooth" waves.

2.2 Motivation

Sleep is absolutely essential for a normal, healthy activity. Studies have shown that for normal functionality of the immune system, sleep is a necessity. It is also essential for maintaining a normal operation of the nervous system and the ability to perform both physically and mentally. In addition, sleep is essential for learning and for normal healthy cell growth. About third of the population suffers from chronic long-term disorders and occasional sleep problems. There are more than 70 different sleep disorders that are classified into three categories: lack of sleep (e.g. insomnia), disturbed sleep (e.g. obstructive sleep apnea) and excessive sleep (e.g. narcolepsy). These disorders can have a very significant effect on our daily life, such as chronic tiredness, difficulty to wake up and fall asleep, unwanted numbing, and even heart diseases.

In most cases, sleep disorders can be easily managed once they are properly diagnosed. One of the modern tools for sleep disorder diagnosis is the EEG test. The test provides a record of the patient's brain wave pattern through the whole night (7-9 hours of data). The EEG monitors various stages of sleep, which are later interpreted by a visual analysis specialist. Such analysis can be difficult, time-consuming, tiresome procedure, and not necessarily accurate. In order to assist in this toilsome process and to achieve a better diagnosis, automatic classifications of EEG sleep pattern must be developed.

The aim of this research is to create a novel method for automatic sleep stage classification, using a multichannel EEG signal. Automatic classification will help the specialists to interpret the EEG signal and to conclude a suitable diagnosis.

2.3 Problem Definition

As mentioned above this research deals with definition and classification of EEG signals. One of the biggest difficulties of neurologists is the interpretation of an EEG signal. Most of the neurologic world still processes the EEG signals manually, by scanning the EEG records visually.

The goal of this research is to solve this problem by offering a method for an automatic EEG signal classification. The specific difficulty that this research tries to deal with is the

detection and classification of different sleep stages in patients who suffer from sleep disorders.

Fig. 2. System's block diagram.

Numerous researches have been done in this field, however most of them are still not sufficient for clinical use. Consequently, this research aims to achieve higher classification accuracy for future clinical use in sleep EEG and in other EEG applications, by using the multichannel analysis approach.

3. Theoretical Overview

3.1 Single and Multichannel Analysis

Signal processing can be divided into two main analyses; single channel analysis and multichannel analysis. The single-channel analysis is very common in the signal processing world. The use of single channel analysis is found in various fields; Medicine (EEG, ECG, EMG etc.), Geophysics and Speech processing. Although the use of single channel analysis for some systems description can produce incorrect system model and get false results, this analysis simplifies the signal processing part within a complicated systems, when the input signal is represented by a scalar $s(t)$. On the other hand the multichannel analysis complicates the computations and the system model. Nevertheless, for several processes the multichannel analysis may offer a much more accurate model. In case of multichannel, the input signal is represented by a d dimensional vector $[s_1(t)\, s_2(t)\, s_3(t) \ldots s_d(t)]^T$, where d represent the number of channels.

The research on sleep stage classification is vastly wide and variant, but almost in all researches the single signal analysis approach is used, (Kerkeni et al., 2005), (Masaaki et al., 2002), (Agerwal & Gotman, 2001), (Estrada et al., 2004), (Krajca et al., 2005), (Van Hese et al., 2001), (Shimada et al., 1998), (Sun et al., 1993). Although the classification system receives an input of more than one EEG channel, the analysis is made per single channel only. Hence, the main goal of this research is to examine sleep stage classification by multichannel analysis of multichannel EEG signal. Fig. 3 & 4 present block diagrams of the discussed signal analysis, fig. 3 demonstrates the single signal analysis and fig. 4 exhibits the multichannel analysis.

The EEG is a digital record of a biological signal that describes the electrical activity of the human brain. The recording is performed by using multiple electrodes (4-128 channels) placed on the scalp, during the sleep. Each electrode that records electrical brain activity contains important information about the neurological activity of the patient during the sleep. The spread of several electrodes on the scalp causes an overlap between multi channel recorded data due to electrodes neighborhood, which in many cases redundant. This neighborhood causes by definition an inter relations between the different sensors. Therefore, when taking this kind of data under consideration, it is much more appropriate to use the multichannel analysis, which considers the relations between the channels and produces a more accurate assumption about the sleep mechanism.

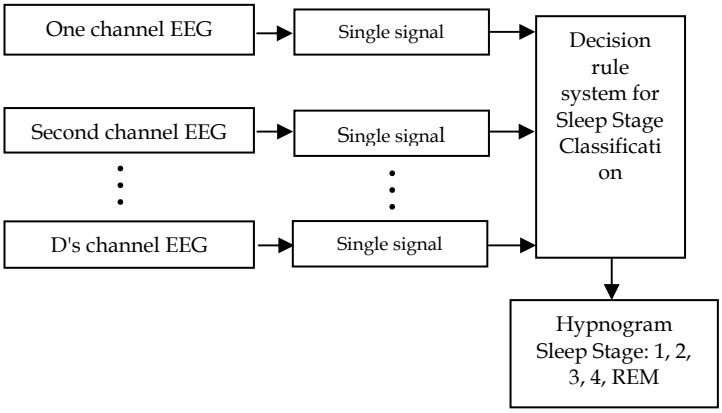


Fig. 3. Block diagram of traditional single analysis.

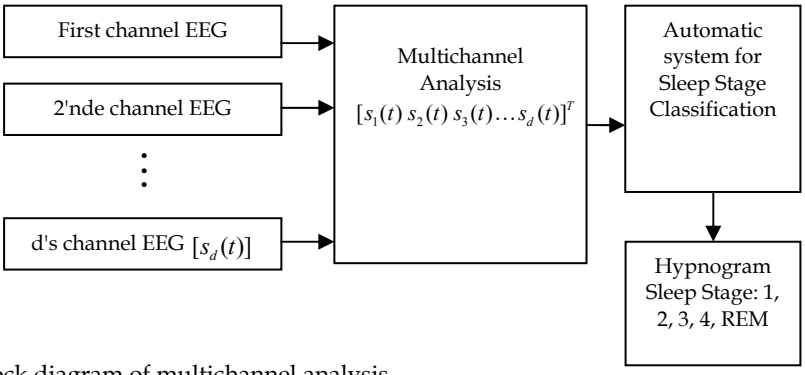


Fig. 4. Block diagram of multichannel analysis.

3.2 Multichannel Analysis

In the previous section the single channel analysis and the necessity of multichannel analysis for EEG signal processing was discussed. The classification method that is presented in this work is based on the multichannel analysis, which will be described in details in the following section. In addition review on other researches in the field will be presented in this section.

3.2.1 Overview

The objective of this work is to classify the EEG signal into the correct sleep stage. For this purpose the EEG signal has to be described by some mathematical model. The most common mathematical model approach in the EEG signal research is the parametrical approach which represents the EEG signal by a specific set of parameters.

There are three main types of parametrical models; the all-pole model known as the Autoregressive (AR) model, the all-zero model known as Moving Average (MA) model and the pole-zero model known as Autoregressive Moving Average (ARMA) model (a mix of AR and MA models) (Makhoul, 1975). These models are in fact filters, the analyzed signal is assumed to be the output of the filter when the input is a white noise.
The most extensively used model, in biomedical signal processing, is the scalar AR model (Makhoul, 1975), (Kay, 1988) and (Priestley, 1989). Several researches in the EEG field showed that the use of scalar AR model can describe the EEG signal in a proper way, yielding a feasible classification. More than 27 years ago, the potential use of the parametrical model for EEG signal analysis, and particularly the scalar AR model was forecasted by (Isaksson et al., 1981) and (Jansen et al., 1981). In (Isaksson et al., 1981) work, discussed the potential that EEG research has and presented information about the parametrical and not parametrical signal analysis in EEG signal. The (Jansen et al., 1981) work is focused on AR model and reviews methods for parameters and model order estimation. E. Estrada and H. Nazeran in their works (Estrada et al., 2004), (Estrada et al., 2005) and (Ebrahimi et al., 2007), attempt to classify the EEG signal into right sleep stages by scalar AR models.
The mentioned studies demonstrate a successful use of scalar AR model for EEG signal in different applications. In the PhD thesis from 1990 (Flomen, 1990), Felix A.Flomen demonstrated the use of AR model for EEG signals and the developing of the multichannel approach for EEG signals analysis, drawing a comparison between them. This work (Flomen, 1990), was one of the pioneers in the MAR model using General Log Likelihood (GLLR) distortion measure, for multichannel analysis signal. Nonetheless, the use of Multichannel AR model (MAR) (Flomen, 1990), (Kay, 1988), (Priestley, 1989) for EEG signal is extremely rare. In (Andreson et al., 1998) work, the multichannel analysis for modeling the EEG signal is used. By modeling the multichannel EEG signal using MAR model, (Andreson et al., 1998) tried to find a satisfying solution for the "Mental Tasks Model and Classification" problem. Furthermore, (Andreson et al., 1998) proved that MAR model for EEG signal provides not only satisfying classification results, but better results than provided by the scalar AR model. The use of MAR model for EEG signal is still not extensive, however, based on the mentioned researches it is clear that the use of MAR model can help with the multichannel classification problem for the EEG signal. Therefore, this research examines sleep stages classification problem, by using the MAR model as a basic EEG signal model.

3.2.2 Multichannel AR Model

The following paragraph will explain in details the MAR model chosen for the multichannel EEG signal in this research.
The basic assumption for the MAR model analysis is that the analyzed signal is assumed to be stationary. Therefore, the MAR model is defined for each EEG signal $\underline{s}(n)$, of duration T, which is assumed to be stationary.

The d dimensional EEG signal $\underline{s}(n)$ is defined as:

$$\underline{s}(n) = [s_1(n)\ s_2(n) \ldots s_d(n)]^T, n = 1 \ldots N \tag{1}$$

Where N is the number of samples per signal duration T.

By the d - dimensional MAR model, the signal $\underline{s}(n)$ is given as a linear combination of past observations and some random input $\underline{u}(n)$,as presented in fig. 5:

$$\underline{s}(n) = -\sum_{k=1}^{p} \underline{A}(k)\underline{s}(n-k) + \underline{G}\underline{u}(n) \tag{2}$$

Where $\underline{G}$ is the gain factor, p is the model order and $\underline{A}(k), k = 1,..,p$ are the $d \times d$ matrices coefficients of the MAR model.
The matrices coefficients $\underline{A}(k), k = 1,..,p$ defined as:

$$\underline{A}(k) = \begin{bmatrix} a_{11}(k) & a_{12}(k) & \dots & a_{1d}(k) \\ a_{21}(k) & a_{22}(k) & \dots & a_{2d}(k) \\ \vdots & \vdots & \ddots & \vdots \\ a_{d1}(k) & a_{d2}(k) & \dots & a_{dd}(k) \end{bmatrix}, k = 1,\dots,d \tag{3}$$

This model is an all-pole model that can be presented in the z plan as the transfer function $H(z)$:

$$H(z) = \frac{\underline{G}}{1 + \sum_{k=1}^{p} \underline{A}(k) z^{-k}} \tag{4}$$

From Eq. (4), fig. 5 and fig. 6, it is evident that MAR is a filter, when the input $\underline{u}(n)$ is a white noise signal and the output signal is $\underline{s}(n)$.

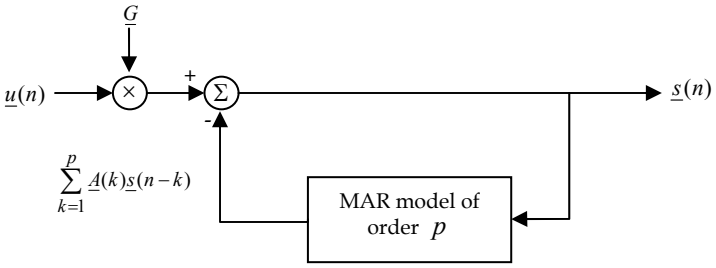


Fig. 5. All-pole model in the time domain

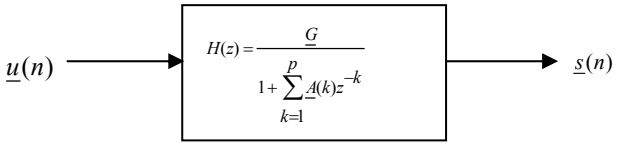


Fig. 6. All-pole model in the frequency domain

The input signal $\underline{u}(n)$ is a totally unknown biological signal, actually it is considered as inaccessible signal, therefore the signal $\underline{s}(n)$ can be linearly predicted only approximately

by (2) and it is defined as:

$$\tilde{\underline{s}}(n) = -\sum_{k=1}^{p} \underline{A}(k)\underline{s}(n-k) \tag{5}$$

Then the error between the actual value $\underline{s}(n)$ and the predicted value $\tilde{\underline{s}}(n)$ is given by:

$$\underline{\varepsilon}(n) = \underline{s}(n) - \tilde{\underline{s}}(n) = \underline{s}(n) + \sum_{k=1}^{p} \underline{A}(k)\underline{s}(n-k) \tag{6}$$

Since the assumption that the input $\underline{u}(n)$ is inaccessible, the gain $\underline{G}$ does not participate in the linear prediction of the signal. And so, it is irrelevant to determine a value for $\underline{G}$. However (6) can be rewritten as:

$$\underline{s}(n) = -\sum_{k=1}^{p} \underline{A}(k)\underline{s}(n-k) + \underline{\varepsilon}(n) \tag{7}$$

From (2) and (7) the following can be seen:

$$\underline{G}\underline{u}(n) = \underline{\varepsilon}(n) \tag{8}$$

Meaning, the input signal is proportional to the error signal.
From comparing (6) with (8), we get:

$$\underline{G}\underline{u}(n) = \underline{\varepsilon}(n) = \underline{s}(n) - \tilde{\underline{s}}(n) \tag{9}$$

By squared Eq. (9) and taking the expectation, we receive:

$$E\left\{(\underline{G}\underline{u}(n))^2\right\} = \underline{G}^2 E\left\{\underline{u}^2(n)\right\} = E\left\{\underline{\varepsilon}^2(n)\right\} = E\left\{(\underline{s}(n) - \tilde{\underline{s}}(n))^2\right\} \tag{10}$$

The input $\underline{u}(n)$ is assumed to be a sequence of uncorrelated samples with zero mean and unit variance, i.e. $E\left\{\underline{u}(n)\right\} = 0$, for all n, and $Var\left\{\underline{u}(n)\right\} = 1$. The derived equation is:

$$E\left\{\underline{u}^2(n)\right\} = 1 \tag{11}$$

By placing (11) into (10), we receive:

$$\underline{G}^2 = E\left\{\underline{\varepsilon}^2(n)\right\} = E\left\{(\underline{s}(n) - \tilde{\underline{s}}(n))^2\right\} \tag{12}$$

When (12) can be written as:

$$\begin{aligned} &\underline{G}^2 = E\left\{\underline{\varepsilon}^2(n)\right\} = E\left\{(\underline{s}(n) - \tilde{\underline{s}}(n))^2\right\} = E\left\{(\underline{s}(n) - \tilde{\underline{s}}(n))(\underline{s}(n) - \tilde{\underline{s}}(n))^T\right\} \\ &\Rightarrow E\left\{(\underline{s}(n) - \tilde{\underline{s}}(n))(\underline{s}(n) - \tilde{\underline{s}}(n))^T\right\} = \\ &E\left\{\underline{s}(n)(\underline{s}(n) - \tilde{s}(n))^T\right\} - E\left\{\tilde{\underline{s}}(n)(\underline{s}(n) - \tilde{\underline{s}}(n))^T\right\} \end{aligned} \tag{13}$$

From (13) and (9) we get:

$$\begin{aligned} &E\left\{\underline{s}(n)(\underline{s}(n) - \tilde{\underline{s}}(n))^T\right\} - E\left\{\tilde{s}(n)(s(n) - \tilde{s}(n))^T\right\} = \\ &E\left\{\underline{s}(n)(\underline{s}(n) - \tilde{\underline{s}}(n))^T\right\} - E\left\{\tilde{\underline{s}}(n)\underline{\varepsilon}^T(n)\right\} \end{aligned} \tag{14}$$

By the orthogonality principle, the next expression is valid:

$$E\left\{\underline{\tilde{s}}(n)\underline{\varepsilon}^T(n)\right\}=0 \tag{15}$$

The (12), (14) and (15) yields:

$$\begin{aligned}&\underline{G}^2=E\left\{\underline{\varepsilon}^2(n)\right\}=E\left\{\underline{s}(n)(\underline{s}(n)-\underline{\tilde{s}}(n))^T\right\}\\&\Rightarrow E\left\{\underline{s}(n)(\underline{s}(n)-\underline{\tilde{s}}(n))^T\right\}=E\left\{\underline{s}(n)\underline{s}^T(n)\right\}-E\left\{\underline{s}(n)\underline{\tilde{s}}^T(n)\right\}\end{aligned} \tag{16}$$

Now, by placing (5) into (16) we receive:

$$\underline{G}^2=E\left\{\underline{\varepsilon}^2(n)\right\}=E\left\{\underline{s}(n)\underline{s}^T(n)\right\}+\sum_{k=1}^{p}\underline{A}(k)E\left\{\underline{s}(n)\underline{s}^T(n-k)\right\} \tag{17}$$

When the autocorrelation matrix of lag i is defined as:

$$\underline{R}(i)=E\left\{\underline{s}(n)\underline{s}^T(n-i)\right\} \tag{18}$$

Where every $R(i), i=1,\ldots,p$ is a $d\times d$ matrix.

By placing (18) into (17) we receive the estimation of residual error covariance matrix, as follow:

$$\underline{G}^2=E\left\{\underline{\varepsilon}^2(n)\right\}=\underline{R}(0)+\sum_{k=1}^{p}\underline{A}(k)\underline{R}^T(k) \tag{19}$$

This expression will assist us in the forward MAR parameters and order estimation.

The accuracy of the MAR model depends mainly on the $\underline{A}(k)$ coefficients estimation and the model order *p* definition; therefore it is critical to estimate it as accurate as possible. There are several ways to estimate the coefficients and the model's order. To estimate the coefficients, the Yule-Walker (YW) equations (Kay, 1988), (Wiggins & Robinson, 1965) should be solved. These equations can be solved by the Levinson, Wiggens, Robinson (LWR) algorithm (Wiggins & Robinson, 1965). The optimum order was estimated by Akaike's Information Criterion (AIC) (Kay, 1988), (Priestley, 1989).

3.2.2.1 Yule-Walker equation for coefficients estimation

The $\underline{A}(k)$ coefficients estimation is an extremely important phase at the MAR model creation. The aim is to minimize the prediction error given by (6). By assuming stationarity of the signal $\underline{s}(n)$ and multiplying both sides of (6) by $\underline{s}^T(n-i)$ from the right, we obtain:

$$\underline{\varepsilon}(n)\underline{s}^T(n-i)=\underline{s}(n)\underline{s}^T(n-i)-\underline{\tilde{s}}(n)\underline{s}^T(n-i) \tag{20}$$

Taking expectation from both sides of (20), yields:

$$E\left\{\underline{\varepsilon}(n)\underline{s}^T(n-i)\right\}=E\left\{\underline{s}(n)\underline{s}^T(n-i)-\underline{\tilde{s}}(n)\underline{s}^T(n-i)\right\} \tag{21}$$

By the orthogonality principle, the left side of (21) equals to zero:

$$0=E\left\{\underline{s}(n)\underline{s}^T(n-i)\right\}-E\left\{\underline{\tilde{s}}(n)\underline{s}^T(n-i)\right\} \tag{22}$$

From (5) and (22) we receive the following:

$$0 = E\left\{\underline{s}(n)\underline{s}^T(n-i)\right\} + \sum_{k=1}^{p} \underline{A}(k) E\left\{\underline{s}(n-k)\underline{s}^T(n-i)\right\} \tag{23}$$

The autocorrelation matrix of lag i was defined by Eq. (18) as:

$$\underline{R}(i) = E\left\{\underline{s}(n)\underline{s}^T(n-i)\right\}$$

Therefore, (18) and (23) lead to a set of linear equations known as Yule-Walker equations:

$$0 = \sum_{k=1}^{p} \underline{A}(k)\underline{R}(i-k) + \underline{R}(i) \tag{24}$$

It may be written in the following matrix form:

$$\begin{bmatrix} \underline{R}(0) & \underline{R}(-1) & \dots & \underline{R}(1-p) \\ \underline{R}(1) & \underline{R}(0) & \dots & \underline{R}(2-p) \\ \vdots & \vdots & \ddots & \vdots \\ \underline{R}(p-1) & \underline{R}(p-2) & \dots & \underline{R}(0) \end{bmatrix} \begin{bmatrix} \underline{A}(1) \\ \underline{A}(2) \\ \vdots \\ \underline{A}(p) \end{bmatrix} = - \begin{bmatrix} \underline{R}(1) \\ \underline{R}(2) \\ \vdots \\ \underline{R}(p) \end{bmatrix} \tag{25}$$

We use the fact that $\underline{R}(-i) = \underline{R}^T(i)$, which can be proven by:

$$\begin{aligned} & \underline{R}^T(i) = E^T\left\{\underline{s}(n)\underline{s}^T(n-i)\right\} = E\left\{\underline{s}(n-i)\underline{s}^T(n)\right\} \\ & \Rightarrow E\left\{\underline{s}(n-i)\underline{s}^T(n)\right\} = E\left\{\underline{s}(n)\underline{s}^T(n+i)\right\} = \underline{R}(-i) \\ & \Rightarrow \underline{R}^T(i) = \underline{R}(-i) \end{aligned} \tag{26}$$

When (25) can be written in the matrix form as:

$$\begin{bmatrix} \underline{R}(0) & \underline{R}^T(1) & \dots & \underline{R}^T(p-1) \\ \underline{R}(1) & \underline{R}(0) & \dots & \underline{R}^T(p-2) \\ \vdots & \vdots & \ddots & \vdots \\ \underline{R}(p-1) & \underline{R}(p-2) & \dots & \underline{R}(0) \end{bmatrix} \begin{bmatrix} \underline{A}(1) \\ \underline{A}(2) \\ \vdots \\ \underline{A}(p) \end{bmatrix} = - \begin{bmatrix} \underline{R}(1) \\ \underline{R}(2) \\ \vdots \\ \underline{R}(p) \end{bmatrix} \tag{27}$$

Where the $[R]$ matrix is a Toeplitz block matrix.

For coefficients estimation, the YW equations should be solved. The most efficient and known way to do it is by applying the LWR recursive algorithm (Wiggins & Robinson, 1965). The LWR algorithm is a generalized form of Levinson's single channel case (Makhoul, 1975). At the end of this process we get p autoregressive coefficients $\underline{A}(k)$ matrices of $d \times d$ dimensions, for every recorded EEG signal.

There are two methods for coefficients estimation, the covariance and the autocorrelation methods. This research has used the autocorrelation method, since it is a more convenient and a widespread method. The autocorrelation method leads to a solution based on the LWR algorithm (Makhoul, 1975), (Chen & Gersho, 1998).

3.2.2.2 Model Order estimation by AIC

An important decision to be made in the MAR model is the determination of an optimal order model. Since the order p of the model is apriori unknown, it is to be determined by minimizing the widespread order criteria AIC (Kay, 1988), (Priestley, 1989). The (Aufrichtig

& Pedersen, 1992), (Herrera et al., 1997), (Akin & Kiymik, 2005) and (Palaniappan, 2006) researches deals with challenging issue of AR model order estimation for EEG signals.
The AIC is defined as:

$$AIC(p) = N \ln(\det \sum\nolimits_p) + 2d^2 p \tag{28}$$

When the $\sum_p$ is the estimation of residual error covariance matrix using a p^{th} order that was defined by Eq. (19), meaning:

$$\sum\nolimits_p = \underline{G}^2 = E(\underline{\varepsilon}^2(n)) = \underline{R}(0) + \sum_{k=1}^{p} \underline{A}(k)\underline{R}^T(k) \tag{29}$$

This matrix is a by-product of the LWR algorithm; therefore it is calculated recursively by the algorithm. The aim of AIC is to estimate the optimal order by finding the trade-off between the estimated prediction error matrix and the model order value. The AIC is calculated each time for a range of $p's$, and the selected p yields the minimum AIC.

4. Classification Method

The goal of this research is to classify the EEG signal into different sleep stages, by a multichannel analysis. In this chapter the suggested method will be described. The first section gives a general review about the processes, using a block diagram. The next section broadens the blocks of this diagram (Fig. 7).

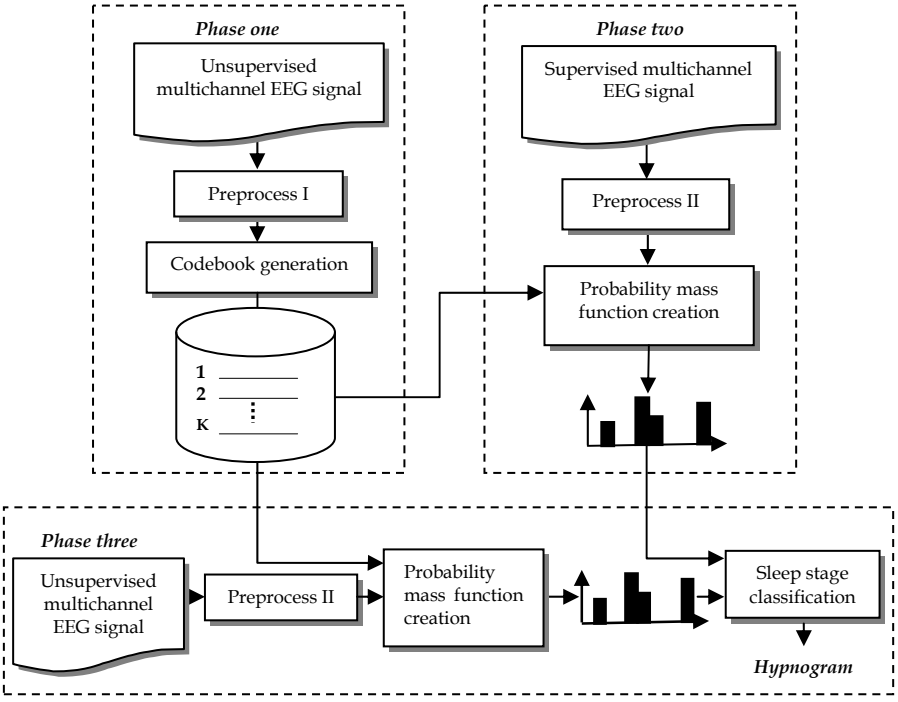


Fig. 7. Block diagram of the proposed classification system.

4.1 Classification System - Block Diagram

The block diagram appearing in fig. 7, describes the classification system that was created in this research. The system consists of three main phases; the first and the second phase are presented as the training phases. The first phase creates K size codebook from unsupervised data. The second phase builds histogram for each of the sleep stages, using supervised EEG signals and the codebook's codewords. The final phase is the classification stage that verifies the performance of the system.

4.2 Preprocess

The classification system composed of three phases (Fig.7), receives as an input some multichannel EEG signals. Every signal that gets into the classification system has to pass through the preprocess step, described by a block diagram in fig. 8. The preprocess step takes the raw EEG signal and makes it ready for the classification system. There are two kinds of preprocess steps: preprocess I and II. Both preprocesses are very similar, the difference between them will be explained in the following chapters. Preprocess I is used in the first phase and preprocess II is used in the second and third phases (Fig. 7). This section will explain in details preprocess I.

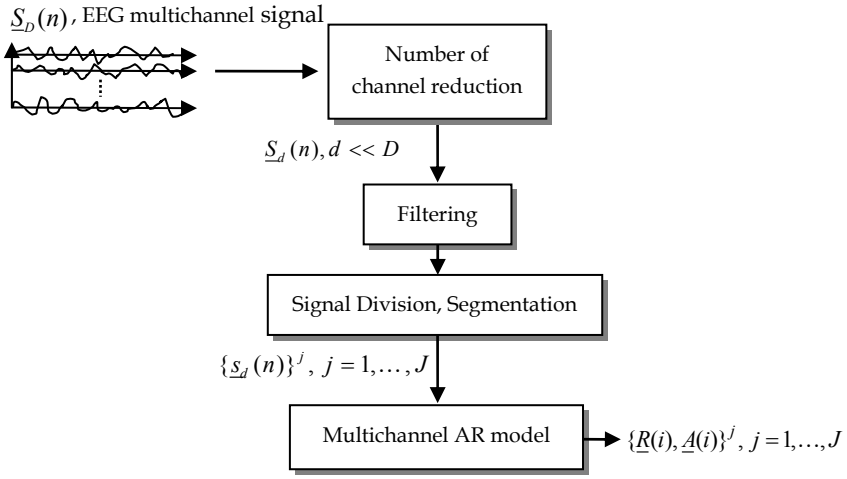


Fig. 8. Preprocess of EEG signal block diagram.

Preprocess I starts with a channel reduction. The raw EEG signal $\underline{S}_D(n)$, can contain up to 128 recorder channels (D) that can cause data redundancy. Therefore, according to the recommendation of an expert neurologist, a sub set of few channels ($d, d << D$) is chosen to represent the originally recorded EEG signal. Following channel reduction, the signals pass thorough an anti aliasing and if necessary a down sampling filter, for noise reduction.
The EEG signal is a stochastic non-stationary multichannel (vector) process. Therefore, the sampled EEG signal has to be divided into fixed length segments, when *j* is the segment index. For further MAR model creation, the segments length N ($n = 1,\ldots,N$) should be short enough in order to be considered stationary. Nevertheless, N should be long enough

to enable accurate feature estimation, meaning enough samples per one coefficient estimation. The next part is the main part of the preprocess I step, the MAR model parameters estimation. The MAR model is described profoundly in chapter 3.2. In this step, the matrix coefficients $\underline{A}(i)$ are calculated for every one of signal segment j. The coefficients are calculated by the LWR recursive algorithm for the MAR model. Each phase in the system receives as an input a set of coefficients $\{\underline{R}(i), \underline{A}(i)\}^j$, where $\underline{R}(i)$ is the autocorrelation matrix and $\underline{A}(i)$ is the matrix coefficients. The autocorrelation matrix $\underline{R}(i)$ is necessary for GLLR (Flomen, 1990) calculation which is part of every phase in the proposed classification system. Therefore, in addition to $\underline{A}(i)$ matrix, the autocorrelation matrix $\underline{R}(i)$ considered as a part of the EEG signal representation.

After the preprocess I, the classification system works only with the coefficient's matrices and has no direct use with the original EEG signal.

The following sections will explain in details the automatic classification system in all three phases.

4.3 Codebook Generation - First Phase

The first phase creates from unsupervised EEG signals a *K* codewords codebook, using the Linde, Buzo, Gray (LBG) algorithm (Linde et al., 1980). The LBG algorithm takes the MAR coefficients $\{\underline{R}(i), \underline{A}(i)\}^j, j = 1,\ldots,J$ that calculated from an unsupervised EEG data in preprocess I, and creates new *K* clusters called codewords. The role of this phase is to present a large amount of data by a reduced amount of representatives called codewords.

First phase block diagram:

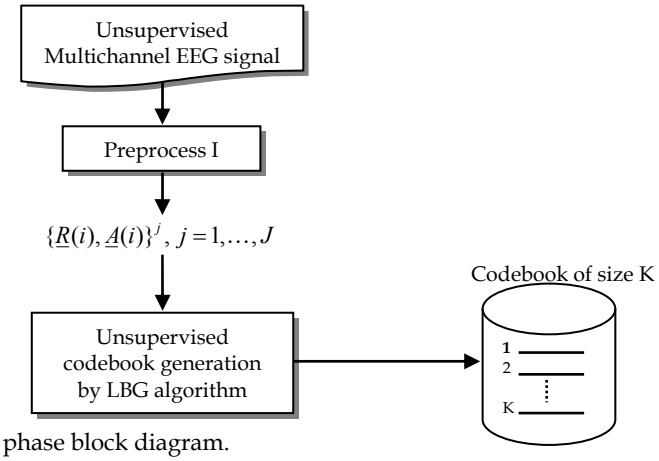


Fig. 9. The first phase block diagram.

As mentioned above, the data used in this phase is an unsupervised data, i.e. the input EEG signal does not pass through visual analysis and is not classified for any sleep stage. The entire unsupervised EEG signals, existing in our data base, pass through the preprocess I step yielding a set of J coefficient's matrices denoted by $\{\underline{R}(i), \underline{A}(i)\}^j$. The J coefficient's

matrices are the input parameters of the clustering algorithm LBG. The aim of the LBG algorithm is to reduce the number of MAR coefficients J, eventually creating a codebook with K ($K << J$) coefficient's matrices $\{\underline{R}(i), \underline{A}(i)\}^k$ as codewords.

The LBG algorithm, like any cluster algorithm, is based on some distortion measure. We used a Generalized Log Likelihood Ratio (GLLR) distortion measure that was first developed by Felix A.Flomen in 1990 (Flomen, 1990), as part of his thesis work. The Log Likelihood Ratio (LLR) (Itakura, 1975), originally proposed by Itakura, is widely used in speech processing application for measuring the dissimilarity between two AR processes.

The LLR measure was already tested on the EEG signal in the past. In (Estrada et al., 2005) and (Ebrahimi et al., 2007) by means of LLR, a similarity between EEG and electro-oculographic (EOG) is measured during different sleep stages. In (Kong et al., 1997), a change in EEG pattern was detected by the LLR and in (Estrada et al., 2004) a similarity between base line EEG segments (sleep stage) with the rest of EEG was measured. The works (Estrada et al., 2005), (Kong et al., 1997) and (Estrada et al., 2004) showed that LLR may be used as a distortion measure in an AR model for EEG signals. We use the LLR in its generalized form, for the multichannel case and it is defined as:

$$D_{GLLR} = \log\left(\frac{\det(\underline{A}_t \underline{R}_r \underline{A}_t^T)}{\det(\underline{A}_r \underline{R}_r \underline{A}_r^T)}\right) \tag{30}$$

D_{GLLR} is the GLLR distortion, $\underline{A}_r$ and $\underline{R}_r$ are the reference AR coefficients, and $\underline{A}_t$ is the tested AR coefficients.

It is important to mention that without the generalized form of LLR distortion measure of Felix A.Flomen (Flomen, 1990) it would be impossible to use the MAR model for classification of the EEG signal by the proposed system.

4.4 System Training - Second Phase

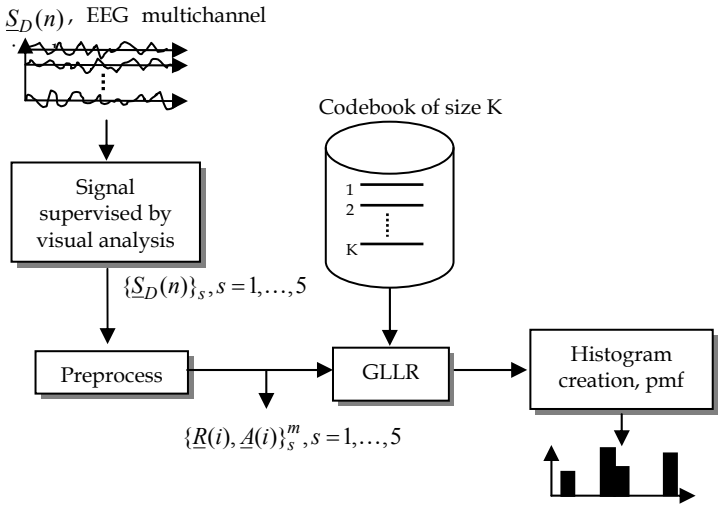


Fig. 10. The second phase block diagram.

Following the codebook creation, the second phase can be carried out. The intention of this phase is to represent each sleep stage by discrete probability mass function (pmf), of K codewords that estimated by histogram. Fig. 10 provides a general look on the training phase.

At first, a new unused and unsupervised EEG signals visually classifieds into a suitable sleep stage. The manual classification of unsupervised EEG signal is performed by an EEG expert. The manually supervised EEG signals clustered into five groups according to the supervised sleep stage. Every supervised EEG signals group is pass through the preprocess II that generates M MAR coefficients. Preprocess II is slightly different from preprocess I, the channel reduction and the filtering is the same, however the segmentation step has been changed according to the new needs of the second phase. In firs the supervised EEG signal divided into one minute duration fragments. Of course every one minute fragment represents only one specific sleep stage. Subsequently, every minute fragments, 60 seconds, were divided into Q ($q_1,..,q_Q$) segments with 50% overlap. When segment's duration is T and samples number N, as it illustrated in fig. 11.

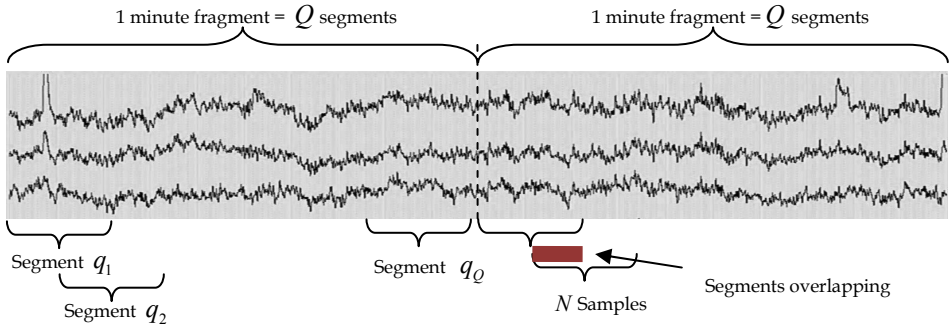


Fig. 11. Classification for every segment of EEG signal.

Preprocess II yields a M set's of $\{\underline{R}(i),\underline{A}(i)\}_s^m$ coefficients for all the segments, when 's' is the sleep stage tag in the range of $s=1,\ldots,5$.

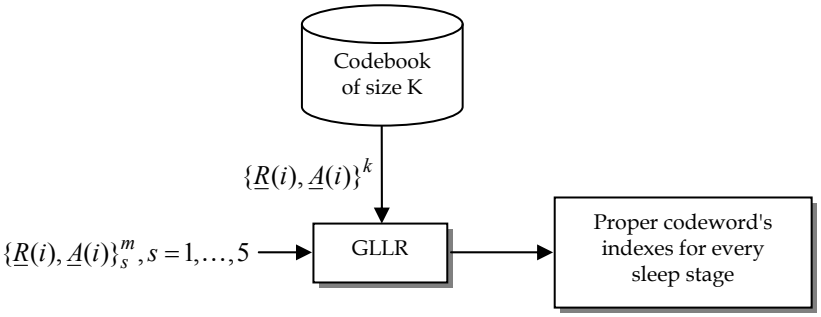


Fig. 12. Block diagram focused on codewords selection for every sleep stage.

After the parameters estimation for all segments of every one minute fragment, the next step can be preformed. The $\{\underline{R}(i), \underline{A}(i)\}_s$ coefficients of the supervised data are compared by the GLLR distortion measure (30) with each of the codeword $\{\underline{R}(i), \underline{A}(i)\}^k$ from the codebook. Fig. 12 illustrates a close-up look of this step in the second phase.

The GLLR distortion measure, D_{GLLR}, is calculated between parameters of segment q and every codeword from the codebook. The codeword that produce the minimum D^k_{GLLR}, is the chosen codeword index to represent the segment, i.e.:

$$Index_{m,s} = \arg\min_{k\in 1,..,K}\{D^k_{GLLR}\},\ m = 1,..,M,\ s = 1,..,5 \tag{31}$$

In other words, for every segment, a codeword is fitted by minimization of D^k_{GLLR}, $k = 1,..K$, resulting its argument i.e. index.

This process repeats for all segments of the supervised signal. At the end of this process we get a set of Q codeword indexes for every minute of the data. Next, for every minute, a histogram from the Q codeword indexes is created and normalized. In fact by this way we define a new random variable x as follow:

Let k be the codeword index from the codebook ($k = 1,..,K$). Let us define a random variable x which indicates which index k has been chosen for some segment q (of duration T) by the $\arg\min_{k\in 1,..,K}\{D^k_{GLLR}\}$.

The distribution of x is given by:

$$\begin{aligned} &\Pr(x = k) = p(k) \quad k = 1,..,K \\ &\sum_{k=1}^{K} p(k) = 1 \end{aligned} \tag{32}$$

By this action we receive a probability mass function (pmf) $\Pr(x = k)$, for random variable x per every minute of data that is estimated by a histogram.

We locate all the pmfs (histogram) of a certain sleep stage and by averaging them we receive a single pmf which represents the codewords distribution for a certain sleep stage. Eventually, a specific pmf $P_s(x = k), s = 1,..,5$ is estimated for every sleep stage. Fig. 13 exhibits the averaging of all pmfs (represented by histograms) ascribed to one sleep stage, and create the pmf of codebook indexes.

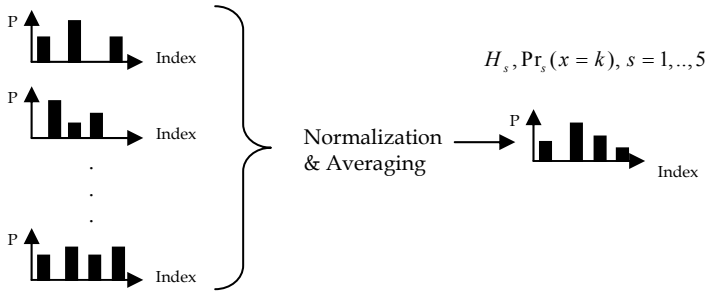


Fig. 13. Histogram for each sleep stage.

The histograms provide the relevant information for the classification phase, i.e. the relation between the coefficients of the unsupervised data to the supervised data for each sleep stage.

4.5 Signal Classification – Third Phase

Previous sections discussed the classification system fundamentals - the training phases. This section will discuss the third phase of the system - the classification of a new, unknown EEG signal.

The classification phase input, is a new set of an unseen EEG test signal. Actually it's the second set of unseen EEG signal used for fair system evaluation. The test signal passes through the preprocess II step, the MAR coefficients are estimated and compared to the codewords of the original codebook by GLLR distortion measure. Histograms created from codewords indexes and compared to the sleep stages histograms (chapter 4.4). By a minimal Kullback-Leibler (KL) divergence between the new signal pmf and all the five stages pmfs the classification is made. Fig. 14 illustrates the classification phase.

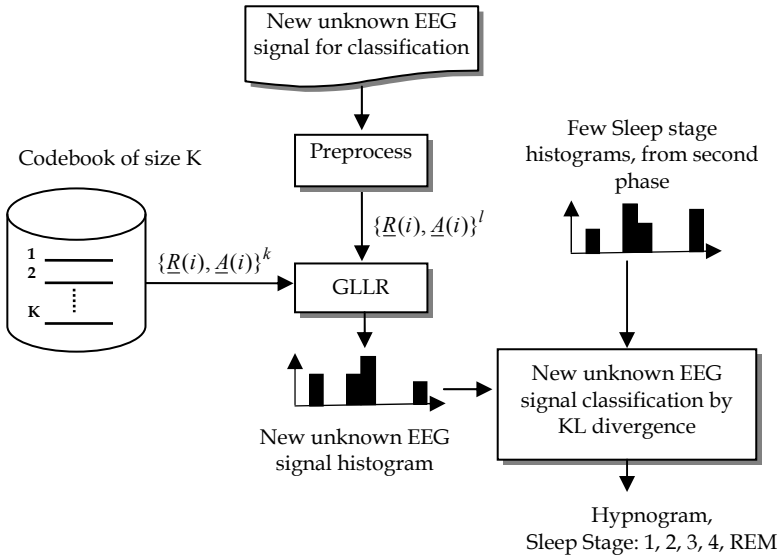


Fig. 14. Block diagram for classification phase.

This phase describes the classification of a new multichannel EEG signal into five different sleep stages. As mentioned in section 4.2, every raw EEG signal entering the classification system first has to pass through the preprocess step in this case preprocess II. Section 4.4 explains that in the preprocess II, the signal is divided into one minute fragments, and every fragment divided once again into Q overlapping segments of N samples (Fig. 12). Following the segmentation, the MAR coefficients are calculated for every segment q. Eventually preprocess II yields L MAR coefficients $\{\underline{R}(i), \underline{A}(i)\}^l$, $l = 1,..,L$, where L is the total number of segments in the new EEG signal.

Considering the preprocess II step, we have L MAR coefficients separated into sets of Q coefficients per every minute. Next, by the $Index_{t,s} = \arg\min_{k\in 1,..,K}\{D^k_{GLLR}\}$, (31), codewords indexes matched for each of the L MAR coefficients. Identically to the second phase (section 4.4), for every minute of the new EEG signal a normalized histograms is created, as can be seen in fig. 15. To be precise, these histograms are the pmf's, $\Pr_t(x=k)$ (32), of the K codewords.

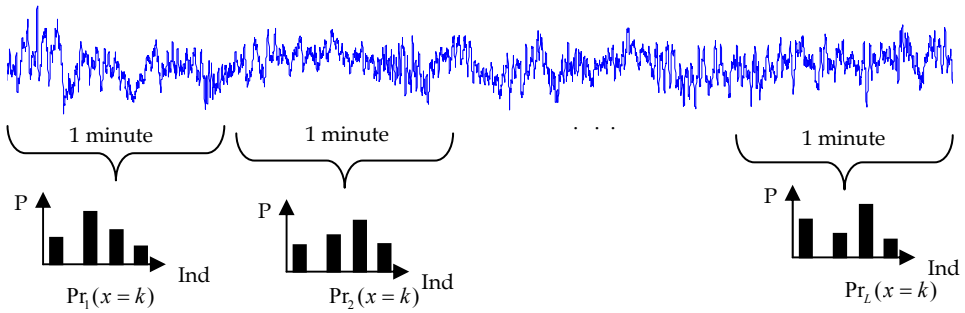


Fig. 15. Histogram per each minute.

In the current situation, there is a histogram per every minute of the new classified signal, and pre trained five histograms for every sleep stage.

The classification strategy is based upon a measure of similarity between two pmfs histograms. Therefore, every histogram of the tested signal has to be compared to each of the five sleep stages histograms that were created during the training phase (phase 2, section 4.4). In other words, some similarity measure has to be determined between the tested pmf $\Pr_t(x=k)$ and the sleep stage reference pmf $\Pr_s(x=k)$.

The Kullback-Leibler (KL) divergence (or relative entropy) (Flomen, 1990), (Cover & Thomas, 1991) is a measure of the difference between two probability distributions, when discrete probability distributions characterized by a pmf. Therefore KL divergence can be used as a similarity measure between pmf $\Pr_t(x=k)$ of the tested EEG signal and the referents sleep stage pmf $\Pr_s(x=k)$. The KL divergence is defined as:

$$D^s_{KL}(i) = \sum_{k=1}^{K} p_t(k)\log\frac{p_t(k)}{p^s_r(k)} = \sum_{k=1}^{K} p_t(k)\log p_t(k) - \sum_{k=1}^{K} p_t(k)\log p^s_r(k) \qquad (33)$$

The $p_t(k)$, is the probability of the tested signal and $p^s_r(k)$ is the sleep stage reference probability. KL divergence measure is not symmetric, always non-negative and is equal to zero only in case of $p_t(k) = p^s_r(k)$. The KL divergence measure calculation occurs between the distribution of the tested signal $\Pr_t(x=k)$ and the distribution of all sleep stages signals, i.e. $\{\Pr_s(x=k)\}_{s=1,..,5}$. The unknown EEG signal is classified according to a minimum KL divergence (maximum similarity) measure between the new signal pmf and all the reference stages pmf. A sleep stage which distribution produces the minimum D^s_{KL} is the classified

one, i.e.:

$$S = \arg\min_{s=1,..,5} \{D_{KL}^{s}\} \tag{34}$$

S represent the sleep stage classification of unknown EEG signal per one minute, where $S = 1,..,5$.

5. System Evaluation

Chapter 5 describes the suggested classification system of EEG signals into sleep stages. The system has been theoretically described and explained; real data producing real classifications has yet to be tested. This chapter will evaluate the proposed system, by testing the classification accuracy on real EEG signals. The goal of this system is to classify a real EEG data, verify its performances, and justify the use of multichannel analysis upon single channel analysis.

This research classifies EEG signal into four sleep stages, rather than five as the theory mentions (steps 1,2,3,4 and REM). Stages 3 and 4 are considered as deep sleep stages, classifieds as the same stage, stage 3&4, rather than two different stages. Brain waves appearing in these two stages differ only in threshold of delta wave presence (chapter 3); therefore it is especially difficult to distinguish between them. Numerous latest researches ((Heiss et al., 2001), (Gerla & Lhotska, 2006), (Virkkala et al., 2007)) in the field of sleep classification considered these two stages as one, a slow wave sleep stage. Moreover, at this stage of research the classification system cannot classify movement or awakeness as well, given that the database contains EEG signals recorded only during sleep. Correction and improvement of these weaknesses should be the source of further research, as mentioned in chapter 6.

5.1 Database - General Information

The data used for system evaluation was taken from the EEG lab for epilepsy study at Soroka hospital. The "data" is a Video EEG test that recorded electrical activity of the brain and in parallel documented subject's functioning by video camera, which contributes to the visual classification phase. Video documentation provides imperative information on EEG recording quality; it tracks when the subject is awake, asleep and moving, etc.

The database this research is based on contains about 30 hours of recorded EEG signals collected during the sleep process. It was collected from 25 subjects of different ages and gender, suffering from epilepsy. Striving to create a global classification system, this research is not interested in testing a certain cut of the population. The original EEG signals recorded nonstop during 24 to 96 hours per subject. For this research, the EEG signals were carefully chosen from the recorded data. The chosen data take only during sleep time and with minimal signal interruption, e.g. moving and other artifacts.

5.2 Framework

After the database presentment, the constants parameters such as channels number, signal sample rate, segments length and model order have to be defined. These constants are the keystone of the classification system, since; every computation will be established on them.

Original EEG signals are recorded through 29 electrodes (channels), in a sampling rate of 256 Hz. These EEG signals have to be processed before entering the classification system, such preprocess is described in general in section 5.2.
Due to redundancy existing in 29 channels, a sub set of five electrodes; Pz, Cz, Pz, T3, T4 $(d = 5)$ has been chosen.

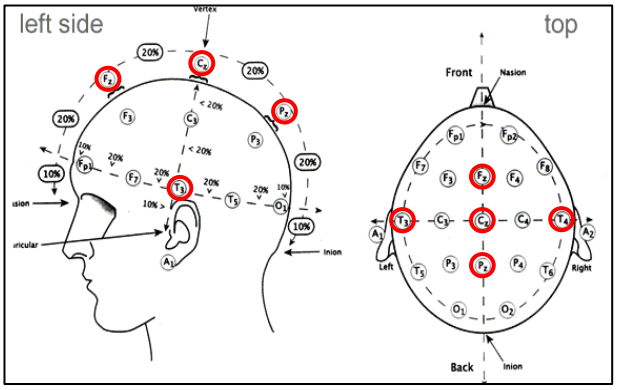


Fig. 16. Electrodes location on the scalp, the red circles mark the chosen electrodes.

Specific channels were selected according to trial and error technique cooperated with neurologist expert recommendation. The preferred channels achieved minimal MAR model order, while keeping segments duration reasonable.
The common frequency range of EEG signal during the sleep is between 0.5 Hz to 25 Hz, therefore a sampling frequency reduction is carried out using anti aliasing filter of 40 Hz, notch filter of 50 Hz (to reduce the power line noise) and additional down sampling to 85 Hz is performed.
The multichannel EEG signal is analyzed per short segments; therefore the next step is segment duration and model order definition. In fact the model order and segment duration are connected. In order to define one of them, the other has to be known. Hence, segment duration and model order determination are a tradeoff. Segment duration has to be long enough for all MAR coefficients estimation, when the MAR coefficients number is defined on the model order, and yet short enough for the stationarity assumption to be valid (as mentioned in section 4.2).
Several segment duration and model order combinations were examined, using trial and error method. The optimal segment duration was 4 seconds (duration T of one segment q) which is 340 samples per segment $(N = 340)$. The optimal model order was determined to be $p = 6$. With these constants the system will have enough samples for all parameters estimation; nevertheless the segment is short enough for stationarity assumption.
The estimation of MAR coefficients and model order is explained theoretically in section 3.2. The MAR coefficients are the foundation of the proposed system, and they are estimated per each segment of EEG signal with 50% overlapping between the segments. Coefficient estimation significantly depends on model order estimation. The suitable model order $p = 6$ is estimated according to AIC (Eq. (28) in section 3.2) of training data. A set of

AIC's is calculated per range of $p's$, from 1 to 30 for every segment of the training data, fig. 17 (a-d). The order producing minimum AIC is determined as segments order. A repetition of the mentioned procedure for all tested segments will produce a set of optimal orders ($p's$). A probability density function of optimal $p's$ is estimated by using a histogram, which will be described in fig. 18. The most probable order p is selected out of the histogram.

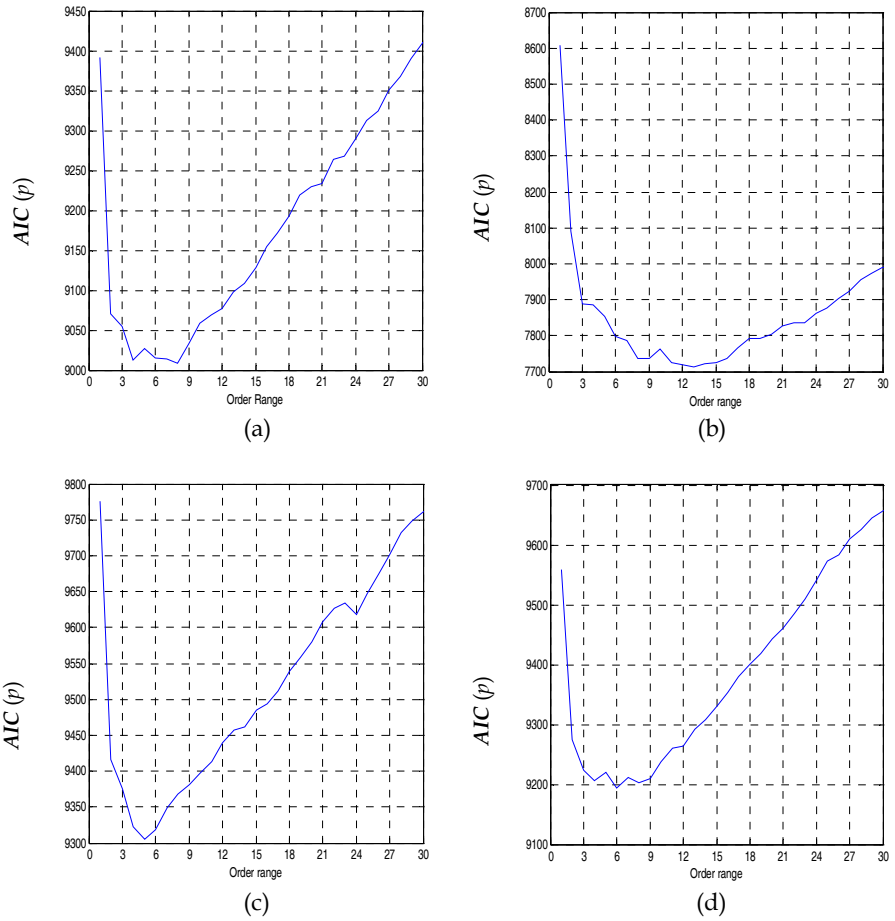


Fig. 17. (a-d).This figure shows an example of segment AIC's calculation for range of orders, 1 to 30.

A known phenomenon is AIC's over determining the model order (Kay, 1988). As a practical result one usually selects a lower order than the most probable order. In fig. 18 it can be seen that the optimal order $p = 7$ is estimated by AIC, the order for this research is

chosen to be $p = 6$. Order $p = 6$ and segment length $N = 340$ are the optimal option of the tradeoff between these parameters under the system limitations.

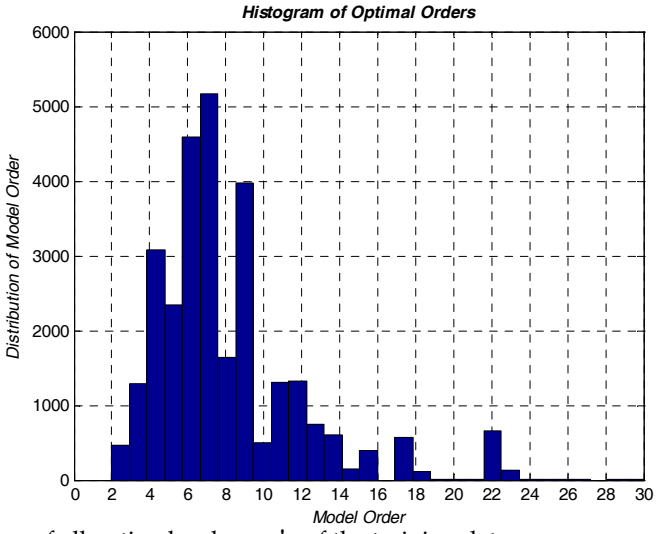


Fig. 18. Histogram of all optimal orders $p's$ of the training data.

5.3 System's Practical Consideration

After defining the constant parameters of the system in section 5.2, the evaluation system could be created and evaluated for both training and test data.

Chapter 4 explained that the training of the system is divided to two phases; the first phase creates the codebook of the MAR coefficients, and the second phase creates a pattern histogram for every sleep stage.

5.3.1 Codebook Generation

The first phase creates a codebook by using 41% of the whole database, i.e. 12.4 hours of recorded EEG signals. These 12.4 hours of data first pass through preprocess I (detailed at section 4.2) and produce nearly 21,576 segments that yielding 21,576 sets of MAR coefficients $\{\underline{R}(i), \underline{A}(i)\}^j, j = 1,\ldots,21,576$. By the LBG cluster algorithm (explained in chapter 4.3) the 21,576 MAR coefficients are quantized into 64 clusters that represent the codebook codeword's. Namely the codebook contains 64 sets of $\{\underline{R}(i), \underline{A}(i)\}^k$ representing all of the training data, when $k = 1,..,64$ ($K = 64$).

Fig. 19 describes schematically the process of codebook generation that will be used by the classification system in all it phases.

Several sizes of codebook were evaluated and tested; 32, 64 and 128 clusters (codewords) were examined for classification accuracy. The 32 and 128 codebook size provided insufficient classification results. Classification results showed that 32 clusters were not enough for training data representation and 128 clusters caused a redundancy in codewords. The choice of 64 clusters produces the best classification accuracy.

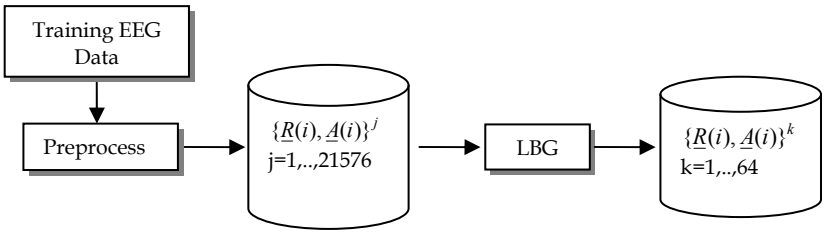


Fig. 19. Block diagram of codebook generation.

5.3.2 Histograms Representing Sleeping Patterns

Phase two is the second part of the training system using 21% of the database i.e. 6.18 hours of recorded EEG signals. Before the beginning of phase 2, the data used in this phase is visually classified into four sleep stages (stage 1, 2, 3&4 and REM). Visual classification is a vital process in the training phase, the entire system creation is established on this classification. In case that the data is classified incorrectly by an EEG expert, the classification will be wrong.

After visual classification, the training data passes through preprocess II (section 4.2) producing 10,753 sets of MAR coefficients $\{\underline{R}(i), \underline{A}(i)\}^j$, $j = 1,..,10,753$ for all 6.18 hours of data. In the framework section (section 5.2) segment duration was set to be 4 sec. Consequently, the number of segments per one minute Q equal to 29, due to 50% overlapping between the 4 sec segments.

As mentioned in section 4.4, indexes of specific codewords were chosen for every MAR coefficient of the tagged data. The codewords have been chosen according to minimal GLLR distortion measure (chapter 4) between the codewords and the coefficients.

For every minute of data, a histogram was created from the chosen codeword indexes (Fig. 19). According to the visual sleep stage classification, all histograms of every sleep stage were summarized and normalized. This action produced pmf ($\Pr_s(x = k) = p_s(k)$) for every sleep stage.

6.1 hours of tagged data were not divided equally between the four stages, the amount of minutes representing each sleep stage is: Stage 1 - 33 minute of tagged data, Stage 2 -134 minute of tagged data, Stage 3 & 4 - 164 minute of tagged data, Stage REM - 40 minute of tagged data.

Sleep stage 1 and REM (sleep stage 5) are very hard to detect in patient's EEG signals; consequently these stages have less data for testing. Sleep stage 1 lasts only five to maximum ten minutes in the beginning of sleep. Unfortunately, in case there is any movement these few minutes (which happens occasionally in the process of falling asleep), the recorded signal has much more noise than a physiological signal and cannot be used as training or testing data. Most people, who take the EEG test in labs, feel uncomfortable during the test and therefore they tend to move while trying to fall asleep. The REM sleep stage occurs when the patient has fallen deeply asleep. As stated above, it is very hard for people to sleep in unfamiliar places in addition to a set of electrodes attached to their heads, therefore the patients do not necessarily get to the REM sleep stage and wake up instead.

The classification accuracy is influenced by the amount of training data, thus it can be expected that the classification of stages 1 and REM will be less accurate than stages 2 and 3

& 4, due to lack in training data. As a result, the four histograms have to be normalized to the same scale for the pmf creation.

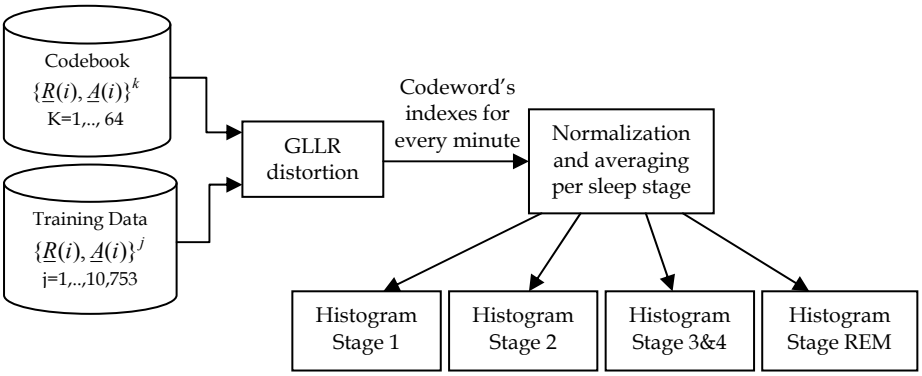


Fig. 20. Block diagram of Histograms creation.

The process described in this section creates four normalized histograms representing the sleep stage pattern by codewords distribution. Fig. 21-24 demonstrates the fourth sleep stage histograms used in this research. These histograms will take part in the classification stage as examples of the known data.

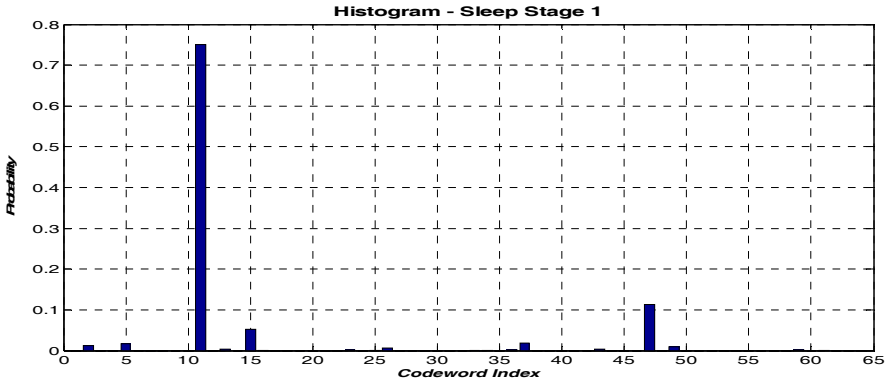


Fig. 21. Histogram of sleep stage 1.

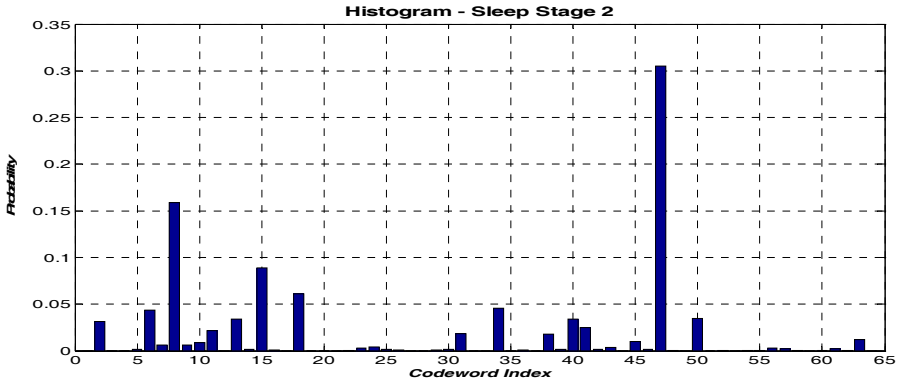


Fig. 22. Histogram of sleep stage 2.

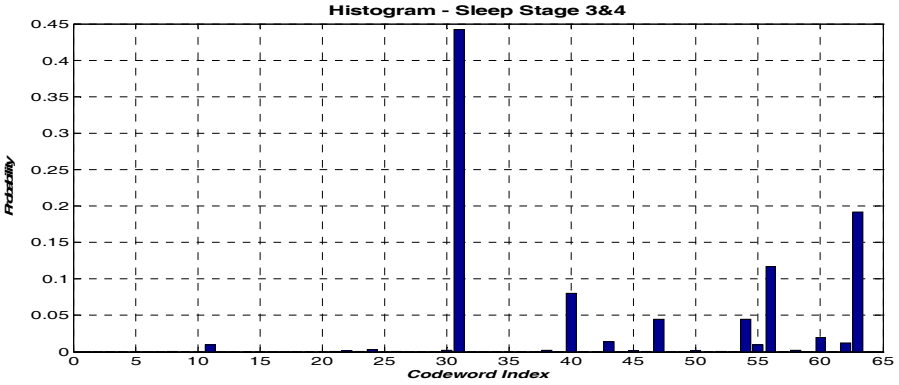


Fig. 23. Histogram of sleep stage 3&4.

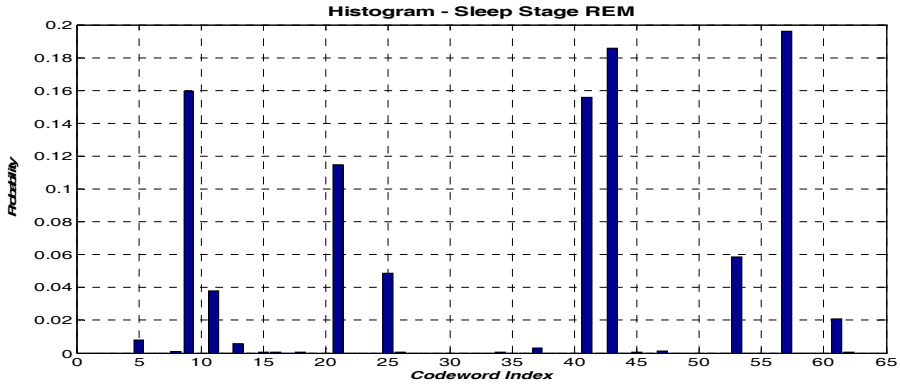


Fig. 24. Histogram of sleep stage 5 (REM).

According to fig. 21-24, the difference among sleep stage histograms is proven. This fact makes the proposed method of EEG signal classification possible.
Fig. 25 presents a 3D graph of all four distributions together. This view emphasizes the difference between the distributions.

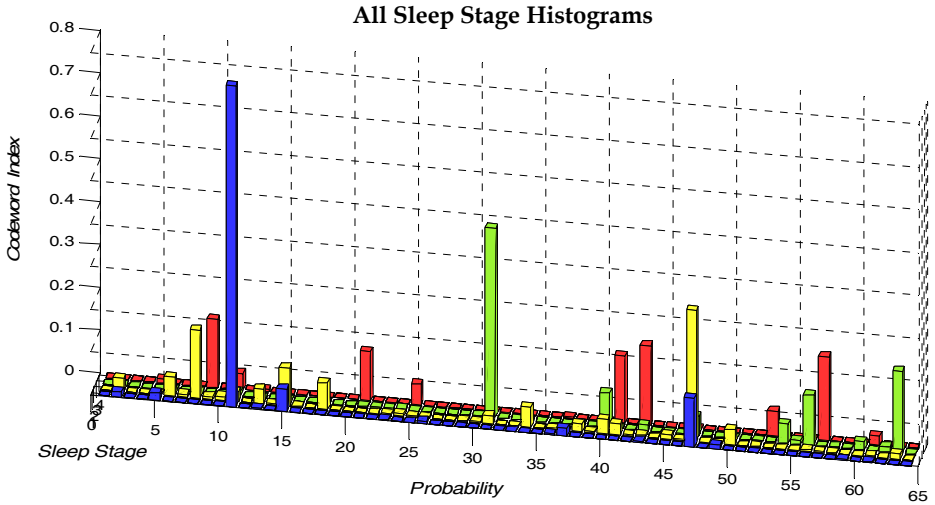


Fig. 25. The histograms of all the sleep stages on one graph.

5.3.3 Phase 3- Signal Classification

The classification of a new signal is performed by comparing the referential sleep stage histograms (from the training phase) and the new signal histograms. As detailed in section4.5, for every minute of new EEG signal a histogram (pmf) is created according to the codebook from phase one. This histogram is compared by the KL divergence to each of the four histograms from phase two.
System evaluation is performed by two EEG databases, one is the training database used during the second phase (21% of entire database) and the other is a completely new set of EEG signals consisting of 11.5 hours.
First, the system is quantitatively tested by the tagged data from the training phase. Since the tagged data is carefully visually classified per every minute, the classification quality is evaluated according to this data. By leave-one-out cross-validation (LOOCV) method all the training data is classified by the suggested classification method. Then, the data is divided into small groups of observations of different stages and subjects. One group is used for system evaluation and the rest of the groups are used as the training data for referential histograms creation, this process is repeated over and over, thus each group in the database is used as the evaluation data. Eventually all the training data is classified by the system, yet without using the same data in training and classification.
The classification system output is a Hypnogram of continuous EEG signal, therefore the second evaluation test the final Hypnogram generation. This test is a more qualitative evaluation, given that the results are presented in graphical form. The tested EEG signal is recorded continuously for several hours from two subjects. As opposed to the basic classification per each minute of the signal, the final classification system classifies the raw

EEG signal into four stages (as mentioned before) and another fifth stage called the "zero" stage. "Zero" stage means, the classification is not good enough for any of the four defined stages, therefore this minute is classified as an exception which the system does not recognize. Such events occur when the subject is moving during the sleep or waking up. If the minimal KL divergence measure is above some determined threshold the analyzed minute will be classified as a "zero" stage. The threshold was determined by applying the trial and error method on all the tested data.
In addition, in order to reduce the noise and other distortions of the signal, a median filter for every three minute with two minutes overlapping is considered, and was formally defined as:

$$Median\_class(i) = class\{(i-1), class(i), class(i+1)\} \tag{5.1}$$

According to neurologist opinion, three minutes smoothing is acceptable and will neither decrease the classification quality nor harming the medical diagnose. Five minutes median filter was also tested, but produced much less accurate classification.

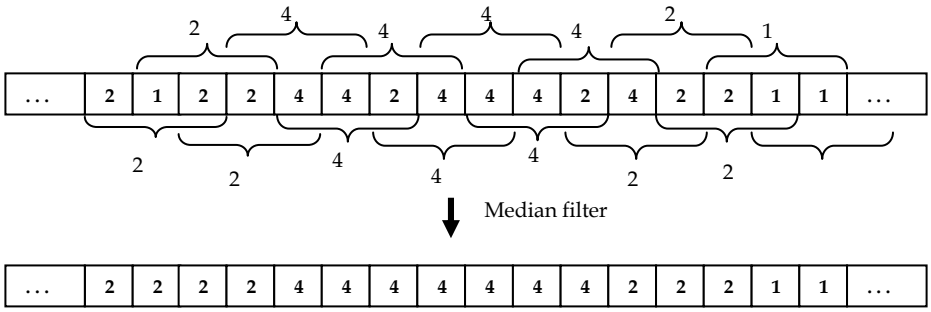


Fig. 26. Stage smoothing.

Fig. 26 demonstrates the performance and the outcome of the median filter per every three minutes. With the help of this technique no data is being wasted, every minute of data has a representation in the Hypnogram, while the Hypnogram is more robust to any distortion (e.g. moving during the sleep).

6. Evaluation Results

In this subchapter the evaluation results of the proposed sleep stage classification method are presented and analyzed.
Visually classified training data was used for system evaluation. The period of one minute was classified by the LOOCV method into four different sleep stages; stage 1, stage 2, stage 3 & 4 and stage 5, when stage 1,2 and 3 & 4 are the NREM stages and stage 5 is the REM stage. In fact, the period of one minute is the second layer of the classification. The first layer is the basic 4 second segment of an EEG signal consisting of one minute period, represented by histogram as depicted in fig. 15.
Table 1 presents the classification performance during the period of one minute when the evaluation data contains 33 minutes of stage 1, 134 minutes of stage 2, 164 minutes of stage 3&4 and 40 minutes of the REM stage.

		The Automatic Sleep Stage Classification				
Evaluation Data		**Stage 1**	**Stage 2**	**Stage 3&4**	**Stage REM**	**%Sensitivity**
	Stage 1	***29***	4	0	0	87.8
	Stage 2	0	***122***	12	0	91
	Stage 3&4	1	4	***156***	3	95.1
	Stage REM	0	1	0	***39***	97.5
	% Specificity	96.6	93.1	92.8	92.8	***93.2***

Table 1. Evaluation results.

In order to evaluate the classification results we used two common statistical measures: specificity and sensitivity. Sensitivity is a statistical measure of how well a classification test identifies a condition, specificity on the other hand is a statistical measure of how well a classification test identifies the negative cases, or cases that do not meet the condition, thus, the two measures complete themselves. Table 1 demonstrates the satisfying classification results of the proposed system. Classification performances of the method are 93.8% specificity and 92.8% sensitivity with an average of 93.2%.

The results confirm the strong potential this method possesses in the field of EEG signal processing. Although the results meet the expectations, there are still inadequacies, the amount of classified and tested data in each stage is uneven what can affect the classification quality and reliability. This fact does not damage the classification performance, yet it bears an influence on the confidence level of the classification system. The following classification test provides proof that the reliability of the classification system is very high.

Evaluation by LOOCV provides quantitative measure of the classification method and proves that the system is reliable for sleep stage classification. The next system test provides more visual evaluation of the classification quality, yet followed by quantitative information.

The classification system receives an input of several hours of EEG signals recorded during the sleep, and the output of the system is the sleep pattern of the signal, a Hypnogram. Eventually, producing a Hypnogram of the continuous sleep signal is the purpose of the system; therefore this test provides Hypnograms of two subjects produced by automatic classification method against the classification of an expert. The following figures (Fig. 27-32) demonstrate these Hypnograms.

Fig. 27-32 shows the results of real continuous EEG signal classification. Fig. 27 demonstrates automatic classification of almost 4½ hours of sleep collected from a single subject "A" and fig. 30 demonstrates automatic classification of nearly 7 hours of sleep of single subject "B", when fig. 28 and 31 present expert's classification of subject "A" and "B", respectively. The Hypnograms created by the automatic classification are one minute resolution and median filtered for every three minutes (as explained in section 5.3.2).

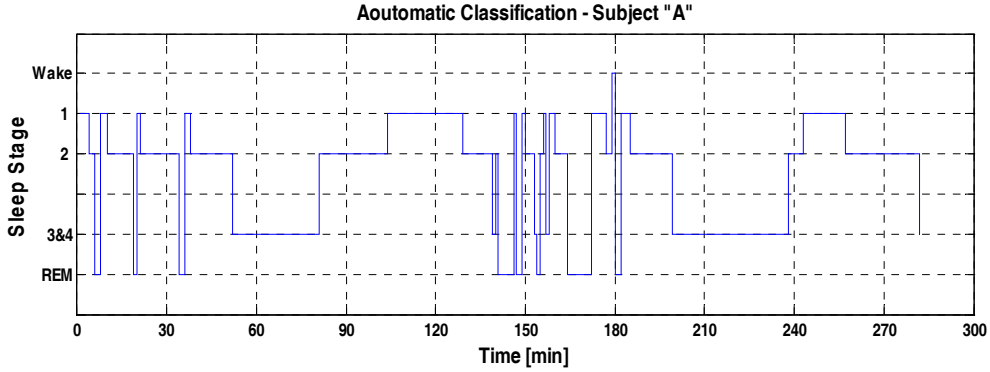


Fig. 27. Hypnogram of subject "A", automatic classificatin.

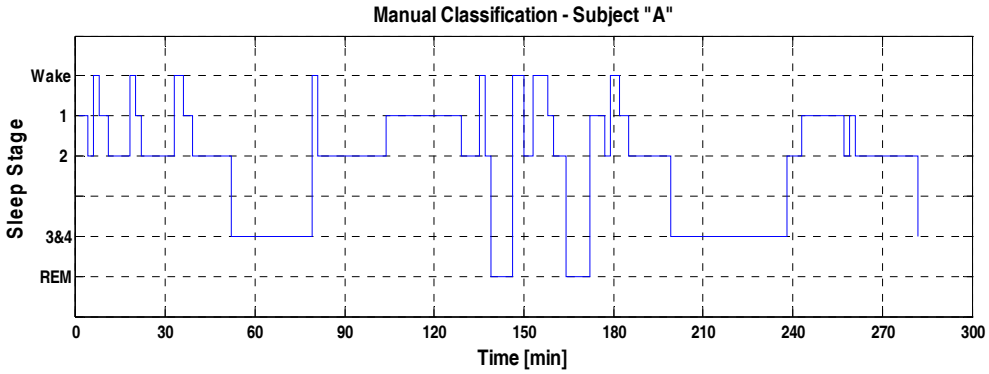


Fig. 28. Hypnogram of subject "A", expers classification.

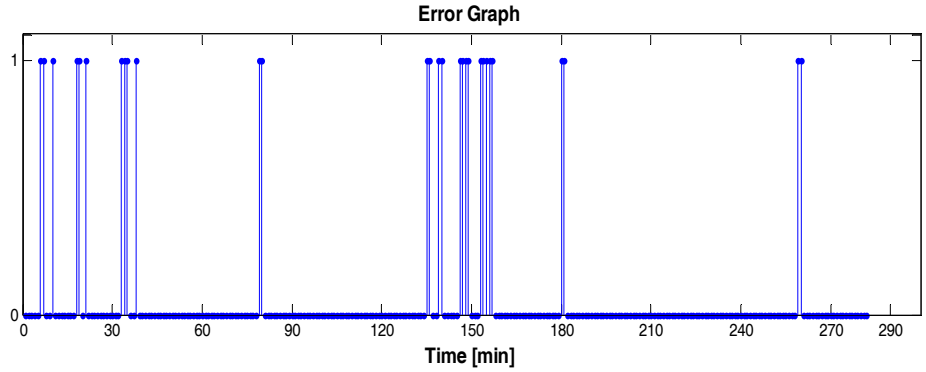


Fig. 29. Error Graph- A diffrence between automatic and expers classification of subject "A".

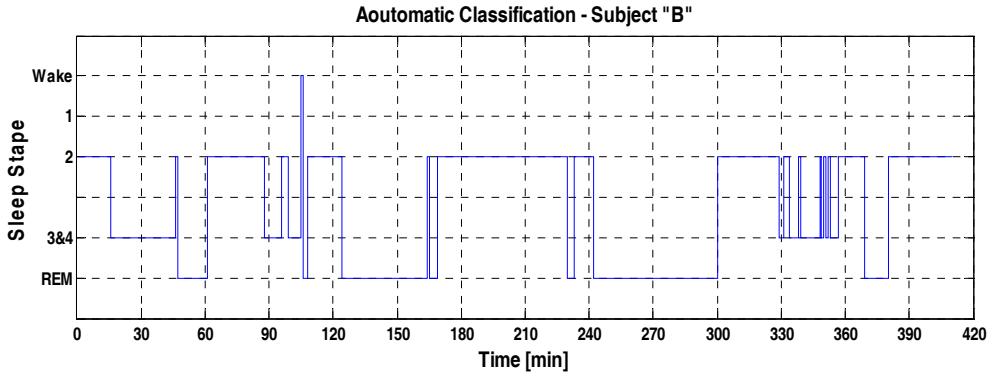


Fig. 30. Hypnogram of subject "B", automatic classification.

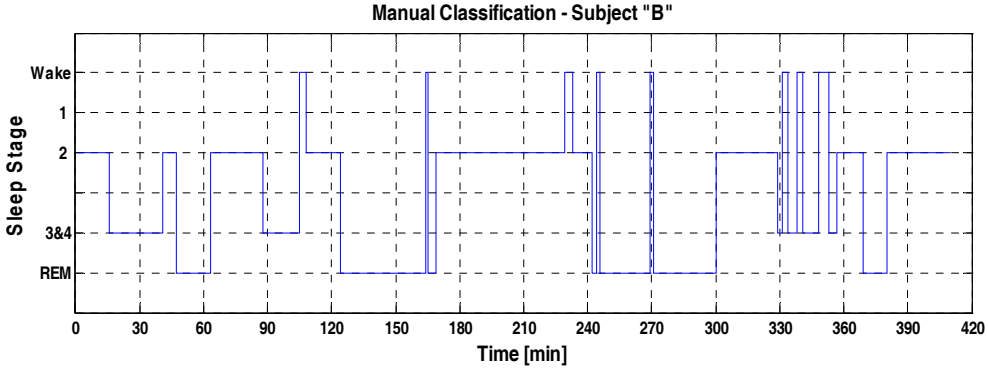


Fig. 31. Hypnogram of subject "B", experts classification.

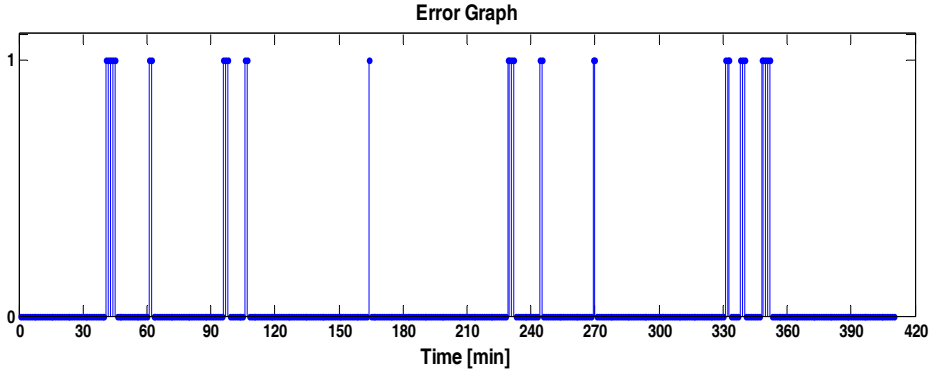


Fig. 32. Error Graph-A diffrence between automatic and expert classification of subject "B".

The difference in qualitative pattern description between the automatic and expert classification can be seen in fig. 27, 28 and 29, 30 representing subjects "A" and "B", respectively. A quantitative difference among automatic and expert's classification for subject "A" and "B" are shown in fig. 29 and 32, respectively. These two graphs present the classification errors per every minute of an EEG signal. Each stem in the graphs is represented by binary method showing the dissimilarity between the classifications. The agreement rate between automatic and expert's classification for subject "A" is 89.8% and for subject "B" is 92.2%.

As can be seen from fig. 27-32 and from the agreement rate, the classifications of the automatic system and of the human eye are very similar. The produced Hypnograms provide quite an accurate over-all aspect of the sleep stage pattern. Although the classification results are very fulfilling there are several dissimilarities between the classifications.

The most noticeable dissimilarities between the automatic and the expert's classification appear in cases the subject is moving during the sleep. The expert classifies these movements as awakening stage ("zero" stage), however the system predominantly classifies it as unstable state expressed in repeated sleep stages changes.

The Hypnograms depicted in fig. 27 and 28 belong to a subject who hardly slept during the recording time. Most of the sleep period subject "A" was in stage 2 and intermediately between stage 1 and alertness. The automatic classification implemented in this research has no ability to classify alertness or movement, therefore these events are expressed as unstable stages and the agreement rate between the system and the experts is relatively low. Apart of the mentioned weakness, the classification system provides very accurate information on sleep patterns of the subjects. For subject "B", the Hypnograms in fig. 30 and 31 have a high agreement. Subject "B" had a smooth and relaxed sleep, therefore the classification of the system is almost similar to the experts opinion. Although subject "B" had some undefined stages (movement during the sleep) the classification accuracy error is less than 8%.

Evaluation results show, that the classification system presented in this research allows very high classification accuracy of 93.2% and can be used on real EEG signals.

7. Conclusions and Discussion

The aim of this research is to develop an automatic classification system based upon parametric multichannel analysis approach. This classification system would classify multichannel EEG signals during sleep into the correct sleep stage. The Classification system created in this research succeeded in classifying EEG signals into the right sleep stage with a high accuracy rate, specificity of 93.8% and sensitivity of 92.8%.

The evaluation results of the research are significant due to the employed rich EEG database. The database includes 30 hours of real EEG signals recorded from 25 different subjects.

The developed system classifies EEG signals into four sleep stages when stage 4 represents both stage 3 and 4 (since they differ only in delta wave percent appearance and are known to be slow waves sleep stages (SWS), or rather deep sleep stages). This assumption is considered conventional by EEG expert (neurologist advisor) and by most of the recent researches (Ebrahimi et al., 2007), (Song et al., 2007), (Heiss et al., 2001), (Virkkala et al., 2007) in sleep.

The weakness of the system comes from the lack of awake stage classification; the awake stage was not a part of the training sleep stages set. This awake stage is not a regular condition stage due to many environmental influences. Therefore, the system does not recognize movement and awakeness in the EEG signal. This fact decreases the classification accuracy of the Hypnogram, a though it does not impair the classification of the other stages. Instead of classifying these actions as an awake stage (stage "zero"), in most cases the system classifies them as noise (unstable stages, as can be seen in fig. 27, chapter 6). Movement and awakeness can be observed in the current system in cases where the Hypnogram behavior is physically feasible and instable. A further research overcoming this weakness should be considered. The awake stage model must include both brain behavior and environmental impacts on the stage.

Sleep stage classification enfolds numerous issues that should be the source for further research. For instance, in the framework of this research we have decided to utilize five EEG channels Pz, Cz, Pz, T3 and T4, however it still has to be examined in what matter the number of channels and their locations affect the system behavior.

In conclusion, the method suggested in this work provides a relatively accurate sleep stage classification (93.2%), by using a multichannel analysis as the basic principle. The genuine multichannel approach of this research, in contrast to the customary researches, turns this research into a very valuable study.

The promising and encouraging results achieved by the multichannel approach for EEG signal classification in this work, emphasize it's the high potential. This approach posses a great aptitude not only in sleep stage classification but also in many other medical fields, including epileptic seizure detection and classification, diverse brain researches - brain computer interface (BCI), and of course classification of other biomedical signal such as ECG, EMG EOG etc.

8. References

Agerwal, R & Gotman, J. (2001). "Computer-Assisted Sleep Staging" , *IEEE TBME*, Vol. 48. 12.

Akin M. & Kiymik, M.K. (2004). "Application of Periodogram and AR Spectral Analysis to EEG Signals", *Journal of Medical Systems*, Vol. 24. 4.

Andreson, C.H.; Stolz, E.A. & Shamsunder, S. (1998). "Multivariate Autoregressive Models for Classification of Spontaneous Electroencephalogram During Mental Task", *IEEE Transactions on biomedical engineering* , Vol. 45. 3.

Aufrichtig, R. & Pedersen, S.B. (1992). "Order Estimation and Model Verification in Autoregressive Modeling of EEG Sleep Recordings" , *IEEE EMBS*, Vol 14, pp. 2653-2654.

Chen, J.H. & Gersho, A. (1998). "Covariance and Autocorrelation Methods for Vector Linear Prediction", *IEEE*.

Cover, T.M. & Thomas, J.A. (1991). *Elements of Information Theory* [ed.] Donald L.Schiling. 2nd. City College of New York : Wiley Series in Telecommunications.

Ebrahimi, F.; Mikaili, M.; Estrada E. & Nazeran, H.(2007). "Assessment of Itakura as a Valuable Feature for Computer-aided Classification of Sleep Stage", *IEEE EMBS*.

Estrada, E.; Nazeran, H.; Nava, P.; Behbehani, K.; Burk J. & Lucas, E. (2004). " EEG Feature Extraction for Classification of Sleep Stages", *IEEE EMBS*.

Estrada, E.; Nazeran, H.; Nava, P.; Behbehani, K.; Burk J. & Lucas, E. (2005). "Itakura Distance: A Useful Similarity Measure between EEG and EOG Signal in Computer-aided Classification of Sleep Stages", *IEEE EMBS*.

Flomen, F.A. (1990). *Distortion/Distance Measures for Coding of Vector Processes Represented by Multichannel AR Mode*l : PH.D Thesis (Hebrew vertion).

Gerla, V. & Lhotska, L. (2006). " Multichannel Analysis of the Newborn EEG Data", *IEEE ITAB*.

Ghosh, J. & Zhong, S. (2002). " HMMs and Coupled HMMs for Multi-channel EEG Classification", *IEEE*.

Heiss, J.E.; Held, C.M.; Estevez, P.A.; Holzmann, C.A. & Perez, J.P. (2001). "Classification of Sleep Stages in Infants: A Neuro Fuzzy Approach", *IEEE EMBS*.

Herrera, R.E.; Sun, M.; Dahl, R.E.; Ryan, N.D. & Sclabassi, R.J. (1997). "Vector Autoregressive Model Selection In Multichannel EEG", *IEEE EMBS*, Vol. 2.

Isaksson, A.; Wennberg, A. & Zetterberg, L.H. (1981). "Computer Analysis of EEG Signal with Parametric Models", *IEEE*, Vol. 69. 4.

Itakura, F. (1975). "Minimum Prediction Residual Principle Applied to Speech Recognition", *IEEE ASSP*, Vol. 23, pp. 67-72.

Jansen, B.J.; Bourne, J.R. & Ward, J.W. (1981). "Autoregressive Estimation of Short Sgment Spectra for Computerized EG Analysis", *IEEE BME*, Vol. 28. 9.

Kay, S.M. (1988). *Modern Spectral Estimation:Theory &Application*. [ed.] Alon V. Oppenheim. Universit of Rhode Island, Kingston : Prentice Hall Signal Processing Series.

Kelly, D.D. (1991). *Part VI. Hypothalamus, Limbic System and Cerebral Cortex. Chapter 40, Physiology of sleep and Dreaming.*

Kerkeni, N.; Alexcandre, F.; Bedoui, M. H.; Bougrain, L. & Dogui, M. (2005). "Neuronal Spectral Analysis of EEG and Expert Knowledge Integration for Automatic Classification of Sleep Stages", *WSEAS Transactions on information science and applications*.

Kong, X.; Lou, X. & Thakor, N.V. (1997). "Detection of EEG Change via a Generaluzed Itakura Distance", *IEEE EMBS*.

Krajca, V.; Petranek, S.; Paul, K.; Matousek, M.; Mohylova J. & Lhotska, L. (2005). "Automatic Detection of Sleep Stage in Neonatal EEG Using the Structural TIme Profiles", *IEEE EMBS*.

Linde, Y.; Buzo, A. & Gray, R.M. (1980). "An Algorithm for Vector Quantizer Design", *IEEE Transaction on Communications*, Vol. 28. 1.

Luo, G.(2007). " Subject-Adaptive Real-Time Sleep Stage Classification Based on Conditional Random Field, *American Medical Informatics Association Annual Symposium*, pp. 488-492.

Makhoul, J. (1975). "Linear Prediction: A Tutorial Review", *IEEE*, Vol. 63, pp. 561-580.

Masaaki, H.; Masaki, K. & Haruaki, Y. (2002). " Automatic Sleep Stage Scorng Based on Waveform Recognition ethod and Decision -Tree Learning", *Systems and Computers in Japan*, Vol. 33.

Pace-Schot, F. E. & Hobson, J. A. (2002). "The Neurobiology of Sleep: Genetics, Cellular Physiology and Subcortical Networks. Nature Reviews Neuroscience", *Nature Reviews Neuroscience*. pp. 591-605. Vol. 3.

Palaniappan, R. (2006). "Towards Optimal Model Order Selection for Autoregressive Spectral Analysis of Mental Tasks Using Genetic Algorithm" , *IJCSNS,* Vol. 6. 1.

Pinero, P.; Garcia, P.; Arco, L.; A lvarez, A.; Garca, M. M. & Bonal, R. (2004). "Sleep stage classifcation using fuzzy sets and machine learning techniques", *ELSELVIER, Neurocomputing*.

Priestley, M.B. (1989). *Probability and Mathematical Statistics: Spectral Analysis and Time Series*, [ed.] Z.W. Brinbaum and E.Lukacs, Academic Press.

Qianli, M.; Xinbao, N.; Jun, W. & Jing, L. (2005). "Sleep-stage Characterization by Nonlinear EEG Analysis using Wavelet-based Multifractal Formalism", *IEEE EMBS*.

Schaibold, M.; Harms, R.; Scholler, B.; Pinnow, I.; Cassel, W.; Perzel, T.; Becker H.F. & Bolz, A. (2003). "Knowledge-Based Automatic Sleep-Stage Recognition Reduction in the Interpretation Variability", *Somnologie*, pp. 59-65.

Shimada, T.; Shiina T. & Saito, Y. (1998). "Sleep Stage Diagnosis System With Neural Network Analysis", *IEEE EMBS*, Vol. 20.

Shimada, T. & Shiina, T. (1999). "Autodetection of Characteristics of Sleep EEG with Integration of Multichannel Information by Neural Networks and Fuzzy Rules", *Systems and Computers in Japan*, Vol. 30. 4.

Song, I. H.; Ji, Y. S. ; Cho, B. K. ; Ku, J. H. ; Chee, Y. J. ; Lee, J. S.; Lee, S. M.; Kim, I. Y. & Kim, S.I. (2007). "Multifractal Analysis of Sleep EEG Dynamics in Humans", *IEEE EMBS*.

Sun, M.; Ryan, N.D; Dahl, R.E.; Hsin, H.; lyengar, S. & Sclabassi, R.J. (1993). "A Neural Network Systtem For Automatic Classificatin of Sleep Stages", *IEEE.*Wanli, M. &

Van Hese, P.; Philips, W.; De Koninck, J.; Van de Walle R. & Lemahieu, I. (2001). "Automatic Detection of Sleep Stage Stage Using the EEG", *IEEE EMBS*.

Virkkala, J.; Hasan, J.; Varri, A.; Himanen, S.L. & Muller, K. (2007). "Automatic Sleep Stage Classification Using Two-Channel Electrooculography", *Journal of Neuroscience Methods*.

Wiggins, R.A. & Robinson, E.A. (1965). "Recursive Solution to the Multichannel Filtering Problem", *Journal of Geophysical Research*, Vol. 70. 8.

Zumsteg, D.; Hungerbühler, H. & Wieser, H. G. (2004). *Atlas of Adult Electroencephalography*, Hippocampus Verlag GmbH, 2004.

7

P300-Based Speller Brain-Computer Interface

Reza Fazel-Rezai
Department of Electrical Engineering, University of North Dakota
USA

1. Introduction

In this chapter, recent advances in the P300-based speller brain-computer interface (BCI) are discussed. Brain Computer Interface (BCI) technology provides a direct interface between the brain and a computer for people with severe movement impairments. The goal of BCI is to liberate these individuals and to enable them to perform many activities of daily living thus improving their quality of life and allowing them more independence to play a more productive role in society and to reduce social costs. A BCI involves monitoring conscious brain electrical activity, via electroencephalogram (EEG) signals, and detecting characteristics of brain signal patterns, via digital signal processing algorithms, that the user generates in order to communicate with the outside world.

A BCI system has an input, an output, and a pattern recognition algorithm that maps the input to the output (Wolpaw et al., 2002). Based on the type of input, there are four categories for a BCI system. A P300-based BCI is one of them. The focus in this chapter is the P300-based BCI. The P300 evoked potential speller was introduced by Farwell and Donchin (Farwell & Donchin, 1988). This paradigm has been the most widely used P300-based speller and a benchmark for many research groups.

Several shortcomings of existing P300-based BCIs have been identified, and many research groups have tried to overcome those shortcomings. However, more progress in resolving many of the challenges currently experienced in P300-based speller should be made to move BCI into the realm of practicality and to take it outside research laboratories into practical applications.

This chapter is organized as follows. First, a brief introduction to a BCI system will be given. Then, different types of BCI systems and P300-based speller BCI will be introduced. Details of a P300-based speller BCI including signal processing methods, sensitivity parameters, habituation, applications, and different variations of the system will be explained in detail. Then, recent advances including a new P300-based speller paradigm will be presented and results will be discussed.

2. Brain Computer Interface (BCI)

BCI technology involves monitoring conscious brain electrical activity, via EEG signals, and detecting characteristics of EEG patterns, via digital signal processing algorithms, that the user generates to interact with environment. It has the potential to enable the physically

disabled people to perform many activities. Therefore, it can improve their quality of life and productivity, allow them more independence, and reduce social costs. The challenge with BCI, however, is to extract the relevant patterns from the EEG signals produced by the brain each second.

EEG signals (typically 64 channels) can be recorded from the scalp or brain surface using a specific system of electrode placement called the International 10-20 system (Jasper, 1957). EEG signals are voltage changes of tens of microvolts at frequencies ranging from below 1 Hz to about 50 Hz. A BCI system has different components; an input, a pattern recognition algorithm, an output (Wolpaw et al., 2002). In addition, it has a protocol that determines the interaction of the BCI system with the user. A simple schematic diagram of a BCI system is shown in Figure 1.

The input of a BCI system is the EEG signal. EEG signals can be recorded invasively (i.e., intracortical) or non-invasively (i.e., surface EEG). The EEG signals can be the results of an evoked or a spontaneous signal. Evoked EEG is due to a stimulus and spontaneous EEG is a brain activity rhythm that is not the result of any stimulation. Pattern recognition components of a BCI system are the parts of the system that analyze EEG to extract a signature from EEG signals as shown in Figure 1. This includes signal processing methods for preprocessing (e.g., noise removal), feature extraction, classification, and post processing methods. Many different approaches from linear to nonlinear have been used (Mirghasemi et al., 2006). The output of a BCI system is connected to a computer, or a device that is supposed to be controlled using the BCI system. Spelling devices, i.e., virtual keyboard, is one of them. The operating protocol of a BCI system determines how the system interact with the environment, e.g., if the system supposed to work continuously or not and how the system should be turned off.

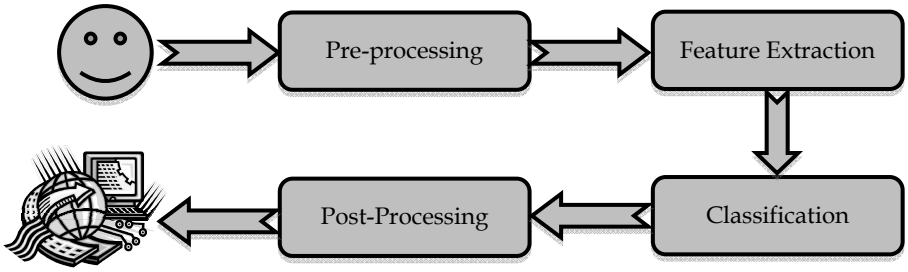


Fig. 1. A BCI system receives recorded EEG signals and extract signatures of the brain activity using several signal processing algorithm to interact with outside environment.

3. Different Brain Computer Interface (BCI) Systems

Based on the type of input, components, output, and protocol, BCI systems are categorized in different groups. Various properties in EEG can be used as the input of a BCI system such as rhythmic brain activity, event-related potentials, event-related desynchronization (ERD)

and event-related synchronization (ERS). In general, four different types of brain activity reflected in EEG signals have been mostly used as the BCI system inputs: 1) visual-evoked potentials, 2) slow cortical potentials, 3) mu and beta rhythms, and 4) the P300 component of Event-Related Potentials, or ERPs) (Wolpaw et al., 1991; Blankertz et al., 2004; Pfurtscheller et al., 1993; Scherer et al., 2004; Birch & Mason, 2000; Kostov & Kostov, 2000).

3.1 Rhythmic brain activity

Normal brain waves show different rhythmic activity based on the level of consciousness. This happens not only during the sleep (different sleep stages) but it can be seen in EEG during the waking state. These rhythms are affected by internal parameters such as subject's actions and thoughts and external factors such as fatigue and stress. For example, when a subject plans to move an object (without performing the task), a particular rhythm can be generated in EEG signal. The fact that mere thoughts affect the brain rhythms can be used as the input for a BCI system. The frequency bands from low to high frequencies are called delta (δ), theta (θ), alpha (α), beta (β), and gamma (γ) as shown in Table 1.

Band	Frequency Range
δ	0.5-4 Hz
θ	4-7.5 Hz
α	8-13 Hz
β	14-26 Hz
γ	> 30 Hz

Table 1. EEG rhythmic frequency ranges.

EEG waves below 4 Hz (usually 0.5-3.5 Hz) belong to the delta (δ) waves. Infants show irregular delta activity of 2-3.5 Hz in the waking state. However, delta waves are only seen in deep sleep in adults and they have not been used in BCI systems. Theta (θ) waves are between 4 and 7.5 Hz. The alpha (α) rhythm is in the range of 8-13 Hz and is seen with eyes closed and under conditions of physical relaxation. Mu (8-12 Hz) and beta (14-26 Hz) rhythms are mostly used in BCI systems. In people who are awake, the primary sensory or motor cortical areas often display 8–12 Hz EEG activity when they are not engaged in processing sensory input or producing motor output. This idling activity is called the mu rhythm when focused over the somatosensory or motor cortex. These mu rhythms are usually associated with 14–26 Hz beta rhythms. While some beta rhythms are harmonics of mu rhythms, some are separable from them by topography and/or timing, and thus are independent EEG features. It has been shown that subjects were able to raise and lower mu rhythm waves by thinking about activities such as relaxing (raising mu activity) and lifting weights (decreasing mu activity) (Pfurtscheller, 1993). In addition, it has been shown that mu rhythm is modulated by the expression of self-generated movement and observation and imagination of movement (Pineda et al., 2000).

3.2 Event-Related Potentials (ERPs)

Event-related potentials (ERPs) are changes in the EEG that occur in response to a stimulus. ERP is a voltage fluctuation in the EEG induced within the brain that is time-locked to a sensory, motor, or cognitive event. The stimulus can be visual, auditory, or somatosensory.

ERPs provide a different indication of how the stimulus is processed. They are measures to reflect the responses of the brain to events in the external or internal environment of the organism. Although they are not easy to be detected, they have wide usage for clinical-diagnostic and research purposes. It should be noted that whereas the amplitude of ongoing EEG activity is in the range of 50 microvolts, the amplitude of the ERP is much smaller in the range of 5 to 10 microvolts. The "peaks" and "valleys" in the ERP are termed components, with positive deflections labelled with "P" and negative deflections labelled with "N". The N400, for example, is the negative peak near 400 milliseconds after a stimulus onset. The P300 component is the fourth major positive peak. The P300 component of ERP was first reported by Sutton (Sutton et al., 1965). The P300 component of ERP is a positive peak at about 300ms after stimulus. It is elicited with a simple, two-stimulus discrimination task. This procedure has been dubbed the "oddball paradigm", whereby two stimuli are presented in a random series such that one of them (the "oddball") occurs relatively infrequently. A user is asked to distinguish between the stimuli by mentally noting each oddball stimulus, which creates a P300 wave, and ignoring the standard stimulus. Unfortunately, the P300 has a relatively small amplitude (5-10 microvolts), and cannot be readily seen in an EEG signal (10-100 microvolts). The averaging process is used to decrease the influence of the background EEG, but it typically takes 30 to 50 trials before the contribution of the background EEG subsides and a good, clean representation of the P300 is obtained.

4. P300-BCI Speller

The most widely used P300-based BCI is a spelling device designed by Farwell and Donchin in 1988 (Farwell & Donchin, 1988). Since then, it has been a benchmark for P300-based BCI systems in the research laboratory. In this speller, a matrix of six by six symbols, comprising all 26 letters of the alphabet and 10 digits (0-9), is presented to the user on a computer screen (Figure 2). The rows and columns of the matrix are intensified successively for 100ms in a random order. After intensification of a row/column, the matrix is blank for 75ms. At any given moment, the user focuses on the symbol he/she wishes to communicate, and mentally counts the number of times the selected symbol flashes. In response to the counting of this oddball stimulus, the row and column of the selected symbol elicit a P300 wave, while the other 10 rows and columns do not. Detection of the P300 makes it possible to match the responses to one of the rows and one of the columns, and thus identify the target symbol.

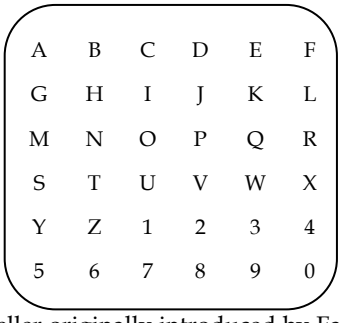


Fig. 2. The P300-based BCI speller originally introduced by Farwell and Donchin (Farwell & Donchin, 1988).

One of the greatest advantages of the P300-based BCI is that it does not require intensive user training, as P300 is one of the brain's "built-in" functions. However, P300 detection for real-time applications is not easy. Although signal-processing methods play a crucial role in BCI design, they cannot solve every problem. While they can enhance the signal-to-noise ratio, extract proper features, and classify them into different groups, they cannot directly address the impact of changes in the signal itself. There are several issues to be addressed before any P300-based BCI can be taken outside the research laboratory and put to practical application:
A) Transfer rate: Typically, many channels and several features should be used. In addition, the ensemble averaging of a large number of trials is required to improve the signal-to-noise ratio, because the P300 is buried in the ongoing EEG (Donchin et al., 2000).
B) Accuracy: For real-time applications, the P300-based BCI accuracy should be very high to be reliable to be used in real-world environment.
C) EEG pattern variability: EEG signal patterns change due to factors such as motivation, frustration, level of attention, fatigue, mental state, learning, and other nonstationarities that exist in the brain. In addition, different users might provide different EEG patterns.

4.1 Recent Advances in the P300-BCI Speller

Recently, many research groups have been trying to increase both transfer rate and accuracy, which are interdependent (Serby et al., 2005). The following parameters have to be addressed by a BCI system:
(a) Attentional blink: Attentional blink occurs if the interval between two targets is less than 500ms (Raymond et al., 1992). In such a case, the first target is correctly identified, while the second is not detected at all. This can be a source of P300 speller error if a non-target row/column near the target attracts attention by flashing less than 500ms before the target is flashed.
(b) Repetition blindness: If two identical targets in a stream of non-targets are flashed at intervals of less than 500ms, the second target may be missed (Kanwisher, 1987). This can be a source of error whenever a target row (column) is flashed less than 500ms after flashing a target column (row). Although they are intensified in random sequences, because the total intensification plus blank time is 175 ms, there are several cases where the interval between two targets is less than 500ms. This is another source of error.
(c) Target probability: It has been shown that the P300 amplitude is related to the probability of oddball occurrence (target row/column flashing) (Donchin et al., 2000). The less probable the oddball event, the larger the P300 amplitude. Although, in this paradigm, the probability of creating the P300 wave is 0.17 (2/12), the P300 amplitude can be increased by decreasing the oddball probability.
(d) Habituation: Attention decreases with repeated presentation of the same stimulus. When the user loses focus on the target character, the P300 is not elicited, and the system error increases with time.
(e) Row/Column error sensitivity: In the P300-based speller, the target symbol is the intersection of the target row and column, both of which, therefore, have to be detected correctly. This creates a sensitive system to error in detecting target row and column.
Fazel-Rezai (Fazel-Rezai, 2007) performed an analysis to reveal potential human error in the P300-based BCI system. In this work, two data sets were used. The first dataset was from BCI 2003 competition provided by Blankertz (Blankertz, et al, 2004), Wadsworth Center,

Albany, NY. This dataset was selected, because the results can be compared with the results of other works and repeated by other research groups. Features were extracted from averaged Mexican hat wavelet coefficients. More details of the feature extraction can be found in (Fazel-Rezai & Peters, 2005; Fazel-Rezai & Ramanna, 2005; Ramanna & Fazel-Rezai, 2006). It should be noted that since the objective of analysis was to reveal the possible errors and compare speller paradigms, no learning set was used.

In the first data set, BCI competition (Blankertz, et al, 2004), EEG signals were recorded from 64 electrodes; however, they used only Fz, Pz, Cz, C1, and C2 channels. Row/column intensifications (100msec) were block randomized in blocks of 12. Sets of 12 intensifications were repeated 15 times for each character. Each sequence of 15 sets of intensifications was followed by a 2.5 s period, and during this time the matrix was blank. This period informed the user that this character was completed and to focus on the next character in the word that was displayed on the top of the screen. In other words, for each character, 180 entries of feature values are stored 90 for row intensification and 90 for column intensification. For example, "A" is recognized only when row 7 and column 1 features indicate a P300 component. The signals were digitized at 240Hz and collected from one subject in two sessions. Each session consisted of a number of runs. In each run, the subject focused attention on a series of characters. Target words presented to the subject were: BOWL, CAT, DOG, FISH, FISH, GLOVE, HAT, HAT, RAT, SHOES, and WATER. Note that words FISH and HAT were presented two times. The total number of target characters in these words is 42. The results of analysis is shown in Table 1.

Word	BOWL				CAT			DOG			FISH				FISH				GLOVE				
Target	B	O	W	L	C	A	T	D	O	G	F	I	S	H	F	I	S	H	G	L	O	V	E
Detected	B	O	W	F	C	A	T	D	O	G	F	H	M	H	E	D	S	A	G	H	7	V	E
E(Col)	0	0	0	0	0	0	0	0	0	0	0	-1	0	0	-1	1	0	-1	0	-4	0	0	0
E(Row)	0	0	0	-1	0	0	0	0	0	0	0	0	-1	0	0	-1	0	-1	0	0	3	0	0

Word	HAT			HAT			RAT			SHOES					WATER				
Target	H	A	T	H	A	T	R	A	T	S	H	O	E	S	W	A	T	E	R
Detected	H	B	T	L	A	T	R	A	Z	S	H	A	Q	M	S	A	T	E	R
E(Col)	0	1	0	4	0	0	0	0	0	0	0	-2	0	0	-4	0	0	0	0
E(Row)	0	0	0	0	0	0	0	0	1	0	0	-2	2	-1	0	0	0	0	0

Table 1. Target character and detected character in the target words presented to subject. The shaded cells show when an error occurred. Differences between column/row of the target and that of the detected characters are shown in the fourth/fifth row of the table, respectively.

The first and second rows in Table 1 show the target words and characters for the first data sets, respectively. The detected character is shown in the third row. Differences between detected columns/rows and target columns/rows are shown in the fourth and fifth rows of the table. If the target row or column is detected correctly, zero is shown. If the detected row is at the bottom or top of the target row, a positive or negative number is shown respectively. In a similar notation, if the detected column is at the right or left of the target column a positive or negative number is shown, respectively. Shaded cells in this table show if there is an error in detection. From total 42 target characters, 27 characters (65%) were detected correctly and 15 characters (35%), shaded in the table, were not detected correctly because their row or column (or both) was identified wrongly. Since no training set was used, low accuracy results were expected. In this study, error cases were analyzed to search for a consistent pattern for the error.

Figure 3 shows another representation of the row and column differences. The focus in this representation is how close the detected character is to the target character. It shows that 27 times target character was detected correctly. However, any other nonzero number shows an error in detection (they are shaded cells). In Figure 3, for example, number "3", located on the top of "27", indicates that "three times" the detected row was one row above target row while the column was detected correctly or number "2", located at the far left in the figure and four cells away from 27, indicates that "two times" a column was detected which was on the left side of target column and four columns away from it, while the row was detected correctly.

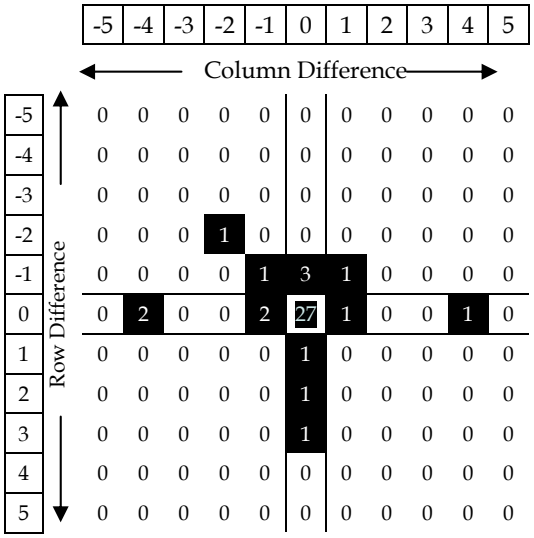


Fig. 3. Error in detecting target row and column as a function of the distance between target and detected row and column. In this figure the target row/column is considered at the center. The shaded cells indicate when an error occurs and the value of the cells show the difference between target and detected row or column.

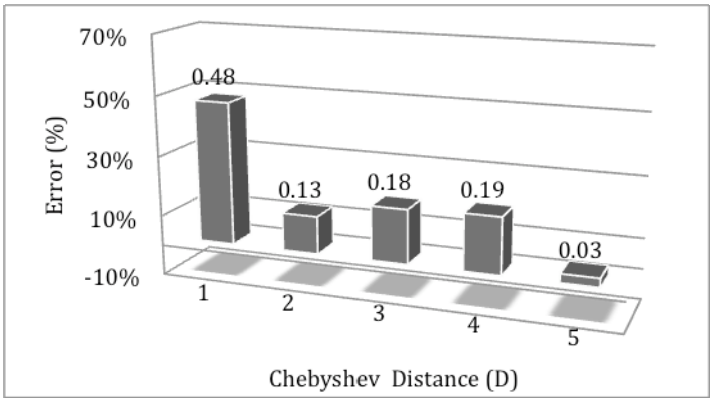


Fig. 4. Percentage of error for Farwell-Donchin paradigm as a function of distance between target (center cell) and detected row/column for both data sets. The Chebyshev distance was considered.

Using the Chebyshev distance, the distance between detected row/column and target row/column was calculated as follows:

$$= \lim_{\to\infty}\left(\sum_{=} | - |^{}\right)^{} = \quad (| - |) \tag{1}$$

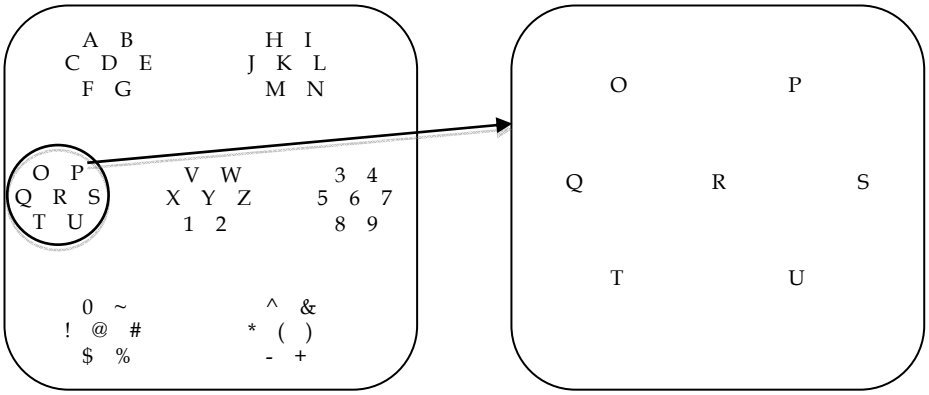
where and represent the locations of the target (center) and the wrongly detected characters, respectively. Percentage of the error cases for different values of D is shown in Figure 4.

It can be concluded that the majority of the error cases in detecting a character are when the adjacent row/column to the target row/column (D=1) was detected (48%). This can be attributed to subject's attention where the subject is unable to focus precisely only on the target letter and flashing of the adjacent row/column elicits the P300. It confirms the hypothesis that intensification of adjacent rows and columns to the desired character creates unwanted P300.

4.2 Region Based P300-Basd BCI

To overcome the problem described in the pervious section and several other shortcomings, an innovative P300-based speller, in which the screen is divided into seven regions (Figure 5) was proposed (Fazel-Rezai, 2008). In addition to 26 alphabet letters and 10 digits (0-9), 13 special characters (@,# %, *, /, …) are included, for a total of 49 symbols. The target symbol is recognized at two levels. First, the 49 symbols are placed in seven regions (seven symbols per region), and the regions are flashed in a random sequence. The user focuses on the region with the target symbol. Second, after the target region has been detected using the P300 wave, the seven symbols in the target region will be distributed among the seven regions (one symbol per region). Flashing these seven regions again will enable the identification of the target symbol. The challenges mentioned in the pervious section (a-e) were addressed as follows:

(a) Attentional blink: In the region-based system, the regions can be widely spaced on the screen, to reduce the error caused by attentional blink.
(b) Repetition blindness: Although the regions are intensified in random order, the time between intensifications in a region will be set for more than 500ms, to eliminate the repetition blindness error.
(c) Target probability: The probability of the oddball in this paradigm is 14% (1/7) - less than that for the Farwell and Donchin paradigm. Lower oddball probability increases P300 amplitude, and has two benefits; it decreases the number of trials required for each character, thereby increasing the transfer rate, and it makes P300 detection easier, which increases the accuracy.
(d) Habituation: Habituation in the new speller can be reduced by changing the region locations on the screen, their background, and other visual effects that create a novelty in the paradigm. This novelty increases the user's attention for the later presentation of a new symbol stimulus.
(e) Row/column error sensitivity: This speller has no row/column error sensitivity. Although it has a two-level error sensitivity, a "backup" word can be displayed (as the eighth region) and flashed at the beginning of the second level. If the region is not detected correctly at the first level, the user can go back to the first level by focusing on a region that is considered for "backup" (not shown in Figure 5).
Furthermore, the new region-based P300 BCI speller has several new features:
(f) Number of symbols: The total number of symbols has increased from 36 to 49.
(g) User selection: The user can choose several parameters, including region locations, region shapes, and background colors.
(h) Adaptive detection: This is one of the most important features of the proposed P300-based BCI. The objective is to design a system which adapts to the user's performance. In general, when a user has significant response (e.g., when the P300 amplitude is high), fewer trial repetitions, channels, and features are needed for a correct decision (P300 detection).



(a) (b)

Fig. 5. The region-based paradigm. a) The first level of intensification, each group contains up to 7 characters. b) One of the regions is expanded in the second level.

To evaluate the new paradigm, a data set was recorded from 10 subjects (20-29 years, 2 females). Before each experiment the procedure was explained to the subjects and they signed the informed consent approved by the Ethics Board. Subjects were seated on a chair in front of the screen and asked to be relaxed and avoid moving as much as possible during the experiments. Two paradigms were presented to all subjects. They were asked to spell "P3A". Therefore, EEG signals for total of 30 characters for each paradigm were collected. They were recorded across midline recording sites of Fz, Cz, and Pz referenced to the left mastoid with a forehead ground. EEG signals were sampled at 500 Hz.

The results for the new data set and for the Farwell-Donchin paradigm and the new paradigm are shown in Table 2 and Table 3, respectively (Abhari & Fazel-Rezai, 2006). Table 2 shows in 22 error cases out of 60 cases. However, this error was reduced to 11 out of 60 cases in a region-based paradigm (Table 3), that is 50% improvement in the accuracy.

Target character →	P		3		A		Total Error
Target column/row →	4	3	5	5	1	1	
Subject	Detected column/row						
1	4	3	5	5	5	4	2
2	1	3	6	5	1	1	2
3	4	5	1	1	1	1	3
4	4	3	1	5	1	1	1
5	3	4	5	5	1	1	2
6	6	3	2	5	1	2	3
7	5	2	5	5	6	1	3
8	1	3	6	5	1	4	3
9	4	3	5	5	4	4	2
10	4	3	4	5	1	1	1
Total Error	5	3	6	1	3	4	22

Table 2. Detected and target regions in spelling "P3A" using Donchin-Farwell paradigm.

Target character →	P		3		A		Total Error
Target region →	3	2	5	1	1	1	
Subject	Detected region						
1	4	5	5	1	1	1	2
2	3	3	5	1	1	1	1
3	3	6	5	1	1	1	1
4	3	2	5	1	1	1	0
5	3	2	1	1	1	1	1
6	1	2	5	1	1	1	1
7	3	2	4	1	2	1	2
8	3	4	4	1	1	1	2
9	3	2	5	1	1	1	0
10	3	2	1	1	1	1	1
Total Error	2	4	4	0	1	0	11

Table 3. Detected and target regions in spelling "P3A" using 7-region paradigm.

5. Conclusion

In this chapter, different components and types of BCI systems were discussed briefly and P300-based BCI was discussed in more details. In order to reveal the possible sources of error in the P300-BCI paradigm, an analysis was presented. It showed that in the Farwell-Donchin paradigm, the majority of error cases happened when a row/column adjacent to the target row/column is detected. Therefore, in the design of any new P300 speller paradigm this perceptual error should be considered. Then, a new paradigm which is based on dividing computer screen to different regions was presented. Experimental results of two paradigms confirmed that the accuracy in the region-based paradigm could be increased. In addition to higher accuracy, the 7-region paradigm has several advantages. 1) The number of characters has been increased from 36 to 49. This gives more flexibility to subjects in spelling a word. 2) The oddball probability in flashing a row/column in Farwell-Donchin is 2/12 or 1/6. This is reduced to 1/7 in the new paradigm. It has been shown that the lower probability in oddball paradigm results in higher amplitude in the P300 (Donchin et al., 2000). Therefore, in the new paradigm, P300 amplitude is larger and the detection of the P300 would be easier. This means that smaller number of flashing is required for a successful P300 detection that results in a faster P300 speller.

6. References

Wolpaw, J.R.; Birbaumer, N.; McFarland, D.J.; Pfurtscheller, G. & Vaughan, T.M. (2002). Brain-computer interfaces for communication and control. *Clinical Neurophysiology*, Vol. 113, pp. 767-791.

Jasper, H.H. (1957). The ten-twenty electrode system of the international federation. *Electroencephalogram and Clinical Neurophysiology*, Vol. 10, pp. 371-375.

Wolpaw, J.R.; McFarland, D.J.; Neat, G.W. & Forneris, C.A. (1991). An EEG-based brain-computer interface for cursor control. *Electroenceph Clin Neurophys*, Vol. 78, pp. 252-259.

Blankertz, B.; Müller K.R.; Curio G.; Vaughan T.M.; Schalk G.; Wolpaw J.R.; Schlögl A.; Neuper C.; Pfurtscheller G.; Hinterberger T. Schröder M. & Birbaumer N. (2004). The BCI competition 2003: progress and perspectives in detection and discrimination of EEG single trials. *IEEE Trans. Biomed. Eng.*, Vol. 51, pp. 1044-1052.

Pfurtscheller, G.; Flotzinger, D. & Kalcher, J. (1993). Brain computer interface-a new communication device for handicapped persons. *J. of Microcomputer Applications*, Vol. 16, pp. 293-299.

Scherer, R.; Müller G.R.; Neuper C.; Graimann B. & Pfurtscheller G. (2004). An asynchronously controlled EEG-based virtual keyboard: improvement of the spelling rate. *IEEE Trans Biomed. Eng.*, Vol. 51, No. 6, pp. 979-985.

Birch, G.E. & Mason, S.G. (2000). Brain–computer interface research at the Neil Squire Foundation. *IEEE Trans. Rehab. Eng.*, Vol. 8, No. 2, pp. 193–195.

Kostov, A. & Kostov, M. (2000). Parallel man–machine training in development of EEG-based cursor control. *IEEE Trans. Rehab. Eng.*, Vol. 8, pp. 203–205.

Sutton, S.; Braren, M.; Zubin, J. & John E.R. (1965). Information delivery and the sensory evoked potential," *Science*, Vol. 155, pp. 1436–1439.

Farwell, L.A. & Donchin, E. (1988). Talking off the top of the head: toward a mental prosthesis utilizing event-related brain potentials. *Electroenceph. Clin. Neurophysiol.*, pp. 510-523.

Donchin, E.; Spencer, K.M. & Wijesinge, R. (2000). The mental prosthesis: Assessing the speed of a P300-based brain-computer interface. *IEEE Trans. Rehab. Eng.*, Vol. 8, pp. 174–179.

Fazel-Rezai, R. & Peters, J.F. (2005), P300 wave feature extraction: preliminary results. *Proceedings of the Canadian Conference on Electrical and Computer Engineering*, pp. 376-379.

Fazel-Rezai, R. & Ramanna, S. (2005). Brain signals: feature extraction and classification using rough set methods, In: Rough sets, fuzzy sets, data mining and granular computing, D. Slezak et al. (Eds), pp. 709-718, Vol. 2, Lecture Notes in Artificial Intelligence, Springer-Verlag.

Fazel Rezai, R. & Abhari, K. (2008). A comparison between a matrix-based and a region-based P300 speller paradigms for brain-computer interface. *Proceedings of the 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 1147-1150, Vancouver, Canada.

Fazel-Rezai, R. (2007). Human error in P300 speller paradigm for brain-computer interface. *Proceedings of the 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 2516-2519, Lyon, France.

Ramanna, S. & Fazel-Rezai, R. (2006). A robust P300 detection based on rough sets *Transactions on Rough Sets*, vol. 5, pp. 207-223.

Mirghasemi, H.; Shamsollahi, M.B. & Fazel-Rezai, R. (2006). Analysis of classifiers in P300 recognition. *Proceedings of the 28th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 6205-6208, New York, USA.

Mirghasemi, H., Fazel-Rezai R. & Shamsollahi, M.B. (2006). Assessment of preprocessing methods on classifiers used in the P300 speller paradigm. *Proceedings of the 28th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 1319-1322, New York, USA.

Abhari, K. & Fazel-Rezai, R. (2006). P300-based speller paradigms for brain-computer interface. *Proceedings of the 29th Canadian Medical and Biological Engineering Conference*, Vancouver, Canada.

Raymond, J.E. et al. (1992). Temporary suppression of visual processing in a RVSP task: An attentional blink*? J. of Experimental Psychology: Human Perception and Performance*, pp. 849-860.

Kanwisher, N.G. (1987). Repetition blindness: Type recognition without token individuation. *Cognition*, Vol. 7, pp.117- 143.

Pineda, J.A; Allison B.Z. & Vankov A. (2000). The effects of self-movement, observation, and imagination on μ rhythms and readiness potentials (RP's): Toward a brain-computer interface (BCI). *IEEE Transactions on Rehabilitation Engineering*, Vol.8, No. 2, pp.219–222.

Serby, H.; Yom-Tov, E. & Inbar, G.F. (2005). An improved P300-based brain-computer interface. *IEEE Trans Neural System Rehabilitation Eng.* Vol. 13, No. 1, pp. 89-98.

Seller E.W. & Donchin E. (2006). A P300-based brain–computer interface: Initial tests by ALS patients. *Clinical Neurophysiology,* Vol. 117, pp. 538–548.

8

Alterations in Sleep Electroencephalography and Heart Rate Variability During the Obstructive Sleep Apnoea and Hypopnoea

Dean Cvetkovic[1,*], Haslaile Abdullah[2], Elif Derya Übeyli[3], Gerard Holland[4] and Irena Cosic[5]

**Corresponding author –*

[1]RMIT University; School of Electrical and Computer Engineering; GPO Box 2476V Melbourne VIC 3001, Australia;
e-mail: dean.cvetkovic@rmit.edu.au

[2]RMIT University; School of Electrical and Computer Engineering; GPO Box 2476V Melbourne VIC 3001, Australia;
email: s3209714@student.rmit.edu.au

[3]TOBB Economics and Technology University; Faculty of Engineering, Department of Electrical and Electronics Engineering; 06530 Söğütözü, Ankara, Turkey;
e-mail: edubeyli@etu.edu.tr

[4]St. Luke's Hospital; Sleep Centre; Sydney NSW, Australia;
e-mail: gholland@slc.org.au

[5]RMIT University; Science, Engineering and Technology; GPO Box 2476V Melbourne, VIC 3001, Australia;
e-mail: irena.cosic@rmit.edu.au

1. Introduction

Sleep is a vital physiological function and high quality sleep is essential for maintaining the good health. Sleep disorders however are amongst the most common disorders suffered by humans and it is rare for most people to regularly enjoy the amount of quality sleep they need. The behavioural and social causes of sleep disorders are typically the result of modern lifestyle, which are usually linked to Obstructive Sleep Apnoea Hypopnea Syndrome (OSAHS). Healthcare professionals and sleep researchers are currently looking for ways to improve the clinical diagnosis of OSAH sufferers. OSAH means "cessation of breath" [Vgontzas, 1999]. Due to this cessation of breath, the sufferer might experience related

changes in the electrical activity of the *brain and heart* [Roche, et al., 1999; Cvetkovic et al., 2009; Dingli, et al., 2003; Jurysta, et al., 2006; Task Force of the European Society of Cardiology and the North American Society of Pacing and Electrophysiology, 1996]. This activity recorded from electroencephalographic (EEG) and electrocardiographic (ECG) signals might differ in patients suffering from sleep apnoea. The analysis and tools used in sleep scoring and detection till now are based on the conventional sleep staging from whole night polysomnography (PSG) introduced by Rechtschaffen and Kales [Rechtschaffen, et al., 1968] which covere only a limited part of sleep-related EEG phenomena. A considerable number of computerised scoring systems have been introduced to attain a more standardised system of sleep EEG evaluation. In sleep research, the analysis of sleep microstructure has been recognised. The sleep EEG, in particular during the OSAH episodes often contains linear and non-linear information as well as stochastic components (noise). The separation and evaluation of these signal components remains a problem not entirely solved. Hence, new approaches to the detection and evaluation of sleep EEG transients during the OSAH episodes are required. The application of non-linear time series analysis method to sleep EEG was carried in the framework of the chaos hypothesis and characterized as the Lyapunov exponent [Fell, et al., 1993; Fell, et al., 1996; Übeyli, 2006]. Positive Lyapunov exponents have indicated that the EEG may result from a low-dimensional chaotic process. Lyapunov exponents can serve as clinically useful parameters. Estimation of the Lyapunov exponents is computationally more demanding, but estimates of these parameters are more readily interpreted with respect to the presence of chaos, as positive Lyapunov exponents are the hallmark of chaos [Haykin & Li, 1995; Abarbanel, et al., 1991]. Eigenvector methods are used for estimating frequencies and powers of signals from noise-corrupted measurements. These methods are based on an eigen-decomposition of the correlation matrix of the noise–corrupted signal. Even when the signal-to-noise ratio (SNR) is low, the eigenvector methods produce frequency spectra of high resolution. These methods are best suited to signals that can be assumed to be composed of several specific sinusoids buried in noise [Schmidt, 1986; Akay, et al., 1990; Übeyli, 2008; Übeyli, et al., 2007; Cvetkovic et al., 2009]. In this study, the eigenvector's multiple signal classification (MUSIC) method, of the linear time series characteristic, was used for estimating frequencies and powers of EEG signals from noise-corrupted measurements.

The use of Heart Rate Variability (HRV) and EEG signals for detecting cardiovascular disease and sleep disorder has been proposed to overcome the drawback of PSG method. The HRV parameters derived from frequency domain analysis of ECG R-R Intervals (RRI) have been used to assess an Autonomic Nervous System (ANS) which play an important role in the control of cardiac activity like heart rate and rhythms. The Low Frequency (LF) and High Frequency (HF) bands are used to reflect the activation of ANS subsystem; the sympathetic and parasympathetic, respectively. Whereas, the HRV's Very Low Frequency (VLF) is believed to reflect thermoregulation mechanism. The LF and HF ratio (LF/HF) is normally used as the marker of sympathovagal balances [Task Force of the European Society of Cardiology and the North American Society of Pacing and Electrophysiology, 1996]. Previous studies [Vaughn, et al., 1995; Bonnet, et al., 1997; Vanoli, et al., 1995; Jurysta, et al., 2003] have suggested that HRV vary with sleep stages; the LF component is dominant during Wake stage and gradually decreases in Non-Rapid Eye Movement (NREM) and peaks again in Rapid Eye Movement (REM) stage. An opposite trend was reported for the

HF component. Few studies have investigated the relationship between specific EEG frequency bands with the HRV parameters in healthy patients and in the sleep apnoea hypopnoea sufferers [Ako, et al., 2003; Yang, et al., 2002; Brandenberger, et al., 2001; Ehrhart, et al., 2000].

The aim of this study is two-fold. Firstly, it is set to investigate any possible changes in the human EEG activity due to OSAH occurrences by applying the non-linear and linear time series methods. Secondly, it is set to assess the correlation between EEG frequency bands and HRV in normal and sleep apnoea human clinical patients at different sleep stages. The parameters obtained from this correlation can be used to distinguish the healthy and patients suffering from sleep apnoea and hypopnoea syndromes. The first part of this study consisted of investigating the EEG activity throughout all sleep stages for one patient only. The second part of the study involved a much larger sample size of healthy and unhealthy patients to investigate the changes and possible correlation of EEG and HRV activity.

2. Data Description

Subjects
For the first part of this study, one human patient (age: 32, sex: male, weight: 96kg, height: 1.76m), diagnosed with the Apnoea Hypopnoea Index (AHI) of 5.1, was recruited for an overnight PSG recording at St. Lukes Hospital (Sydney, NSW, Australia). The second part of the study consisted of 8 healthy (5 males and 3 females) and 11 sleep apnoea patients (9 males and 2 females). Descriptive clinical features and sleep parameters of healthy and sleep disorder patients are presented in Table 1.

	Healthy	Sleep Apnoea
Age (years)	48.13 ± 10.52	50.64 ± 11.39
BMI (kg/m²)	27.01 ± 2.94	32.92 ± 5.30
AHI	2.75 ± 1.22	48.97 ± 27.52
Total sleep time (min)	379.83 ± 59.57	393.83 ± 33.01
Sleep latency (min)	26.45 ± 30.59	23.25 ± 23.89
REM latency (min)	211.36 ± 89.81	161.52 ± 39.13
Sleep efficiency (%)	82.70 ± 8.03	87.89 ± 8.31
Wake (%)	16.35 ± 8.31	11.32 ± 8.21
Stage 1 (%)	7.5 ± 6.10	6.47 ± 3.34
Stage 2 (%)	45.83 ± 5.42	52.57 ± 11.22
Stage 3 (%)	12.51 ± 3.32	10.30 ± 7.82
Stage 4 (%)	1.84 ± 3.61	2.52 ± 4.36
REM (%)	15.01 ± 7.48	16.03 ± 5.03

Table 1. Patient clinical features and sleep scoring parameters.

Experimental Protocol
An eight-hour standard clinical sleep PSG was recorded with sampling frequency of 256 Hz using Bio-Logic System and Adults Sleepscan Vision Analysis (Bio-Logic Corp, USA). Surface electrodes were placed on the scalp's surface (C3, C4 and O2; 10-20 system) and referenced to bridged left and right mastoid to record the EEG activity. Two channels were

used to record the eye movements, with one electrode placed 1 cm above and slightly lateral to the outer canthus of one eye and the second electrode recording the potentials from an electrode 1 cm below and slightly lateral to the outer canthus of the other eye. The other electrodes recorded the EMG from the muscle areas on and beneath the chin, ECG (using lead-II across the chest area), nasal and oral airflow, snoring sounds, breathing effort (measured at the chest and abdomen), oxymetry, actigraphy recording body positioning and leg movements (right and left anterior tibialis).

The respiratory signals of apnoea and hypopnoea events were evaluated using American Academy of Sleep Medicine (AASM) Criteria [American Academy of Sleep Medicine Task Force, 1999] and visually scored by the sleep technician from 30 second epochs. The one patient sleep analysis reported 64.5% sleep efficiency and 91.8% in sleep maintenance with 132 min spent in wake (W) stage, 30 min in stage 1 (S1), 125 in stage 2 (S2), 17 min in stage 3 (S3), 49 min in stage 4 (S4), 26 min in stage REM, 220 min non-REM and 3 min in movement time [Cvetkovic, et al., 2007]. The hypopnoea segments used in this investigation were only visually detected in S2.

3. Signal Processing Methods

Nonlinear Measures

The Lyapunov exponents are a quantitative measure for distinguishing among the various types of orbits based upon their sensitive dependence on the initial conditions, and are used to determine the stability of any steady-state behaviour, including chaotic solutions. The reason why chaotic systems show a periodic dynamics is that phase space trajectories that have nearly identical initial states will separate from each other at an exponentially increasing rate captured by the so called Lyapunov exponent [Haykin et al., 1995; Abarbanel, et al., 1991]. This is defined as follows. Consider two (usually the nearest) neighbouring points in phase space at time 0 and at time t, distances of the points in the *i-th* direction being $\|\delta x_i(0)\|$ and $\|\delta x_i(t)\|$, respectively. The Lyapunov exponent is then defined by the average growth rate λ_i of the initial distance,

$$\frac{\|\delta x_i(t)\|}{\|\delta x_i(0)\|} = 2^{\lambda_i t} (t \to \infty) \text{ or} \tag{1}$$

$$\lambda_i = \lim_{t\to\infty} \frac{1}{t} \log_2 \frac{\|\delta x_i(t)\|}{\|\delta x_i(0)\|} \tag{2}$$

The existence of a positive Lyapunov exponent indicates chaos. This shows that any neighbouring points with infinitesimal differences at the initial state are abruptly separate from each other in the *i -th* direction. In other words, even if the initial states are close, the final states are much different. This phenomenon is sometimes called sensitive dependence on initial conditions. Numerous methods for calculating the Lyapunov exponents have been developed in the past decade. Generally, the Lyapunov exponents can be estimated either from the equations of motion of the dynamic system (if it is known), or from the observed time series. The latter is what is of interest due to its direct relation to the work in this chapter. The idea is based on the well-known technique of state space reconstruction with delay coordinates to build a system with Lyapunov exponents identical to that of the original system from which our measurements have been observed. Generally, Lyapunov

exponents can be extracted from observed signals in two different ways. The first is based on the idea of following the time-evolution of nearby points in the state space. This method provides an estimation of the largest Lyapunov exponent only. The second method is based on the estimation of local Jacobi matrices and is capable of estimating all the Lyapunov exponents. Vectors of all the Lyapunov exponents for particular systems are often called their Lyapunov spectra [Haykin et al., 1995; Abarbanel, et al., 1991].

Linear Measures

Eigenvector methods are used for estimating frequencies and powers of signals from noise-corrupted measurements. A number of eigenvector methods have been applied by authors [Akay, et al., 1990; Übeyli, 2008; Übeyli, et al., 2007; Cvetkovic et al., 2009], such as Pisarenko, Multiple Signal Classification (MUSIC) and Minimum-Norm. The polynomial $A(f)$ which contains zeros on the unit circle can then be used to estimate the power spectral density (PSD):

$$A(f) = \sum_{k=0}^{m} a_k e^{-j2\pi fk} \tag{3}$$

where $A(f)$ represents the desired polynomial, a_k represents coefficients of the desired polynomial, and m represents the order of the eigenfilter, $A(f)$.

The polynomial can also be expressed in terms of the autocorrelation matrix R of the input signal. Assuming that the noise is white:

$$R = E\left\{x(n)^* \cdot x(n)^T\right\} = SPS^{\#} + \sigma v^2 I \tag{4}$$

where $x(n)$ is observed signal, S represents the signal direction matrix of dimension $(m+1)\times L$ and L is the dimension of the signal subspace, R is the autocorrelation matrix of dimension $(m+1)\times(m+1)$, P is the signal power matrix of dimension $(L)\times(L)$, σv^2 represents the noise power, * represents the complex conjugate, I is the identity matrix, # represents the complex conjugate transposed, T shows the matrix transposed. S, the signal direction matrix is expressed as:

$S = \left[Sw_1 \quad Sw_2 \quad \cdots \quad Sw_L\right]$

where w_1, w_2, $\cdots$, w_L represent the signal frequencies:

$Sw_i = \left[1 \quad e^{jwi} \quad e^{j2wi} \quad \cdots \quad e^{jmwi}\right]^T \qquad i = 1,2,\cdots,L$.

In practical applications, it is common to construct the estimated autocorrelation matrix $\hat{R}$ from the autocorrelation lags:

$$\hat{R}(k) = \frac{1}{N}\sum_{n=0}^{N-1-k} x(n+k)\cdot x(n) \qquad k = 0,1,\cdots,m \tag{5}$$

where k is the autocorrelation lag index and N is the number of the signal samples.
Then, the estimated autocorrelation matrix becomes:

$$\hat{R}(k)=\begin{bmatrix} \hat{R}(0) & \hat{R}(1) & \hat{R}(2) & \cdots & \hat{R}(m) \\ \hat{R}(1) & \hat{R}(0) & \hat{R}(1) & \cdots & \hat{R}(m-1) \\ \hat{R}(2) & \hat{R}(1) & \hat{R}(0) & \cdots & \hat{R}(m-2) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \hat{R}(m) & \hat{R}(m-1) & \cdots & \cdots & \hat{R}(0) \end{bmatrix} \tag{6}$$

Multiplying by the eigenvector of the autocorrelation matrix a, equation (4) can be rewritten as:

$$\hat{R}a = SPS^{\#}a + \sigma v^2 a \tag{7}$$

where a represents the eigenvector of the estimated autocorrelation matrix $\hat{R}$ and a is expressed as:

$$\left[a_0, \quad a_1, \quad \cdots, \quad a_m\right]^T.$$

In principle, under the assumption of white noise all noise subspace eigenvalues should be equal:

$$\lambda_1 = \lambda_2 = \cdots = \lambda_K = \sigma v^2$$

where λ_i represents the noise subspace eigenvalues, $i = 1,2,\cdots,K$ and K represents the dimension of the noise subspace.

The MUSIC method has been applied in this particular analysis, which is a noise subspace frequency estimator and eliminates the effects of spurious zeros by using the averaged spectra of all of the eigenvectors corresponding to the noise subspace. The resultant PSD is determined from:

$$P_{MUSIC}(f) = \frac{1}{\frac{1}{K}\sum_{i=0}^{K-1}\left|A_i(f)\right|^2} \tag{8}$$

where K represents the dimension of noise subspace, $A_i(f)$ represents the desired polynomial that corresponds to all the eigenvectors of the noise subspace [Schmidt, 1986; Akay, et al., 1990; Übeyli, 2008; Übeyli and Cvetkovic, 2007; Cvetkovic et al., 2007, 2009].

EEG and HRV Signal Processing and Analysis

The EEG, EOG and ECG data was processed and analysed using Matlab software (Mathworks, USA). The EEG-EOG correction was applied based on the regression analysis, together with source noise removal using the 50 Hz notch filter. The EEG data was processed in 30 sec segments.

A five-minute ECG data window of free movement artefacts for each sleep stage was visually selected for the HRV analysis for each patient. Signal processing algorithm implemented in Matlab, identified a QRS complex for extraction of RR Intervals (RRI) which was based on Hilbert transformation [Benitez, et al., 2001]. RRI was re-sampled at 4 Hz using Berger algorithm [Berger, et al., 1986]. Further non-parametric spectral analysis based on Fast Fourier Transform (FFT) with no windowing function was performed according to Task Force [Task Force of the European Society of Cardiology and the North American

Society of Pacing and Electrophysiology, 1996]. The absolute and normalised spectral power within each frequency band was computed using trapezoidal integration of the area under spectral curve. The HRV frequency bands are as follows: very low frequency (VLF: ≤ 0.04 Hz), low frequency (LF: 0.04-0.15 Hz) and high frequency (HF: 0.15-0.4 Hz). The normalised value was calculated as LFnu=LF/(Total power – VLF) and HFnu= HF/(Total power – VLF).

From the raw EEG signals, the same five-minute segments used for processing ECG signals were applied for spectral analysis. For the present analysis, EEGs' C3-A2 derivation was used due to our earlier study which has shown significant changes of EEG activity in an apnoea-hypopnoea patient [Cvetkovic et al., 2009]. The power spectral of 10x30-second epochs for each segment was computed using FFT with no windowing points. The estimated power is grouped into five EEG frequency bands: delta (0.5-4.5 Hz), theta (5-8.5 Hz), alpha (9-12.5 Hz), sigma (13-16.5 Hz) and beta (17-30 Hz). Initially, an absolute power was computed followed by the relative power which was derived by dividing the power within each band by the total power (0.5-30 Hz). For further analysis, the relative powers were averaged every five-minutes.

4. Results

Average Lyapunov Exponent Method
Multiple Wilcoxon (matched-pairs) signed-ranks non parametric tests were performed to analyse the pre and during-hypopnoea average Lyapunov exponents for individual EEG electrodes (C3, C4 and O2), using SPSS 17 (SPSS Inc., USA). For the average Lyapunov exponent results, a significant decrease (z=-2.934, p<0.0003) was revealed from pre-hypopnoea (mean=5.806, SD=0.434) to during-hypopnoea (mean=5.425, SD=0.539) at C3, as shown in Table 3 and Figure 1A) and 1B). It was evident for C3 electrode that throughout the 11 epochs, the average Lyapunov exponent values were significantly lower during the hypopoea in comparison to pre-hypopnoea. The other significant decrease (z=-2.312, p<0.021) from pre-hypopnoea (mean=5.664, SD=0.468) to during-hypopnoea (mean=5.445, SD=0.492) was evident at O2 electrode (see Table 3 and Figure 1B). Similar for C3 electrode, at O2, the average Lyapunov exponent values were lower during-hypopnoea (throughout the 11 epochs except for epoch 10). A *post hoc* analysis with alpha rate correction of p<0.017 was calculated using Bonferroni test for multiple average Lyapunov exponent values. According to this alpha rate correction, the only significant difference recognised was at C3 electrode of (p<0.0003).

MUSIC Method
For the MUSIC results, a significant increase (z=-2.045, p<0.041) was revealed from pre-hypopnoea (mean=0.005, SD=0.004) to during-hypopnoea (mean=0.009, SD=0.009) at O2 and sigma EEG band, as shown in Table 2 and Figure 2A). Throughout the 11 epochs, the average MUSIC values were significantly higher during the hypopoea in comparison to pre-hypopnoea (except for epochs 2 and 6) (Figure 2A)). The similar significant increase (z=-2.233, p<0.026) from pre-hypopnoea (mean=0.004, SD=0.002) to during-hypopnoea (mean=0.005, SD=0.003) was evident at the same O2 electrode and beta EEG band (see Table 2 and Figure 2B)).

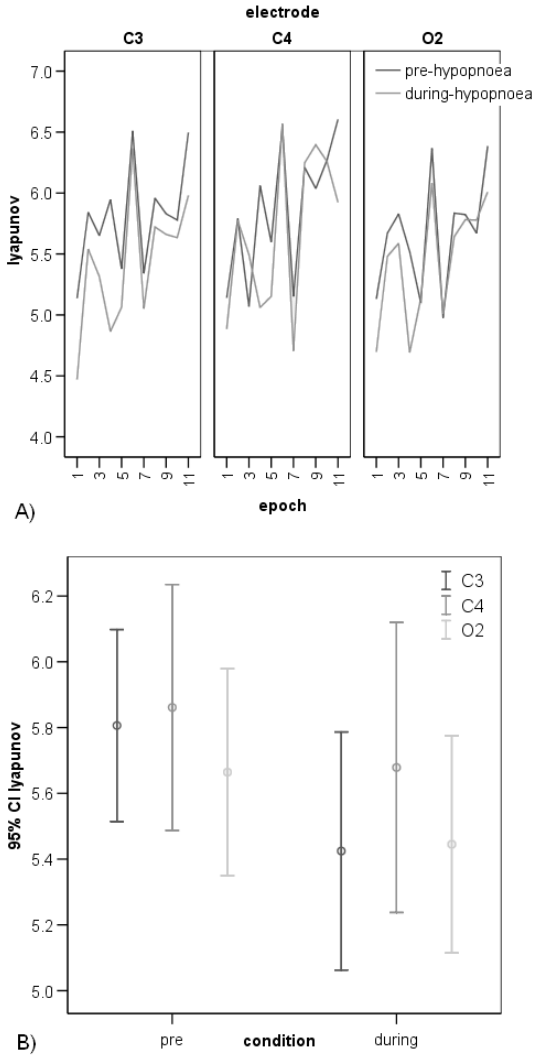


Fig. 1. The comparison of A) pre and during hypopnoea average Lyapunov exponents (y-axis) vs. 1-11 epoch (x-axis) for each EEG electrode (C3, C4 and O2) is represented. B) The confidence intervals of Lyapunov exponents (y-axis) vs. pre and during-hypopnoea conditions (x-axis) for each EEG electrode (C3, C4 and O2) is illustrated.

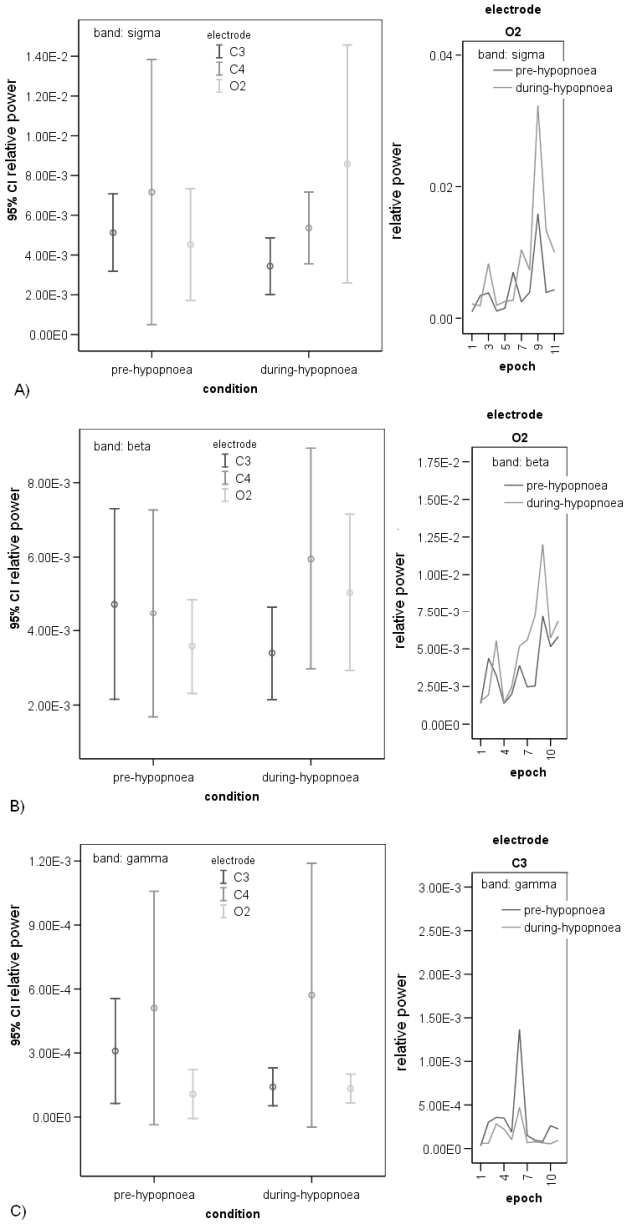


Fig. 2. The comparison of pre and during hypopnoea relative MUSIC powers (y-axis) vs. 1-11 epoch (x-axis) for specified EEG electrode is represented together with the confidence intervals of relative MUSIC powers (y-axis) vs. pre and during-hypopnoea conditions (x-axis) for each EEG electrode (C3, C4 and O2) at A) sigma, B) beta and C) gamma bands.

A significant decrease was revealed at gamma EEG band at C3 electrode (z=-2.667, p<0.008) from pre-hypopnoea (mean=0.0003, SD=0.0003) to during-hypopnoea (mean=0.0001, SD=0.0001) (see Table 2 and Figure 2B)). Epoch 1 only showed a slight inconstancy in this significant decrease throughout all the epochs. According to a *post hoc* analysis with the Bonferroni test alpha rate correction of p<0.0028, calculated for multiple average MUSIC values, none of the significant differences were recognized as actually being significant. Therefore, no significant finding could be concluded from the MUSIC method.

Linear MUSIC	Z Asymp. Sig. (p)	delta	theta	alpha	sigma	beta	gamma
C3	Z	-1.156	-1.156	-0.356	-1.511	-0.711	**-2.667**
	p	0.248	0.248	0.722	0.131	0.477	**0.008**
C4	Z	-0.622	-1.156	-0.178	-0.711	-1.334	-1.067
	p	0.534	0.248	0.859	0.477	0.182	0.286
O2	Z	-0.445	-1.067	-0.089	-2.045	-2.233	-1.778
	p	0.657	0.286	0.929	0.041	0.026	0.075

Table 2. Wilcoxon (matched-pairs) signed-ranks test analysis of pre and during-hypopnoea average MUSIC values for individual EEG electrodes (C3, C4 and O2) and bands (delta, theta, alpha, sigma, beta and gamma).

Non-linear	Z Asymp. Sig. (p)	Lyapunov
C3	Z	***-2.934***
	p	***0.003***
C4	Z	-1.156
	p	0.248
O2	Z	-2.312
	p	0.021

Table 3. Wilcoxon (matched-pairs) signed-ranks test analysis of pre and during-hypopnoea average Lyapunov exponents for individual EEG electrodes (C3, C4 and O2).

Statistical Correlation Between EEG and HRV Results

The Mann-Whitney, a non-parametric test was used to asses the differences of HRV parameters and EEG frequency bands with sleep stages *between the healthy and the sleep apnoea* patients. To study the relationship between HRV parameters and EEG frequency bands in different sleep stages, Pearson's correlation was applied in both groups. An alpha level of 0.05 was used for statistical test using SPSS 17 (SPSS Inc., USA) software.

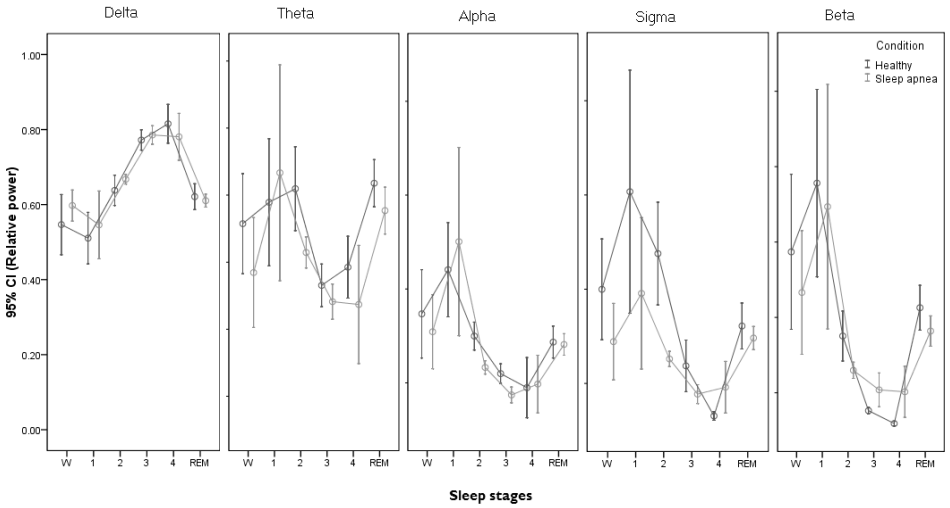


Fig. 3. Comparison of mean relative power and 95% confidence intervals (CI) of five EEG frequency bands (delta, theta, alpha, sigma and beta) versus sleep stages in the healthy and sleep apnoea groups.

Figure 3 shows that only delta EEG increased from Wake to Stage 4 and decreased in REM. Other EEG frequency bands indicated an increase from Wake to Stage 1, further decreased throughout sleep Stages 1-4 and increased again in REM. In sleep apnoea group, the relative power of theta EEG was lower compared to the healthy group during Wake ($z=1.98$, $p=0.048$), Stage 2 ($z=-2.63$, $p=0.008$) and Stage 4 ($z=-2.15$, $p=0.032$) (Figure 3). In addition to Stage 2 ($z=-4.19$, $p<0.001$), the relative power in alpha EEG of sleep apnoea group was lower in Stage 3 ($z=-4.26$, $p<0.001$). Likewise, the difference of relative power in sigma EEG for the sleep apnoea and the healthy groups were observed in Wake ($z=-2.22$, $p=0.026$), Stage 2 ($z=-5.13$, $p<0.001$) and Stage 3 ($z=-2.37$, $p=0.018$). Conversely, the relative power of beta EEG in the sleep apnoea group was higher than the healthy group during Stage 3 and lower during Stage 2 ($z=-2.95$, $p=0.003$) and REM ($z=-2.58$, $p=0.01$).

In the healthy group, the LFnu and LF/HF ratio continuously decreased from Wake to Stage 4 and peaked during REM while HFnu had the converse effect (Figure 4). In the sleep apnoea group, the LFnu was higher than the healthy group during Stage 3 ($z=-3.11$, $p=0.002$). Despite the fact that the sleep apnoea group had similar trends in HFnu component, its activity was slightly lower during Stage 3 ($z=-2.42$, $p=0.016$). The sleep apnoea group also revealed a higher LF/HF activity during Stage 2 ($z=-2.13$, $p=0.033$) and Stage 3 ($z=-2.93$, $p=0.003$) in comparison to the healthy group (Figure 4).

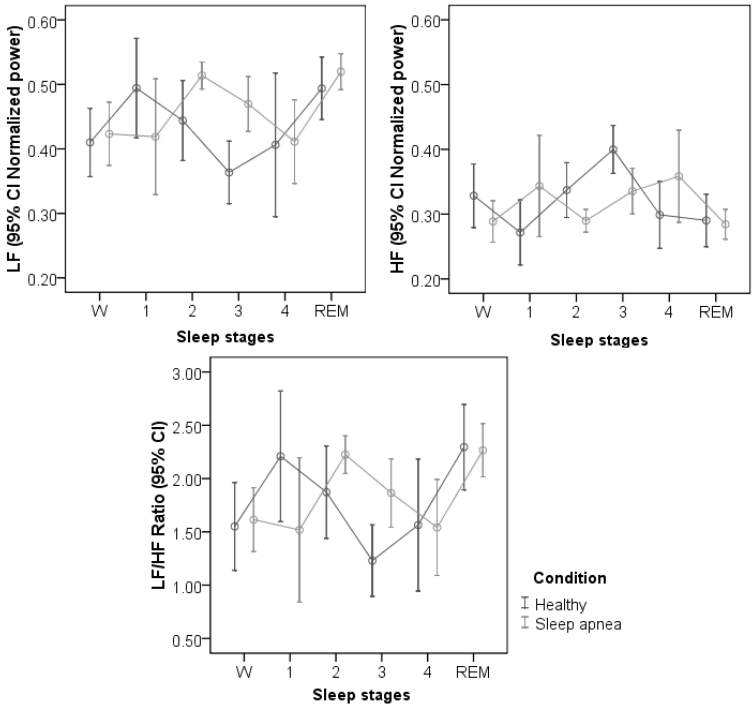


Fig. 4. Mean and 95% confidence intervals (CI) normalized power of low frequency (LF), high frequency (HF) and the ratio between LF and HF of HRV versus sleep stages in normal and patients with sleep apnoea.

The relation between the five frequency bands of EEG and HRV parameters for each sleep stage in both groups were summarized in Table 4. In the healthy groups, delta EEG negatively correlated with LF/HF and LFnu during sleep Stage 3, 4 and REM and positively correlated with HFnu in Stage 3 and 4. In contrast to healthy group, the LF/HF, LFnu and HFnu in sleep apnoea group correlated with delta EEG only in Stage 3. Theta EEG only correlated with LFnu (Stage 1) in the healthy group and with HFnu (Stage 1-2) in the sleep apnoea group. Correlation between HRV parameters and alpha EEG did not exist in the sleep apnoea group however it was only correlated with LFnu for the healthy group. Interestingly, for sleep apnoea group, sigma and beta EEG showed positive correlation with HFnu and negative correlation in LF/HF and LFnu both in Stage 2 and REM. Similar correlation trends for sigma and beta EEG bands was observed but only in LFnu and HFnu in the healthy group.

A post hoc analysis with Bonferroni alpha rate correction of $p<0.002$ revealed that in the healthy group, only delta EEG was negatively correlated with LF/HF during Stage 4. Whereas a positive correlation was observed in the sleep apnoea patients. Sigma and beta EEG also showed a significant difference for LF/HF and LFnu during Stage 2 and REM respectively in the sleep apnoea group.

		delta		theta		alpha		sigma		beta	
		H	S	H	S	H	S	H	S	H	S
LF/HF	W	0.10	0.12	0.08	-0.13	-0.05	-0.08	-0.11	-0.09	-0.18	-0.12
	S1	0.02	0.09	0.29	-0.19	0.02	-0.25	-0.14	-0.05	-0.30	0.23
	S2	0.09	-0.04	-0.20	-0.12	-0.32*	-0.14	-0.25	-0.31**	-0.15	-0.33**
	S3	-0.36*	0.29*	-0.06	0.12	0.05	-0.07	-0.22	-0.04	-0.18	-0.17
	S4	-0.88**	0.14	-0.17	0.01	0.21	-0.06	0.28	-0.09	0.25	-0.08
	REM	-0.29*	-0.04	-0.06	0.05	0.11	-0.15	-0.18	-0.32*	-0.16	-0.31*
LFnu	W	0.21	0.23	0.22	-0.01	-0.04	-0.24	0.01	0.03	-0.13	-0.04
	S1	-0.07	-0.09	0.48*	-0.13	0.26	-0.17	0.02	0.13	-0.16	0.26
	S2	0.03	-0.05	-0.09	-0.06	-0.20	-0.13	-0.37*	-0.33*	-0.09	-0.43*
	S3	-0.25	0.39**	-0.11	0.20	0.06	-0.18	-0.29*	-0.13	-0.29*	-0.25
	S4	-0.79*	0.22	-0.41	-0.10	0.63*	-0.18	0.20	-0.24	0.15	-0.24
	REM	-0.30*	-0.01	-0.02	0.08	0.14	-0.19	-0.10	-0.32**	-0.08	-0.32**
HFnu	W	-0.18	-0.12	0.06	0.08	0.29*	0.21	0.17	0.03	0.14	-0.03
	S1	-0.25	-0.22	-0.07	0.47*	0.21	0.25	0.28	0.15	0.41*	0.07
	S2	-0.14	0.13	0.11	0.26*	0.20	0.13	0.31*	0.29*	0.15	0.35*
	S3	0.39*	-0.27*	0.01	-0.08	-0.04	0.04	0.09	-0.01	0.19	0.09
	S4	0.74*	-0.01	0.07	-0.20	-0.18	-0.16	-0.20	-0.17	-0.15	-0.16
	REM	0.20	0.05	0.05	-0.11	-0.18	0.05	0.15	0.24*	0.16	0.21*

Table 4. Correlation coefficients between EEG frequency bands and HRV parameters throughout different sleep stages in the healthy (H) and sleep apnoea (S) groups.
*alpha level at $p<0.05$, ** alpha rate correction at $p<0.001$.

5. Discussion & Conclusion

The evaluation of sleep EEG transients from low-dimensional chaotic process during the OSAH episodes was conducted using the established non-linear time series analysis method, the positive Lyapunov exponent [Fell, et al., 1993; Fell, et al., 1996]. Whereas, the eigenvector's multiple signal classification (MUSIC) method, of the linear time series characteristic, was used for estimating frequencies and powers of EEG signals [Übeyli and Cvetkovic, 2007; Übeyli, 2008; Cvetkovic et al., 2007, 2009]. Previous studies investigated the analysis of the EEG signals using Lyapunov exponents, which were used as inputs of the multilayer perceptron neural networks (MLPNNs) [Übeyli, 2006], and multiclass support vector machine (SVM) [Übeyli, 2008]. Similar research demonstrated the performances of Lyapunov exponents and eigenvector methods in representing the EEG signals [Andrzejak, et al., 2001]. The power levels of the power spectral density estimates (PSDs) obtained by the eigenvector methods can be used to represent the features of the PPG, ECG, EEG signals [Übeyli and Cvetkovic, 2007].

The statistical results from this first part of our one-patient EEG sleep study needed to consider the Bonferroni test alpha rate correction which depended on the number of factors involved in the multiple tests. The reason MUSIC method revealed corrected non-significant findings was due to six EEG bands and three electrodes (18 factors). Whereas, for the Lyapunov alpha correction, only three factors were considered. As the result, the corrected alpha rate was much lower ($p<0.017$) and the only 'true' significant finding was revealed at the left central region (C3). Therefore, the results from this first part of study indicated

significant changes in the human EEG activity due to OSAH occurrences by applying the non-linear series methods at C3 electrode.

The results from the second part of this study confirmed the existence of association between the HRV parameters and delta EEG frequency band in normal patients which varies with sleep stages as reported in previous studies [Jurysta, et al., 2003; Ako, et al., 2003; Yang, et al., 2002; Brandenberger, et al., 2001]. The power extracted from HRV and EEG bands was computed using spectral analysis based on Fast Fourier Transform (FFT). Our results showed that in healthy group, delta EEG which often prevails in deep sleep was inversely correlated with LFnu and LF/HF and positively correlated with HFnu suggesting a decrease in sympathetic activity and an increase in parasympathetic activity. The present study also found an increase in LFnu and LF/HF particularly during Stage 2 in sleep apnoea patients compared with the healthy. This finding was in line with study by [Roche, et al., 1999; Cvetkovic, et al., 2009] which suggested an increase in sympathetic and parasympathetic activity around apnoea-hypopnoea episodes which occurred mostly during NREM sleep. As reported by [Svanborg, et al., 1996], increased during NREM apnoea. Our results revealed that delta EEG and LFnu positively correlated, observed in sleep apnoea group during Stage 3. In addition, beta and sigma also showed a negative association with the LFnu and LF/HF parameters. This association observed during Stage 2 and REM may be due to predominance of cardiac sympathetic during apnoea-hypopnoea episodes.

In conclusion, our study elucidates a significant correlation between HRV activity and EEG frequency bands particularly in delta, beta and sigma in sleep apnoea group. Further studies using non-linear methods for EEG and ECG feature extraction are necessary to verify this association.

6. Acknowledgement

The authors gratefully acknowledge Dr. Kwok Yan and Mr. Gerard Holland from St. Luke's Hospital (Sleep Centre) in Sydney (NSW, Australia), for providing continuous consulting, sleep monitoring and scoring input to our sleep research. The authors also thank Dr. Namunu Maddage from RMIT University for the assistance in signal processing. This work was supported in part by the Ministry of Higher Education, Malaysia and University Technology of Malaysia.

7. References

Abarbanel, H.D.I.; Brown, R. & Kennel, M.B. (1991). Lyapunov exponents in chaotic systems: their importance and their evaluation using observed data. *International Journal of Modern Physics B,* Vol. 5, No. 9, 1347-1375, ISSN: 0217-9792.

Akay, M; Semmlow, J.L.; Welkowitz, W.; Bauer, M.D. & Kostis, J.B. (1990). Noninvasive detection of coronary stenoses before and after angioplasty using eigenvector methods, *IEEE Transactions on Biomedical Engineering,* Vol. 37, No. 11, 1095-1104, ISSN:0018-9294.

Ako, M.; Kawara, T.; Uchida, S.; Miyazaki, S.; Nishihara, K.; Mukai, J.; Hirao, K.; Ako, J. & Okubo, Y. (2003). Correlation between electroencephalographic and heart rate variability during sleep, *Psychiatry and Clinical Neuroscience*, Vol. 53, pp. 59–65, ISSN: 1323-1316.

American Academy of Sleep Medicine Task Force. (1999). Sleep-related breathing disorders in adults: Recommendations for syndrome definition and measurement techniques in clinical research, *Sleep,* Vol. 22, pp. 667–688, ISSN: 0161-8105.

Andrzejak, R.G.; Lehnertz, K.; Mormann, F.; Rieke, C.; David, P. & Elger, C.E. (2001). Indications of nonlinear deterministic and finite-dimensional structures in time seires of brain electrical activity: dependence on recording region and brain state, *Physical Review E,* Vol. 64, 061907-1-061907-8, ISSN: 1539-3755.

Benitez, D.; Gaydecki, P.A.; Zaidi A. & Fitzpatrick, A.P. (2001). The use of the Hilbert transform in ECG signal analysis, *Computers in Biology and Medicine,* Vol. 31, pp. 399–406, ISSN: 0010-4825.

Berger, R.D.; Akselrod, S.; Gordon, D. & Cohen, R.J. (1986). An efficient algorithm for spectral analysis of heart rate variability, *IEEE Transaction on Biomedical Engineering,* Vol. 33, pp. 900-904, ISSN: 0018-9294.

Bonnet, M. H. & Arand, D. L. (1997). Heart rate variability: sleep stage, time of night and arousal influences. *Electroencephalography and Clinical Neurophysioogyl.* Vol. 102, pp. 390-396, ISSN: 0013-4694.

Brandenberger, G.; Ehrhart, J.; Piquard, F. & Simon, C. (2001). Inverse coupling between ultradian oscillations in delta wave activity and heart rate variability during sleep, *Clinical Neurophysiology,* Vol. 112, pp. 992–996, ISSN: 1388-2457.

Cvetkovic, D.; Holland, G. & Cosic, I. (2007). The relationship between EEG, PPG and oxymetric signal responses during the OSAH events: A pilot sleep study, *Sleep and Biological Rhythms*, Vol. 5, Supp. 1, A42, ISSN: 1446-9235.

Cvetkovic, D.; Übeyli, E.D.; Holland, G. & Cosic, I. (2009). Alterations in sleep EEG activity during the hypopnoea episodes, *Journal of Medical Systems*, (doi: 10.1007/s10916-009-9261-1), ISSN: 0148-5598.

Dingli, K.; Assimakopoulos, T.; Wraith, P.K.; Fietze, I.; Witt, C. & Douglas, N.J. (2003). Spectral oscillations of RR intervals in sleep apnoea/hypopnoea syndrome patients, *European Respiratory JournaJ*, Vol. 22, pp. 943–950, ISSN: 0903-1936.

Ehrhart, J.; Toussaint, M.; Simon, C;, Gronfier, C.; Luthringer, R. & Brandenberger, G. (2000). Alpha activity and cardiac correlates: three types of relationships during nocturnal sleep, *Clinical Neurophysiology*, Vol. 111, pp. 940–946, ISSN: 1388-2457.

Fell, J.; Roschke, J. & Beckmann, P. (1993). Deterministic chaos and the first positive Lyapunov exponent: a nonlinear analysis of the human electroencephalogram during sleep. *Biol. Cybern.,* Vol. 69, 139-146, ISSN: 0340-1200.

Fell, J.; Roschke, J.; Mann, K. & Schaffner, C. (1996). Discrimination of sleep stages: a comparison between spectral and nonlinear EEG measures. *Electroencephalography and Clinical Neurophysiology,* Vol. 98, 401-410, ISSN:0013-4694.

Haykin, S. & Li, X.B. (1995). Analysis of EEG signals using Lyapunov exponents. *Proceedings of the IEEE,* Vol. 83, No. 1, 95-122, ISSN: 0018-9219.

Jurysta, F.; Lanquart, J.P.; van de Borne, P.; Migeotte, P.F.; Dumont, M.; Degaute, J.P. & Linkowski, P. (2006). The link between cardiac autonomic activity and sleep delta power is altered in men with sleep apnea-hypopnea syndrome, *The American*

Journal of Physiology-Regulatory, Integrative and Comparative Physiology, Vol. 291, pp.1165-1171, ISSN: 0363-6119.

Jurysta, F.; van de Borne, P.; Migeotte, P.F.; Dumont, M.; Lanquart, J.P.; Degaute, J.P. & Linkowski, P. (2003). A study of the dynamic interactions between sleep EEG and heart rate variability in healthy young men, *Clinical Neurophysiology.* Vol. 114, pp. 2146-2155, ISSN: 1388-2457.

Rechtschaffen, A. & Kales, A. (1968). *A manual of standardized terminology techniques and scoring system for sleep stages of human subjects*, Washington, DC: US Government Printing Office; Publication No. 204.

Roche, F.; Gaspoz, J.M.; Court-Fortune, I.; Minini, P.; Pichot, V.; Duverney, D.; Costes, F.; Lacour, J.R. & Barthélémy, J.C. (1999). Screening of obstructive sleep apnea syndrome by heart rate variability analysis, *Circulation,* Vol. 100, pp.1411–1415, ISSN: 1524-4539.

Schmidt, R.O. (1986). Multiple emitter location and signal parameter estimation. *IEEE Transactions on Antennas and Propagation,* Vol. AP-34, No. 3, 276-280, ISSN:0018-926X.

Svanborg, E. & Guilleminaul, C. (1996). EEG frequency changes during sleep apneas, *Sleep*, Vol. 19, pp. 248-254, ISSN: 0161-8105.

Task Force of the European Society of Cardiology and the North American Society of Pacing and Electrophysiology. (1996). Heart rate variability: standards of measurement, physiological interpretation and clinical use, *Circulation*, Vol. 93, pp.1043-1065, ISSN: 1524-4539.

Übeyli, E.D. (2006). Analysis of EEG signals using Lyapunov exponents. *Neural Network World,* Vol. 16, No. 3, 257-273, ISSN: 1210-0552.

Übeyli, E.D. (2008). Analysis of EEG signals by combining eigenvector methods and multiclass support vector machines, *Computers in Biology and Medicine,* Vol. 38, No. 1, 14-22, ISSN: 0010-4825.

Übeyli, E.D.; Cvetkovic, D. & Cosic, I. (2007). Eigenvector methods for analysis of human PPG, ECG, and EEG signals, *Proceedings of 29th Annual International Conference IEEE Engineering in Medicine and Biology Society (EMBS),* pp. 3304-3307, ISSN: 1557-170X ISBN: 978-1-4244-0787-3, Lyon, France, Sep. 2007, IEEE.

Vanoli, E.; Adamson, P.B.; Ba-Lin; Pinna, G.D.; Lazzara, R. & Orr, W.C. (1995). Heart rate variability during specific sleep stages, *Circulation*, Vol. 91, pp. 1918-1922, ISSN: 1524-4539.

Vaughn, B.V.; Quint, S.R.; Messenheimer, J.A. & Robertson, K.R. (1995). Heart period variability in sleep, *Electroencephalography and Clinical Neurophysiology.* Vol. 94, pp.155-162, ISSN: 0013-4694.

Vgontzas, A.N. & Kales, A. (1999). Sleep and its disorders. *Annual Review of Medicine,* Vol. 50, 387-400, ISSN: 0066-4219.

Yang, C.; Lai, C.W.; Lai, H.Y. & Kuo, T.B.J. (2002). Relationship between electroencephalogram slow-wave magnitude and heart rate variability during sleep in humans, *Neuroscience Letter,* Vol. 329, pp. 213–216, ISSN: 0304-3940.

9

Flexible implantable thin film neural electrodes

Sami Myllymaa, Katja Myllymaa and Reijo Lappalainen
University of Kuopio, Department of Physics
Finland

1. Introduction

Implantable neural electrodes are widely used tools in basic neuroscience to study the behavior and function of the central nervous system at the tissue and cellular level. There is also a growing interest in the clinical applications of stimulating and recording neural electrodes as well as implantable neural prostheses in which the microelectrodes act as a functional element that connects neurons, i.e. the electrically active cells of the nervous system to the electronic circuitry. Several implantable neural prostheses such as cardiac pacemakers, cochlear implants and deep brain stimulators have been already successfully commercialized. Cochlear implant is a surgically implanted electronic device that utilizes an electrode array inserted into the scala tympani of the cochlea to stimulate the auditory nerve through the bone, enabling the deaf to hear sounds. Deep brain stimulators consist of intracortically implanted electrodes through which a high frequency electrical stimulation is delivered to a targeted region of the brain; they are used in treating several neurological disorders, e.g. Parkinson's disease. Intracranial electrodes consisting of depth electrodes (probes) and subdural electrodes are used to measure local field potentials and occasionally spikes in some clinical cases such as in the presurgical evaluation of patients who are candidates for epilepsy or brain tumor surgery.

Despite the major advances in microsystem technologies and microelectronics, most human neural electrodes are handcrafted from platinum foils and silicone sheet and therefore the devices are quite bulky. Miniaturized electrode arrays with high-channel densities are common in animal models, but their human counterparts are mainly being investigated or at best undergoing clinical trials. Microfabrication techniques, e.g. photolithography and thin film deposition could be useful to obtain multichannel microelectrodes with a high channel density. This would result in a higher spatial resolution which is necessary for accurate recording and stimulation of tiny nervous structures. One major advantage of using well-proven lithographic and thin films techniques is the high reproducibility and the possibility to manufacture many electrodes in a single batch. Additionally modern nanotechnology provides many tools and opportunities to improve the long-term functionality and biocompatibility of implantable electrodes. These include the use of surface modifying thin films of high purity materials and controlled nanotopography and/or microtopography which can be achieved for example by laser ablation techniques.

In this chapter, we provide a description of requirements related to implantable electrode materials. And then we discuss the state of the art of the electrodes used in intracranial

recordings and stimulation, taking into account not only neurophysiological research but also their clinical applications. We will depict the fabrication process and performance of our flexible polyimide-based microelectrodes as a tool for multi-channel cortical surface recordings. We critically review the corresponding other flexible sensor approaches and compare our fabrication methods, material choices and *in vivo* and *in vitro* performance of our electrodes to those of other research groups. We will also discuss the future trends and strategies to prolong functional life span and to minimize the immune response to implanted electrodes.

2. Requirements for materials used in implantable electrodes

The requirements and demands placed on the materials used in invasive neural electrodes and prostheses are understandingly very high. There are a number of important factors that must be taken into account when selecting materials for electrodes and for the substrate/encapsulation layer to ensure optimal performance and longevity of the neural implant. The ideal neural implant material must be biocompatible without triggering any vigorous local or generalized host response or allergic reactions. The electrode-tissue impedance must be stable and low enough to enable long-term reliable measurements/stimulations. Additionally, the materials should be compatible with magnetic resonance imaging and they should be visible radiographically (Geddes & Roeder, 2003). In this section, we will discuss these requirements by describing materials typically used particularly concentrating on their electrochemical properties. Biocompatibility aspects will be discussed in section 5.

2.1 Electrode materials

The key function of recording electrodes is to convert the ionic current around the electrode site into electron current. On the other hand, stimulation electrodes deliver the electrical charge to brain tissue aiming to stimulate excitable cells and tissue. Electrodes can be roughly divided into polarizable and non-polarizable electrodes. Ideally polarizable electrode passes current from the electrode to an electrolytic solution by changing the charge distribution within the solution near the electrode. This kind of electrode acts like a capacitor, and only the displacement current crosses the interface. In contrast, ideally a non-polarizable electrode allows the current to pass freely through the electrode-electrolytic interface without changing the charge distribution in the electrolytic solution.

Two identical bioelectrodes in contact with electrolyte can be modeled in an equivalent circuit model as presented in Fig. 1A. Electrode polarization impedance consists of Warburg (polarization) resistance R_w and capacitance C_w. The Faradic resistance, R_f as a parallel with Warburg components, provides a route for direct current to pass through the interface. R_s is the resistance of the electrolyte solution. Polarization reactance ($X_w = 1/2\pi fC_w$) as well as R_w are both frequency dependent: they decrease as ($1/\sqrt{f}$) with increasing frequency. At a sufficient high frequency, the impedance between two electrodes in saline asymptotically approaches the value of R_s (Fig. 1B). R_w, C_w and R_f are also dependent on the geometry of the electrode material, electrode size and current density.

A silver-silver chloride electrode (Ag/AgCl) is almost perfectly non-polarizable, has a small and stable offset potential and therefore is widely utilized as the surface (skin) biopotential electrode, but it cannot be used as an implantable electrode because of the rapid dissolution

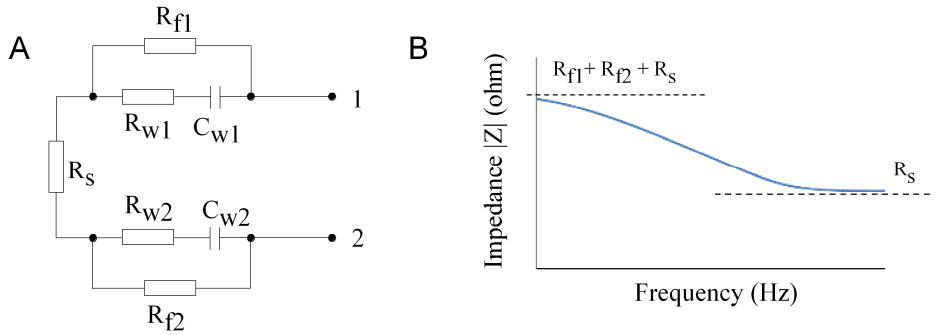


Fig. 1. (A) The equivalent circuit model for a two identical bioelectrodes immersed in saline. (B) Typical electrode-electrolyte impedance graph measured between the terminals 1 and 2 as a function of frequency

of silver. Furthermore, AgCl as well as pure Ag has a toxic effect on neural tissue (Yuen et al., 1987). Noble metals such as platinum (Pt), gold (Au) and iridium (Ir) have been the most commonly used electrode materials in neural prosthetic applications due to their excellent corrosion resistance and biocompatibility properties (Geddes & Roeder, 2003). They are highly polarizable electrodes meaning that the ionic current in tissue can give rise to an electric current in the metal electrode, and almost no charge carriers pass through the interface. Current densities for recording purposes are so low that the interface has a linear current-voltage relationship. If one wishes stimulation, then higher current densities are required to elicit neural excitation reactions resulting in Faradaic current flow, in which charge carriers pass through the interface. Irreversible electrochemical reactions can occur and change the chemical composition of the electrode-electrolyte interface. The limit for a maximum reversible charge injection depends on the electrode material as well as its shape, size and the stimulation waveform.

Pt is clinically utilized as an electrode material in various neural applications such as in cardiac pacemakers, cochlear implants and subdural strip and grid electrodes. Pt suffers from corrosion only slightly in both recording and stimulating use. Pt is commonly used as an inert reference electrode in electrochemistry since it is at the extreme of the galvanic series. Au is often used in recording electrodes but it suffers from corrosion if used for stimulating purposes. On the contrary, Ir is mainly utilized as a stimulating electrode. The electrode properties of Ir can be further developed by electrochemical activation that leads to the adhesion of a porous and very stable oxide layer on the electrode surface. The property of enhanced corrosion resistance of activated iridium oxide (AIROF) is particularly important in stimulating electrodes to avoid metal dissolution during electric pulsing. The AIROF has a high charge delivery capacity and therefore it is well suited for use in stimulating electrodes. Activation also reduces the electrode impedance due to the porosity. Another new potential electrode candidate for use in stimulating electrodes is titanium nitride (TiN). The high charge delivery capacity of TiN is based on its high effective surface area due to the columnar structure of this material. TiN is already being clinically utilized in pacemakers. Optically transparent indium tin oxide (ITO) is an extensively employed electrode material used in electrochemistry, but at present its use in neurophysiology is virtually limited to *in vitro* microelectrode arrays for recording electrical signals from

neurons. The compatibility of ITO for chronic *in vivo* recording is not completely understood although recent investigations have shown promising results in terms of cell adhesion and proliferation (Bogner et al., 2006) as well as protein adsorption (Selvakumaran et al., 2008). Correspondingly, the electrically conductive hard ceramic, glassy carbon is an extensively applied electrode material in electroanalytical chemistry, but it is rarely used as an implantable electrode.

2.2 Substrate and encapsulation materials

The substrate and encapsulation materials need to be biocompatible, biostable and possess good dielectric properties. Good biocompatibility means that material must not evoke a toxic, allergic or immunologic reaction. Biostability means that the implant material must not be susceptible to attack by biological fluids or any metabolic substances. The surface area of substrate/encapsulation material is much larger compared to areas of the active electrode sites. Therefore substrate and insulator materials have an important role in long-term stability and functioning of neural implants.

A substrate material that has been widely utilized for neural interfaces in the central nervous system is silicon (Si). Silicon dioxide (SiO_2) and silicon nitride (Si_3N_4) are typical insulators used with silicon base. However, there is a mismatch between the mechanical properties of the soft brain tissue and the rigid silicon and thus causes many adverse effects such as tissue damage, inflammatory reactions and scar formation (Polikov et al., 2005). Therefore, there has been a growing interest toward developing polymer-based implant materials that could be flexible enough to mimic biological tissue and to reduce mechanical damage but not evoking any adverse tissue reactions. Additionally, their composition and surface properties can be modified in a variety of ways to make by them more biocompatible. Polymers can be employed as both substrate and/or encapsulation materials. Different flexible polymer materials such as polyimide, benzocyclobutene (BCB), polydimethylsiloxane (PDMS), parylene and epoxy resins have been recently investigated as substrate/encapsulating materials in implantable neural interfaces (Cheung, 2007). Polyimide has been used as an insulator in microelectronics because of its numerous excellent properties: excellent resistance to solvents, strong adhesion to metals and metal oxides and good dielectric properties. The volume resistivity and dielectric strength of polyimide are comparable to silicon and typical silicon insulation materials, such as SiO_2 and Si_3N_4 (Stieglitz et al., 2000). When used as a neural implant material, polyimide has also appropriate mechanical properties and its biocompatibility has been demonstrated in many studies *in vitro* and *in vivo* (Stieglitz et al., 2000; Richardson et al., 1993). Table 1 summarizes some important properties of polyimide and PDMS in comparison to Si and SiO_2. The major disadvantage of polyimide is its permeability to environmental moisture and ions that could be crucial in terms of short circuiting and reducing the electrode's life span. BCB has been a widely used material in microelectronics due to its low moisture absorption, good dielectric properties and chemical stability. Recently BCB has been demonstrated to be a potential material for use in implantable neural interfaces (Lee et al., 2004a). Parylene and silicone rubber (PDMS) are known to be biocompatible materials and they are approved for use as an implanted material in medical devices (i.e. FDA approved). Parylene is a thermoplastic polymer that can be vapor-deposited at room temperature to form pin-hole free barrier coatings that are stress-free, chemically and biologically inert and stable as well as being

minimally permeable to moisture. Paralyne as well as PDMS are both hydrophobic and optically transparent materials and thus they can be used in optical elements.

	Units	PDMS	Polyimide	Silicon	Silicon dioxide
Tensile strength	kg/mm^2	0.4-1.1	13.5	1225	5-14
Elongation	%	140	15	-	-
Density	g/cm^3	1.11	1.45	2.1	2.3
Young's modulus	kg/mm^2	0.04 – 0.09	245	$20 \cdot 10^3$	$7 \cdot 10^3$
Dielectric constant at 1 kHz, 50 % RH	relative (ε_r)	2.70	3.3	12	3.9
Dielectric strength at 1 KHz, 50 % RH	V/cm	$1.4 \cdot 10^5$	$2 \cdot 10^6$	$3 \cdot 10^5$	$6 \cdot 10^6$
Volume resistivity	$\Omega \cdot cm$	$2.9 \cdot 10^{14}$	10^{16}	-	10^{16}

Table 1. Some physical properties of PDMS (Sylgard 184, Dow Corning Inc.) and polyimide (Pyralin PI 2525, HD Microsystems GmbH) in comparison to silicon and silicon dioxide. Values are taken from the datasheets for Sylgard 184 (Dow Corning Inc.) and for Pyralin PI 2525 (HD Microsystems GmbH) and from Stieglitz et al., 2000; Mata et al., 2005; Armani et al., 1999

3. Recent recording and stimulating brain electrodes in clinical use

Human electroencephalography (EEG) recordings are typically performed with different numbers of electrodes attached to the scalp and measuring the potential differences between electrodes. However, source localization of scalp-EEG is rather poor. To achieve better spatial resolution, it is advantageous to place the electrodes closer to the brain tissue. Both penetrating (intracortical) electrode probes and non-penetrating electrode arrays placed on the surface of cerebellum cortex have been utilized to measure neuronal action potentials and local field potentials from the brain. Compared to human scalp-EEG, intracranial recordings are advantageous in terms of increased spatial resolution, signal amplitude level and reduced noise level meaning that one can acquire more detailed maps directly of brain. However, the use of penetrating electrode probes in humans is limited to the areas destined for surgical resection because of the risk of damaging the brain tissue, whereas flexible subdural strip and grid electrodes placed on the surface of cortex are used as a less invasive method in some clinical cases, e.g. in mapping cerebellum cortex in patients undergoing epilepsy or brain tumor surgery. The signals recorded with subdural electrodes are considerably higher in amplitude because they are much closer to the source of the activity, and are separated from it by only a relatively high conductivity media (CSF, brain parenchyma). Subdural electrodes are virtually free of artifacts (e.g. not affected by electrical signals originating from muscles) that are seen in scalp-EEG, and thus they yield a much higher signal-to-noise ratio (Nair et al., 2008). Typical commercially available subdural strip and grid electrodes consist of stainless steel or platinum foil contacts with exposed diameter of about 2-3 mm embedded in flexible biomedical-grade silicone plates (e.g., AD Tech

Medical Instrument Corp., Racine, WI, USA; PMT Corp. Chanhassen, MN, USA). Epidurally implanted surface electrodes as well as penetrating intracortical probes are also utilized for stimulating purposes in humans. Motor cortex stimulation (MCS) is a relative new technique in which surface electrodes are used at low current densities to stimulate the motor area (Brodman 4) in treating chronic neuropathic pain (Rasche et al., 2006). Deep brain stimulators consist of intracortically implanted electrodes through which a high frequency electrical stimulation is delivered to targeted regions of the brain, for example being used in treating severe neurological disorders, e.g. in Parkinson's disease.
Although modern microfabrication techniques possess a great potential to develop novel high-density microelectrode arrays for recording of brain activity or electrical stimulation of neural tissue, and these kinds of microarrays are extensively utilized in animal studies, no microelectrode array suitable for use with human patients is on the market today. For example, different lithographic and thin film techniques could be utilized in the development of substrate-integrated electrodes with higher spatial resolution than the present subdural electrodes in clinical use.

4. Microelectrode arrays in animal models

Neurophysiologists have traditionally performed chronic cortical recordings in animal models with insulated microwires and wire bundles (Nicolelis et al., 1997; Williams et al., 1999). The first chronic recordings from freely moving animals with stainless steel microwire electrodes were performed in the 1950s (Strumwasser, 1958). The development of photolithographic, silicon etching and thin-film deposition techniques provided novel approaches in the development of improved recording and stimulating systems used in experiments with animal models (Wise, 2005). Silicon has been the most widely used substrate material because of the well-established precise microfabrication methods which have been developed for integrated circuits. The main advantage of microfabricated electrode arrays compared to microwire bundles is the precise control of the electrode sizes and the separations between electrodes. Batch-processing also means that there can be large volume and low-cost production of identical microsensors. Michigan probes (Najafi & Wise, 1986) and Utah arrays (Campbell et al., 1991) are perhaps the best-known Si-based microelectrodes. The typical Michigan arrays are single-shank or multi-shank Si penetrating probes where the electrodes are placed along the length of the shank allowing the measurement of neuronal activity at various depths of the brain tissue. The three-dimensional Utah array consists of 100 conductive Si needles (length 1.5 mm) with platinum coated tips. However, due to mechanical mismatch between the rigid Si (Young's modulus ~ 170 GPa) and soft brain tissue (~3 kPa) and subsequent adverse tissue responses, there has been a growing interest in developing polymer-based implants that could be flexible enough to mimic biological tissue and to reduce adverse tissue reactions.

4.1 Flexible polymer-based microelectrode arrays

Different flexible polymer materials such as polyimide, BCB, PDMS, parylene-C and epoxy resins have been recently investigated as substrate materials in implantable neural interfaces (Cheung, 2007). These polymers are also utilized in coating silicon-based implants providing both a chemical barrier and an electrical insulation layer. The flexibility of the polymer may decrease the mismatch between the mechanical properties of nervous tissue and the rigid

implant and thus reduce the risk of tissue damage and inflammation reactions due to electrode micromotion.

Several polyimide-based microelectrode arrays (MEA), i.e. polyimide-metal-polyimide sandwich structures have been developed in the field of neural engineering for various applications (Cheung, 2007). Typically these approaches utilize polyimide as a base material with a thickness of a few tens of micrometers. A thin film of suitable electrode material, such as Au, Pt or Ir is coated on the top of polyimide by using physical vapour deposition techniques, e.g. sputtering or evaporation, and then they are structured photolithographically. A thin layer (a few micrometers) of insulating material is structured on the top of the metallization layer leaving the desired areas (i.e. electrode sites) open, whereas the interconnecting lines are completely covered to avoid short-circuits in a moist environment containing salt ions. These kinds of flexible microelectrode arrays have been developed for studying brain slices *in vitro* (Boppart et al., 1992), cortical surface field potential recordings (Owens et al., 1995; Hollenberg et al., 2006; Takahashi et al., 2003; Myllymaa et al., 2008b; 2009) and intracortical multiunit neural activity recordings (Rousche et al., 2001; Cheung et al., 2007; Mercanzini et al., 2008) and for action potential recording in nerve and muscle tissues *in vivo* (González & Rodríguez 1997; Rodríguez et al., 2000; Spence et al., 2007). Stieglitz et al. have also fabricated different kinds of polyimide-based microdevices (e.g. sieve and cuff electrodes) for interfacing with regenerating peripheral nerves (Stieglitz et al., 2000; Stieglitz et al., 1997; 2001). Some recently published papers focusing on the development of polymer-based neural microelectrodes for recording brain activity and stimulating the brain tissue are summarized in Table 2.

reference	substrate (S)/ insulator (I) material and their thicknesses	electrode material and its deposition method	electrode size and number of electrodes	electrode impedance measured *in vitro* at 1 kHz	application (performed *in vivo* studies)
Owens et al., 1995	S: polyimide 2 µm I: polyimide 2 µm	evaporated Cr (15 nm) + Au (300 nm) + electroplated Pt black	40 x 40 µm^2 24 electrodes (lithographic patterning)	11-16 kΩ	acute evoked potential recording from the surface of ferret cortex
Rousche et al., 2001	S: polyimide 10 µm I: polyimide 10 µm	evaporated Cr (25 nm) + Au (200 nm)	30 x 30 µm^2 (lithographic patterning)	1837 ± 197.3 kΩ	intracortical recording with bioactive capability
Takahashi et al., 2003	S: polyimide 25 µm I: polyimide < 1 µm	Cr (thin) + Au (200 nm)	80 x 80 µm^2 69 electrodes (lithographic patterning)	330 ± 65 kΩ	acute evoked potential recording from the surface of rat cortex
Lee et al., 2004b	S: BCB 10 µm I: BCB 10 µm	Au (200 nm)	20 x 20 µm^2 3 electrodes (lithographic patterning)	~1200 kΩ	intracortical recording with micro-fluidic channels

Takeuchi et al., 2005	S: parylene 5 µm I: parylene 5 µm	Au (1 µm)	different geometries and numbers of electrodes	~100 kΩ (single Au electrode inside the channel)	intracortical recording with micro-fluidic channels
Kitzmiller et al., 2006	PDMS	evaporated Pt (200 nm)	200 x 200 µm^2 8 electrodes (lithographic patterning)	not mentioned	acute evoked potential recording from the surface of pig cortex
Hollenberg et al., 2006; Yeager et al., 2008	S: polyimide 25/50 µm I: SU-8 epoxy 13 µm	sputtered Ti-W (5 nm) + Au (300 nm)	Ø: 150 µm (round) 64 electrodes (lithographic patterning)	225 ± 90 kΩ	acute and chronic evoked potential recording from the surface of rat cortex
Molina-Luna et al., 2007; Hosp et al., 2008	S: polyimide 7 µm I: polyimide 7 µm	Ti (10 nm) + Au (300 nm) + TiN (800 nm)	Ø: 100 µm (round) 72 electrodes (lithographic patterning)	5-10 kΩ	epidural stimulation mapping of rat motor cortex
Cheung et al., 2007; Mercanzini et al., 2008	S: polyimide 20 µm I: polyimide 1.5 µm + 1.5 µm	sputtered Ti (50 nm) + Pt (200 nm)	Ø: 25 µm (round) 16 electrodes (in two layers)	~1000 kΩ	acute and chronic intracortical recordings into the rat cortex
Myllymaa et al., 2008b; 2009	S: polyimide 30 µm I: polyimide 3 µm	sputtered Ti (20 nm) + Pt (200 nm)	Ø: 100-200 µm (round) 8-16 electrodes (lithographic patterning)	25.5 ± 2.0 kΩ	acute and chronic recording evoked potentials from the surface of rat cortex

Table 2. Microfabricated polymer-based neural electrodes for recording brain activity and stimulating brain tissue (summary of recent publications)

4.2 Development of microelectrode arrays (MEA) for cortical surface recordings

Our sensor research is focused on the development of polymer-based thin film electrodes and electrode arrays which can be used for recording of biosignals as well as for stimulation of excitable tissue. The fabrication and testing of flexible microelectrode array that is suitable for multichannel cortical surface recordings in rats is described in detail elsewhere (Myllymaa et al., 2008b; 2009). Here we discuss the major aspects of the development and fabrication process as well as the results from *in vitro* and *in vivo* testing. The fabrication process was developed and iterated to a fully functional solution during a two year trial period when suitable substrate material, substrate thickness, electrode material, sensor layout and insulation material were optimized (Myllymaa et al., 2008a).

4.2.1 Array fabrication

The fabrication process of the microelectrode array is based on creating magnetron sputter deposited (AJA Inc., Stiletto Serie ST20I), lithographically patterned stable Pt thin films between two biocompatible polyimide layers. The fabrication steps of arrays are schematically illustrated in Fig. 2. The fabrication was implemented on top of microscope glass slides, 2″ x 3″ (Logitech), to guarantee a rigid support during fabrication. As a base layer we used a 25 µm thick Kapton HN film (DuPont) together with spin coated Pyralin PI 2525 (HD Microsystems GmbH) polyimide layer (Fig. 2A). A negative photoresist (ma-N 1420, Micro resist technology GmbH) was spun on the base layer and patterned using 365 nm UV exposure and wet developer chemicals (Fig. 2B). Thin films of Ti (20 nm) and Pt (200 nm) were DC sputtered onto a surface of the base layer (Fig. 2C) and structured with a lift-off technique (Fig. 2D). Ti was used beneath the Pt to improve the adhesion between the polyimide and the metal layer. Then, photosensitive polyimide PI-2771 (HD Microsystems GmbH) was spin coated on an array pattern (Fig. 2E) and patterned with alignment of metallization layer. About a 3 µm thick insulation layer was formed on all positions except for the areas of the electrodes and connection pads (Fig. 2F). Finally, the array was detached from the glass slide (Fig. 2G).

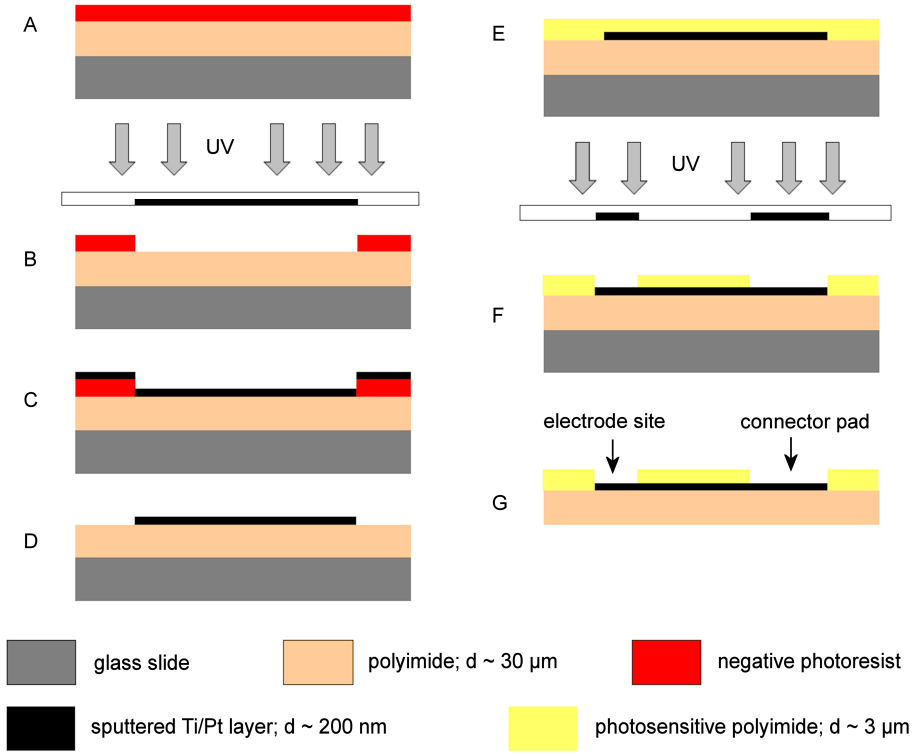


Fig. 2. Fabrication steps of flexible microelectrode arrays. See details in the text

The developed arrays consisted of either 8 or 16 round-shaped Pt-electrodes (round, Ø 100-200 µm) in an area of about 2 mm x 2 mm at the end of polyimide ribbon. Thin film connector pads at one of the edges of polyimide foil were designed to fit into a 16-channel 0.5 mm pitch zero-insertion-force (ZIF) connector (JST Ltd., Halesworth, UK). The ZIF connector was soldered to a thin (0.2 mm) printed circuit board (PCB) adapter. The PCB contains also a 16-channel surface mount microsocket (CLM-serie, Samtec Inc., New Albany, IN, USA) through which the electrical signals from the electrodes are transferred to the preamplifier and further to the recording instrumentation. The prototype of flexible microelectrode array with connector board as well as a schematic view of the array cross-section is shown in Fig. 3.

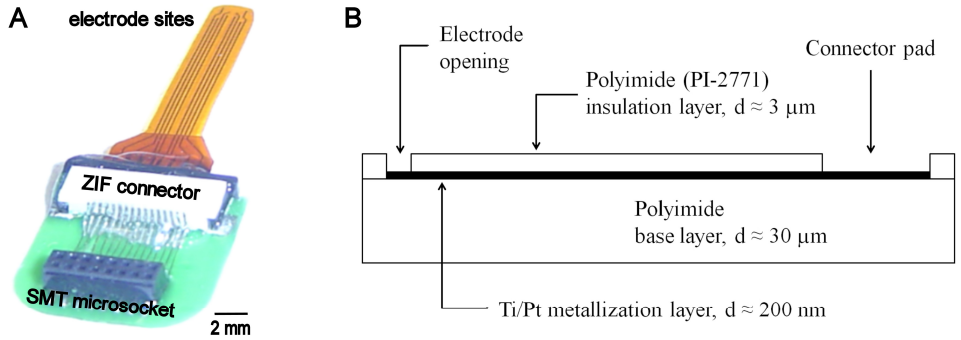


Fig. 3. (A) The developed microelectrode array. The array is connected via a printed circuit board consisting of a ZIF type connector and a SMT microsocket to the recording instrumentation. (B) A schematic illustration of the array cross-section (not to scale, d: thickness)

The use of non-photosensitive polyimide as an insulation layer has been very common in neural interfaces (González & Rodríguez 1997; Cheung et al., 2007; Mercanzini et al., 2008). In contrast, the fabrication process described above utilizes photosensitive polyimide grade that remarkably simplifies the array fabrication i.e. it can be patterned directly by UV exposure and developer chemicals. Thus, no photoresist and/or hard mask as an etch template is needed as is the case with standard non-photosensitive polyimides. Aluminum is a commonly used hard mask material in polyimide dry etching. It is possible that aluminum residuals may remain on the final device after the fabrication process and further come into contact with neural tissue and these ions can be very toxic and cause adverse tissue reactions (Wennberg, 1994). This risk can be eliminated by using photosensitive polyimide. The SEM images (Fig 4) demonstrate that it is possible to pattern the photosensitive polyimide (PI-2771) layer with an almost vertical edge profile without any residuals.

While the excellent biocompatibility of polyimide has been proven in many studies (Richardson et al., 1993; Stieglitz et al., 2000; Seo et al., 2004), these exists only one report focusing on the assessment of cytotoxicy of photosensitive polyimide grade (Sun et al., 2008) involving different chemistry and solvent processes. Therefore, further investigations are

needed to ensure the biocompatibility of photosensitive polyimide grades that there will be no adverse, long term effects and that the material is suitable for chronic implantable use.

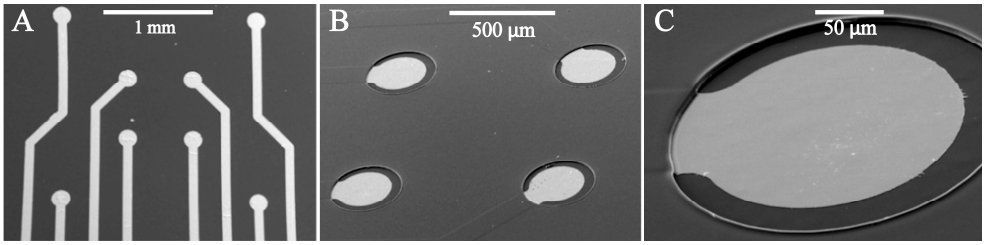


Fig. 4. SEM-image of Ti/Pt metallization layer viewed at the end of recording sites (A) and electrode opening(s) at two magnifications (B-C). A thin insulating layer (photosensitive polyimide) is patterned with alignment of the metallization layer

4.2.2 Electrochemical characterization

Electrochemical impedance spectroscopy (EIS) is a commonly used method to characterize the electrochemical properties of microelectrodes, i.e. to assess the recording capabilities of the electrode for neural recording experiments. In a typical two electrode cell system, the microelectrode is immersed in physiological saline solution (0.9 % NaCl) and a small sinusoidal perturbation voltage (5-50 mV) is applied between the microelectrode and the counter electrode (typically a noble metal, e.g. Pt) having a much larger surface area. A small perturbation voltage is needed to ensure that a linear current-voltage response is obtained at each frequency. Measurements are usually taken over a wide frequency region (e.g. 0.1 Hz – 100 kHz) at room temperature. The induced current and its phase are recorded. Typical impedance spectra, measured with a Solartron 1260 impedance gain/phase analyzer (Solartron Analytical, Farnborough, UK) for Pt and Au microelectrodes with diameter of 200 µm is presented in Fig. 5.

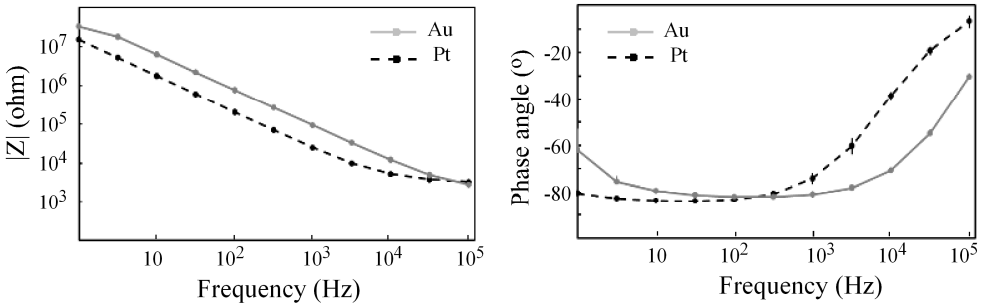


Fig. 5. Impedance magnitude and phase for Pt and Au microelectrodes with a diameter of 200 µm

The EIS results demonstrated that our microelectrodes possessed suitable electrochemical properties for neural recordings. Pt has a much lower impedance level compared to Au especially at low frequencies (below 1 kHz) which are the most relevant in any recordings of cortical signals.

4.2.3 *In vivo* testing

The performance of our microelectrode arrays was tested in acute and chronic recordings in Wistar rats. All animal tests were conducted in accordance with the Council of Europe guidelines and approved by the Institutional Animal Care and Use committee and the State Provincial Office of Eastern Finland.

During the recording sessions, the signals from the electrodes was passed through a preamplifier (Neuralynx Inc., Bozeman, MT, USA) and led into the main amplifier (Grass Instruments, West Warwick, RI, USA) with the data acquisition PC running DataWave SciWorks. Both electrical current and auditory stimuli were used. The square wave current stimuli were of 1 ms duration, and the stimuli of auditory were 10 ms in duration, generated by a WPI (Aston, Stevenage, Herts, UK) isolated current source and led into a piezo buzzer on the front paw of the rat. A stimulus set consisted of paired pulses (a gating paradigm) and the trials were averaged in sets of 25.

Wistar rats were anesthetized with 1.2 – 1.5 g/kg urethane for acute recordings and with a mixture of 0.5 mg/kg medetomidine (Domitor) and 75 mg/kg ketamine (Ketalar) intraperitoneally for implantation for chronic recordings. The rats were placed in a stereotaxic apparatus. Holes for the microelectrode array and for reference electrodes were drilled on the same side of the skull. The microelectrode array was inserted on the dura over the parietal cortex. A stainless steel (SS) screw was used as a reference (ground) electrode placed at A 1.0, L 1.0 with respect to bregma, i.e. 2-3 mm in front of bregma. SS screws also acted as an anchor when the array was cemented onto the skull in the chronic recordings. After anesthesia, the rat was aroused by administration of the antagonist atimepazole (Antisedan), after the surgery. The rats were allowed to recover for a minimum of 7 days before the first experiments. In the chronic recordings, the implanted rat was connected to the recording apparatus and it was either allowed to move freely in the recording box or was gently immobilized by holding it securely in a towel.

The arrays demonstrated excellent flexibility and mechanical strength during handling and implantation onto the surface of rat cortex. During the acute recording session, the microelectrode array was capable of yielding stable readings as observed in the form of the standard response parameters, such as latencies, onsets and decays of the main components in the voltage traces. An example of the somatosensory evoked potential (SEP) recording, obtained 4 hours after the onset of anesthesia is presented in Fig. 6.

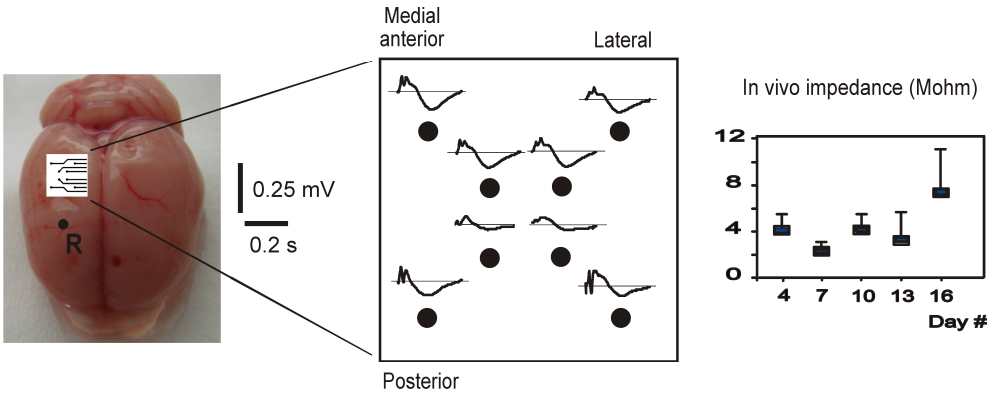


Fig. 6. An example of the *in vivo* somatosensory evoked potential recording in the rat parietal cortex. Profiles of averaged evoked potentials using the 8-channel MEA are shown above the corresponding recording site (black dots). The current stimuli were delivered to the left paw of the rat. The position of MEA and reference screw electrode ("R") is depicted on the brain of a rat after the experiment and its surgical removal. Right graph: average electrode impedance (mean and standard deviation, n = 8) measured *in vivo* with three-day intervals. In chronic recordings, the electrode was capable of yielding signals comparable with the presented acute signals for approximately two weeks after array implantation. Note the impedance jump on day 16, coinciding with a decay in the responses (not shown)

The SEP recordings obtained with our array are comparable in their signal-to-noise ratios with previously published recordings with various electrode materials, such as platinum (Hayton et al., 1999), silver (Kalliomäki et al., 1998) and tungsten (Ureshi et al., 2004). Recordings are also comparable to previous studies with microelectrode arrays (Stett et al., 2003). In chronic recordings, the electrode was capable of capturing a biologically meaningful signal (comparable to signals in the acute experiment) for approximately two weeks after array implantation. However, after 16 days, the responses decayed due to a variety of mechanical and technical reasons. The signal amplitude decreased at the same time as the electrode impedance doubled (Fig. 6). Although no macroscopic scar formation was seen in the examination of the brain after the array was removed, this kind of increase in *in vivo* impedance is compatible with a thickening of the dura, a common finding in chronic experiments, and microscopic growth of non-conductive fibrous tissue around the electrode, shrinking the freely exposed surface area of the recording site.

Our current research is concentrating on improving biocompatibility issues, the array construction and implantation techniques in order to prolong the functional lifetime. The goals of the work include development of nanostructured electrode surfaces by utilizing ultra short pulsed laser deposition (USPLD) techniques or electrolytic deposition methods to improve the electrochemical functionality of electrodes as well as to incorporate functional coatings and/or biological compounds to polyimide layer to reduce the growth of scar tissue in the vicinity of array.

5. Biocompatibility testing

Despite the very challenging microfabrication processes used in electrode fabrication, electrode failures due technical reasons (short-circuits, broken signal transmission lines etc.) are rare, and the stable, long-term functioning of neural implants is mainly determined by the tissue response to chronically implanted electrodes. A general immune activation of the brain in response to the presence of a foreign body implant has been commonly believed to be the main reason for signal deterioration of chronically implanted electrodes (Nicolelis et al., 2003; Edell et al., 1992; Schmidt et al., 1993). Biocompatibility aspects of neural implant materials will therefore be very crucial in the development of new generation chronic implants. In this section, we summarize some *in vivo* implantation and *in vitro* cell culture studies, concentrating particularly on these potential flexible thin film electrode materials that were already discussed in this chapter, i.e. Pt, Au, Ir, polyimide, SU-8, PDMS etc.

5.1 *In vivo* implantation studies

The tissue response to neural implant materials has been widely reported in the literature. Stensaas and Stensaas implanted 27 different materials, consisting of metals and insulators, into the cortices of rabbits. After 30 days, the histological examination revealed no adverse reaction in response to gold, platinum and tungsten (Stensaas & Stensaas, 1978). Moreover, subdural implantations of two common insulators, parylene-C and polyimide (PI-2555), into the cerebral cortex of the cat evoked only minimal tissue reactions after a 16 week implantation period. All neurons situated beneath the implant site appeared to be normal. Tissue reactions to polyimide and parylene-C were comparable to pure platinum (negative control) whereas Ag-AgCl (positive control) had triggered chronic inflammatory reactions (Yuen et al., 1987). In another study, polyesterimide-coated gold wires implanted subdurally on rabbit cortex showed no evidence of toxicity. There was only slight gliosis, and there were no inflammatory reactions present at 16 weeks after implantation (Yuen & Agnew, 1995). On the other hand, polyimide-platinum electrodes implanted to the rat scientic nerve for 3 months were demonstrated to induce a mild scar response and local inflammation reactions, though these were limited to a small area around the electrode (Lago et al., 2007).
Unfortunately all artificial materials implanted in the CNS have an inconvenient tendency to induce significant glial scar tissue formation (Polikov et al., 2005). This gliotic scar tissue is attributable mainly to glial cells such as astrocytes and microglia and it poses one of the greatest challenges in the field of neural prosthetics. The formation of scar tissue can cause serious impairment of implant performance due to decreased local density of neurons and the formation of an encapsulation layer that increases electrode impedance and lowers the signal amplitudes. Astrocytes that account for 30-65 % of glial cell population in the CNS are able to secrete 8-10 nm diameter intermediate filaments of polymerized glial fibrillary acid protein (GFAP). When the electrode is implanted into the CNS, the astrocytes become activated and they transform into reactive phenotype which produces much larger size GFAP filaments and there is also enhanced cell proliferation and migration capacity. GFAP is the most commonly used astrocyte specific cell marker with which to assess the level of scar formation after electrode implantation. The brain slices of rats used in the chronic experiments are in cultured with anti-GFAP staining antibodies. GFAP- positive cells are counted around the implantation area of the cortex in order to estimate the extent of astrocyte activation and gliosis. In previous studies with probe-type electrodes, it has been

shown that the astrocyte response to brain injury can be divided into two phases, the early response that occurs immediately after electrode implantation and the long-term chronic response. It has been reported that the number of astrocytes and microglia is significantly increased in the area surrounding the silicon probes (Szarowski et al., 2003; Turner et al., 1999) within a few hours after insertion. The extent of this early response is dependent on probe size, shape and surface roughness (Szarowski et al., 2003). The long-term response, starting approximately one week after electrode implantation (Norton et al., 1992), consists of a compact glial scar tissue formation surrounding the electrode and which ultimately isolates the microelectrodes from neurons/neural tissue elevating the impedance and causing signal deterioration (Nicolelis et al., 2003; Edell et al., 1992; Schmidt et al., 1993).

5.2 *In vitro* cytotoxicity

In vitro cytotoxicity tests play an important role in evaluating the biocompatibility properties of potential novel materials. The international standard "Biological evaluation of medical devices: tests for *in vitro* cytotoxicity" (ISO 10993-5) describes the test methods to assess the *in vitro* cytotoxicity of implant materials and medical devices. The mammalian cell line is selected according to the intended application. Cells are seeded in contact with a tested material and/or with an extract of a test material and different parameters are evaluated such as cell adherence, adhesion, proliferation, morphological changes and metabolism changes.

Kotzar et al. (2002) performed an *in vitro* cytotoxicity evaluation of a wide variety of materials used in microelectro-mechanical systems (MEMS) including Si, thermal oxide, n-doped polysilicon, Si_3N_4, Ti, SU-8 and silicon carbide. Minimal cytotoxicity was demonstrated for these materials.

Polyimides have been shown to be non-cytotoxic in many *in vitro* studies (Richardson et al., 1993; Stieglitz et al., 2000). Stieglitz et al. (2000) studied the cytotoxicity of three commercial polyimide grades (Pyralin PI 2611, PI 2556, PI 2566, HD Microsystems) and reported excellent biocompatibility for PI 2611 and PI 2556 and good results for PI 2566. This last polyimide, since it is fluorinated, differs from the others with respect to its chemical structure. Furthermore, Lee et al. (2004b) reported that fibroblast cells attached, spread out and grew on polyimide surfaces in a corresponding manner to the behaviour of control cells growing on the surface of polystyrene. However the previously tested polyimides are mainly non-photosensitive. In a recent study, Sun et al. (2008) assessed the biocompatibility of a photosensitive polyimide for use in an implantable medical device. They concluded that the photosensitive polyimide (Fujifilm Durimide 7020) was also noncytotoxic and the fibroblast (L929) cell adhesion, morphology, and spreading was even enhanced on the photosensitive grade than one non-photosensitive grade (HD Microsystem PI-2611).

Figure 7 presents fibroblast cells (BHK-21) cultured on polyimide (Pyralin PI-2525, HD Microsystems GmbH), high-density polyethylene (negative control) and latex rubber (positive control) after 24 h incubation period. Only a few cells are seen on cytotoxic latex rubber whereas the cell density on polyimide is even higher than on PE demonstrating the excellent biocompatibility of this polyimide grade.

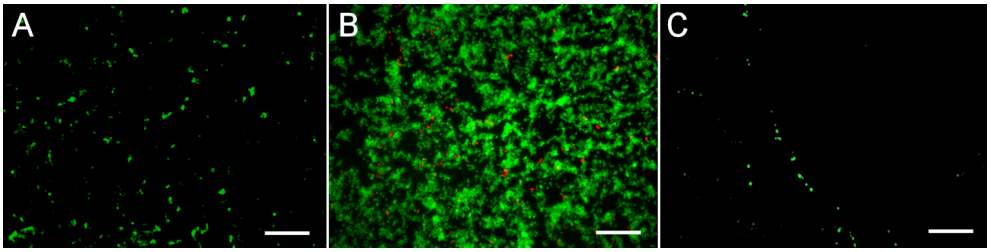


Fig. 7. Live/dead staining of BHK-21 fibroblast cells cultured on the surfaces of (A) polyethylene (negative control), (B) polyimide (C) latex rubber (positive control) after a 24 hour incubation period. Confocal laser microscope image, viable cells: green, dead cells: red, scale bar = 200 μm

The optimal electrode site should support neuronal cell attachment and growth. Thanawala et al. (2007) studied the biocompatibility of metal thin films *in vitro* cell culture studies. The results showed that iridium oxide and platinum thin films are biocompatible and non-toxic for neural cell (cortical rat neurons) growth. Furthermore, Pt films were reported to be superior to iridium oxide films in their ability to support cell attachment.

6. Future trends and strategies to improve the long-term performance of implanted electrodes

Surface modifications of materials intended for implantation are needed in order to improve the functionality and biocompatibility of implanted electrodes, i.e. to reduce the formation of the glial scar, prevent inflammatory reactions and extend the functional lifetime of the implanted devices, and this is an area in which there is intensive worldwide research. It must be noted that there are different requirements for different parts of a neural implant. Recording electrode sites should remain bare to achieve good signal output whereas encapsulation materials should be allowed to be covered by tissue cells like fibroblasts in order to allow them to co-exist in the human body. In this section, we will concentrate on future trends and strategies to improve the long-term performance of neural implants by dividing them into two categories: material science and bioactive molecule strategies.

6.1 Material science strategies

It is well-known that the there is an inverse relationship between electrode impedance and its surface area. Therefore, it is advantageous to develop microelectrodes with high nanoscale surface topography since in this way it will be possible to achieved high effective surface area without increasing the geometrical surface area. Electrodeposition of platinum (Pt black) is a classic technique which can increase surface roughness and lower electrode impedance (de Haro et al., 2002). Other common surface modification techniques used to increase nanoscale roughness are wet etching of gold and activating of iridium to form a nanoporous iridium oxide layer. Different etching and machining techniques (e.g. reactive ion etching, laser ablation etc.) can be also used to form different micro/nanostructures, e.g. grooves and wells onto the surface of the substrate upon which the electrode material can then be deposited using traditional thin film techniques. On the other hand, USPLD (Phipps, 2007; Eason, 2007) can be used both to deposit thin films and to create a nanostructured

surface in a well-controlled and straightforward manner without the need for several process steps. An example of nanostructured Pt surface from our experiments is shown in Fig. 8. The size of the features is about 30 nm in this case. Furthermore, it is possible to combine micro/nanomachining using ultra short pulsed laser ablation and deposition to create well controlled and high quality surface textures like grooves, wells, holes etc.

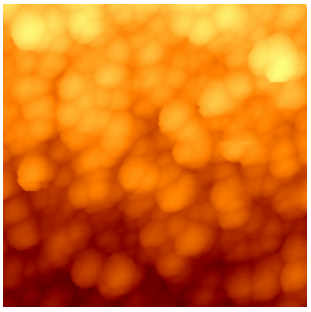


Fig. 8. AFM image of nanostructured Pt surface deposited using the USPLD technique. The average size of Pt particles is 30 nm

It has been also shown that electrode surface morphology has a strong influence on neurocompatibility. Enhanced functions of neurons have been demonstrated on biomaterials structured at the nanoscale (Bayliss et al., 1999; Raffa et al., 2007). Raffa et al. (2007) used focused ion beam technology as a nanometric precision machining technique to modify the surface morphology of tungsten, and studied the effect of the nanofeatures on adherence and adhesion of PC12 neural cells. They concluded that the ability of PC12 cells to adhere to the surface is strongly linked to the dimensions of nanofeatures and the feature size of 25 nm is superior than 270 nm material. Positive interactions between neurons and nanostructured silicon surfaces with 10 nm pores have been also shown (Bayliss et al., 1999). However, there is one report of decreased astrocyte adhesion and increased extension of neuritis on nanoporous silicon surfaces (Moxon et al., 2004). Decreased adhesion, reduced proliferation and less long-term functional activation of astrocytes on carbon nanofiber materials (fiber diameter 60 nm) compared to conventional carbon fibres (diameters in range of 125-200 nm) have also been demonstrated (McKenzie et al., 2004). These results provide evidence that nanoscale materials may have a promising future, i.e. they increase effective surface area of electrode and interact positively with neurons but at the same time they reduce the functions of astrocytes leading to decreased glial scar tissue formation.
As discussed before, conventional lithographic techniques can be combined with thin film techniques like evaporation, sputtering or electrodeposition. In spite of their widespread use, for example in the electronics industry, they have some shortcomings in electrode preparation compared to some novel techniques such as USPLD. In this study, we utilized Coldab™ USPLD configurations developed by Picodeon Ltd, Finland. This technique (Amberla et al., 2006) has several advantages in thin film preparation such as high adhesion, smooth or nanostructured surface topography and the possibility to deposit films or particles on heat sensitive substrates. Furthermore, different kinds of materials (polymers, metals or ceramics) can be used as a source to prepare single material films or composites, i.e. multilayer structures or homogeneous or graded composite films. High purity source

materials available in a solid form can be utilized and transferred from the target to the sample in a stoichiometric manner. Due to the significant increase in power produced by picosecond pulsed lasers during recent years, this deposition technique has acquired commercial state, i.e. it can be used for deposition on large surfaces (in dimensions of ten centimetres), not just on small laboratory test samples. Thus USPLD technique holds an evident potential for neural electrode development.

The USPLD is a very effective method to convert a material so it can be deposited as a plasma, atom or molecular beam to produce bioinert, biocompatible, bioactive or bioresorbable films without generating extra heat on the sample surface (Eason, 2007). Therefore, the sample remains normally very close to room temperature and it is possible to deposit thin films on heat sensitive materials like polymers, which can be utilized in flexible electrodes or as a photoresist. Samples containing polymeric materials can be kept at low temperatures during the deposition and thus avoid hardening or flow during deposition. This leads to smooth high quality films and easy lift-off in the case of lithography. Since the target materials remain almost at room temperature, no water cooling is needed which is convenient since different target materials and dimensions can be utilized. Expensive target materials like Pt, Au, Ir or ITO can be used effectively and they remain free from contamination. Since one has many possible material choices, the material combinations can be optimized with respect to signal stability, biocompatibility, cell attachment, antibacterial properties etc. Enhanced surface properties can be achieved by including appropriate surface modifying layers: these can be organic compounds like collagen, fibronectin or even cells.

If USPLD is used in the creation of deposition electrodes, one important advantage compared to several other methods, is that the growing film does not contain microparticles or droplets, which would cause pinholes in the films. This is due to the fact that the laser ablation pulse is so short that the energy penetrates to only very shallow (tens of nanometers) surface layer which is effectively converted to plasma and the target surface remains smooth without any flows or deteriorations. The quality of the film becomes an important issue, for example in the case of insulator films for neural sensors. Based on our tests, these films do not seem to contain any microdefects which could lead to short circuiting especially in liquids. In addition to ceramic insulators like alumina (Al_2O_3), polymers can be deposited with the same technique. The ultrasmoothness of deposited coatings might provide an additional benefit to avoid scar tissue formation, inflammation reactions or activation of glial cells.

6.2 Strategies involving anti-inflammatory agents and bioactive molecules

Several strategies are currently being studied to discover effective methods to minimize the immune response and inflammation around the neural implant as well as to encourage neural growth at electrode sites (Polikov et al., 2005; Zhong & Bellamkonda, 2008). Anti-inflammatory agents, such as dexamethasone (DEX) can be used to minimize the release of inflammatory mediators and to attenuate the formation of a fibrous capsule. DEX and other pharmaceutical agents can be administered in several way e.g. subcutaneous injection (Shain et al., 2003; Spataro et al., 2005) and the local release from electrode coatings of the drugs as incorporated agents (Zhong & Bellamkonda, 2007; Kim & Martin 2006). It is believed that the local delivery of DEX reduces the inflammatory tissue response and prevents the elevation in electrode impedance more effectively compared to peripheral

injection since in the latter case there may be difficulties in the drugs crossing the blood-brain barrier and chronic use of corticosteroids can evoke several side effects, such as myopathy and diabetes (Zhong & Bellamkonda, 2007). Additionally, Si-based microelectrode arrays have been developed with integrated microfluidic drug delivery channels through which the bioactive molecule was released (Retterer et al., 2004; Rathnasingham et al., 2004). In addition, flexible multi-channel polyimide recording probes with incorporated small pores and wells that can be selectively filled with a dextran hydrogel and bioactive species such as nerve growth factor (NGF) (Rousche et al., 2001) as well as devices containing microfluidic channels for drug delivery (Metz et al., 2004) have been developed.

Several bioactive molecules are being investigated as potential binding molecules to promote the adherence and functions of neurons at the electrode site. Instead of using intact cell adhesion proteins such as collagen and fibronectin (Ignatius et al., 1998), new research efforts are concentrated on functionalizing the electrode surfaces with relevant protein sequences such as RGD, YIGSR and IKVAV (Polikov et al., 2005; Kam et al., 2002). It has been demonstrated that functionalized surfaces with these kinds of protein sequences can be used to support neuronal growth or repel glial growth. Furthermore, peptides can be deposited in patterns by using soft lithographic (Xia & Whitesides, 1998) techniques intended to guide cellular outgrowth.

7. Conclusion

Modern microfabrication techniques offer great opportunities in the development of novel implantable neural interfaces for recording and stimulating purposes. Miniaturized electrode arrays with high-channel densities are now in common use in animal models, but their human counterparts are still mainly being investigated or at best undergoing clinical trials. Silicon has been the mostly utilized substrate material despite its rigidity, but recently there has been a growing interest toward polymeric implant materials that could be flexible enough to mimic biological tissue and to reduce adverse tissue reactions. Despite the very challenging microfabrication processes used in electrode fabrication, electrode failures due technical reasons are rare, and the stable, long-term functioning of neural implants is mainly determined by the tissue response to chronically implanted electrodes. Therefore, current investigations are now concentrating on the evaluation of the tissue-electrode interface aiming to improve the functionality and biocompatibility of implanted electrodes, i.e. to reduce the formation of the glial scar, prevent inflammatory reactions and extend the functional lifetime of the implanted microdevices. Further development of implantable neural electrodes promises to deliver novel options to acquire more detailed information of brain functions, to plan more precise surgical operations as well as improving the quality of life for many patients suffering pain or from loss of neural functions.

8. Acknowlegdements

This study was supported by the PhD-programme in Musculoskeletal Diseases and Biomaterials and the Otto A. Malm Foundation. The authors would like to thank the contribution of the following people: Prof. Heikki Tanila and Dr. Kaj Djupsund (*in vivo* experiments), Dr. Virpi Tiitu and Sanna Miettinen (cell culture studies), Dr. Markku Tiitta

(EIS studies), Hannu Korhonen (sputter depositions) and Dr. Ewen MacDonald (language editing). The Microsensor Laboratory of the Savonia University of Applied Sciences is acknowledged for providing photolithographic and AFM imaging facilities. Picodeon Ltd is acknowledged for providing Coldab™ depositions.

9. References

Amberla, T.; Rekow, M.; Köngäs, J.; Asonen, H.; Salminen, T.; Viitanen, N.; Kulmala, M.; Vuoristo, P.; Pessa, M. & Lappalainen, R. (2006). Pulsed laser deposition with a high average power fiber laser. *Proceedings of 49th Annual Technical Conference* (Society of Vacuum Coaters), pp. 79-82, Washington D.C., USA, April 22-27, 2006.

Armani, D.; Liu, C. & Aluru, N. (1999). Re-configurable fluid circuits by PDMS elastomer micromachining. *IEEE International Conference on Micro Electro Mechanical Systems,* pp. 222–227, January 17–21, 1999.

Bayliss, S. C.; Buckberry, L. D.; Fletcher, I. & Tobin, M. J. (1999). The culture of neurons on silicon. *Sensors and Actuators A: Physical*, Vol. 74, No. 1-3, (1999), pp. 139-142.

Bogner, E.; Dominizi, K.; Hagl, P.; Bertagnolli, E.; Wirth, M.; Gabor, F.; Brezna, W. & Wanzenboeck, H. D. (2006). Bridging the gap—Biocompatibility of microelectronic materials. *Acta Biomaterialia*, Vol. 2, No. 2, (2006), pp. 229-237.

Boppart, S. A.; Wheeler, B. C. & Wallace, C. S. (1992). A flexible perforated microelectrode array for extended neural recordings. *IEEE Transactions on Biomedical Engineering*, Vol. 39, No. 1, (1992), pp. 37-42.

Campbell, P. K.; Jones, K. E.; Huber, R. J.; Horch, K.W. & Normann, R. A. (1991). A silicon-based, three-dimensional neural interface: manufacturing processes for an intracortical electrode array. *IEEE Transactions on Biomedical Engineering*, Vol. 38, No. 8, (1991), pp. 758-768.

Cheung, K. C. (2007). Implantable microscale neural interfaces. *Biomedical Microdevices*, Vol. 9, No. 6, (2007), pp. 923-938.

Cheung, K. C.; Renaud, P.; Tanila, H. & Djupsund, K. (2007). Flexible polyimide microelectrode array for in vivo recordings and current source density analysis. *Biosensors and Bioelectronics*, Vol. 22, No. 8, (2007), pp. 1783-1790.

de Haro, C.; Mas, R.; Abadal, G.; Muñoz, J.; Perez-Murano, F. & Domínguez, C. (2002). Electrochemical platinum coatings for improving performance of implantable microelectrode arrays. *Biomaterials*, Vol. 23, No. 23, (2002), pp. 4515-4521.

Eason, R. (2007). *Pulsed Laser Deposition of Thin Films*: *Applications-Led Growth of Functional Materials*, Wiley-Interscience, ISBN: 978-0-470-05211-2, New Jersey.

Edell, D. J.; Toi, V. V.; McNeil, V. M. & Clark, L. D. (1992). Factors influencing the biocompatibility of insertable silicon microshafts in cerebral cortex. *IEEE Transactions on Biomedical Engineering*, Vol. 39, No. 6, (1992), pp. 635-643.

Geddes, L. A. & Roeder, R. (2003). Criteria for the selection of materials for implanted electrodes. *Annals of Biomedical Engineering*, Vol. 31, No. 7, (2003), pp. 879-890.

González, C. & Rodríguez, M. (1997). A flexible perforated microelectrode array probe for action potential recording in nerve and muscle tissues. *Journal of Neuroscience Methods*, Vol. 72, No. 2, (1997), pp. 189-195.

Hayton, S. M.; Kriss, A. & Muller, D. P. (1999). Comparison of the effects of four anaesthetic agents on somatosensory evoked potentials in the rat. *Laboratory Animals*, Vol. 33, No. 3, (1999), pp. 243-251.

Hollenberg, B. A.; Richards, C. D.; Richards, R.; Bahr, D. F. & Rector, D. M. (2006). A MEMS fabricated flexible electrode array for recording surface field potentials. *Journal of Neuroscience Methods*, Vol. 153, No. 1, (2006), pp. 147-153.

Hosp, J.A.; Molina-Luna, K.; Hertler, B.; Atiemo, C.O.; Stett, A. & Luft, A.R. (2008). Thin-film epidural microelectrode arrays for somatosensory and motor cortex mapping in rat. *Journal of Neuroscience Methods*, Vol. 172, No. 2, (2008), pp. 255-262.

Ignatius, M. J.; Sawhney, N.; Gupta, A.; Thibadeau, B. M.; Monteiro, O. R. & Brown, I. G. (1998). Bioactive surface coatings for nanoscale instruments: effects on CNS neurons. *Journal of Biomedical Materials Research*, Vol. 40, No. 2, (1998), pp. 264-274.

Kalliomäki, J.; Luo, X.-L.; Yu, Y.-B. & Schouenborg, J. (1998). Intrathecally applied morphine inhibits nociceptive C fiber input to the primary somatosensory cortex (SI) of the rat. *Pain*, Vol. 77, No. 3, (1998), pp. 323-329.

Kam, L.; Shain, W.; Turner, J. N. & Bizios, R. (2002). Selective adhesion of astrocytes to surfaces modified with immobilized peptides. *Biomaterials*, Vol. 23, No. 2, (2002), pp. 511-515.

Kim, D.-H. & Martin, D. C. (2006). Sustained release of dexamethasone from hydrophilic matrices using PLGA nanoparticles for neural drug delivery. *Biomaterials*, Vol. 27, No. 15, (2006), pp. 3031-3037.

Kitzmiller, J.; Beversdorf, D. & Hansford, D. (2006). Fabrication and testing of microelectrodes for small-field cortical surface recordings. *Biomedical Microdevices*, Vol. 8, No. 1, (2006), pp. 81-85.

Kotzar, G.; Freas, M.; Abel, P.; Fleischman, A.; Roy, S.; Zorman, C.; Moran, J. M. & Melzak, J. (2002). Evaluation of MEMS materials of construction for implantable medical devices. *Biomaterials*, Vol. 23, No. 13, (2002), pp. 2737-2750.

Lago, N.; Yoshida, K.; Koch, K. P. & Navarro, X. (2007). Assessment of biocompatibility of chronically implanted polyimide and platinum intrafascicular electrodes. *IEEE Transactions on Biomedical Engineering*, Vol. 54, No. 2, (2007), pp. 281-290.

Lee, K.; He, J.; Clement, R.; Massia, S. & Kim, B. (2004). Biocompatible benzocyclobutene (BCB)-based neural implants with micro-fluidic channel. *Biosensors and Bioelectronics*, Vol. 20, No. 2, (2004), pp. 404-407, (a).

Lee, K.; Singh, A.; He, J.; Massia, S.; Kim, B. & Raupp, G. (2004). Polyimide based neural implants with stiffness improvement. *Sensors and Actuators B: Chemical*, Vol. 102, No. 1, (2004), pp. 67-72, (b).

Mata, A.; Fleischman, A. J. & Roy, S. (2005). Characterization of polydimethylsiloxane (PDMS) properties for biomedical micro/nanosystems. *Biomedical Microdevices,* Vol. 7, No. 4, (2005), pp. 281-293.

McKenzie, J. L.; Waid, M. C.; Shi, R. & Webster, T. J. (2004). Decreased functions of astrocytes on carbon nanofiber materials. *Biomaterials*, Vol. 25, No. 7-8, (2004), pp. 1309-1317.

Mercanzini, A.; Cheung, K.; Buhl, D. L.; Boers, M.; Maillard, A.; Colin, P.; Bensadoun, J.-C.; Bertsch, A. & Renaud, P. (2008). Demonstration of cortical recording using novel flexible polymer neural probes. *Sensors and Actuators A: Physical,* Vol. 143, No. 1, (2008), pp. 90-96.

Metz, S.; Bertsch, A.; Bertrand, D. & Renaud, P. (2004). Flexible polyimide probes with microelectrodes and embedded microfluidic channels for simultaneous drug delivery and multi-channel monitoring of bioelectric activity. *Biosensors and Bioelectronics*, Vol. 19, No. 10, (2004), pp. 1309-1318.

Molina-Luna, K.; Buitrago, M. M.; Hertler, B.; Schubring, M.; Haiss, F.; Nisch, W.; Schulz, J. B. & Luft, A. R. (2007). Cortical stimulation mapping using epidurally implanted thin-film microelectrode arrays. *Journal of Neuroscience Methods*, Vol. 161, No. 1, (2007), pp. 118-125.

Moxon, K. A.; Kalkhoran, N. M.; Markert, M.; Sambito, M. A.; McKenzie, J. L. & Webster, J. T. (2004). Nanostructured surface modification of ceramic-based microelectrodes to enhance biocompatibility for a direct brain-machine interface. *IEEE Transactions on Biomedical Engineering*, Vol. 51, No. 6, (2004), pp. 881-889

Myllymaa, S.; Myllymaa, K.; Korhonen, H.; Djupsund, K.; Tanila, H. & Lappalainen, R. (2008). Development of flexible thin film microelectrode arrays for neural recordings. *Proceedings of International Federation of Medical and Biological Engineering (IFMBE)*, pp. 286-289, Riga, Latvia, June 2008, Springer, Berlin, (a).

Myllymaa, S.; Myllymaa, K.; Korhonen, H.; Gureviciene, I.; Djupsund, K.; Tanila, H. & Lappalainen, R. (2008). Development of flexible microelectrode arrays for recording cortical surface field potentials. *Proceedings of IEEE Engineering in Medicine and Biology Society*, pp. 3200-3203, Vancouver, British Columbia, Canada, August 2008, (b).

Myllymaa, S.; Myllymaa, K.; Korhonen, H.; Töyräs, J.; Jääskeläinen, J. E.; Djupsund, K.; Tanila, H. & Lappalainen, R. (2009). Fabrication and testing of polyimide-based microelectrode arrays for cortical mapping of evoked potentials. *Biosensors and Bioelectronics*, Vol. 24, No. 10, (2009), pp. 3067-3072.

Nair, D. R.; Burgess, R.; McIntyre, C. C. & Lüders, H. (2008). Chronic subdural electrodes in the management of epilepsy. *Clinical Neurophysiology*, Vol. 119, No. 1, (2008), pp. 11-28.

Najafi, K. & Wise, K. D. (1986). Implantable multielectrode array with on-chip signal processing. *IEEE Journal of Solid-State Circuits*, Vol. 21, No. 6, (1986), pp. 1035-1044.

Nicolelis, M. A. L.; Ghazanfar, A. A.; Faggin, B. M.; Votaw, S. & Oliveira, L. M. O. (1997). Reconstructing the engram: simultaneous, multisite, many single neuron recordings. *Neuron,* Vol. 18, No. 4, (1997), pp. 529-537.

Nicolelis, M. A. L. ; Dimitrov, D.; Carmena, J. M.; Crist, R.; Lehew, G.; Kralik, J. D. & Wise, S. P. (2003). Chronic, multisite, multielectrode recordings in macaque monkeys. *Proceedings of the National Academy of Sciences*, Vol. 100, No. 19, (2003), pp. 11041-11046.

Norton, W. T.; Aquino, D. A.; Hozumi, I.; Chiu, F. C. & Brosnan, C. F. (1992). Quantitative aspects of reactive gliosis: a review. *Neurochemical research*, Vol. 17, No. 9, (1992), pp. 877-885.

Owens, A. L.; Denison, T. J.; Versnel, H.; Rebbert, M.; Peckerar, M. & Shamma, S. A. (1995). Multi-electrode array for measuring evoked potentials from surface of ferret primary auditory cortex. *Journal of Neuroscience Methods*, Vol. 58, No. 1-2, (1995), pp. 209-220.

Phipps, C. (2007). *Laser Ablation and Its Applications*, Springer, ISBN: 978-0-387-30452-6, New York.

Polikov, V. S.; Tresco, P. A. & Reichert, W. M. (2005). Response of brain tissue to chronically implanted neural electrodes. *Journal of Neuroscience Methods*, Vol. 148, No. 1, (2005), pp. 1-18.

Raffa, V.; Pensabene, V.; Menciassi, A. & Dario, P. (2007). Design criteria of neuron/electrode interface. The focused ion beam technology as an analytical method to investigate the effect of electrode surface morphology on neurocompatibility. *Biomedical Microdevices*, Vol. 9, No. 3, (2007), pp. 371-383.

Rasche, D.; Ruppolt, M.; Stippich, C.; Unterberg, A. & Tronnier, V. M. (2006). Motor cortex stimulation for long-term relief of chronic neuropathic pain: a 10 year experience. *Pain*, Vol. 121, No. 1-2, (2006), pp. 43-52.

Rathnasingham, R.; Kipke, D. R.; Bledsoe, S. C., Jr. & McLaren, J. D. (2004). Characterization of implantable microfabricated fluid delivery devices. *IEEE Transactions on Biomedical Engineering*, Vol. 51, No. 1, (2004), pp. 138-145.

Retterer, S. T.; Smith, K. L.; Bjornsson, C. S.; Neeves, K. B.; Spence, A. J. H.; Turner, J. N.; Shain, W. & Isaacson, M. S. (2004). Model neural prostheses with integrated microfluidics: a potential intervention strategy for controlling reactive cell and tissue responses. *IEEE Transactions on Biomedical Engineering*, Vol. 51, No. 11, (2004), pp. 2063-2073.

Richardson, R. R.; Miller, J. A. & Reichert, W. M. (1993). Polyimides as biomaterials: preliminary biocompatibility testing. *Biomaterials*, Vol. 14, No. 8, (1993), pp. 627-635.

Rodríguez, F. J.; Ceballos, D.; Schüttler, M.; Valero, A.; Valderrama, E.; Stieglitz, T. & Navarro, X. (2000). Polyimide cuff electrodes for peripheral nerve stimulation. *Journal of Neuroscience Methods*, Vol. 98, No. 2, (2000), pp. 105-118.

Rousche, P. J.; Pellinen, D. S.; Pivin, D. P. Jr.; Williams, J. C.; Vetter, R. J.; & Kipke, D. R. (2001). Flexible polyimide-based intracortical electrode arrays with bioactive capability. *IEEE Transactions on Biomedical Engineering*, Vol. 48, No. 3, (2001), pp. 361-371.

Schmidt, S.; Horch, K. & Normann, R. (1993). Biocompatibility of silicon-based electrode arrays implanted in feline cortical tissue. *Journal of Biomedical Materials Research*, Vol. 27, No. 11, (1993), pp. 1393-1399.

Selvakumaran, J.; Keddie, J. L.; Ewins, D. J. & Hughes, M. P. (2008). Protein adsorption on materials for recording sites on implantable microelectrodes. *Journal of Materials Science: Materials in Medicine*, Vol. 19, No. 1, (2008), pp. 143-151.

Seo, J.-M.; Kim, S. J.; Chung, H.; Kim, E. T.; Yu, H. G. & Yu, Y. S. (2004). Biocompatibility of polyimide microelectrode array for retinal stimulation. *Materials Science and Engineering: C*, Vol. 24, No. 1-2, (2004), pp. 185-189.

Shain, W.; Spataro, L.; Dilgen, J.; Haverstick, K.; Retterer, S.; Isaacson, M.; Saltzman, M. & Turner, J. N. (2003). Controlling cellular reactive responses around neural prosthetic devices using peripheral and local intervention strategies. *IEEE transactions on neural systems and rehabilitation engineering*, Vol. 11, No. 2, (2003), pp. 186-188.

Spataro, L.; Dilgen, J.; Retterer, S.; Spence, A. J.; Isaacson, M.; Turner, J. N. & Shain, W. (2005). Dexamethasone treatment reduces astroglia responses to inserted neuroprosthetic devices in rat neocortex. *Experimental neurology*, Vol. 194, No. 2, (2005), pp. 289-300.

Spence, A. J.; Neeves, K. B.; Murphy, D.; Sponberg, S.; Land, B. R.; Hoy, R. R. & Isaacson, M. S. (2007). Flexible multielectrodes can resolve multiple muscles in an insect appendage. *Journal of Neuroscience Methods*, Vol. 159, No. 1, (2007), pp. 116-124.

Stensaas, S. S. & Stensaas, L. J. (1978). Histopathological evaluation of materials implanted in the cerebral cortex. *Acta Neuropathologica*, Vol. 41, No. 2, (1978), pp. 145-155.

Stett, A.; Egert, U.; Guenther, E.; Hofmann, F.; Meyer, T.; Nisch, W. & Haemmerle, H. (2003). Biological application of microelectrode arrays in drug discovery and basic research. *Analytical and Bioanalytical Chemistry*, Vol. 377, No. 3, (2003), pp. 486-495.

Stieglitz, T.; Beutel, H. & Meyer, J.-U. (1997). A flexible, light-weight multichannel sieve electrode with integrated cables for interfacing regenerating peripheral nerves. *Sensors and Actuators A: Physical*, Vol. 60, No. 1-3, (1997), pp. 240-243.

Stieglitz, T.; Beutel, H.; Schuettler, M. & Meyer, J.-U. (2000). Micromachined, polyimide-based devices for flexible neural interfaces. *Biomedical Microdevices*, Vol. 2, No. 4, (2000), pp. 283-294.

Stieglitz, T. (2001). Flexible biomedical microdevices with double-sided electrode arrangements for neural applications. *Sensors and Actuators A: Physical*, Vol. 90, No. 3, (2001), pp. 203-211.

Strumwasser, F. (1958). Long-term recording' from single neurons in brain of unrestrained mammals. *Science*, Vol. 127, No. 3296, (1958), pp. 469-470.

Sun, Y.; Lacour, S. P.; Brooks, R. A.; Rushton, N.; Fawcett, J. & Cameron, R. E. (2008). Assessment of the biocompatibility of photosensitive polyimide for implantable medical device use. *Journal of Biomedical Materials Research Part A*, (2008).

Szarowski, D. H.; Andersen, M. D.; Retterer, S.; Spence, A. J.; Isaacson, M.; Craighead, H. G.; Turner, J. N. & Shain, W. (2003). Brain responses to micro-machined silicon devices. *Brain Research*, Vol. 983, No. 1-2, (2003), pp. 23-35.

Takahashi, H.; Ejiri, T.; Nakao, M.; Nakamura, N.; Kaga, K. & Herve, T. (2003). Microelectrode array on folding polyimide ribbon for epidural mapping of functional evoked potentials. *IEEE Transactions on Biomedical Engineering*, Vol. 50, No. 4, (2003), pp. 510-516.

Takeuchi, S.; Ziegler, D.; Yoshida, Y.; Mabuchi, K. & Suzuki, T. (2005). Parylene flexible neural probes integrated with microfluidic channels. *Lab on a Chip*, Vol. 5, (2005), pp. 519-523.

Thanawala, S.; Palyvoda, O.; Georgiev, D. G.; Khan, S. P.; Al-Homoudi, I. A.; Newaz, G. & Auner, G. (2007). A neural cell culture study on thin film electrode materials. *Journal of Materials Science: Materials in Medicine*, Vol. 18, No. 9, (2007), pp. 1745-1752.

Turner, J. N.; Shain, W.; Szarowski, D. H.; Andersen, M.; Martins, S.; Isaacson, M. & Craighead, H. (1999). Cerebral astrocyte response to micromachined silicon implants. *Experimental neurology*, Vol. 156, No. 1, (1999), pp. 33-49.

Ureshi, M.; Matsuura, T. & Kanno, I. (2004). Stimulus frequency dependence of the linear relationship between local cerebral blood flow and field potential evoked by activation of rat somatosensory cortex. *Neuroscience Research*, Vol. 48, No. 2, (2004), pp. 147-153.

Wennberg, A. (1994). Neurotoxic effects of selected metals. *Scandinavian journal of work, environment & health*, Vol. 20 Spec No, (1994), pp. 65-71.

Williams, J. C.; Rennaker, R. L. & Kipke, D. R. (1999). Long-term neural recording characteristics of wire microelectrode arrays implanted in cerebral cortex. *Brain Research Protocols*, Vol. 4, No. 3, (1999), pp. 303-313.

Wise, K. D. (2005). Silicon microsystems for neuroscience and neural prostheses. *IEEE engineering in medicine and biology magazine: the quarterly magazine of the Engineering in Medicine & Biology Society*, Vol. 24, No. 5, (2005), pp. 22-29.

Xia, Y. & Whitesides, G. M. (1998). Soft lithography. *Annual Review of Materials Science,* Vol. 28, No. 1, (1998), pp. 153-184.

Yeager, J.D.; Phillips, D.J.; Rector, D.M. & Bahr, D.F. (2008). Characterization of flexible ECoG electrode arrays for chronic recording in awake rats. *Journal of Neuroscience Methods*, Vol. 173, No. 2, (2008), pp. 279-285.

Yuen, T. G.; Agnew, W. F. & Bullara, L. A. (1987). Tissue response to potential neuroprosthetic materials implanted subdurally. *Biomaterials*, Vol. 8, No. 2, (1987), pp. 138-141.

Yuen, T. G. & Agnew, W. F. (1995). Histological evaluation of polyesterimide-insulated gold wires in brain. *Biomaterials*, Vol. 16, No. 12, (1995), pp. 951-956.

Zhong, Y. & Bellamkonda, R. V. (2007). Dexamethasone-coated neural probes elicit attenuated inflammatory response and neuronal loss compared to uncoated neural probes. *Brain Research*, Vol. 1148, No. 15-27, (2007), pp. 15-27.

Zhong, Y. & Bellamkonda, R. V. (2008). Biomaterials for the central nervous system. *Journal of the Royal Society, Interface / the Royal Society*, Vol. 5, No. 26, (2008), pp. 957-975.

10

Developments in Time-Frequency Analysis of Biomedical Signals and Images Using a Generalized Fourier Synthesis

Robert A. Brown, M. Louis Lauzon and Richard Frayne
McGill University and University of Calgary
Canada

1. Introduction

Quantitative time-frequency analysis was born with the advent of Fourier series analysis in 1806. Since then, the ability to examine the frequency content of a signal has become a critical capability in diverse applications ranging from electrical engineering to neuroscience. Due to the fundamental nature of the time-frequency transform, a great deal of work has been done in the field, and variations on the original Fourier transform (FT) have proliferated (Mihovilovic and Bracewell, 1991; Allen and Mills, 2004; Peyre and Mallat, 2005). While the FT (Allen and Mills, 2004) is an extremely important signal analysis tool, other related transforms, such as the short-time Fourier transform (STFT) (Allen and Mills, 2004), wavelet transform (WT) (Allen and Mills, 2004) and chirplet transform (Mihovilovic and Bracewell, 1991), have been formulated to address shortcomings in the FT when it is applied to certain problems. Considerable research has been undertaken in order to discover the properties of, and efficient algorithms for calculating the most important of these transforms.

The S-transform (ST) (Stockwell *et al.*, 1996; Mansinha *et al.*, 1997) is of interest as it has found several recent applications in medicine including image transmission (Zhu *et al.*, 2004), the study of psychiatric disorders (Jones *et al.*, 2006), early detection of multiple sclerosis lesions (Zhu *et al.*, 2001), identifying genetic abnormalities in brain tumours (Brown *et al.*, 2008), analysis of EEG recordings in epilepsy patients (Khosravani *et al.*, 2005) and analysis of ECG and audio recordings of cardiac abnormalities (Leung *et al.*, 1998). It has also been successfully applied to non-biomedical tasks such as characterizing the behaviour of liquid crystals (Özder *et al.*, 2007), detecting disturbances in electrical power distribution networks (Chilukuri and Dash, 2004), monitoring high altitude wind patterns (Portnyagin *et al.*, 2000) and detecting gravitational waves (Beauville *et al.*, 2005). However, the computational demands of the ST have limited its utility, particularly in clinical medicine (Brown *et al.*, 2005).

In this chapter we consider several of the more prominant transforms: the Fourier transform, short-time Fourier transform, wavelet transform, and S-transform. A general framework for describing linear time-frequency transforms is introduced, simplifying the direct comparison of these techniques. Using insights from this formalism, techniques developed for the Fourier and wavelet transforms are applied to the formulation of a fast discrete S-transform algorithm with greatly diminished computational and storage demands. This transform is much more computationally efficient than the original continuous approximation of the ST (Stockwell *et al*., 1996) and so allows the ST to be used in acute clinical situations as well as allowing more advanced applications than have been investigated to date, including analyzing longer signals and larger images, as well as transforming data with three or more dimensions, *e.g.*, volumes obtained by magnetic resonance (MR) or computed tomography (CT) imaging. Finally, the STFT and ST are demonstrated in an example biomedical application.

The terminology is, unfortunately, inconsistent between the ST, wavelet and FT literatures. Though these inconsistencies will be pointed out when they arise, we will follow the wavelet convention, where the *continuous* transform takes as input a continuous signal and outputs a continuous spectrum, the *discrete approximation* transforms a discrete, sampled signal into a discrete, oversampled spectrum and the *discrete* transform converts a discrete signal into a discrete, critically sampled spectrum. Additionally, the term *fast* will be used to refer to computationally efficient algorithms for computing the discrete transform.

2. Overview of Selected Time-Frequency Transforms

2.1. The Fourier Transform

The Fourier transform converts any signal, $f(t)$, into its frequency spectrum, which represents the signal in terms of infinite complex sinusoids of different frequency, v, and phase:

$$F(v) = \int_{-\infty}^{+\infty} f(t) e^{-i2\pi vt} dt \tag{1}$$

The FT transforms entirely between the amplitude-time signal-space and the amplitude-frequency frequency-space. That is, the spectrum produced by the FT is necessarily global – it represents the average frequency content of the signal (Mansinha *et al*., 1997). For stationary signals, where the frequency content does not change with time, this is ideal. However, most interesting biomedical signals are non-stationary: their frequency content does vary with time. However, the FT provides no information about this important property.

The FT, as with each of the transforms discussed in this section, is generalizable to any number of dimensions. Higher dimensional transforms may be used to analyze images (two-dimensional), volumetric data from tomographic medical scanners (three-dimensional) or volumetric scans over time (four-dimensional). Though the term "time-frequency" is commonly used, implying one-dimensional functions of amplitude versus time, these concepts are generalizable to higher dimensions and other parameters.

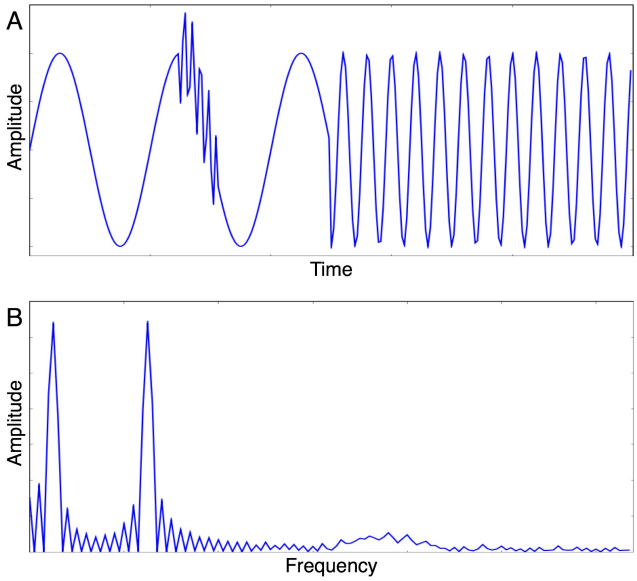


Fig. 1. A sample signal (A) and its Fourier transform (B).

The continuous FT can be calculated analytically according to Eq. (1) for many useful functions but computation of the FT for arbitrarily measured signals requires a discrete formulation. The discrete Fourier transform (Cooley *et al.*, 1969) (DFT) is calculated on a discretely sampled finite signal and provides a discretely sampled finite spectrum.

Simply evaluating the discrete form of Eq. (1) has a compuational complexity of $O(N^2)$. That is, the number of operations required to calculate the DFT grows approximately as the square of the signal length. The fast Fourier transform (Cooley and Tukey, 1965) (FFT) utilizes a divide-and-conquer approach to calculate the DFT more efficiently: it has a computational complexity of $O(N\log N)$. This difference means that computing the FFT of even short signals may be much faster than the DFT, so the FFT is almost universally preferred.

Fig. 1 shows a non-stationary test signal along with its discrete Fourier spectrum, calculated via the FFT algorithm. Note that Fourier spectra are normally complex-valued and include both positive and negative frequencies. For simplicity, figures in this chapter show the absolute value of the spectrum, and the positive-frequency half only. The test signal includes three frequency components: (1) a low frequency for the first half of the signal, (2) a higher frequency for the second half and (3) a very high burst added to the signal in the middle of the low frequency portion. The Fourier spectrum shows strong peaks corresponding to (1) and (2) but (3) is not well detected due to its short duration. Additionally, the sharp transitions between frequencies cause low-amplitude background throughout the spectrum. Note that the Fourier spectrum does not indicate the relative temporal positions of the frequency components.

2.2. The Short-Time Fourier Transform

The Gabor, or short-time Fourier transform (STFT) (Schafer and Rabiner, 1973), Eq. (2), improves Fourier analysis of non-stationary signals by introducing some temporal locality. The signal is divided into a number of partitions by multiplying with a set of window functions, $w(t-\tau)$, where τ indicates the centre of the window. In the case of the Gabor transform, this window is a Gaussian but the STFT allows general windows. In the simplest case, this window may be a boxcar, in effect, partitioning the signal into a set of shorter signals. Each partition is Fourier transformed, yielding the Fourier spectrum for that partition. The local spectra from each partition are combined to form the STFT spectrum, or spectrogram, which can be used to examine changes in frequency content over time.

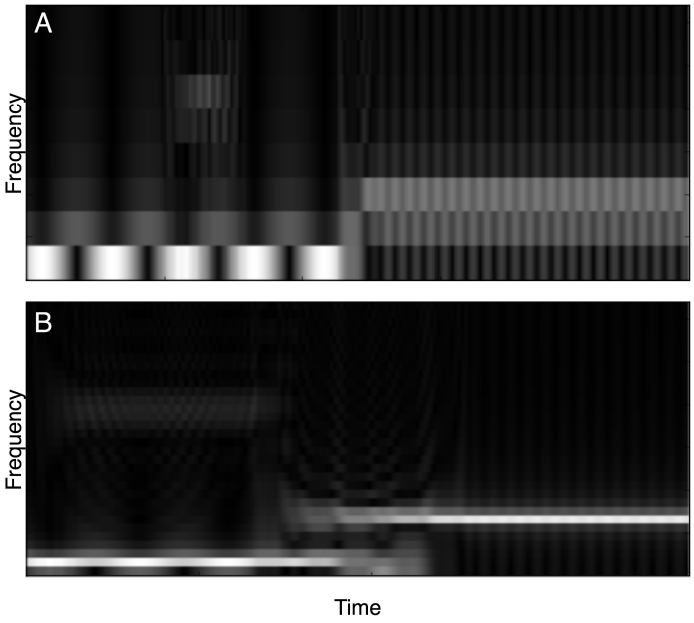


Fig. 2. The STFT of the signal in Fig. 1A with boxcar windows whose widths are 16 samples (A) and 32 samples (B).

$$F(\tau, v) = \int_{-\infty}^{+\infty} f(t) w(t-\tau) e^{-i2\pi vt} dt \tag{2}$$

Fig. 2 shows the STFT spectrum of the test signal in Fig. 1, using boxcar windows of two different widths. The STFT does provide information about which frequencies are present and where they are located, but this information comes at a cost. Narrower windows produce finer time resolution, but each partition is shorter. As with the FT, shorter signals produce spectra with lower frequency resolution. The tradeoff between temporal and frequency resolution is a consequence of the Heisenberg uncertainty principle (Allen and Mills, 2004):

$$\Delta t \Delta v \geq C \tag{3}$$

which states that the joint time and frequency resolution, $\Delta t \cdot \Delta v$, has a lower bound. Additionally, the Shannon sampling theorem (Shannon, 1949) requires that a wavelength be represented by more than two samples and, to avoid artifacts, the window must be wide enough to contain at least one wavelength. This means that lower frequencies are better represented by wider windows (sacrificing time resolution) while high frequencies benefit from narrower windows (sacrificing frequency resolution). In the STFT the window width is fixed so it must be chosen *a priori* to best reflect a particular frequency range of interest.

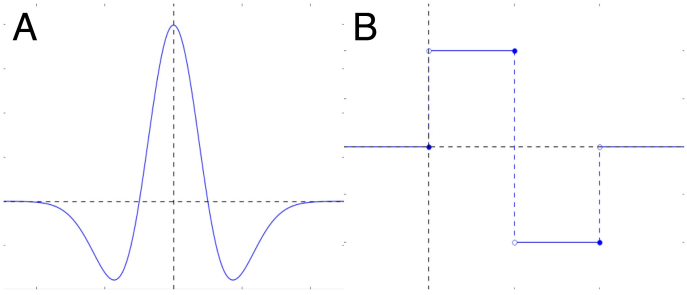


Fig. 3. Examples of two mother wavelets: (A) the continuous Ricker or Mexican hat wavelet and (B) the discrete Haar wavelet.

2.3. The Wavelet Transform

The obvious solution to the window-width dilemma associated with the STFT is to use frequency-adaptive windows, where the width changes depending on the frequency under examination. This feature is known as *progressive resolution* and has been found to provide a more useful time-frequency representation (Daubechies, 1990). Eq. (4) is the wavelet transform (Daubechies, 1990) (WT), which features progressive resolution:

$$\Psi(a,b) = \frac{1}{\sqrt{|a|}} \int_{-\infty}^{\infty} f(t) \psi\left(\frac{t-b}{a}\right) dt \tag{4}$$

where a is the dilation or scale factor (analogous to the reciprocal of frequency) and b is the shift, analogous to τ. The WT describes a signal in terms of shifted and scaled versions of a mother wavelet, $\psi\left(\frac{t-b}{a}\right)$, which is the analog of the complex sinusoidal basis functions used by the FT, but differs in that it is finite in length and is not a simple sinusoid. The finite length of the mother wavelet provides locality in the wavelet spectrum so windowing the signal, as with the STFT, is not necessary. Examples of two common mother wavelets are plotted in Fig. 3.

However, since the mother wavelet is not a sinusoid, the WT spectrum describes a measurement that is only related to frequency, usually referred to as scale, with higher

scales roughly corresponding to lower frequencies and *vice versa*. Additionally, since the mother wavelet is shifted during calculation of the WT, any phase measurements are local; i.e., they do not share a global reference point (Mansinha *et al.*, 1997).

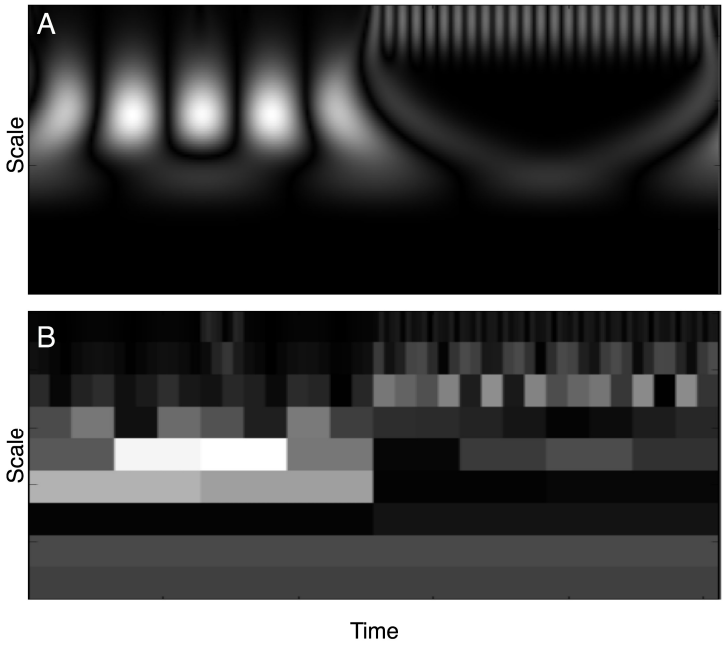


Fig. 4. The continuous Ricker (Mexican hat) wavelet transform (A) and discrete Haar wavelet transform (B) of the signal in Fig. 1A.

Some wavelets, such as the Ricker wavelet (Fig. 4A) or the Morlet wavelet, do not have well behaved discrete formulations and must be calculated using a discrete approximation of the continuous wavelet transform (CWT). This continous approximation is generally difficult to calculate and only practical for short signals of low dimension. However, many mother wavelets yield transforms that have discrete forms and can be calculated via the computationally efficient discrete wavelet transform (DWT). Some wavelets, such as the Haar (Allen and Mills, 2004), Fig. 4B, have a computational complexity of $O(N)$, even faster than the FFT (Beylkin *et al.*, 1991).

2.4. The S-Transform

The S-transform (Stockwell *et al.*, 1996; Mansinha *et al.*, 1997) (ST) combines features of the STFT and WT.

The ST is given by:

$$S(\tau,v)=\int_{-\infty}^{+\infty} f(t)\frac{|v|}{\sqrt{2\pi}}e^{-\frac{(t-\tau)^2v^2}{2}}e^{-i2\pi vt}dt \tag{5}$$

which can be interpreted as an STFT that utilizes a frequency-adaptive, Gaussian window, providing progressive resolution. Alternatively, the ST can be derived as a phase correction to the Morlet wavelet, yielding a wavelet-like transform that provides frequency and globally referenced phase information. The ST of the test signal in Fig. 1 is shown in Fig. 5.

Unfortunately, these advantages come at a cost. The Morlet wavelet must be calculated via the inefficient CWT. Similarly, the continuous approximation of the ST in Eq. (5) has a computational complexity of $O(N^3)$. A more efficient algorithm, however, is described in Eq. (6), in which the ST is calculated from the Fourier transform of the signal (Stockwell *et al.*, 1996):

$$S(\tau,v)=\int_{-\infty}^{+\infty} F(\mu+v)e^{-\frac{2\pi^2\mu^2}{v^2}}e^{i2\pi\tau\mu}d\mu \tag{6}$$

where $F(\mu+v)$ is the Fourier transform of the signal, and it is multiplied by a Gaussian and the inverse Fourier transform kernel. The integration is over frequency, μ. In this form, the ST can be calculated using the FFT but this algorithm is still $O(N^2\log N)$. Additionally, the ST requires $O(N^2)$ units of storage for the transform result, while the DFT and DWT require only $O(N)$. For a 256×256 pixel, 8-bit complex-valued image, which requires 128 kilobytes of storage, the DFT or DWT will occupy no more space than the original signal. But, the ST will require 8 gigabytes of storage space. Either the compuational complexity, memory requirements or both quickly make calculation of the ST for larger signals prohibitive. Addressing these problems is a prerequisite for most clinical applications and also for practical research using the ST.

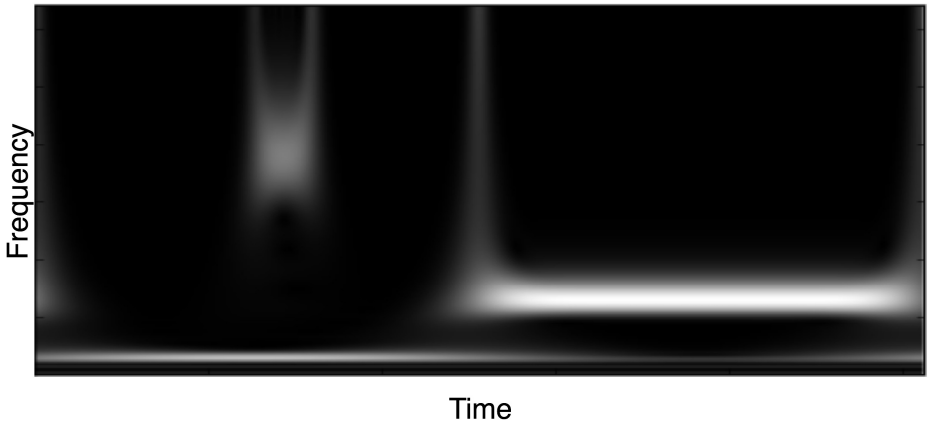


Fig. 5. The ST of the signal in Fig. 1A.

Further discussion of the transforms covered in this section, along with illustrative biomedical examples, can be found in (Zhu *et al.*, 2003).

3. General Transform

Though the different transforms, particularly the Fourier and wavelet transforms, are often considered to be distinct entities, they have many similarities. To aid comparison of the ST with other transforms and help translate techniques developed for one transform to be used with another, we present several common transforms in a unified context. Previous investigators have noted the similarities between the FT, STFT and wavelet transform and the utility of representing them in a common context. To this end, generalized transforms that describe all three have been constructed (Mallat, 1998; Qin and Zhong, 2004). However, to our knowledge, previous generalized formalisms do not explicitly specify separate kernel and window functions. Separating the two better illustrates the relationships between the transforms, particularly when the ST is included.

The ST itself has been generalized (Pinnegar and Mansinha, 2003):

$$S(\tau,v)=\int_{-\infty}^{+\infty} f(t)w(t-\tau,v)e^{-i2\pi vt}dt \tag{7}$$

This generalized S-transform (GST) admits windows of arbitrary shape. It may additionally be argued that $w(t-\tau,\nu)$ can be defined such that the window does not depend on the parameter ν. The result is a fixed window width for all frequencies, and the transform becomes a STFT. However, the presence of ν in the parameter list is limiting and we prefer the following more general notation:

$$S(\tau,v)=\int_{-\infty}^{+\infty} f(t)w(t-\tau,\sigma)e^{-i2\pi vt}dt \tag{8}$$

where σ may be chosen to be equal to ν, to perform an ST, or may be a constant, producing an STFT. In the latter case, if $w(t-\tau,\sigma)=1$, the transform is an FT. Thus, Eq. (8) is a general Fourier-family transform (GFT), describing each of the transforms that utilize the Fourier kernel.

3.1. Extension to the Wavelet Transform

The wavelet transform, though it accomplishes a broadly similar task, at first glance appears to be very disctinct from the Fourier-like time-frequency transforms. The WT uses basis functions that are finite and can assume various shapes, many of which look very unusual compared to the sinusoids described by the Fourier kernel. However, when the basis function is decomposed into its separate kernel and window functions, the WT can be united with the Fourier-based transforms.

Consider the wavelet transform, defined in Eq. (4). Let $g(t)=\frac{\psi(t)}{e^{-i2\pi t}}$, that is, a version of the mother wavelet divided by a phase ramp. For a shifted and scaled wavelet, this becomes:

$$g\left(\frac{t-b}{a}\right)=\frac{\psi\left(\frac{t-b}{a}\right)}{e^{-\frac{i2\pi(t-b)}{a}}} \tag{9}$$

Rearranging and substituting into Eq. (4) yields:

$$\Psi(a,b)=\frac{1}{\sqrt{|a|}}\int_{-\infty}^{\infty} f(t)g\left(\frac{t-b}{a}\right)e^{-\frac{i2\pi(t-b)}{a}}dt \tag{10}$$

The complex exponential term can be expanded into two terms, one of which is similar to the familiar Fourier kernel:

$$\Psi(a,b)=\frac{1}{\sqrt{|a|}}\int_{-\infty}^{\infty} f(t)g\left(\frac{t-b}{a}\right)e^{\frac{i2\pi b}{a}}e^{-\frac{i2\pi t}{a}}dt \tag{11}$$

Letting $\tau = b$, $\nu=\frac{1}{a}$, and $S(\nu,\tau)=\Psi\left(\frac{1}{\nu},\tau\right)$, this becomes:

$$\Psi(a,b)=S(v,\tau)=\sqrt{|v|}\int_{-\infty}^{\infty} f(t)g\left(v\lfloor t-\tau\rfloor\right)e^{i2\pi v\tau}e^{-i2\pi vt}dt \tag{12}$$

Finally, letting $w(t-\tau,\nu)=g(\nu[t-\tau])\sqrt{|\nu|}e^{i2\pi\nu\tau}$, Eq. (12) becomes the GFT, Eq. (8), with σ=ν. Substituting the Fourier-style variables into Eq. (9), rearranging and simplifying gives the window function in terms of the mother wavelet:

$$w(t,\tau,v)=\sqrt{|v|}\psi\left(v\lfloor t-\tau\rfloor\right)e^{i2\pi vt} \tag{13}$$

Thus, the wavelet transform can also be described as a GFT.

4. The Fast S-Transform

Though calculating a discrete approximation of a continuous transform is useful, as with the continuous wavelet and S-transforms, a fully discrete approach makes optimal use of knowledge of the sampling process applied to the signal to decrease the computational and memory resources required. In this section a discrete fast S-transform (FST) (Brown and Frayne, 2008) is developed by utilizing properties that apply to all of the discrete versions of transforms described by the GFT, Eq. (8).

A sampled signal has two important features – the sampling period, Δt, and the number of samples, N. Multiplying these two values gives the total signal length, W_t. Sampling the signal and limiting it to finite length imposes several limitations on the transformed spectrum. The Fourier transform is the simplest case. The DFT of a signal is limited to the same number of samples, N, as the original signal, conserving the information content of the signal. The highest frequency that can be represented, ν_{max}, is the Nyquist frequency, which is half the signal sampling frequency, $\frac{1}{2\Delta t}$. The sampling period of the frequency spectrum, $\Delta\nu$, is the reciprocal of the signal length, $1/W_t$.

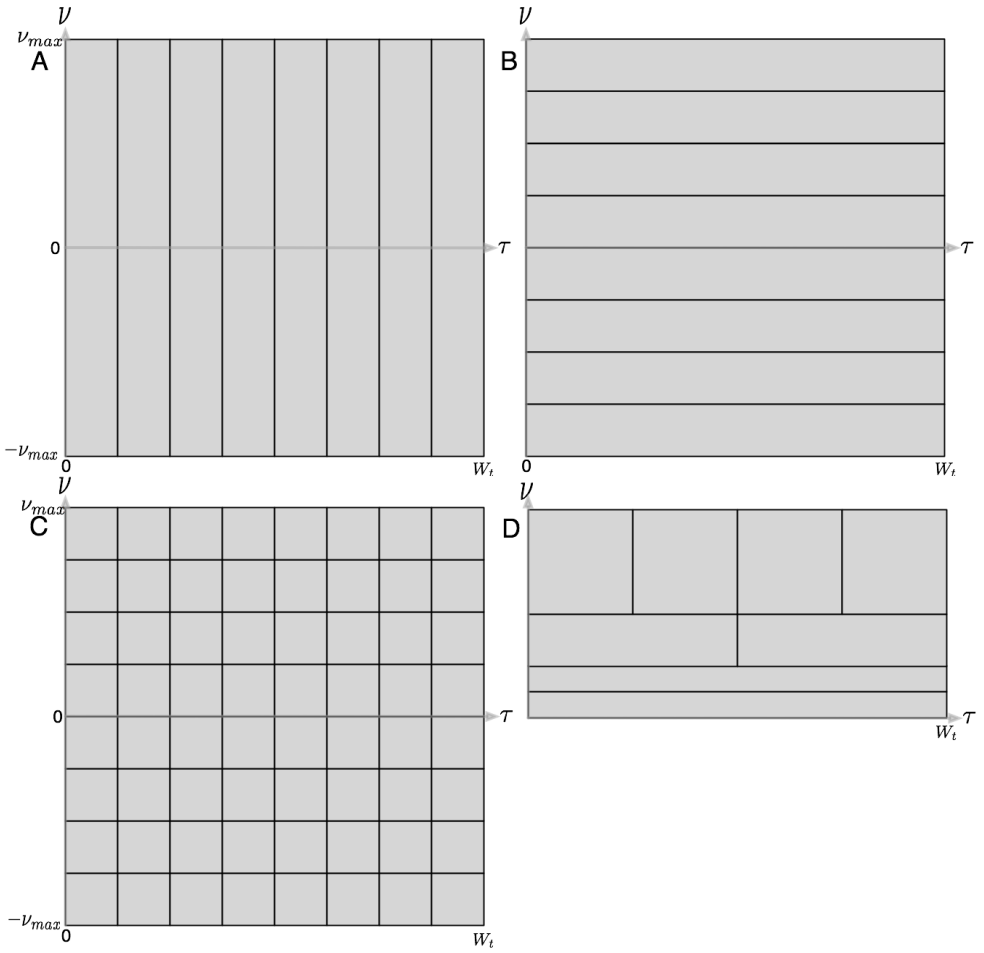


Fig. 6. ν-τ sampling scheme for (A): a uniformly sampled signal with N = 8, (B): the Fourier transform, (C): the conventional S-transform and (D): a discrete Wavelet transform.

Ideally, the result of the continuous S-transform of a one-dimensional signal is $S(\tau,\nu)$, a spectrum with both time and frequency axes. A fast ST must sample the τ-ν plane sufficiently such that the transform can be inverted (*i.e.*, without loss of information) but also avoid unnecessary oversampling. If the original signal contains N points, we possess N independent pieces of information. Since information cannot be created by a transform (and must be conserved by an invertible transform) an efficient ST will produce an N-point spectrum, as does the DFT. The original definition of the ST produces a spectrum with N^2 points. Therefore, for all discrete signals where $N>1$, the continuous ST is oversampled by a factor of N.

In addition, the ST varies temporal and frequency resolution for different frequencies under investigation: higher frequency/lower time resolution at low frequencies and the converse at higher frequencies, but this is not reflected in the uniform N^2-point τ-ν plane sampling scheme. According to the uncertainty principle, at higher frequencies the FST should produce τ-ν samples that have a lesser resolution along the frequency axis and greater resolution along the time axis, in analogy to the DWT. The cost of this oversampling is evident in the increased computational complexity and memory requirements of the ST.

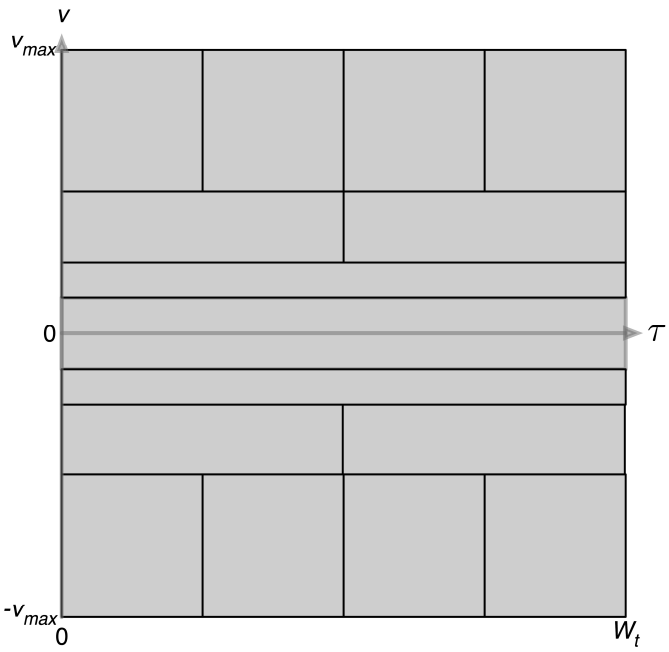


Fig. 7. Sampling scheme for the discrete S-transform of a complex 8-sample signal.

The dyadic sampling scheme used by discrete wavelet transforms provides a progressive sampling scheme that matches underlying resolution changes. In light of the similarities between the DWT and ST illustrated by the GFT, a dyadic sampling scheme can be used to construct a discrete ST. In the case of the ST of a complex signal, a double dyadic scheme is

necessary to cover both the positive and negative frequency ranges. In this arrangement, time resolution is at a minimum for low frequencies, both positive and negative, and increases with the absolute value of frequency. For comparision, the τ-ν plane sampling schemes for the signal, FT, continuous (*i.e.*, conventional) ST and DWT are shown in Fig. 6. The double dyadic scheme of the FST is illustrated in Fig. 7, and a particular algorithm for calculating the FST with this type of sampling scheme is presented in Algorithm 1. The result from Algorithm 1 is presented for a sample signal in Fig. 8.

As might be expected from the GFT formalism, Algorithm 1 is very similar to a filterbank DWT algorithm. High- and low-pass filters, applied in the frequency domain, divide the signal into high- and-low frequency halves (often called "detail" and "approximation", respectively, in the wavelet literature). The high frequency portion is then multiplied by the necessary windowed kernel functions. The low frequency portion forms the input to the next iteration of the algorithm, where it is further decomposed. This simple arrangement produces a dyadic scale with a strict power of two pattern, but can be modified by adjusting the filters to produce finer or coarser frequency domain sampling. However, care must be taken to appropriately modify the time domain sampling to match, and never to violate the Nyquist criterion.

The Gaussian windows of the ST are effectively infinite in extent but when calculating the transform of a finite length signal, the Gaussians are necessarily multiplied by a boxcar window the width of the signal itself. Therefore, the actual window for any finite length signal is a Gaussian multiplied by a boxcar. This situation is particularly apparent at lower frequencies, where the Gaussians are wider and may still be of appreciable amplitude when they are clipped at the edges of the signal. In the discrete approximation of the continuous ST, the Gaussian window is scaled with frequency but the boxcar is not. This contrasts with the Morlet wavelet, which, using the general formalism of Eq. (8), can also be defined with a window that is a composite of a Gaussian and a boxcar. However, in the Morlet wavelet, the Gaussian and boxcar are scaled together. This joint scaling is also inherent in the FST algorithm. Scaling of both parts of the window function is a key refinement, as it both produces more consistent windows and significantly decreases the computational complexity of the FST.

It is only necessary to compute the sums in step 5 of Algorithm 1 for non-zero points (those that are inside the boxcar window). The boxcar must always be wide enough to contain at least one entire wavelength, but the width does not need to be a multiple of the wavelength. This effectively decreases W_t: the full signal length is required only for calculating the DC component, and shorter portions are used at higher frequencies. Since we are downsampling in step 4, Δt is smallest at high frequencies and becomes progressively larger at lower frequencies and so the summation operation in step 5 will always be over a constant number of points. The examples in this paper use 4 points, which produces a slightly oversampled, but smoother, result. This reduces the complexity of step 5, which is nested inside two FOR loops, from $O(N)$ to constant complexity: $O(C)$. As an additional benefit, adjusting the width of the boxcar window greatly reduces the number of kernels and windows that must be pre-calculated in steps 2 and 3 of Algorithm 1, since the kernels

and windows remain essentially constant while the signal's length and resolution are manipulated.

Algorithm 1: The Discrete S-Transform with 2x Oversampling

1. Calculate the Fourier transform of the signal, $H\left(\frac{n}{NT}\right)$

2. Pre-calculate the required kernel functions:

$\phi^{+} = (kT, \frac{n}{NT}) = e^{-\frac{i2\pi kn}{N}}$ and $\phi^{-} = (kT, \frac{n}{NT}) = e^{\frac{i2\pi kn}{N}}$

3. Pre-calculate the window functions: $w(kT, \frac{n}{NT})$

FOR n in $\{\frac{N}{2}, \frac{N}{4}, \frac{N}{8}, ..., 4, 2, 1\}$ DO:

4. Band pass filter $H(\kappa)$:

$$H'(\kappa) = H(\kappa) \text{ where } \frac{n}{2} < |\kappa| \le n$$

and inverse FT to obtain $h'(t)$.

FOR every point j in $h'(t)$ DO:

5. Calculate the transform samples:

$$S(jT, \frac{3n}{4NT}) = \sum_{k=0}^{N-1} h'(kT) \cdot w(kT - T, \left|\frac{3n}{4NT}\right|) \cdot \phi^{+}(kT, \frac{3n}{4NT})$$

$$S(jT, -\frac{3n}{4NT}) = \sum_{k=0}^{N-1} h'(kT) \cdot w(kT - T, \left|\frac{3n}{4NT}\right|) \cdot \phi^{-}(kT, \frac{3n}{4NT})$$

END FOR

END FOR

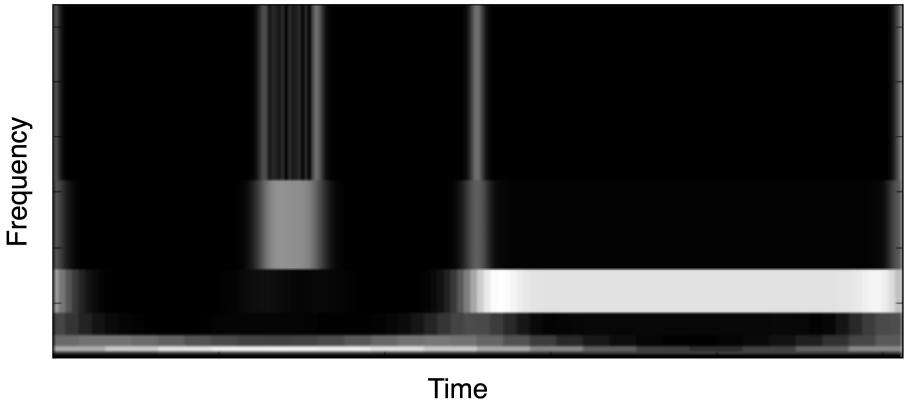


Fig. 8. The fast discrete ST of the signal in Fig. 1A.

The computational complexity of the FST algorithm is $O(NlogN)$ – the same as that of the Fourier transform. The storage requirements are $O(N)$, like the FFT and discrete wavelet transforms (Brown and Frayne, 2008).

5. The Inverse Fast S-Transform

A discrete version of the S-transform should be invertible by the same procedure as the inverse continuous ST: summation of the transform space over the τ-axis, producing the Fourier spectrum, which can then be inverse discrete or fast Fourier transformed to obtain the original signal. The inverse discrete and fast Fourier transforms require a coefficient for each integer value of the frequency index variable ν from $-\frac{\nu_{\max}}{2}$ to $+\frac{\nu_{\max}}{2}$. Since the FST uses an octave (*i.e.*, dyadic) system along the ν-axis, the missing coefficients must be calculated from known points and *a priori* information. Note that the following derivation uses the general definition for the window, $w(t-\tau,\sigma)$. In the specific case of the FST, $\sigma = \nu_p$.

Consider a single line of the GFT spectrum for ν fixed at some value ν_p: $l(\tau) = S(\tau, \nu_p)$. Then, from Eq. (8):

$$l(\tau) = S(\tau, v_p) = \int_{-\infty}^{+\infty} f(t) w(t-\tau,\sigma) e^{-i2\pi v_p t} dt \tag{14}$$

The Fourier transform of $l(\tau)$, $L(\nu')$ is:

$$L(v') = \int_{-\infty}^{\infty} \left\{ \int_{-\infty}^{+\infty} f(t) w(t-\tau,\sigma) e^{-i2\pi v_p t} dt \right\} e^{-i2\pi\tau v'} d\tau \tag{15}$$

Rearranging terms gives:

$$L(v') = \int_{-\infty}^{+\infty} f(t)e^{-i2\pi v_p t}dt \int_{-\infty}^{\infty} w(t-\tau,\sigma)e^{-i2\pi\tau v'}d\tau \tag{16}$$

Evaluating the second integral using the Fourier shift theorem, this becomes:

$$L(v') = \int_{-\infty}^{+\infty} f(t)e^{-i2\pi v_p t}dt\left[W^*(v',\sigma)e^{-i2\pi t v'}\right] \tag{17}$$

where $W^*(\nu',\sigma)$ is the inverse Fourier transform of the window. In the common case where the window function is real and even the inverse and forward FT are identical, in which case $W^*(\nu',\sigma)$ is interchangeable with $W(\nu',\sigma)$, the forward Fourier transform of the window. This can be rearranged to give:

$$L(v') = W^*(v',\sigma)\int_{-\infty}^{+\infty} f(t)e^{-i2\pi v_p t}e^{-i2\pi t v'}dt \tag{18}$$

Evaluating the integral, which is a Fourier transform, gives:

$$L(v') = W^*(v',\sigma)F(v'+v_p) \tag{19}$$

Finally, after rearranging:

$$F(v'+v_p) = \frac{L(v')}{W^*(v',\sigma)} \tag{20}$$

It is clear that, in the continuous case, any Fourier coefficient can be obtained from the Fourier transform of any fixed $\nu = \nu_p$ line of the GFT spectrum. In the discrete, fast transform case, as calculated by Algorithm 1, recall that the Fourier spectrum is band pass filtered during computation of $S(\tau,\nu_p)$. This means that the Fourier spectrum retrieved from Eq. (20) will also be filtered. However, as each $\nu = \nu_p$ line results from a *different* band pass filtering, the full Fourier spectrum can be reconstructed from the $S(\tau,\nu_p)$ spectrum via Eq. (20). It is then a simple matter to perform an inverse FFT and reconstruct the original signal. Note that the inversion procedure for the continuous approximation of the ST (Mansinha *et al.*, 1997), summation over the τ-axis, is equivalent to applying Eq. (20) to each line but discarding all but the DC component, $F(\nu'+\nu_p)$ where $\nu' = 0$.

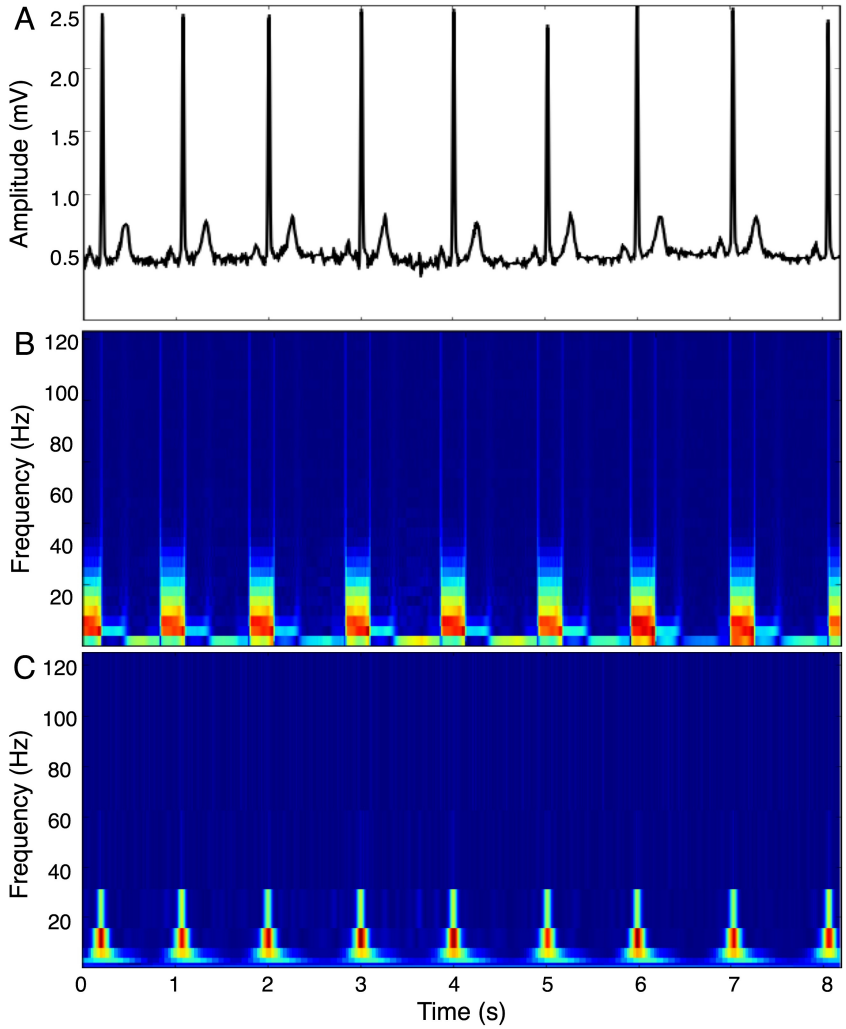


Fig. 9. A short segment of an electrocardiogram (A), its STFT (B) and FST (C). Red indicates high power while blue indicates low power.

6. Biomedical Example

Fig. 9 shows a sample biomedical signal, along with its STFT and FST. The signal (Fig. 9A) is a short electrocardiogram (ECG) recording from a publicly available subset of the European ST-T Database (Physionet, http://pysionet.org), consisting of the first 2048 samples from the V4 lead of record e0103, a 62 year old male subject complaining of mixed angina. Although the STFT (Fig. 9B) and the FST (Fig. 9C) both provide spectra on the time-frequency plane, their respective temporal-frequency resolutions are very different. The window width of the STFT is fixed for all frequencies, in this case at 64 points, yielding the same combination of

time and frequency resolution in all parts within the spectrum. In contrast, the FST demonstrates progressive resolution, trading decreased frequency resolution for increased time resolution at higher frequencies.

This difference is most obvious in the major spectral features corresponding to the R-waves (the dominant peaks in the ECG signal, which are associated with depolarization of the ventricles). In the STFT spectrum, localisation of the R-wave peak is limited by the time resolution of the transform. In this case, the peak can only be approximately located. Note also that the low frequency features corresponding to the other characteristic ECG components that occur between R-waves appear in the lowest frequency band of the STFT. Using this window, these features are not distinguishable from the DC, or zero-frequency component. In order to show these features, the DC component was removed from the signal before transforming.

In the FST spectrum, the R-wave can be localized much more precisely by examining higher frequencies, between 20 and 40 Hz, where the time resolution is much better. At the same time, increased frequency resolution at lower frequencies in the FST spectrum has allowed low frequency features to be better resolved and separated from the DC component.

The example signal in Fig. 9 consists of few samples and it can be transformed in a reasonable time even with the inefficient continuous approximation of the ST: the ST takes 1.6 s and 64 MB compared to 1.7 ms and 2 KB for the FST (2.5 GHz Intel Core 2 Duo, one core only). However, the full ECG signal consists of almost two million samples. Higher dimensional datasets, including medical images and volumes, can be even larger. ST and FST computation times and memory requirements for a few biomedical signals are compared in Table 1.

		ST		FST	
	Samples	Time	Memory	Time	Memory
MR Image	256×256	40 min	64 GB	40 ms	1 MB
CT Image	1024×1024	9 days	16 TB	0.8 s	16 MB
ECG	2^{21}	37 days	64 TB	1.7 s	32 MB
MR Volume	$256 \times 256 \times 64$	156 days	256 TB	3.6 s	64 MB
Visible Human (male) CT	$512 \times 512 \times 1871$	7.6 thousand years	3 EB	9 minutes	7 GB

Table 1. Approximate computation times and memory requirements for (i) the continuous approximation of the ST and (ii) the FST of various biomedical signals. These estimates are based on computations using one core of a 2.5 GHz Intel Core 2 Duo processor.

7. Conclusions

In this chapter we have defined a generalized framework that describes time-frequency transforms, including the familiar Fourier and wavelet transforms, in unified terms. Using

the generalized framework as a guide, we examined the ST, a transform that has proven to be particularly useful in biomedical and medical applications as well as in non-medical fields. A discrete fast implementation of the ST, the FST, was derived, which has a computational complexity of $O(NlogN)$ and memory complexity of $O(N)$, a significant improvement on the continuous approximation of the ST computational complexity of $O(N^2logN)$ and storage complexity of $O(N^2)$. This decrease in complexity allows calculation of the FST, with modest resources, of signals of more than 2^{16} points, *i.e.*, images larger than 256x256 pixels, volumes, and higher dimensional datasets of non-trivial size. The increased efficiency and wider applicability of the FST allows it to be considered for more applications, including those that have strict size or time limitations such as compression, progressive image transmission or acute care medical image analysis.

8. References

Allen, R. L. & Mills, D. W. (2004). *Signal Analysis,* IEEE Press, 0-471-23441-9, Piscataway, NJ

Beauville, F.; Bizouard, M. A.; Bosi, L.; Brady, P.; Brocco, L.; Brown, D.; Buskulic, D.; Chatterji, S.; Christensen, N.; Clapson, A. C.; Fairhurst, S.; Grosjean, D.; Guidi, G.; Hello, P.; Katsavounidis, E.; Knight, M.; Lazzarini, A.; Marion, F.; Mours, B.; Ricci, F.; Viceré, A. & Zanolin, M. (2005). A first comparison of search methods for gravitational wave bursts using LIGO and VIRGO simulated data. *Classical and Quantum Gravity*, 22, (2005) S1293-S1301

Beylkin, G.; Coifman, R. & Rokhlin, V. (1991). Fast wavelet transforms and numerical algorithms I. *Communications on Pure and Applied Mathematics*, 44, 2, (1991) 141-183

Brown, R.; Zhu, H. & Mitchell, J. R. (2005). Distributed vector processing of a new local multi-scale Fourier transform for medical imaging applications. *IEEE Transactions on Medical Imaging*, 24, 5, (2005) 689-691

Brown, R. A. & Frayne, R. A fast discrete S-transform for biomedical signal processing, *IEEE Engineering in Medicine and Biology Conference*, Vancouver, BC, 2008

Brown, R. A.; Zlatescu, M. C.; Sijben, A.; Roldan, G.; Easaw, J.; Forsyth, P. A.; Parney, I.; Sevick, R. A.; Yan, E.; Demetrick, D.; Schiff, D.; Cairncross, J. G. & Mitchell, J. R. (2008). The use of magnetic resonance imaging to noninvasively detect genetic signatures in oligodendroglioma. *Clinical Cancer Research*, 14, (2008) 2357-2362

Chilukuri, M. V. & Dash, P. K. (2004). Multiresolution S-transform-based fuzzy recognition system for power quality events. *IEEE Transactions on Power Delivery*, 19, (2004) 323-330

Cooley, J. W. & Tukey, J. W. (1965). An algorithm for the machine calculation of complex Fourier series. *Mathematics of Computation*, 19, (1965) 297-301

Cooley, J. W.; Lewis, P. A. W. & Welch, P. D. (1969). The finite Fourier transform. *IEEE Transactions on Audio and Electroacoustics*, 17, 2, (1969) 77-85

Daubechies, I. (1990). The wavelet transform, time-frequency localization and signal analysis. *IEEE Transactions on Information Theory*, 36, 5, (1990) 961-1005

Jones, K. A.; Porjesz, B.; Chorlian, D.; Rangaswamy, M.; Kamarajan, C.; Padmanabhapillai, A.; Stimus, A. & Begleiter, H. (2006). S-transform time-frequency analysis of p300 reveals deficits in individuals diagnosed with alcoholism. *Clinical Neurophysiology*, 117, 10, (2006) 2128-2143

Khosravani, H.; Pinnegar, C. R.; Mitchell, J. R.; Bardakjian, B. L.; Federico, P. & Carlen, P. (2005). Increased high-frequency oscillations precede in vitro low Mg2+ seizures. *Epilepsia*, 46, 8, (2005) 1188-1197

Leung, T. S.; White, P. R.; Cook, J.; Collis, W. B.; Brown, E. & Salmon, A. P. (1998). Analysis of the second heart sound for diagnosis of paediatric heart disease. *IEE Proceedings on Science, Measurement and Technology*, 145, 6, (1998) 285-290

Mallat, S. G. (1998). *A Wavelet Tour of Signal Processing,* Academic Press,

Mansinha, L.; Stockwell, R. & Lowe, R. (1997). Pattern analysis with two-dimensional spectral localisation: Applications of two-dimensional S-transforms. *Physica A*, 239, (1997) 286-295

Mihovilovic, D. & Bracewell, R. N. (1991). Adaptive chirplet representation of signals in the time-frequency plane. *Electronics Letters*, 27, 13, (1991) 1159-1161

Özder, S.; Coşkun, E.; Köysal, O. & Kocahan, Ö. (2007). Determination of birefringence dispersion in nematic liquid crystals by using an S-transform. *Optics Letters*, 32, (2007) 2001-2003

Peyre, G. & Mallat, S. (2005). Surface compression with geometric bandelets. *ACM Transactions on Graphics*, 24, 3, (2005) 601-608

Pinnegar, C. R. & Mansinha, L. (2003). The S-transform with windows of arbitrary and varying shape. *Geophysics*, 68, (2003) 381-385

Portnyagin, Y. I.; Forbes, J.; Merzlyakov, E. G.; Makarov, N. A. & Palo, S. E. (2000). Intradiurnal wind variations observed in the lower thermosphere over the south pole. *Annales Geophysicae*, 18, (2000) 547-554

Qin, S. R. & Zhong, Y. M. (2004). Research on the unified mathematical model for FT, STFT and WT and its applications. *Mechanical Systems and Signal Processing*, 18, 6, (2004) 1335-1347

Schafer, R. W. & Rabiner, L. R. (1973). Design and simulation of a speech analysis-synthesis system based on short-time Fourier analysis. *IEEE Transactions on Audio and Electroacoustics*, AU-21, 3, (1973) 165-174

Shannon, C. E. (1949). Communication in the presence of noise. *Proceedings of the I.R.E.*, 37, 1, (1949) 10-21

Stockwell, R. G.; Mansinha, L. & Lowe, R. P. (1996). Localization of the complex spectrum: The S transform. *IEEE Transactions on Signal Processing*, 44, 4, (1996) 998-1001

Zhu, H.; Mayer, G.; Mansinha, L.; Law, A. G.; Archibald, C. J.; Metz, L. & Mitchell, J. R. Space-local spectral texture map based on MR images of MS patients, *MS: Clinical and Laboratory Research, ACTRIMS*, Chicago, 2001

Zhu, H.; Goodyear, B. G.; Lauzon, M. L.; Brown, R. A.; Mayer, G. S.; Law, A. G.; Mansinha, L. & Mitchell, J. R. (2003). A new local multiscale Fourier analysis for medical imaging. *Medical Physics*, 30, 6, (2003) 1134-1141

Zhu, H.; Brown, R. A.; Villanueva, R. J.; Villanueava-Oller, J.; Lauzon, M. L.; Mitchell, J. R. & Law, A. G. (2004). Progressive imaging: S-transform order. *Australian and New Zealand Industrial and Applied Mathematics Journal*, 45, (2004) C1002-1016

11

Automatic Counting of *Aedes aegypti* Eggs in Images of Ovitraps

Carlos A.B. Mello[1], Wellington P. dos Santos[1], Marco A.B. Rodrigues[2], Ana Lúcia B. Candeias[3], Cristine M.G. Gusmão[1] and Nara M. Portela[1]
[1]*Polytechnic School of Pernambuco, University of Pernambuco*
[2]*Department of Electronic and Systems, Federal University of Pernambuco*
[3]*Department of Cartographic Engineering, Federal University of Pernambuco*
Brazil

1. Introduction

Dengue is a disease caused by a virus and transmitted by the *Aedes aegypti* mosquito. The *Aedes aegypti* appeared in Africa (probably in the northeast region) and it was spread there to Asia and Americas, mainly through the maritime traffic. In Brazil, it arrived in the 18th century with the boats that carried slaves, since the eggs of the mosquito can resist without contact with the water for up to one year.

Aedes aegypti is a very efficient disseminator of human pathogens as a result of evolutionary adaptations to frequent haematophagy as well as to the colonization of countless types of habitats, associated with environmental and cultural factors that favor the proliferation of this mosquito in urban ecosystems (Regis et al., 2008). In average, each *Aedes aegypti* lives around 30 days and the female puts between 130 and 200 eggs in each gonadotrophic cycle. It is able to lay their eggs repeatedly along its life, if copulate with the male at least once. The sperm is stored in its spermathecae (a reservoir present inside of its reproductive system). After the mosquito acquires the dengue virus, the female becomes a permanent vector of the disease and may even pass to his successors, who have already born infected.

The eggs are not put directly in the water; they are placed millimeters above the surface, in places such as: empty cans and bottles, tires, gutters, pots of plants or any other place that can store rain water. When it rains and the water level rises, coming into contact with the eggs, the eggs hatch in about 30 minutes. Within 8 days the mosquito can complete its life cycle from egg, to larvae, to pupae and to an adult flying mosquito.

These mosquitoes are responsible for one of the most difficult public health problems in tropical and semi-tropical world: the epidemic proliferation of dengue, a viral disease that, in its most dangerous form, dengue hemorrhagic fever, can even cause death of affected human beings (Perich et al., 2003). In the absence of an effective preventive vaccine, effective treatment or chemoprophylaxis etiologic, the only way to reduce the dengue proliferation is the reduction of the potential breeding containers. This means the involvement of vector control personnel, several public administration sectors, social organizations, productive

sectors and the general community that indirectly contribute to the increasing number of breeding containers (Perich et al., 2003) (Dibo, et al. 2005) (Regis et al. 2008). The early detection to outbreak diseases such as dengue is important to enable shares of research and monitoring by the agencies of public health, which reinforces the need for surveillance systems. The routinely employed method to monitor *Aedes aegypti* population in Vector Control Programs of Brazilian states is larval surveillance in potential breeding containers, which enables the attainment of entomological indicators such as the Premise, Container and Breteau Indexes (Dibo et al., 2005). In non-infected municipalities, this Program recommends the use of larvitraps, mat-black containers (in fact, sections of tires) containing 1 liter of water that are checked on a weekly basis, aiming at detecting foci of *Aedes aegypti* (Dibo et al., 2005).

One of the most efficient methods available for mosquito detection and monitoring is the use of ovitraps, which consist of black containers that are partially filled with tap water holding vertical wooden paddles with one rough side. Ovitraps are sensitive, fast, and economic to determine the presence of egg-laying females of *Aedes aegypti* (Dibo et al., 2005) (Gama, Eiras & Resende, 2007).

To generate important statistics and furnish government agencies and vector control programs information enough to project official actions and programs to develop and increase the control of dengue mosquitoes, it is very important to count the number of *Aedes aegypti* eggs present in ovitraps. This counting is usually performed in a manual, visual and non-automatic form. To aid the control of dengue proliferation, this work approaches the development of automatic methods to count the number of eggs in ovitraps images using image processing, particularly color segmentation and mathematical morphology-based non-linear filters.

In Recife, Brazil, the research on dengue is made mainly by the The Aggeu Magalhaes Research Center (CPqAM). This research is part of a project called SAPIO, granted by FINEP, that aims the development of new technologies for dengue control, surveillance and information dissemination.

This Chapter is organized as follows: next Section describes the images acquired and the algorithms developed to perform automatic counting of *Aedes aegypti* eggs in ovitraps. Following, the results are related and analyzed in Section III. In Section IV it is presented conclusions and performed some commentaries on our results.

2. Material and Methods

For this experiment, we used a digital camera with: 7.2 Megapixels resolution, LCD 2.5'', 4.5 times Optical Zoom and LEICA DC Vario Elmarit lens. The ovitrap was digitized with about 700 dpi resolution and 4 times optical zoom. This process generated a true color digital image of 3,072 *versus* 2,304 pixels which was split into sub-images for the experiments. The amount of eggs in each one of these sub-images is acquired by visual inspection allowing an easy comparison with our new proposal. Figure 1 presents some sample sub-images used in the tests and the amount of eggs in each one of them. The same figure also presents a zooming into some *Aedes aegypti* eggs. The images are digitized in RGB color system due to the camera features.

One of the problems of an automatic counting method is the segmentation of the images (Parker, 1997). A segmentation algorithm divides an image into its relevant objects. As the

concept of object can be different from image to image, segmentation is not a simple task. Classical segmentation algorithms can be found as watershed (Dougherty & Lotufo, 2003) and quadtree decomposition (Gonzalez & Woods, 2007). These techniques however are well-known to produce over-segmentation, *i.e.*, they find more objects than it is needed. Figure 2 presents the segmentation of the image presented in Figure 1-top using watershed (Figure 2-top) and quadtree decomposition (Figure 2-bottom).

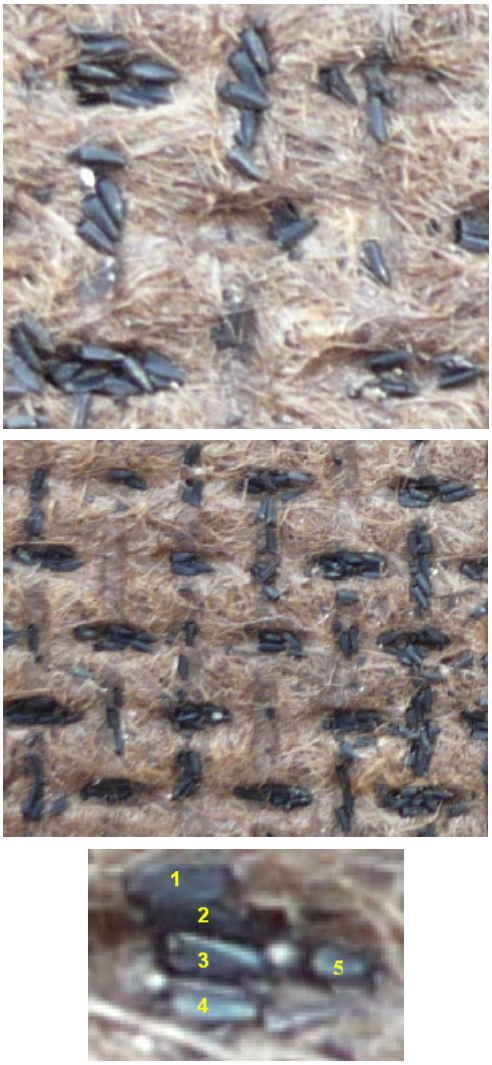


Fig. 1. Samples of an ovitrap with: (top) 34 eggs and (center) 111 eggs (bottom). A zooming into a group of five eggs (labeled in yellow).

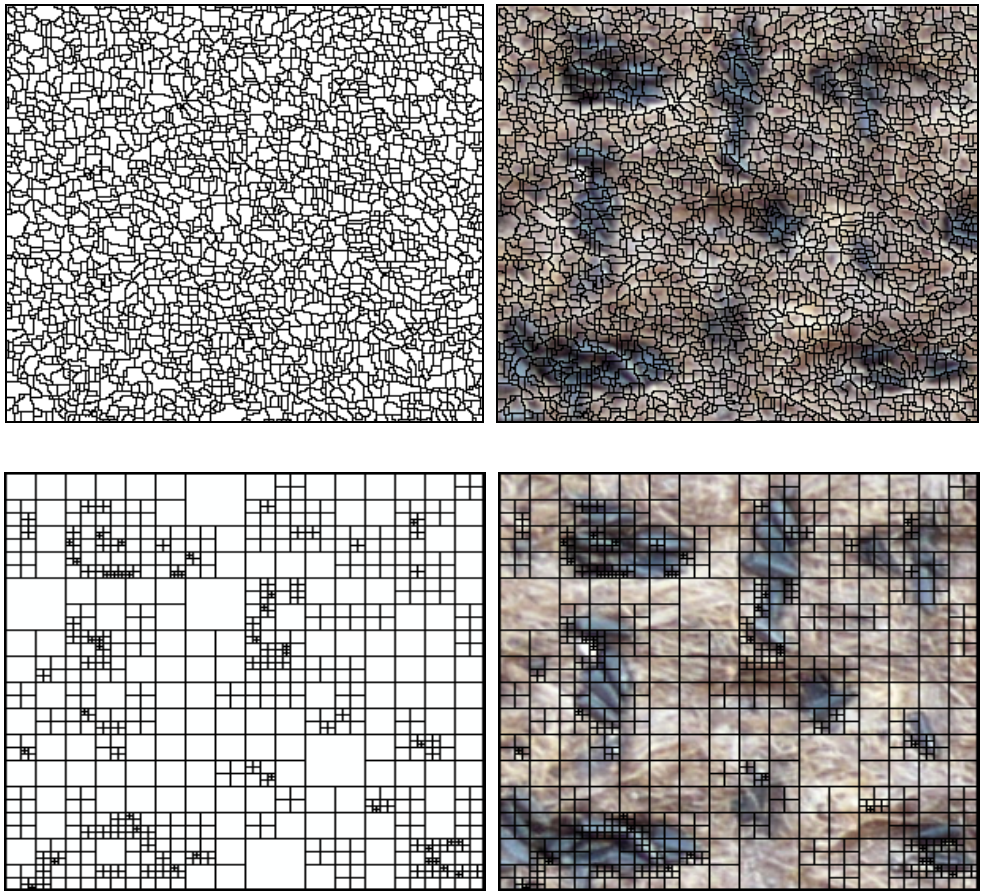


Fig. 2. Segmentation results using (top) watershed and (bottom) quadtree decomposition.

Other texture based methods groups areas that are considered similar based on some feature as texture analysis. The image is broken into k windows. Each window w_r, $1 \leq r \leq k$, is compared with the others. If the windows w_i and w_j satisfies some criteria, they are considered the same texture. In this case, the criteria is a fidelity index. This technique was previously presented in (Mello & Mascaro, 2006) for mammographic image analysis and it uses a fidelity index (Wang & Bovik, 2002) to merge each windows. Figure 3 presents the 36 different areas found in the sample image of Figure 1-top based on texture features. Different colors are used to represent each different area; similar areas must appear in different regions of the image. The problem of over-segmentation can be noticed.

Two methods are proposed for the automatic counting. Each one of them is based on a different color space model. Next, we will detail both of them.

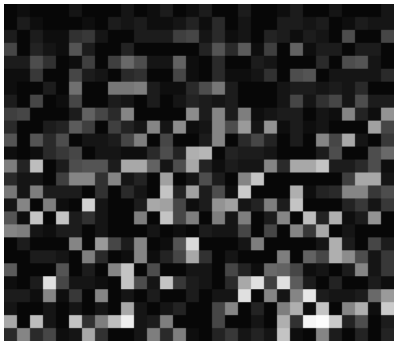


Fig. 3. Segmentation of image of Figure 1-top using a fidelity index to merge similar regions (each different color represents a different region).

3.1 First Method

Trying to achieve more difference between the eggs and the trap, the images are converted from RGB to HSL color model (*Hue*, *Saturation* and *Lightness*). One can see in Figure 4 the HSL components of the image shown in Figure 1-top.

From these three components, the hue is the one that contains information about the color tone. For example a hue value of 240 is related to several blue tones. It is evaluated as (Ballard & Brown, 1982):

$$hue = \cos^{-1}\left(\frac{(((r-g)+(r-b))/2)}{\sqrt{(r-g)^2+(r-g)(g-b)}}\right) \quad (1)$$

where r, g and b are the values of the red, green and blue components for a given color. As can be seen in Figure 4-left, the hue does not retain information about most part of the ovitrap itself.

The hue image is then binarized using Huang thresholding algorithm (Huang & Wang, 1995). Other thresholding algorithms (Sezgin & Sankur, 2004) were tested but Huang's and Li-Lee's algorithms produced the best results.

Huang algorithm uses a function *E(t)* that is applied to all possible threshold values (*t*). When the smallest value of the function is found the corresponding value of *t* is chosen as the threshold. The function E(t) is defined as:

$$E(t) = 1/(m.n) \sum Hf\,(Ux\,(g)).h(g) \quad (2)$$

where:
g = gray level
h(g) = histogram level for a given gray level
n = number of lines in the picture
m = number of columns in the picture
Ux and *Hf* are two values found from a pre-defined set of equations.

The Li-Lee method is an entropy-based technique. It finds the threshold value by minimizing the cross entropy between the image and its segmented version.
To understand this algorithm, we must understand the concept of maximum entropy and from it, minimum cross entropy. The principal of maximum entropy allows us to choose the solution that outputs greatest entropy. It has been shown through experiments that distributions that have greater entropy have higher multiplicity, therefore being more likely to be observed.
Cross entropy measures the theoretical distance between two distributions: *P={p1, p2, p3,…, pn}* and *Q={q1,q2,q3,…,qn}* by:

$$D\ (Q,\ P) = \sum q_k \log_2 q_k / p_k \tag{3}$$

The minimum cross entropy method can be seen as an extension of the maximum cross entropy method, assigning initial estimated values for all p_i when no information is available.
This algorithm functions in a way that it is considered as a reconstruction process of the image distribution. The segmented image *g(x,y)* will be constructed in the following manner:

$$g(x,y) = \begin{cases} u1, & f(x,y) < t \\ u2 & f(x,y) \geq t \end{cases} \tag{4}$$

The segmented image *g(x,y)* will be solely determined by the function *f(x,y)* which has two unknown variables, *μ1, μ2,* and *t.* A criteria function must be used to determine the best values for these variables so that they best satisfy *f.* The criteria function in this method is the cross entropy. It will be associated with the functions shown above, finding the threshold *t* and the final image *g(x,y).*
We opted by Huang's algorithm because of the lower processing time. Figure 5-left shows the bi-level version of the hue image of Figure 4-left. There are still several parts considered objects in the image as they are converted into white.

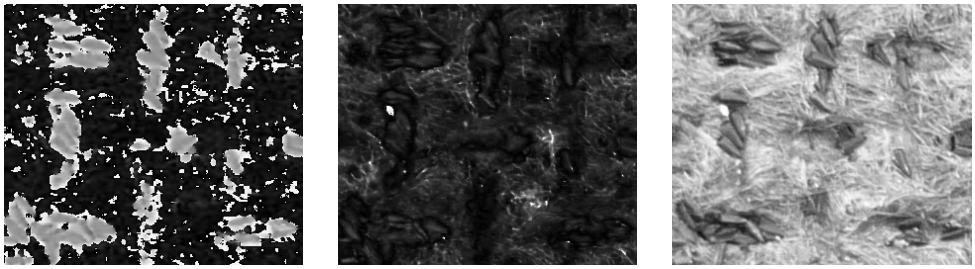


Fig. 4. HSL components of the image in Figure 1-top: (left) Hue components; (center) Saturation and (right) Lightness.

With the bi-level image, a connected components algorithm is applied to label the connected regions of the image (Shapiro & Stockman, 2001). This algorithm puts a different label at each connected white area of the image. With this labeling, it is possible to evaluate each connected area. Small areas can be deleted as it could not contain an egg. Our experiments defined that every area with less than 100 pixels should be deleted. This can be seen in

Figure 5 - right where it is presented the image of Figure 5 - left after the reduction of its white areas.

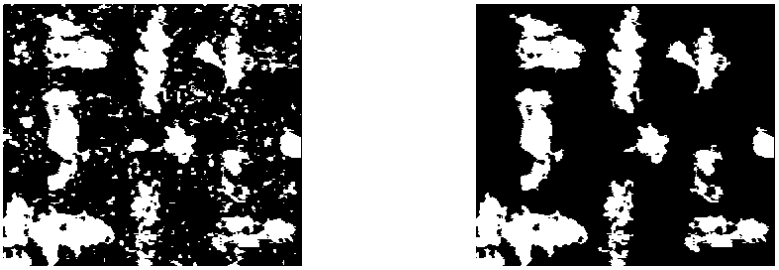


Fig. 5. (left) Hue image after binarized by Huang's thresholding algorithm and (right) bi-level hue image after elimination of small connected areas.

After this, the image is filtered using the morphological operation of closing (Parker, 1997). For this purpose, a structural element is defined in the form of an egg. As the eggs are disposed in different positions along the ovitrap, it was selected a sample egg with average size and with a small inclination angle. To avoid the loss of small eggs in the counting process, the image used as structural element has its original size reduced from 18 x 30 pixels to 8 x 13 pixels. Figure 6 presents the original image (left-top) and the final structural element (left-bottom).
The result of the closing operation applied to the image presented in Figure 5-right is shown in Figure 6-right. The areas with eggs are now more delimited.

Fig. 6. Egg identification. (a) Average egg that was used to define (b) the structural element. (c) Image of Figure 5-right after application of the closing operator with the structural element of Figure 6-left-bottom.

For the final stage, we considered that an egg occupies an area of 170 pixels. So, the number of eggs is the total amount of white pixels divided by this average area. In this case, the method registered an amount of 33 eggs against the correct value of 34 eggs that the image contains.

3.2 Second Method

The second approach used in this work is based on converting RGB sub-images to YIQ ones and, finally, segmenting band I and counting mosquitoes eggs using a standard labeling

algorithm (Gomes, Velho & Frery, 2008). YIQ color base transformation is given as follows (Ballard & Brown, 1982):

$$\begin{bmatrix} Y \\ I \\ Q \end{bmatrix} = \begin{bmatrix} 0.299 & 0.587 & 0.114 \\ 0.596 & -0.275 & -0.321 \\ 0.212 & -0.523 & 0.311 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix}. \quad (5)$$

The segmentation of band I can be performed in two ways: 1) using limiarization with a fixed threshold of 200; 2) binarization using k-means clustering method (Haykin, 1998), with 3 inputs, 4 classes, learning rate of 0.1 and a maximum of 200 iterations. To perform eggs counting, it is considered an average-sized mosquito egg of 220 pixels. Such a difference of size (220 pixels against about 250 of the first method) is due to different segmentation methods and the absence of the application of mathematical morphology-based filters in this method, once there is no structural element.

Figure 7-top-left presents an RGB color sub-image of an ovitrap, where Figure 7-top-right shows the RGB composition of its YIQ version. Figure 7-bottom-left presents segmentation results using the fixed threshold approach, whereas Figure 7-bottom-right shows the k-means based method. It can be seen in Figure 8 and Figure 9, respectively, the gray level version and histogram of band I of the image presented in Figure 7-top-right. In this example, the proposed algorithm counted 34 eggs, exactly the same number of eggs got by the visual non-automatic counting process.

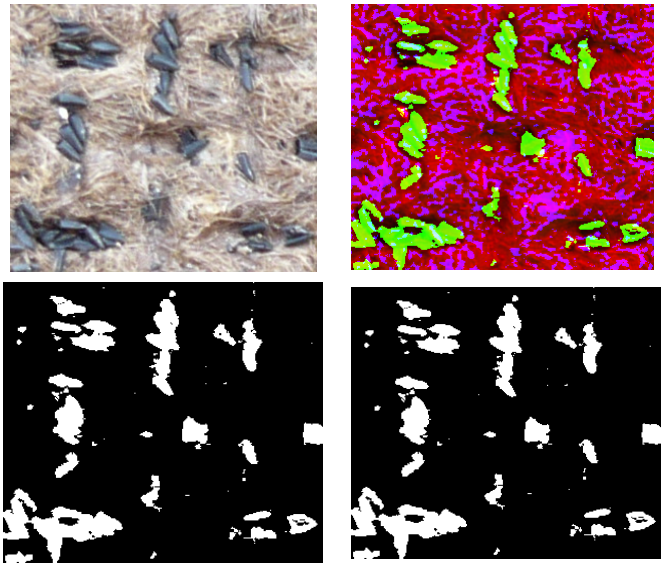
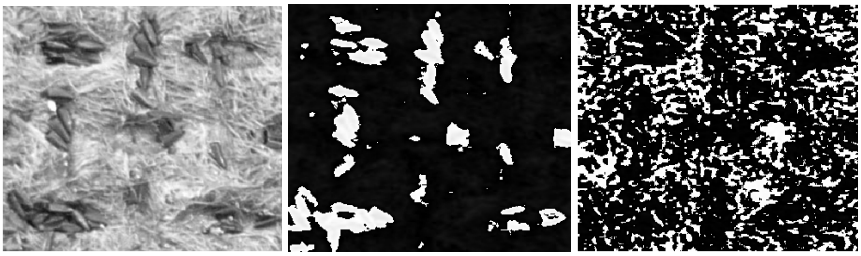


Fig. 7. (top-left) RGB color sub-image of an ovitrap, (top-right) RGB composition of its YIQ representation, and (bottom-left) segmentation of its band I with a fixed limiar of 200 and (bottom-right) a k-means classifier.

4. Discussion and Results

In Table 1, it is presented the results of the two methods applied to another six samples, including an image without eggs. The image labeled as '3' in this Table is the image previously presented in Figure 1-right with 111 eggs.

Figure 8 presents bands of YIQ and Figure 9 shows the histogram of I band. It can be seen that the histogram presents two maximum: one near of zero and other one near of 255 (mosquito eggs).



(a)Y (b) I (c) Q

Fig. 8. YIQ model of sub-image on Figure 7-top-left.

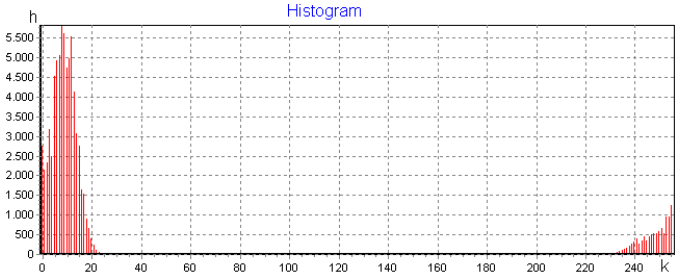


Fig. 9. Histogram of Band I of the RGB sub-image presented in Figure 7-top-left.

Image	Correct Amount of Eggs	Estimated Amount of Eggs by the Proposed Algorithms		
		Method 1	Method 2 Fixed Threshold	Method 2 k-Means
1	22	25	29	20
2	8	10	10	6
3	111	111	113	107
4	30	26	32	28
5	19	19	21	19
6	0	0	0	0

Table 1. Counting results using the proposed methods

The first method reached a maximum error of 25% in the second image where there is a difference of two eggs. But in average the error was about 6.66% which is acceptable in comparison with a non-automatic method.
The version of the second method with segmentation based on unsupervised classification of the YIQ image furnish better results than the other, based on binarization of band I with a fixed threshold of 200. This result is due to the use of all 3 bands of YIQ image, despite the fact that band I is perfectly feasible to be used as input to segmentation algorithm, as can be observed in Figure 7 and Figure 8. The second method using a fixed threshold value for binarization achieved an average error rate of 7.33%, while the use of k-Means produced an average error of 7.84%.
Both methods achieved very satisfactory results. Other color spaces must be tried in search for better responses.

5. Acknowledgments

This research is partially sponsored by FINEP-Brazil (ref. 0787/07, grant 01.08.0396.00). The authors are grateful to Aggeu Magalhães Research Center for the concession of the ovitraps used in this study.

6. References

Perich, M.J., Kardec,A., Braga, I.A., Portal, I.F., Burge,R., Zeichner, B.C., Brogdon, W.A., Wirtz, R.A. Field evaluation of a lethal ovitrap against dengue vectors in Brazil, *Medical and Veterinary Ontomology*, Vol. 17, (2007), pp. 205-210, ISSN 0037-8682.

Dibo, M.R., Chiaravalloti-Neto, F., Battigaglia, M., Mondini, A., Favaro, E.A., Barbosa, A.A.C., Glasser, C.M. Identification of the best ovitrap installation sites for gravid Aedes (Stegomyia) aegypti in residences in Mirassol, state of São Paulo, Brazil, *Mem. Inst. Oswaldo Cruz*, Vol. 100, No. 4, (2005), pp. 339-343.

Regis, L; Souza, W.; Furtado, A.; Fonseca, C.; Silveira , J. C.; Ribeiro, P.; Santos, M.A.M.; Carvalho, M. S.; and Monteiro, A.M.. An Entomological Surveillance System Based on Open Spatial Information for Participative Dengue Control, Anais da Academia Brasileira de Ciências, (2009) (accept for publication).

Regis, L.; Monteiro, A. M.; Santos, M. A. M.; Silveira, J. C.; Furtado, A. F.; Acioli, R. V.; Santos, G. M.; Nakazawa, M.; Carvalho, M. S.; Jr, P. J. R.; and Souza, W. V. Developing new approaches for detecting and preventing Aedes aegypti population outbreaks: bases for surveilance, alert and control system. Memórias do Instituto Oswaldo Cruz, Vol. 103, No. 1, (2008), pp. 50-59.

Gama, R.A., Eiras, A.E., Resende, M.C. Efeito da ovitrampa letal na longevidade de fêmeas de Aedes aegypti (Diptera: Culicidae), *Revista da Sociedade Brasileira de Medicina Tropical*, Vol. 40, No. 6, (November 2007), pp. 640-642, ISSN 0037-8682. (in portuguese)

Parker, J.R. (1997). *Algorithms for Image Processing and Computer Vision*, John Wiley and Sons, ISBN 0471140562, New York.

Dougherty, E.R., and Lotufo, R.A. (2003). *Hands-on Morphological Image Processing*, SPIE Publications, ISBN 081944720X, New Jersey.

Gonzalez, R.C. and Woods, R. (2007). *Digital Image Processing*, Prentice-Hall, 3rd Edition, ISBN 013168728X, New Jersey.

Mello, C.A.B. and Mascaro, A.A. Image Fidelity Index Applied to Digital Mammography Texture Segmentation, *Conferencia Latinoamericana de Informática (CLEI)*, Chile, August 2006, pp. 1-6.

Wang, Z.; Bovik, A.C. (2002). A Universal Image Quality Index. *IEEE Signal Processing Letters*, Vol. 9, No. 3, (March 2002), pp. 81-84, ISSN 10709908.

Ballard, D.H., and Brown, C.M. (1982). *Computer Vision*. Prentice-Hall, ISBN 0131653164, New York.

Huang, L.K., and Wang, M.J. Image Thresholding by Minimizing the Measures of Fuzziness, *Pattern Recognition*, Vol 28, No. 1, (January 1995), pp. 41-51, ISSN 00313203.

Sezgin, M.; Sankur, B. (2004). Survey over image thresholding techniques and quantitative performance evaluation, *Journal of Electronic Imaging*, Vol. 1, No.13, (January 2004), pp. 146-165, ISSN 10179909.

Shapiro, L. and Stockman, G.C. (2001). *Computer Vision*. Prentice Hall, ISBN 0130307963, New Jersey.

Gomes, J., Velho, L., Frery, A. and Levy, S. (2008). *Image Processing for Computer Graphics and Vision*, Springer, ISBN 1848001924, New York.

Haykin, S. (1998). *Neural Networks: A Comprehensive Foundation*, Prentice Hall, ISBN 0132733501, New Jersey.

12

Hyperspectral Imaging: a New Modality in Surgery

Hamed Akbari and Yukio Kosugi
Tokyo Institute of Technology
Japan

1. Introduction

Nowadays medical diagnosis is principally supported by the imaging techniques. Several imaging modalities such as magnetic resonance imaging (MRI), computed tomography (CT), ultrasonography, Doppler scanning, and nuclear imaging have completely expanded medical imaging field. Recently Hyperspectral imaging (HSI), has emerged as a new member of the family of the medical imaging modalities. HSI provides a powerful tool for non-invasive tissue analyses. This technology is able to capture both the spatial and spectral data of an organ or tissue in one snapshot. In other words, the imaging system produces many narrow band images at different wavelengths. Not similar to conventional three-channel color cameras and other filter-based imaging systems, this system captures full neighboring spectral data with spectral and spatial information (Akbari et al., 2008a).

HSI can visualize invisible wavelength regions and bring them to the human vision range. Pervious decades, hyperspectral imaging was a complicated technique that was employed in satellite or aircraft systems. However, this technology has been customized to a compact imaging and spectroscopy tool with potential applications in medicine. In fact, Hyperspectral imaging has already been applied in the medical field. Kellicut et al. utilized HSI to quantitatively evaluate the tissue oxygen saturation in patients with peripheral vascular disease (Kellicut et al., 2004). Khaodhiar et al. employed this imaging technique to predict and follow healing in foot ulcers of diabetic patients (Khaodhiar et al., 2007). HSI was used to diagnose hemorrhagic shock (Cancio et al., 2006), to detect chronic mesenteric ischemia during endoscopy (Friedland et al., 2007), and to detect skin cancer in mice (Martin et al., 2006).

Hyperspectral imaging captures reliable data for the surgeons in the operating room with a convenient instrument. It shows a greater sensitivity for detecting a residual tumor tissue than current surgical tissue sampling techniques (Freeman et al., 2005). Monteiro et al. employed this technique to enhance the regions covered with a layer of blood during surgeries (Monteiro et al., 2006). HSI is utilized to detect ischemic regions of the intestine during surgery (Akbari et al., 2008b). A hyperspectral imaging endoscope is used for the early detection of dysplasia and cancer in lung epithelia (Lindsley et al., 2004). Zuzak et al. coupled a surgical laparoscope for conventional minimally invasive surgical procedures with a near-infrared hyperspectral imaging system to help guide laparoscopic surgeons to visualize biliary anatomy (Zuzak et al., 2007).

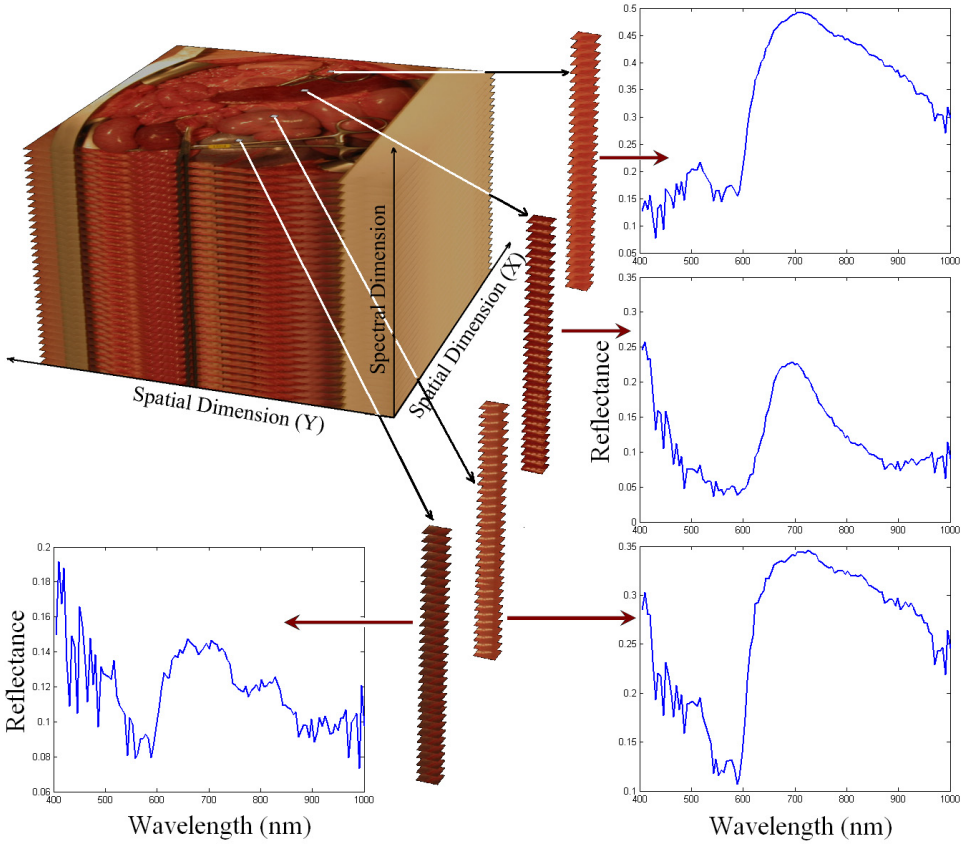


Fig. 1. A schematic view of a hyperspectral image of pig's abdomen is shown. The spectral graph of the average spectrum from the pig's peritoneum, spleen, colon, and urinary bladder are shown in the four graphs. The graph depicts the reflectance for each wavelength in that region.

This chapter presents the application of hyperspectral imaging as a visual supporting tool to detect different organs and tissues during surgeries. Diagnosis of abnormal tissue or tissues which are not in natural anatomic location is an important concern in surgeries. Hyperspectral imaging can be a useful technology for finding ectopic tissues and diagnosis of tissue abnormalities such as intestinal ischemia. The ectopic or heterotopic tissue, particularly when missed, can cause complications in patients. Intestinal ischemia results from a variety of disorders that cause insufficient blood flow to the intestine. The type and prognosis of ischemic injury depend on the blood vessels involved, the underlying medical condition, and the swiftness with which the problem is brought to medical attention for diagnosis and treatment.

2. Hyperspectral Imaging

Hyperspectral imaging captures and analyses data from across the electromagnetic spectrum. This technology extends the human vision that just sees visible light. Hyperspectral imaging can visualize the visible light as well as near-infrared to infrared.
The difference between hyperspectral and multi-spectral imaging is usually defined according to the number of spectral bands. Multi-spectral image contains tens of bands. However, hyperspectral image contains hundreds to thousands of bands. Hyperspectral images are captured by one sensor that captures a set of contiguous bands. However, multi-spectral is a set of spectral bands that are not typically contiguous and can be captured by multiple sensors. Figure 1 shows a schematic view of the hyperspectral image.

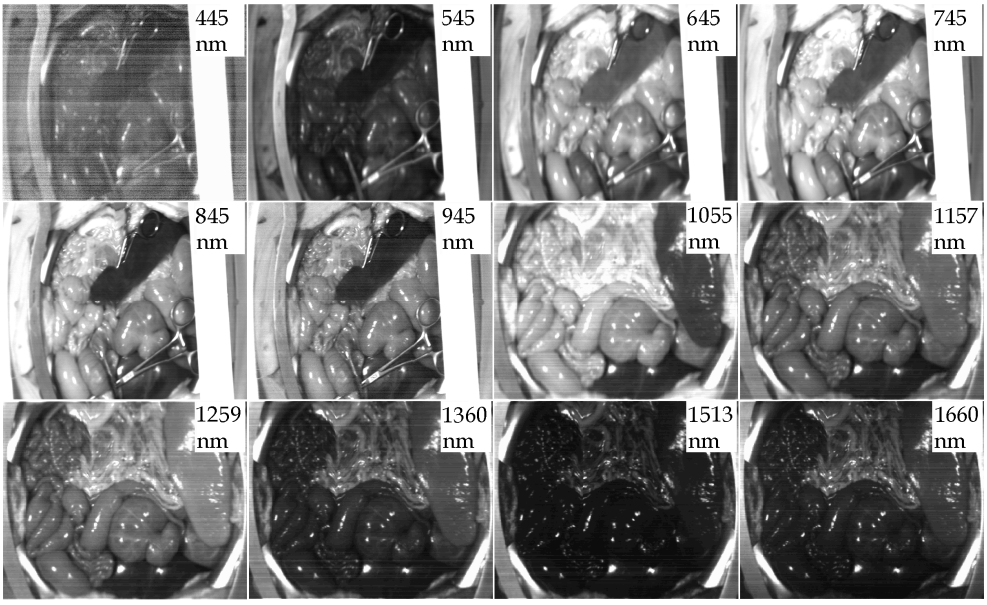


Fig. 2. Several images at different wavelengths during an abdominal surgery on pigs.

Two hyperspectral sensors were used to capture the image data for segmentation of abdominal organs and detection of intestinal ischemic tissue: an ImSpector N17E and a V10E, manufactured by Spectral Imaging Ltd., Oulu, Finland. The hyperspectral systems have more than hundred bands and high spectral resolution. Hyperspectral sensors generate a two-dimensional spatial image along a third spectral dimension. Each pixel in the hyperspectral image has a sequence of reflectance in different spectral wavelengths that can display the spectral signature of that pixel. Since there are a large number of data for each image, artificial neural networks and machine classifiers along with wavelet compression were used to segment the images. The technique was evaluated during the abdominal surgeries on two pigs under general anesthesia. In this research, using the hyperspectral cameras, the spectral signatures for abdominal organs has been created. Using these signatures, the abdominal view through a large incision is segmented. Figure 2 shows the

images in different wavelengths that are captured using two cameras in a large-incision view during the abdominal surgery on the pigs.

2.1 Hyperspectral sensors

The hyperspectral sensors are instruments for capturing many images in different adjacent wavelengths of an illuminated region corresponding to the entrance slit. The main components of a hyperspectral camera are shown in Figure 3. As light sources, two halogen lamps illuminate the object to be captured. The camera's objective lens collects the radiation from the object and projects an image on the entrance slit plane. The slit determines the instantaneous field of imaging in spatial directions. The radiation from the slit is projected to the prism-grating-prism (PGP) components (ImSpector optics). Therefore, the propagation angle of the radiation changes depending on its wavelength. Then it is focused on the matrix detector. Every object's point is represented on the matrix detector by a series of monochromatic points that makes a continuous spectrum in the direction of the spectral axis. For example, an instantaneous image of the AB line is captured as the lines A1B1, A2B2,…, AnBn, where "n" is the wavelength band number.

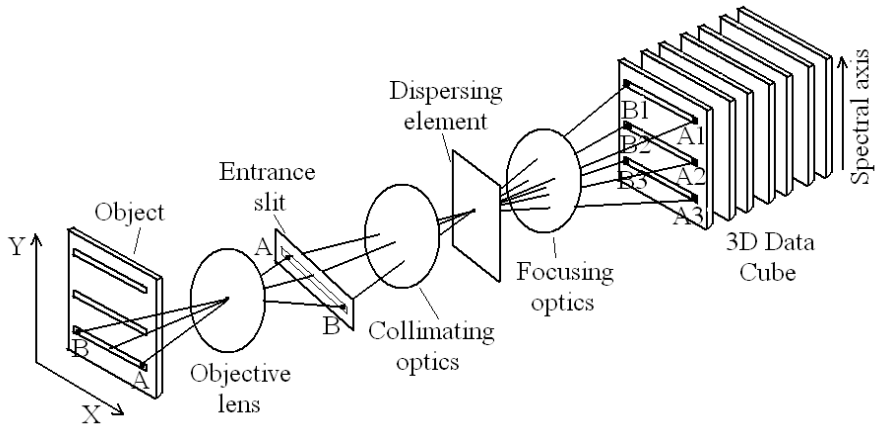


Fig. 3. Design of the hyperspectral imaging sensor (Aikio, 2001)

Whisk-broom scanning, filtered imaging, and push-broom are three main designs for hyperspectral cameras. Whisk-broom scanner captures the spectral dimension pixel-to-pixel. Filtered camera captures two spatial dimensions and temporally samples the spectral dimension. The capturing technique of ImSpector sensors is a push-broom scanning. In this type of imaging spectrometer, the entrance slit limits the imaging field. The 2D detector matrix instantaneously captures the spectral dimension and one spatial dimension. The second spatial dimension is generated by scanning the object. By moving the camera's field of view relative to the object, the second spatial dimension is created. Figure 4 shows a schematic of the imaging techniques.

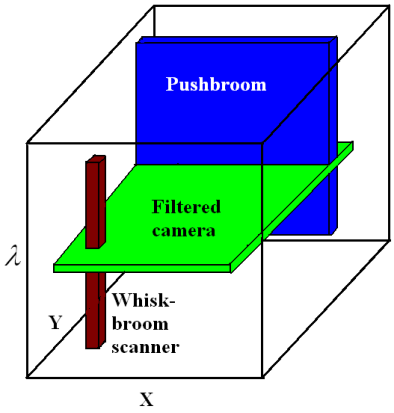


Fig. 4. A schematic of the hyperspectral image capturing techniques (Aikio, 2001)

The PGP is composed of a special grating, two prisms, and an aperture stop. The special grating is located between two prisms and the aperture stop is set in contact with the grating. Short and long pass filters are usually placed between the grating and the prisms, eliminating unwanted wavebands and changing the spectral response. In this technique, since the filters are incorporated inside a PGP, the reflections from their surfaces can be eliminated. Unlike a direct vision prism where the dispersion is a non-linear small dispersion, the diffraction grating in the PGP supplies a large linear dispersion (Aikio, 2001). Using two cameras, the ImSpector N17E and the V10E, the wavelength range of 400 - 1700 nm may be captured. The V10E model captures the spectral range of 400 - 1000 nm, a dispersion of 97.5 nm/mm, and a spectral resolution of 5 nm (with a 30 µm slit). The N17E sensor captures the spectral range of 900 - 1700 nm, a dispersion of 110nm/mm, and a spectral resolution of 5 nm (with a 30 µm slit). All the wavelengths will be passed for only the small region of the object that is exactly in front of the entrance slit. By shifting the sensor between subsequent images, ultimately all parts of the object and all corresponding wavelengths are captured. Therefore, for each wavelength, a monochromatic spectral image can be constructed from the hyperspectral image set.

2.2 Capturing setup

The ImSpector sensors capture the images using the push-broom scanning technique. Therefore, to generate the second spatial dimension the object must be scanned i.e. the second spatial dimension is captured by moving the camera's field of view relative to the object. The linear actuator, a ROBO Cylinder Slider (model RCS-SM-A-100-H-1000-T1-S), is used to move the camera. This actuator is controlled by an XSEL-J-1-100A-N3-EEE-2-1 type controller. The actuator and controller are manufactured by IAI Corporation, Japan. The actuator works with a ball screw drive system, a 100 W motor, and an absolute incremental encoder. The actuator has an 84.9 N rated thrust, a 1-1000 mm/s speed, a ±0.02 mm positioning repeatability, and a maximum backlash of 0.05 mm. The actuator is connected to the controller by two cables: the encoder cable and the motor cable. The movement and velocity are adjusted by a setting tool that is connected to the controller. The actuator moves the camera with a constant velocity.

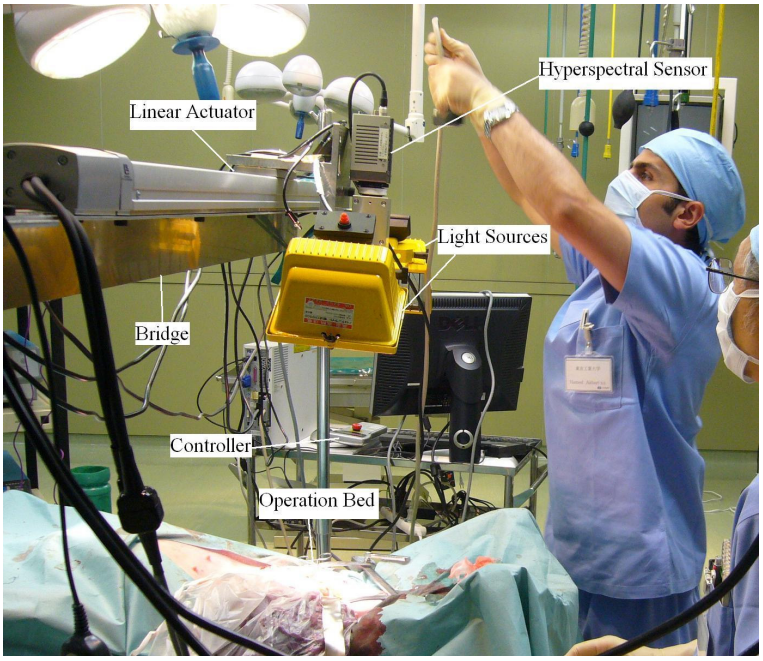


Fig. 5. The acquisition setup

The acquisition setup consists of a pair of 500 W halogen lamps with diffusing reflectors as the light sources and the computer-controlled linear actuator. The linear actuator is fixed on a bridge installed over the surgical bed and the camera has been calibrated and fixed on the frame. Therefore, the distance between the lens and the abdomen is constant and a fairly uniform illumination of the subject is provided by using the two halogen lamps. Figure 5 shows the acquisition setup.

2.3 Data normalization

The captured data should be normalized to treat the spectral non-uniformity of the illumination device. The raw data are changed by illumination conditions and temperature. Therefore, the radiance data were normalized to yield the radiance of the specimen.

White reference and dark current are two data that should be captured for normalization. White reference is the spectrum acquired by the hyperspectral sensor corresponding to the white reference and dark current is a dark image acquired by the system in the absence of light. Figure 6 shows a spectral signature and corresponding white reference and dark current. White reference is used to show the maximum reflectance in each wavelength. Dark current spectroscopy is used to address the defects by calculating the peaks in the dark current spectrum with temperature. To perform this pre-processing step, the radiance of a standard reference white board placed in the scene and the dark current are measured by keeping the camera shutter closed.

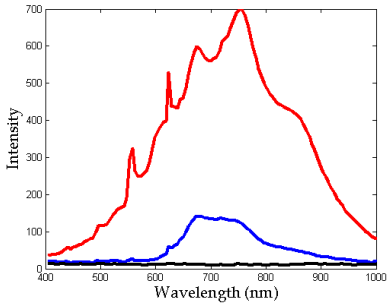


Fig. 6. A spectral signature in blue and corresponding white reference in red and dark current in black.

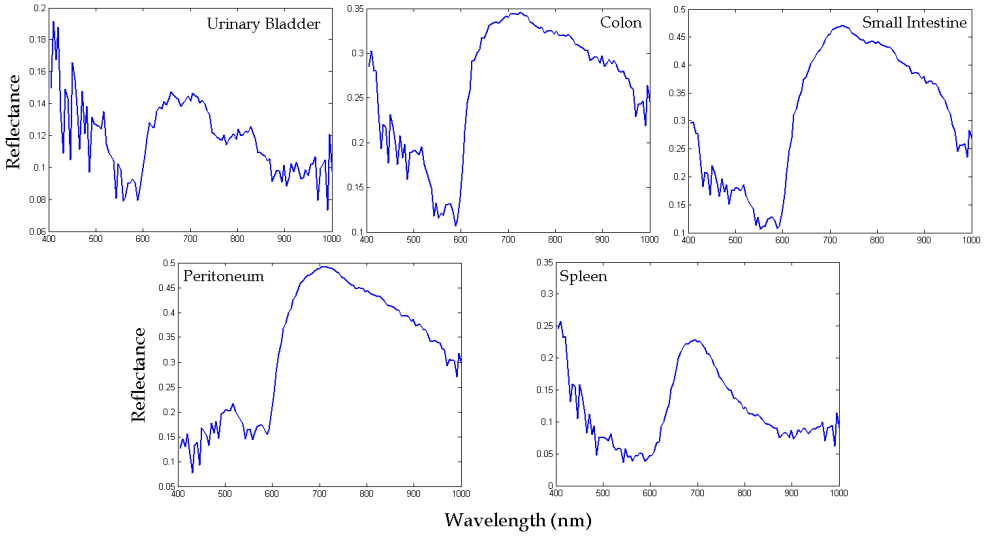


Fig. 7. Reflectance spectra using visible and near infrared camera: the horizontal axis shows different wavelengths in nanometers, and the vertical axis shows the reflectance.

Then the raw data are corrected to reflectance using the following equation:

$$R(\lambda) = \frac{I_{raw}(\lambda) - I_{dark}(\lambda)}{I_{white}(\lambda) - I_{dark}(\lambda)} \quad (1)$$

where $R(\lambda)$ is the calculated reflectance value for each wavelength, $I_{raw}(\lambda)$ is the raw data radiance value of a given pixel, and $I_{dark}(\lambda)$ and $I_{white}(\lambda)$ are, respectively, the dark current and the white board radiance acquired for each line and spectral band of the sensor. The dark current and white reference in the ImSpector N17E sensor is separately captured and included in the main *.raw data file. However, in the ImSpector V10E, it was captured in a separate file in *.drk format. The white reference board should be placed in the

capturing field when the ImSpector V10E is used. However, the dark current should be captured separately. Figure 7 and Figure 8 show the reflectance spectra of the abdominal organs.

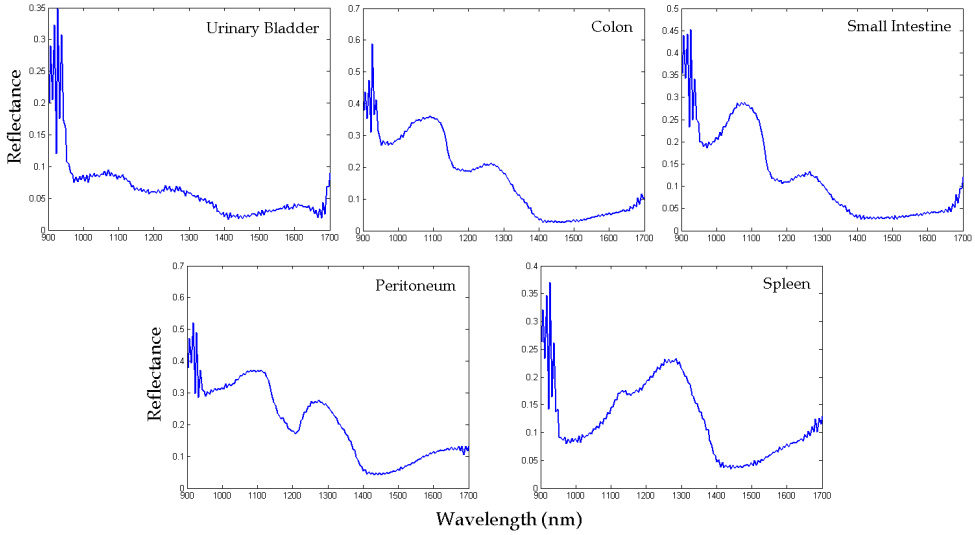


Fig. 8. Reflectance spectra using near infrared and infrared camera: the horizontal axis shows different wavelengths in nanometers, and the vertical axis shows the reflectance.

3. Segmentation of Abdominal Organs

Due to the ambiguity between the organ and its adjacent tissues, it is difficult to segment the organs and tissues during surgeries. Due to the movements of the object, dynamic situations such as in live and/or moving subjects will worsen the detection (Liu et al., 2007). In special situations such as anatomic variations, ectopic tissues, and tissue abnormalities, this problem becomes more challenging. Hyperspectral imaging is used to segment the abdominal organs during the surgeries on two pigs. Two approaches are utilized to classify the hyperspectral data. In the first approach, the data are compressed via wavelet decomposition then classified using learning vector quantization (LVQ) (Akbari et al., 2008a). In the second approach, the data are classified by a support vector machine (SVM) (Akbari et al., 2009).

3.1 Normalized difference indexes

Hyperspectral images may be visualized in a real-time manner using the normalized difference index (NDI). This is a simple method to enhance organs or tissues in hyperspectral data. NDI has been employed in many research studies to estimate chlorophyll content (Richardson et al. 2002), to evaluate the effects of nitrogen fertilization treatments (Moran et al. 2000), to estimate water content (Datt et al., 2003), and to estimate the yields of salt- and water-stressed forages (Poss et al., 2006).

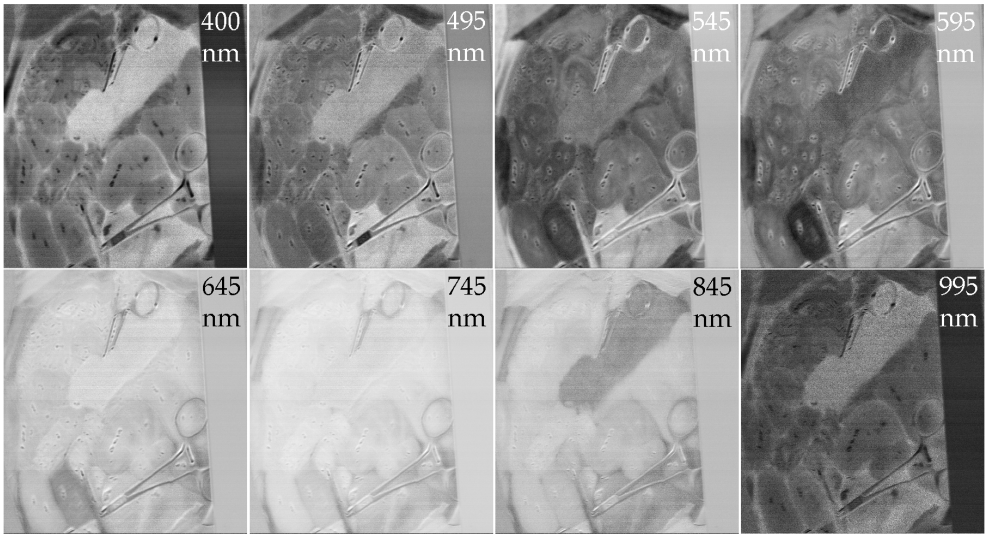


Fig. 9. Eight sample images using the proposed NDI at different wavelengths using visible and near infrared camera (400-1000 nm).

Many combinations of the reflectance and intensity were evaluated to find the appropriate NDI. Each NDI can enhance one or several organs. Several combinations of wavelengths were selected to enhance the difference of organs or tissues. The following equation is applied to calculate the NDI in the hyperspectral data in 400-1000 nm:

$$\mathrm{NDI}(\lambda) = \frac{\mathrm{I}(\lambda) - \mathrm{I}(945\mathrm{nm})}{\mathrm{I}(\lambda) + \mathrm{I}(945\mathrm{nm})} \tag{2}$$

where $\mathrm{NDI}(\lambda)$ is the normalized difference index in the wavelength λ and $\mathrm{I}(\lambda)$ is the intensity of a given pixel in the wavelength λ. Figure 9 shows this normalized difference index images in some sample wavelengths. The equation that is utilized to calculate the NDI in 900-1700 nm hyperspectral data is as follows:

$$\mathrm{NDI}(\lambda) = \frac{\mathrm{R}(\lambda) - \mathrm{R}(1660\mathrm{nm})}{\mathrm{R}(\lambda) + \mathrm{R}(1660\mathrm{nm})} \tag{3}$$

where $\mathrm{NDI}(\lambda)$ is the normalized difference index in the wavelength λ and $\mathrm{R}(\lambda)$ is the intensity of a given pixel in the wavelength λ. Figure 10 shows this normalized difference index images in some sample wavelengths. Although this technique is a fast method for visualization, it does not result in precise segmentation in the image processing. Therefore, for the image segmentation, the hyperspectral data were processed by the image processing techniques.

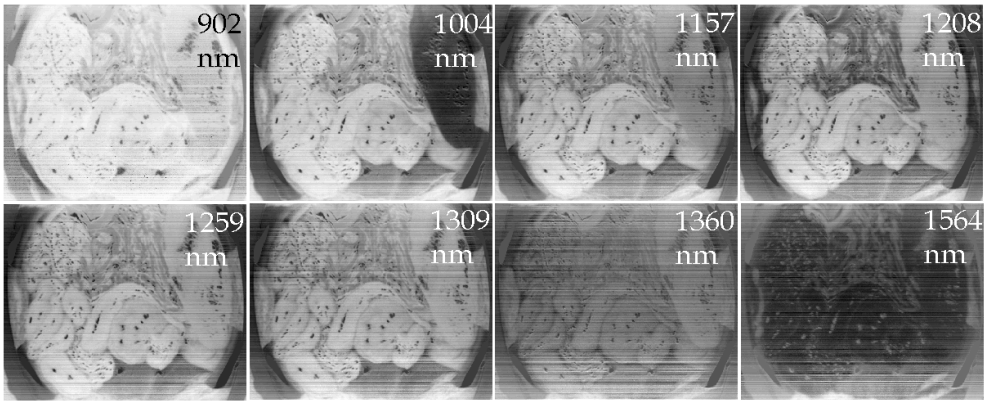


Fig. 10. Eight sample images using the proposed NDI at different wavelengths using near infrared and infrared camera

3.2 Wavelet compression and LVQ classification

Since there is a large quantity of data for each image, it is better to compress the data before processing. In this study, a wavelet transform is used for data compression and LVQ is used to segment the image. Wavelet transform may be used as a type of signal compression for compressing the spectral data. The elements of a signal can be represented by a smaller amount of data. The wavelet transform produces as many coefficients as there are data in the signal, then these coefficients can be compressed. The information is statistically concentrated in just a few coefficients. The wavelet compression is based on the concept that the regular signal component can be accurately approximated using a small number of approximation coefficients and some of the detail coefficients (Chui, 1993; Daubechies, 1992).

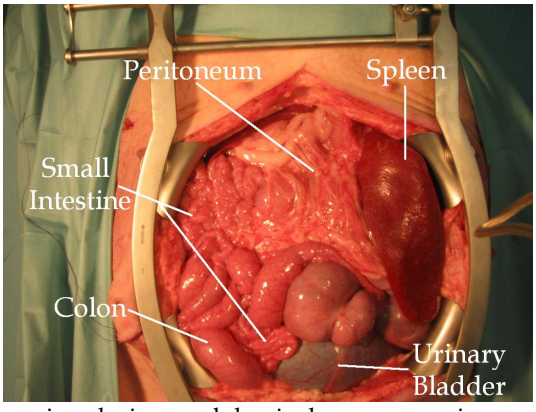


Fig. 11. A large-incision view during an abdominal surgery on pigs

Self-organizing networks can learn to detect regularities and correlations in their input and adapt their future responses to that input accordingly. The neurons of competitive networks

learn to recognize groups of similar input vectors. LVQ is a method for training competitive layers in a supervised manner (Kohonen, 1987). The wavelet-based compressed spectral signatures are the input vectors. The abdominal organs are assigned to be the output of the neural network. The input vectors are correlated to one of seven classes corresponding to the spleen, peritoneum, urinary bladder, small intestine, colon, background, and ambiguous regions. After classification, the pixels which were detected as ambiguous pixels were labeled in the post-processing steps. Figure 11 shows a large-incision view during an abdominal surgery on a pig.

3.3 Support vector machines (SVMs)

Hyperspectral image classification using SVMs has shown superior performance to the other available classification methods (Camps-Valls & Bruzzone, 2005) (Camps-Valls et al., 2004) (Melgani & Bruzzone, 2004) (Huang et al., 2002) (Brown et al., 2000). Multilayer perceptron (MLP) and radial basis function neural networks (RBFNNs) are successful approaches to classify hyperspectral data. However, the high number of spectral bands results in the Hughes phenomenon (Hughes, 1968). Support vector machines (SVMs) can efficiently handle large input spaces or noisy samples (Camps-Valls & Bruzzone, 2005). SVMs use a small number of exemplars selected from the tutorial dataset to enhance the generalization ability. The SVMs are supervised classifiers that have a pair of margin zones on both sides of the discriminate function. The SVM is a popular classifier based on statistical learning theory as proposed by Vapnik (Vapnik, 1995; Brown et al., 2000). The training phase tries to maximize the margin of hyperplane classifier with respect to the training data.

Since the spectral data are not linearly separable, the kernel method is used. Kernel-based methods map data from an original input feature space to a kernel feature space of a higher dimensionality and then solve a linear problem in that space. The Least Squares SVM (LS-SVM), a new version of the SVM, is used for classification (Bao & Liu, 2006; Camps-Valls & Bruzzone, 2005; Liu et al., 2007). A convex quadratic program (QP) solves the classification problem in the SVMs. In LS-SVMs, instead of inequality constraints, a two-norm with equality is applied (Suykens & Vandewalle, 1999). Therefore, instead of a QP problem in dual space, a set of linear equations is obtained. The SVM tries to find a large margin for classification. However, the LS-SVM looks for a ridge regression for classification with binary targets. The selection of hyperparameters is not as problematic and the size of the matrix involved in the QP problem is also directly proportional to the number of training points (Van Gestel et al., 2004). The optimization function of the SVM is modified as follows:

$$\underset{w,b,e}{\text{Min}}\ f(w,e)=\frac{1}{2}w^{T}w+\gamma\frac{1}{2}\sum_{i=1}^{N}e_{i}^{2} \tag{4}$$

subject to the equality constraints

$$y_{i}[w^{T}\varphi(x_{i})+b]=1-e_{i}, \qquad i=1,...,N \tag{5}$$

where w is the weighting vector, b is the bias term, e is for misclassifications, and γ is the tuning parameter. This constrained optimization problem can be solved by determining the saddle points in the Lagrange functional as:

$$L(w,b,e;\alpha) = f(w,b,e) - \sum_{i=1}^{N} \alpha_i \{y_i [w^T \varphi(x_i) + b] - 1 + e_i\} \quad (6)$$

where $\alpha_i \in R$ are Lagrange multipliers that can be positive or negative in the LS-SVM formulation. It is possible to choose many types of kernel functions including linear, polynomial, radial basis function (RBF), multilayer perceptron (MLP) with one hidden layer, and splines. The RBF kernel used in this study was as follows:

$$K(x, x_i) = \exp\{-\|x - x_i\|_2^2 / \sigma^2\} \quad (7)$$

where σ is constant.

Multi-class categorization problems are represented by a set of binary classifiers. To prepare a set of input/target pairs for training, 100 pixels of data from each region in the surgical hyperspectral images are captured. The SVMs are applied one by one to the image for each class, and each pixel was labeled as an organ (Akbari et al., 2009).

3.4 Experimental results

The experiment was done on two pigs under general anesthesia. A large incision was created on the abdomen, and the internal organs were explored. Vital signs were evaluated during the surgery to assure constant oxygen delivery to the organs. Nine hyperspectral images by the ImSpector N17E and seven hyperspectral images by the ImSpector V10E were captured. The actuator velocity was set such that the resolutions of the two spatial dimensions were equal. The performance (i.e. the quality of detection) was evaluated with respect to the hand-created maps produced by a medical doctor and by using anatomical data.

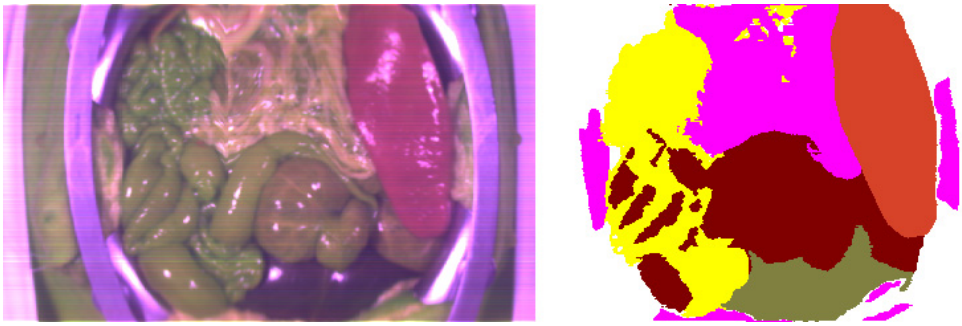


Fig. 12. The RGB image is made using three channels of near-infrared and infrared hyperspectral camera (900-1700 nm) is shown on the left side. Using LVQ method, the segmented image can be viewed on the right side. Spleen is shown in red, peritoneum in pink, urinary bladder in olive, colon in brown, and small intestine in yellow (Akbari et al., 2008a).

The hand-created maps were used as reference maps in calculating the detection rates of the method. Performance criteria for organ or tissue detection were the false negative rate (FNR) and the false positive rate (FPR), which were calculated for each organ. When a pixel was

not detected as an organ or tissue pixel, the detection was considered a false negative if the pixel was a pixel of that organ on the hand-created map. The FNR for an organ was defined as the number of false negative pixels divided by the total number of the organ pixels on the hand-created map. When a pixel was detected as an organ pixel, the detection was a false positive if the pixel was not an organ pixel on the hand-created map. The FPR was defined as the number of false positive pixels divided by the total number of non-organ pixels on the hand-created map. The pixels that were ambiguous and that the medical doctor could not label as an organ were not considered in our calculation. Figure 12 shows a segmented image using the LVQ method. The numerical results of the FPR and FNR for each organ and a comparison between LVQ and SVM methods (Akbari et al., 2008a; Akbari et al., 2009) are given in Table 1.

Camera & method	Organs	Spleen	Urinary Bladder	Peritoneum	Colon	Small Intestine
V10E (SVM)	FPR	3.9%	3.7%	5.3%	5.1%	8.7%
	FNR	4.5%	5.6%	7.3%	6.4%	7.2%
N17E (SVM)	FPR	1.1%	1.2%	4.3%	1.2%	7.3%
	FNR	1.3%	0.7%	5.1%	9.5%	2.7%
N17E (LVQ)	FPR	0.5%	1.3%	6.3%	1.2%	12.3%
	FNR	1.3%	1.4%	7.1%	15%	2.7%

Table 1. The evaluation results and comparison (Akbari et al., 2008a; Akbari et al., 2009)

The peritoneum has the highest value in visible and invisible wavelengths. The higher fat content of this tissue could be a possible explanation. In most spectral regions, the colon has the second highest reflectance value, after the peritoneum. In the colon, the adventitia forms small pouches filled with fatty tissue along the colon. The special histology and the fact that the urinary bladder is hollow inside, can explain the lowest spectral reflectance measured for this organ (Junqueira and Carneiro, 2005).

4. Intestinal Ischemia

Intestinal ischemia results from a variety of disorders that cause insufficient blood flow to the intestinal tract. The intestine like other live organs requires oxygen and other vital substances. These essential substances are delivered by arteries and carbon dioxide and other disposable substances are removed by veins. Intestinal ischemia results from decreasing the blood flow of the intestine to a critical point that delivery of oxygen is compromised. This problem results in intestinal dysfunction and ultimately necrosis. The prognosis of ischemic injuries depends on the quickness that the problem is brought to medical attention for diagnosis and treatment (Rosenthal & Brandt, 2007). Ischemia can be regional and limited to a small part of the intestine, or it may be more extensive. The intestinal ischemia may result from a shortage in blood passage through an artery or vein. There are several ways in which arterial or venous flows can be restricted: an embolus, a thrombus, or a poor blood flow through an artery or vein because of spasm in the blood vessel or clinical interventions (Rosenthal & Brandt, 2007).
Hyperspectral imaging may provide reliable data in near real-time with a convenient device for the surgeon in the operating room to diagnose the intestinal ischemia. In this section,

using the hyperspectral camera (900-1700nm), the spectral signatures for intestine, ischemic intestine and abdominal organs have been created. Using these signatures, the abdominal view through a large incision is segmented. Wavelet transform is used as the compression method and the SVM is used for classification.

4.1 Material and methods

ImSpector N17E is used to capture the hyperspectral data. The data are normalized to address the problem of spectral non-uniformity of the illumination device and influence of the dark current. The image digital numbers are normalized to yield the radiance of the specimen. The white reference and dark current were measured and raw data was normalized to these values as described in section 2.3.

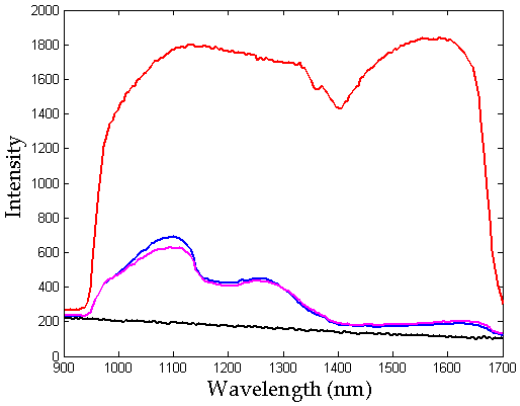


Fig. 13. The spectral signature of normal intestine, ischemic intestine, white reference, and dark current are shown in magnet, blue, red, and black, respectively.

The hyperspectral data are compressed using wavelet transform. Then the normal and ischemic loops of the intestine are segmented using SVM. The comparison of the spectral signatures of normal and ischemic regions of the intestine demonstrates a maximum difference in 1029-1136nm (see Figure 13). Since the main difference between normal and ischemic intestine is in the mentioned wavelength region, for discriminating the normal and ischemic tissues, these twenty two bands are used without compression. Some pixels which were lost because of glare are detected in post-processing. Since most of missed pixels were located at the mid portion of organs an image fill function is utilized as a post processing step. The hyperspectral images are compressed using wavelet transform. Each spectral signal is decomposed choosing the db3 (Daubechies-3) wavelet with level 2 compression (i.e. 1/4 compression). The compressed data are classified using SVM. Since the training data are not linearly separable, the kernel method is used in the study. The wavelet-based compressed pixel signatures are the input of SVM, and each input vector is to be assigned to one of two classes (intestine and non-intestine). In the next step, twenty two elements (1029-1136nm bands) of the original spectral data are the input vectors, and each input vector is to be assigned to one of ischemic or normal classes.

4.2 Experimental results

To perform the experiment, a pig was anesthetized. A large incision was created on the abdomen and intestine and other abdominal organs were explored. Vital signs were controlled during the surgery to guarantee a fairly constant oxygen delivery to the organs. An intestinal segment and the vessels supplying this segment were clamped for 6 minutes and the image was captured. The ImSpector N17E is fixed on the computer controlled linear actuator that was installed on a bridge over the surgical bed. The performance of the method was evaluated for detection of intestine and ischemic intestine. The evaluation was performed for the quality of detection in respect to hand-created maps. The hand-created maps are used as the reference maps in calculating the detection rates of the method. Performance criteria for intestine and ischemic intestine detection are false negative rate (FNR) and false positive rate (FPR). Figure 14 shows the ischemic intestinal pixels that are detected using the proposed method.

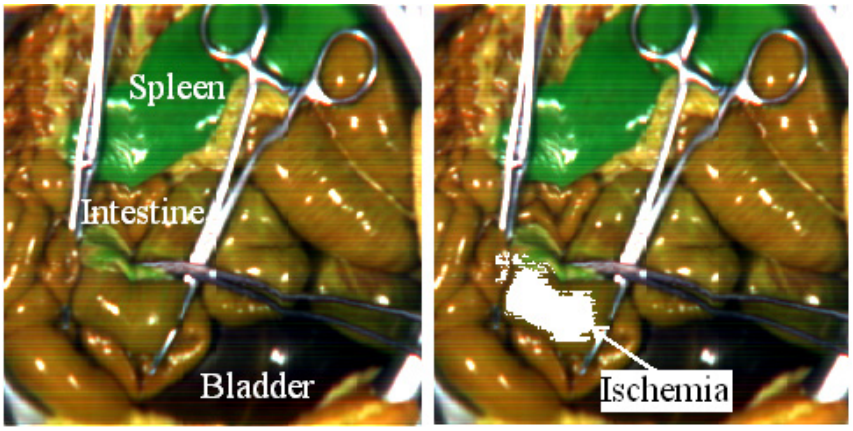


Fig. 14. An RGB image is made using three channels of the hyperspectral image. The detected ischemic intestinal tissue via the proposed method is shown with white (Akbari et al., 2008b).

In the first step, the algorithm detects intestinal pixels. When a pixel is not detected as an intestine pixel, the detection is a false negative if the pixel is a pixel of intestine on the hand-created map. FNR is defined as the number of false negative pixels divided by the total number of the non-intestine pixels on the hand-created map. When a pixel is not detected as an ischemic intestine pixel, the detection is a false negative if the pixel is a pixel of ischemic intestine on the hand-created map. FNR is defined as the number of false negative pixels divided by the total number of the normal intestine pixels on the hand-created map. In the second step, the ischemic intestinal pixels are detected. When a pixel is detected as an intestine pixel, the detection is a false positive if the pixel is not an intestine pixel on the hand-created map. FPR is defined as the number of false positive divided by the total number of intestine pixels on the hand-created map. When a pixel is detected as an ischemic intestine pixel, the detection is a false positive if the pixel is not an ischemic intestine pixel on the hand-created map. FPR is defined as the number of false positive divided by the total number of ischemic intestine pixels on the hand-created map. The ambiguous pixels that the

medical doctor can not label are eliminated in the calculation. The numerical results are given in Table 2 (Akbari et al., 2008b).

	Intestine	Ischemic Intestine
FPR	4.3%	2.3%
FNR	2.7%	9.7%

Table 2. The evaluation results of intestinal tissue and ischemic intestinal tissue detection (Akbari et al., 2008b).

5. Conclusions

This chapter described a new imaging method of hyperspectral imaging as a visual supporting tool during surgeries. Spectral signatures of various organs are presented and difference between normal and ischemic intestinal tissues is extracted. Large quantities of data in hyperspectral images can be processed to extend the range of wavelengths from visible to near infra and infra red wavelengths. This extension of the surgeon's vision would be a significant breakthrough. Capturing and visualizing the optical data of human organs and tissues can provide useful information for physicians and surgeons. This previously unseen information can be analyzed and displayed in an appropriate visual format. Hyperspectral imaging allows surgeons to less invasively examine a vast area without actually touching or removing tissue. A merit of this technique is the ability to both spatially and spectrally determine the differences among variant tissues or organs in surgery. The image-processing algorithms can incorporate detailed classification procedures that would be used for region extraction and identification of organs or tissues. Utilizing this technology in surgery will allow a novel exploration of anatomy and pathology, and may offer hope as a new tool for detection of tissue abnormalities.

6. References

Aikio, M. (2001). *Hyperspectral prism-grating-prism imaging spectrograph*, Espoo, Technical Research Centre of Finland, VTT publications, ISBN 951–38–5850–2, Finland

Akbari, H.; Kosugi, Y.; Kojima, K. & Tanaka, N. (2009). Hyperspectral image segmentation and its application in abdominal surgery, *International Journal of Functional Informatics and Personalized Medicine*, (Vol. 2, No. 2, pp. 201-216, ISSN (Online) 1756-2112 - ISSN (Print) 1756-2104)

Akbari, H.; Kosugi, Y.; Kojima, K. & Tanaka, N. (2008a). Wavelet-based Compression and Segmentation of Hyperspectral Images in Surgery, *Springer Lecture Notes in Computer Science (LNCS)*, Vol. 5125, pp. 142-149, ISSN 0302-9743

Akbari, H.; Kosugi, Y.; Kojima, K. & Tanaka, N. (2008b). Hyperspectral Imaging and Diagnosis of Intestinal Ischemia, *Proceedings of the 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 1238-1241, ISBN 978-1-4244-1814-5, Canada, August 2008, Vancouver

Bao, Y. & Liu Z. (2006). A fast grid search method in support vector regression forecasting time series, *LNCS 4224*, pp. 504-511, ISSN 0302-9743

Brown, M.; Lewis, H.G. & Gunn, S.R. (2000). Linear Spectral Mixture Models and Support Vector Machines for Remote Sensing, *IEEE Trans. Geosci. Remote Sens.* Vol. 38, No. 5, pp. 2346-2360, ISSN 0196-2892

Camps-Valls, G. & Bruzzone, L. (2005). Kernel-based methods for hyperspectral image classification, *IEEE Trans. Geosci. Remote Sens.,* Vol. 43, pp. 1351–1362, ISSN 0196-2892

Camps-Valls, G.; Gomez-Chova, L.; Calpe-Maravilla, J.; Martin-Guerrero, J. D.; Soria-Olivas, E.; Alonso-Chorda, L. & Moreno, J. (2004). Robust support vector method for hyperspectral data classification and knowledge discovery, *IEEE Trans. Geosci. Remote Sens.*, Vol. 42, No. 7, pp. 1530–1542, ISSN: 0196-2892

Cancio, L.C.; Batchinsky, A.I.; Mansfield, J.R.; Panasyuk, S.; Hetz, K.; Martini, D.; Jordan, B.S.; Tracey, B. & Freeman, J.E. (2006). Hyperspectral Imaging: A New Approach to the Diagnosis of Hemorrhagic Shock, *J. Trauma-Injury Infect. Crit. Care*, Vol. 60, No. 5, pp. 1087-1095, ISSN 1079-6061

Chui, C.K. (1993). *Wavelets: a tutorial in theory and applications,* Academic Press Professional, Inc., ISBN 0-12-174590-2, San Diego

Datt, B.; McVicar, T.R.; van Niel, T.G.; Jupp, D.L.B. & Pearlman, J.S. (2003). Preprocessing EO-1 hyperion hyperspectral data to support the application of agricultural indexes, *IEEE Trans. Geosci. Remote Sens.* Vol. 41, pp. 1246-1259, ISSN 0196-2892

Daubechies, I. (1992). *Ten lectures on wavelets*. Society for Industrial and Applied Mathematics, ISBN 0-89871-274-2, Philadelphia

Freeman, J.E.; Panasyuk, S.; Rogers, A.E.; Yang, S. & Lew, R. (2005). Advantages of intraoperative medical hyperspectral imaging (MHSI) for the evaluation of the breast cancer resection bed for residual tumor, *J. Clin. Oncol.*, Vol. 23, No. 16S, Part I of II (June 1 Supplement), p. 709, ISSN 1527-7755

Friedland, S.; Benaron, D.; Coogan, S.; Sze, D.Y. & Soetikno, R. (2007). Diagnosis of chronic mesenteric ischemia by visible light spectroscopy during endoscopy, *Gastrointestinal Endoscopy*, Vol. 65, No. 2, pp. 294-300, ISSN 0016-5107

Huang, C.; Davis, L. S.& Townshend, J. R. (2002). An assessment of support vector machines for land cover classification, *Int. J. Remote Sens.*, Vol. 23, No. 4, pp. 725–749, ISSN 0143-1161

Hughes, G. E. (1968). On the mean accuracy of statistical pattern Recognizers, *IEEE Trans. Inf. Theory*, Vol. 14, pp. 55–63, ISSN 0018-9448

Junqueira, L.C. & Carneiro, J. (2005). *Basic Histology: Text & Atlas*, McGraw-Hill Companies, ISBN-10 0071378294, USA

Kellicut, D.C.; Weiswasser, J.M.; Arora, S.; Freeman, J.E.; Lew, R.A.; Shuman, C.; Mansfield, J.R.& Sidawy, A.N. (2004). Emerging Technology: Hyperspectral Imaging, *Perspectives in Vascular Surgery and Endovascular Therapy*, Vol. 16, No. 1, pp. 53-57, ISSN 1531-0035

Khaodhiar, L.; Dinh, T.; Schomacker, K.T.; Panasyuk, S.V.; Freeman, J.E.; Lew, R.; Vo, T.; Panasyuk, A.A.; Lima, C.; Giurini, J.M.; Lyons, T.E.& Veves, A (2007). The Use of Medical Hyperspectral Technology to Evaluate Microcirculatory Changes in Diabetic Foot Ulcers and to Predict Clinical Outcomes, *Diabetes Care*, Vol. 30, No. 4, pp. 903-910, ISSN 1935-5548

Kohonen, T. (1987). *Self-Organization and Associative Memory*, Springer-verlag, ISBN 0-387-18314-0 2nd ed., Newyork

Lindsley, E.H.; Wachman, E.S. & Farkas, D.L. (2004). The hyperspectral imaging endoscope: a new tool for in vivo cancer detection, *Proceedings of the SPIE*, Vol. 5322, pp. 75-82, ISSN 0277-786X, USA, January 2004, San Jose

Liu, Z.; Yan, J.; Zhang, D. & Li, Q. (2007). Automated tongue segmentation in hyperspectral images for medicine, *Appl. Optics,* Vol. 46, No. 34, pp. 8328-8334, ISSN 0003-6935

Martin, M.E.; Wabuyele, M.B.; Chen, K.; Kasili, P.; Panjehpour, M.; Phan, M.; Overholt, B.; Cunningham, G.; Wilson, D.; Denovo, R.C.& Vo-dinh, T. (2006). Development of an advanced hyperspectral imaging (HSI) system with applications for cancer detection, *Annals of Biomedical Engineering*, Vol. 34, No. 6, pp. 1061–1068, ISSN 1521-6047

Melgani, F. & Bruzzone, L. (2004). Classification of hyperspectral remote sensing images with support vector machines, *IEEE Trans. Geosci. Remote Sens.*, Vol. 42, No. 8, pp. 1778–1790, ISSN 0196-2892

Monteiro, S.T.; Uto, K.; Kosugi, Y.; Kobayashi, N. & Watanabe, E. (2006). Optimization of infrared spectral manipulation for surgical visual aid, *Journal of Japan Society of Computer Aided Surgery*, Vol. 8, No. 1, pp. 33-38, ISSN 1344-9486

Moran, J.A.; Mitchell, A.K.; Goodmanson, G & Stockburger, K.A. (2000). Differentiation among effects of nitrogen fertilization treatments on conifer seedlings by foliar reflectance: a comparison of methods, *Tree Physiol.*, Vol. 20, pp. 1113–1120, ISSN 1758-4469

Poss, J. A.; Russell, W. B. & Grieve, C.M. (2006). Estimating yields of salt- and water-stressed forages with remote sensing in the, visible and near infrared, *J. Environ. Qual.*, Vol. 35, pp. 1060-1071, ISSN 1537-2537

Richardson, A.D.; Duigan S. P. & G. P Berlyn (2002). An evaluation of noninvasive methods to estimate foliar chlorophyll content, *New Phytol.*, Vol. 153, pp. 185-194, ISSN 0028-646X

Rosenthal, L. S. & Brandt L. J. (2007). *Intestinal Ischemia,* Albert Einstein College of Medicine, Montefiore Medical Center, The American College of Gastroenterology, Available: http://www.gi.org/patients/gihealth/pdf/ischemia.pdf

Suykens, J. A. K. & Vandewalle, J. (1999). Least squares support vector machine classifiers, *Neural Processing Letters,* Vol. 9, pp. 293-300, ISSN 1370-4621

Van Gestel, T.; Suykens, J.A.K.; Baesens, B.; Viaene, S.; Vanthienen, J.; Dedene, G.; De Moor, B. & Vandewalle, J. (2004). Benchmarking Least Squares Support Vector Machine Classifiers, *Machine Learning*, Vol. 54, No. 1, pp. 5-32, ISSN 0885-6125

Vapnik, V.N. (1995). *The nature of statistical learning theory,* Springer-Verlag, ISBN-10 0387987800, Berlin

Zuzak, K.J.; Naik, S.C.; Alexandrakis, G.; Hawkins, D.; Behbehani, K. & Livingston, E.H. (2007). Characterization of a near-infrared laparoscopic hyperspectral imaging system for minimally invasive surgery, *Anal. Chem.*, Vol. 79, pp. 4709-4715, ISSN 0003-2700

Dialectical Classification of MR Images for the Evaluation of Alzheimer's Disease

Wellington Pinheiro dos Santos,
Escola Politécnica de Pernambuco, Universidade de Pernambuco
Brazil

Francisco Marcos de Assis,
Departamento de Engenharia Elétrica, Universidade Federal de Campina Grande
Brazil

Ricardo Emmanuel de Souza,
Departamento de Física, Universidade Federal de Pernambuco
Brazil

Plínio Bezerra dos Santos Filho
Department of Physics, North Carolina State University
USA

1. Introduction

Alzheimer's disease is the most common cause of dementia, both in senile and presenile individuals, observing the gradual progress of the disease as the individual becomes older (Ewers et al., 2006). The major manifestation of Alzheimer's disease is the diminution of the cognitive functions with gradual loss of memory, including psychological, neurological and behavioral symptoms indicating the decline of the daily life activities as a whole. Alzheimer's disease is characterized by the reduction of gray matter and the growth of cerebral sulci. However, the white matter is also affected, although the relation between Alzheimer's disease and white matter is still unknown (Friman et al., 2006).

Acquisition of diffusion-weighted magnetic resonance images (DW-MR images) turns possible the visualization of the dilation of the lateral ventriculi temporal corni, enhancing the augment of sulci, related to the advance of Alzheimer's disease (Haacke et al., 1999). Therefore, volumetrical measuring of cerebral structures is very important for diagnosis and evaluation of the progress of diseases like Alzheimer's (Ewers et al., 2006), especially the measuring of the volumes occupied by sulci and lateral ventriculi, turning possible the addition of quantitative information to the qualitative information expressed by the DW-MR images (Hayasaka et al., 2006).

Usually, the evaluation of the progress of Alzheimer's disease using image analysis of DW-MR images is performed after acquiring at least three images of each slice of interest,

generated using the sequence spin-echo Stejskal-Tanner with different diffusion exponents, where one of the exponents is 0 s/mm², that is, a T2-weighted spin-echo image (Haacke et al., 1999). Then, a fourth image is calculated: the Apparent Diffusion Coefficient Map, or ADC map, where each pixel is associated to the corresponding apparent diffusion coefficient of the associated voxel: the brighter the pixels, the greater the corresponding apparent diffusion coefficients (Haacke et al., 1999).

The dialectical conception of reality is a kind of philosophical investigative method for analyzing processes present in nature and in human societies. Its origins are connected to the philosophy of the ancient civilizations of Greece, China and India, closely connected to the thoughts of Heraclite, Plato, and the philosophies of Confucionism, Buddhism, and Zen. As a general analysis method, dialectics has experienced considerable progress due to the development of German Philosophy in the 19th century, with Hegel's dialectics and, in the 20th century, the works of Marx, Engels, and Gramsci. All those philosophers produced seminal works on the dynamics of contradictions in nature and class-based societies, giving rise to the Historical Materialism (Marx, 1980; Engels, 1975; Gramsci, 1992a; Gramsci1992b; Bobbio, 1990).

The dialectical method of Historical Materialism is a tool for studying systems by considering the dynamics of their contradictions, as dynamic processes with intertwined phases of *evolution* and *revolutionary crisis*. It has inspired us to conceive an evolutionary computational intelligent method for classification that is able to solve problems commonly approached by neural networks and genetic algorithms.

Each of the most common paradigms of Computational Intelligence, namely neural networks, evolutionary computing, and culture-inspired algorithms, has its basis in a kind of theory intended to be of general application, but in fact very incomplete; e.g. the neural networks approach is based on a certain model of the brain; evolutionary computing is based on Darwin's theory; and cultural-inspired algorithms are based on the study of populations, such as those of ant colonies.

However, it is important to note that it is not necessarily the case (and indeed it may be impossible) that the theories an algorithm are based on have to be complete. For example, neural networks utilize a well-known incomplete model of the neurons. This is a strong reason for investigating the use of Philosophy as a source of inspiration for developing computational intelligent methods and models to apply in several areas, such as pattern recognition.

Thornley and Gibb discussed the application of Dialectics to understand more clearly the paradoxical and conceptually contradictory discipline of information retrieval (Thornley & Gibb, 2007), while Rosser Jr. attempted to use some aspects of Dialectics in nonlinear dynamics, comparing some aspects of Marx and Engel's dialectical method with concepts of Catastrophe Theory, Emergent Dynamics Complexity and Chaos Theory (Rosser Jr., 2000). However, there are no works proposing a mathematical approach to establish the fundamentals of Dialectics as a tool for constructing computational intelligent methods.

This work presents the Objective Dialectical Method (ODM), which is an evolutionary computational intelligent method, and the Objective Dialectical Classifier (ODC), an instance of ODM that operates as a non-supervised self-organized map dedicated to pattern recognition and classification. ODM is based on the dynamics of contradictions among dialectical poles. In the task of classification, each class is considered as a dialectical pole. Such poles are involved in pole struggles and affected by revolutionary crises, when some

poles may disappear or be absorbed by other ones. New poles can emerge following periods of revolutionary crisis. Such a process of pole struggle and revolutionary crisis tends to a stable system, e.g. a system corresponding to the clusterization of the original data.
This chapter presents a relatively new approach to evaluate the progress of Alzheimer's disease: once the ADC map usually presents pixels with considerable intensities in regions not occupied by the head of patient, a degree of uncertainty can also be considered in the pixels inside the sample. Furthermore, the ADC map is very sensitive to noisy images (Haacke et al., 1999; Santos et al., 2007). Therefore, in this case study, images are used to compose a multispectral image, where each diffusion-weighted image is considered as a spectral band in a synthetic multispectral image. This multispectral image is classified using the Objective Dialectical Classifier, a new classification method based on Dialectics as defined in the Philosophy of Praxis.

2. Materials and Methods

2.1 DW-MR Images and ADC Maps

The DW-MR images used in this work were acquired from the clinical images database of the Laboratory of MR Images, at the Department of Physics of Universidade Federal de Pernambuco, Recife, Brazil. This database is composed by clinical images acquired from Alzheimer's volunteers, using clinical 1.5 T MR imaging systems. We used 60 cerebral DW-MR images corresponding to male patients with Alzheimer's disease. To perform the training of the proposed analysis, we chose the MR images corresponding to the 13th slice, showing the temporal corni of the lateral ventriculi, to furnish a better evaluation for specialists and facilitate to stablish correlations between data generated by the computational tool and *a priori* specialist knowledge.

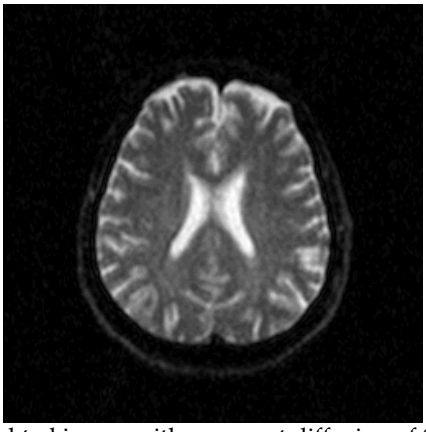


Fig. 1. Axial diffusion-weighted image with exponent diffusion of 0 s/mm^2

An image can be considered as a mathematical function, where its domain is a region of the plane of the integers, called grid, and its counterdomain is the set of the possible values occupied by the pixels corresponding to each position on the grid.

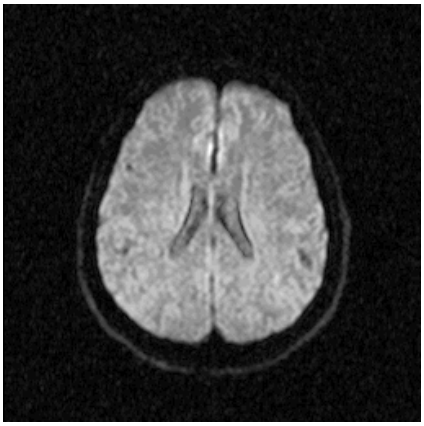


Fig. 2. Axial diffusion-weighted image with exponent diffusion of 500 s/mm²

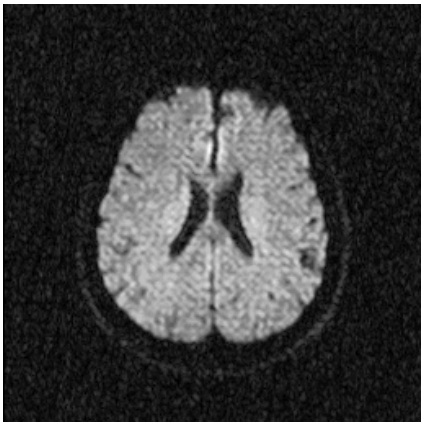


Fig. 3. Axial diffusion-weighted image with exponent diffusion of 1000 s/mm²

Let $f_i : S \to W$ be the set of the diffusion-weighted MR images, where $1 \leq i \leq 3$, $S \subseteq \mathbf{Z}^2$ is the grid of the image f_i, where $W \subseteq \mathbf{R}$ is its codomain. The synthetic multispectral image $f : S \to W^3$ composed by the MR images of the figures 1, 2 and 3 is given by:

$$f(\mathbf{u}) = (f_1(\mathbf{u}), f_2(\mathbf{u}), f_3(\mathbf{u}))^T \quad [1]$$

where $\mathbf{u} \in S$ is the position of the pixel in the image f, and f_1, f_2 and f_3 are the diffusion-weighted MR images. Considering that each pixel $f_i(\mathbf{u})$ is approximately proportional to the signal of the corresponding voxel as follows (Castano-Moraga et al., 2006):

$$f_i(\mathbf{u}) = K\rho(\mathbf{u})e^{-T_E/T_2(\mathbf{u})}e^{-b_i D_i(\mathbf{u})}, \quad [2]$$

where $D_i(\mathbf{u})$ is the nuclear spin diffusion coefficient measured after the i-th experiment, associated to the voxel mapped in the pixel in position $\mathbf{u}$; $\rho(\mathbf{u})$ is the nuclear spin density

in the voxel; K is a constant of proportionality; $T_2(\mathbf{u})$ is the transversal relaxation time in the voxel; T_E is the echo time and b_i is the diffusion exponent, given by (Haacke et al., 1999):

$$b_i = \gamma^2 G_i^2 T_E^3 / 3, \quad [3]$$

where γ is the gyromagnetic ratio and G_i is the gradient applied during the experiment i. Figures 1, 2 and 3 show images with diffusion exponents 0 s/mm², 500 s/mm² and 1000 s/mm², respectively.

The analysis of DW-MR images is often performed using the resulting ADC map $f_{\text{ADC}} : S \to W$, which is calculated as follows (Basser, 2002):

$$f_{\text{ADC}}(\mathbf{u}) = \frac{C}{b_2} \ln\left(\frac{f_1(\mathbf{u})}{f_2(\mathbf{u})}\right) + \frac{C}{b_3} \ln\left(\frac{f_1(\mathbf{u})}{f_3(\mathbf{u})}\right), \quad [4]$$

where C is a constant of proportionality.

Considering n experiments, we can generalize equation 4 as follows:

$$f_{\text{ADC}}(\mathbf{u}) = \sum_{i=2}^{n} \frac{C}{b_i} \ln\left(\frac{f_1(\mathbf{u})}{f_i(\mathbf{u})}\right). \quad [5]$$

Thus, the ADC map is given by:

$$f_{\text{ADC}}(\mathbf{u}) = C\bar{D}(\mathbf{u}), \quad [6]$$

where $\bar{D}(\mathbf{u})$ is an ensemble average of the diffusion coefficient $D(\mathbf{u})$ (Fillard et al., 2006).

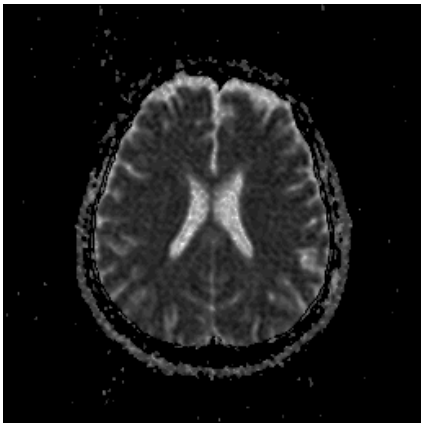


Fig. 4. ADC map calculated from the three diffusion images

Therefore, pixels of the ADC map are proportional to diffusion coefficients in the corresponding voxels. In figure 4 can be seen several artifacts associated to presence of noise. In regions of image where signal-to-noise ratio is poor (let us say, $s/n \approx 1$), the ADC map produces artifacts as consequence of the calculation of logarithms (see equations 4 and 5). Consequently, pixels of the ADC map not necessarily correspond to diffusion coefficients but *apparent* diffusion coefficients, once several pixels indicate high diffusion rates in voxels in empty areas or in very solid areas, e.g. bone in the cranial box, as can be seen in figure 4. This fact generates a considerable degree of uncertainty about the values inside brain area.

In this work we present an alternative to the analysis of the ADC map: the multispectral analysis of the image $f : S \to W^3$ using methods based on neural networks as an alternative that could be easily extended to other diffusion-weighted images than cerebral ones. The proposed analysis is performed using the Objective Dialectical Classifier, presented in the following section.

2.2 Classification using the Objective Dialectical Method

Objective Dialectical Classifiers (ODC) are an adaptation of Dialectics, as defined in the Philosophy of Praxis, to tasks of classification (Gramsci, 1992a; Gramsci, 1992b). This means that the feature vectors are mounted and considered as vectors of conditions. Specifically, once they are applied to the inputs of the dialectical system, their coordinates will affect the dynamics of the contradictions among the integrating dialectical poles. Hence, the integrating poles model the recognized classes at the task of non-supervised classification.

Therefore, an ODC is in fact an adaptable and evolutionary-based non-supervised classifier where, instead of supposing a predetermined number of classes, we can set an initial number of classes (dialectical poles) and, as the historical phases happen (as a result of pole struggles and revolutionary crises), some classes are eliminated, others are absorbed, and a few others are generated. At the end of the training process, the system presents a number of statistically significant classes present in the training set and, therefore, a feasible classifier associated to the final state of the dialectical system.

To accelerate the convergence of the dialectical classifier, we have removed the operator of pole generation, present at the revolutionary crises. However, it could be beneficial to the classification method, once such operator is a kind of diversity generator operator. The solution found can then be compared to other sort of evolutionary-based image classifiers.

The following algorithm is a possible implementation of the training process of the objective dialectical classifier, used in this work:

1 Set the following initial parameters:
  1.1 Number of historical phases, n_P;
  1.2 Length of each historical phase, n_H;
  1.3 Desired final number of poles, $n_{C,f}$;
  1.4 Step of each historical phase, $0 < \eta(0) < 1$;
  1.5 Maximum crisis, $0 \leq \chi_{\max} \leq 1$;
  1.6 Initial number of poles $\#\Omega(0) = n_C(0)$, defining the initial set of poles: $\Omega(0) = \{C_1(0), C_2(0), \ldots, C_{n_C(0)}(0)\}$.
2 Set the following thresholds:
  2.1 Minimum force, $0 \leq f_{\min} \leq 1$;
  2.2 Minimum contradiction, $0 \leq \delta_{\min} \leq 1$.
3 Initialize the weights $w_{i,j}(0)$, where $1 \leq i \leq n_C(0)$ and $1 \leq j \leq n$.
4 Let $\#\Omega(t)$ be the cardinality of $\Omega(t)$, repeat until n_P iterations or $\#\Omega(t) = n_{C,f}$:
  4.1 Repeat until n_H iterations:
    4.1.1 Initialize the measures of force $f_i = 0$, for $1 \leq i \leq n_C(t)$.
    4.1.2 For all vectors of conditions $\mathbf{x} = (x_1, x_2, \ldots, x_n)^T$

of the input set $\Psi = \{\mathbf{x}^{(l)}\}_{l=1}^{L}$, repeat:

4.1.2.1 Compute the values of the anticontradiction functions:

$g_i(\mathbf{x}) = e^{-||\mathbf{x}-\mathbf{w}_i||}$,

where $1 \leq i \leq n_C(t)$.

4.1.2.2 Calculate $g_{\max}$:

$g_{\max} = \max\{g_1(\mathbf{x}), g_2(\mathbf{x}), \ldots, g_{n_C(t)}(\mathbf{x})\}$.

4.1.2.3 Calculate the index $k(t)$ of the winner class:

$g_i = g_{\max} \Rightarrow k(t) = i$.

4.1.2.4 Adjust the weights of the winner pole:

$$w_{i,j}(t+1) = \begin{cases} w'_{i,j}(t), & i = k(t) \\ w_{i,j}(t), & i \neq k(t) \end{cases},$$

where

$w'_{i,j}(t) = w_{i,j}(t) + \eta(t)(x_j(t) - w_{i,j}(t))$.

4.1.2.5 Update the measure of force of the integrating poles:

$$f_i(t+1) = \begin{cases} f_i(t) + 1, & i = k(t) \\ f_i(t), & i \neq k(t) \end{cases}.$$

4.1.3 Quantitative changing: $\Omega(t+1) = \Omega(t)$.

4.2 Calculate the normalized measures of force:

$\bar{f}_i(t) = \frac{f_i(t)}{\max\{f_j(t)\}_{j=1}^{n_C(t)}}$,

for $1 \leq i \leq n_C(t)$.

4.3 Compute the contradictions:

$\delta_{i,j} = 1 - g_i(\mathbf{w}_j)$,

where $2 \leq j \leq n_C(t)$, $1 \leq i < j$, and find the maximum contradiction

$\delta_{\max} = \max\{\delta_{i,j}, i \neq j\}$,

for $j = 2, 3, \ldots, n_C(t)$ and $i = 1, 2, \ldots, j-1$.

4.4 Qualitative changing: compute the new set of poles, $\Omega(t+1)$:

$\bar{f}_i(t) > f_{\min} \Rightarrow C_i(t) \in \Omega(t+1)$,

where $1 \leq i \leq n_C(t)$ and

$\delta_{i,j} \geq \delta_{\min} \Rightarrow C_i(t), C_j(t) \in \Omega(t+1)$,

$\delta_{i,j} < \delta_{\min} \Rightarrow C_i(t) \in \Omega(t+1)$,

$\delta_{i,j} = \delta_{\max} \Rightarrow C_q \in \Omega(t+1)$,

where $2 \leq j \leq n_C(t)$, $1 \leq i < j$, $q = n_C(t) + 1$, and

$$w_{q,k}(t+1) = \begin{cases} w_{i,k}(t+1), & k \bmod 2 = 1 \\ w_{j,k}(t+1), & k \bmod 2 = 0 \end{cases},$$

for $k = 1, 2, \ldots, n$.

4.5 Add the crisis effect to the weights of the new integrating poles of the dialectical system:

$w_{i,j}(t+2) = w_{i,j}(t+1) + \chi_{\max} G(0,1)$,

for $1 \leq i \leq n_C(t+1)$, $1 \leq j \leq n$ and $\Omega(t+2) = \Omega(t+1)$.

Once the training process is complete, ODC behavior occurs in the same way as any non-supervised classification method. This is clear if we analyze the training process when $n_P = n_H = 1$. This transforms the ODC into a k-means method, for instance.
The classification is performed in the following way: given a set of input conditions

$$\mathbf{x} = (x_1, x_2, \ldots, x_n)^T,$$

if the dialectical system reaches stabilization when $\Omega = \{C_1, C_2, \ldots, C_{n_C}\}$, then we apply the following classification rule:

$$g_k(\mathbf{x}) = \max\{g_1(\mathbf{x}), g_2(\mathbf{x}), \ldots, g_{n_C}(\mathbf{x})\} \Rightarrow \mathbf{x} \in C_k,$$

where $1 \leq k \leq n_C$.

4. Discussion and Results

The ground-truth image was built by the use of a two-degree polynomial network to classify the multispectral image. The training set was assembled using anatomic information obtained from T1, T2 and spin density MR images.
The ODC was trained using an initial system of 10 integrating classes, affected by 3 input conditions, studied during 5 historical 100-length phases, with an initial historical step $\eta_0 = 0.1$. At the stages of revolutionary crisis we considered a minimum measure of force of 0.01, minimum contradiction of 0.25 and maximum crisis of 0.25. The stop criterion was the final number of classes, in our case, 4 classes. The input conditions are the values of pixels on each of the 3 bands.

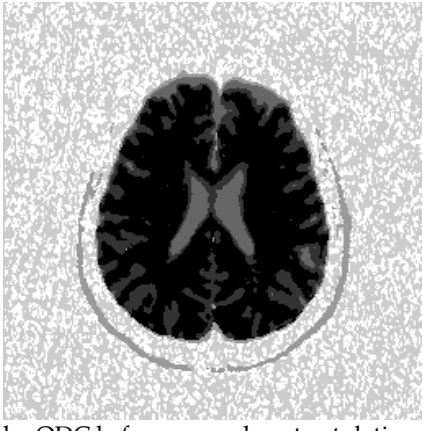


Fig. 5. Classification result by ODC before manual post-rotulation

ODC training resulted in 6 classes, reduced to 4 classes after a manual post-rotulation that merged the 3 classes external to the cranium, i.e. background, noise and cranial box, into a single class, namely background. This post-rotulation was performed manually because the 3 populations are statistically different and only conceptually can they be reunited in a unique class. Figures 5 and 6 show the resulting classification by ODC before and after manual post-rotulation, respectively.

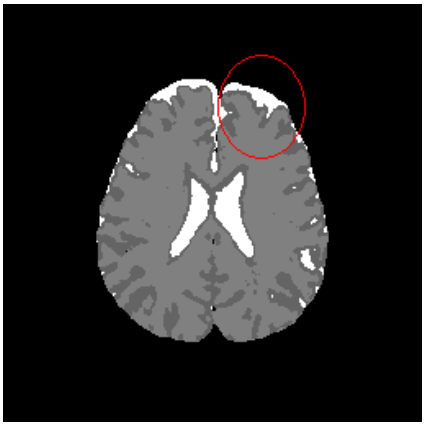


Fig. 6. Classification result by ODC after manual post-rotulation. White areas are indication of cerebrospinal fluid, once gray and dark gray areas indicate white and gray matter, respectively. The damaged area is emphisized.

From Figure 6 we can see that ODC was able to make a distinction between white and gray matter, the latter present in the interface between cerebrospinal fluid and white matter. Notice that an increased damaged area is highlighted. The classification fidelity was measured using the morphological similarity index, with structure element square 3×3, and Wang's index (Wang & Bovik, 2002), yielding 0.9877 and 0.9841, respectively.

The objective dialectical classifier could identify statistically significant classes in situations where the initial number of classes is not well known. It makes possible the detection of relevant classes and even singularities beyond the initial prediction made by the medical specialist. It is also able to aid the medical specialist to measure the volumes of interest, in an attempt to establish a correlation of such measuring with the advance of neurodegenerative diseases, such as Alzheimer's, and to differentiate significant alterations in the values of the measured diffusion coefficients. ODC can qualitatively and quantitatively improve the analysis of the human medical specialist.

The objective dialectical classifier can be used in problems where the number of statistically significant classes is not well known, or in problems where we need to find a sub-optimum clustering map to be used for classification. The task of finding a suboptimum clustering map is empirical, once it is necessary to analyze the behavior of the training process as a function of the several parameters of the method, namely the minimum force, the minimum contradiction, the initial number of classes, the number of historical phases, the duration and the historical step of each historical phase, that is, all the initial parameters of the proposed segmentation algorithm. Nevertheless, it is important to emphasize that, as the number of initial parameters is given, the classification performance of the dialectical classifiers is highly dependent on these initial parameters.

The objective dialectical method starts a new family of evolutionary methods inspired in the Philosophy, especially the Philosophy of Praxis, which can be used to solve both classical and new image analysis problems, such as the one presented in our case study, that is, biomedical image analysis and processing.

We conclude by stating that philosophical thought is a great source of inspiration for constructing new computational intelligent methods highly applicable to Biomedical Engineering problems, since we are simply returning to our original source of knowledge: Philosophy as an important tool to a better understanding of nature and ourselves in a larger sense.

5. References

Ewers, M.; Teipel, S.J.; Dietrich, O.; Schönberg, S.O.; Jessen, F.; Heun, R.; Scheltens, P.; van de Pol, L.; Freymann, N.R.; Moeller, H.J. & Hampela, H. (2006). Multicenter assessment of reliability of cranial MRI. *Neurobiology of Aging*, 27, 1051-1059

Friman, O.; Farnebäck, G. & Westin, C.F. (2006). A bayesian approach for stochastic white matter tractography. *IEEE Transactions on Medical Imaging*, 25, 8, 965-978

Haacke, E.M.; Brown, R.W.; Thompson, M.R. & Venkatesan, R. (1999). *Magnetic Resonance Imaging: Physical Principles and Sequence Design,* Wiley-Liss

Marx, K. (1980). Critique of Hegel's dialectics and philosophy. In *Economic and Philosophic Manuscripts of 1844*, International Publishers

Engels, F. (1975). The role played by labor in the transition from ape to man. In *Collected Works of Karl Marx and Frederik Engels*, International Publishers

Gramsci, A. (1992). Introduction to the Study of Philosophy and Historical Materialism. In *Prison Notebooks*, Columbia University

Gramsci, A. (1992). Some Problems in the Study of the Philosophy of Praxis. In *Prison Notebooks*, Columbia University

Bobbio, N. (1990). *Saggi su Gramsci*, Feltrinelli, Milano

Thornley, C. & Gibb, F. (2007). A dialectical approach to information retrieval. *Journal of Documentation*, 63, 5, 755-764

Rosser Jr, J.B. (2000). Aspects of dialectics and nonlinear dynamics. *Cambridge Journal of Economics*, 24, 3, 311-324

Hayasaka, S.; Du, A.T.; Duarte, A.; Kornak, J.; Jahng, G.H.; Weiner, M.W. & Schuff, N. (2006). A non-parametric approach for co-analysis of multi-modal brain imaging data: Application to Alzheimer's disease. *NeuroImage*, 30, 768-779

Santos, W.P.; Souza, R.E. & Santos-Filho, P.B. (2007). Evaluation of Alzheimer's disease by analysis of MR images using multilayer perceptrons and Kohonen SOM classifiers as an alternative to the ADC maps. *In 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, Lyon, France, EMBS-IEEE

Castano-Moraga, C.A.; Lenglet, C.; Deriche, R. & Ruiz-Alzola, J. (2006). A Fast and Rigorous Anisotropic Smoothing Method for DT-MRI. In *Proceedings of the ISBI 2006*, CS-IEEE

Basser, P.J. (2002). Diffusion-Tensor MRI: Theory, Experimental Design, and Data Analysis. In *Procedings of the 2nd Joint EMBS BMES Conference*, 1165-1166, Houston, USA,, EMBS-IEEE-BMES

Fillard, P.; Arsigny, V.; Pennec, X. & Ayache, N. (2006). Clinical DT-MRI estimations, smoothing and fiber tracking with log-Euclidean metrics. In *Proceedings of the ISBI 2006*, CS-IEEE

Wang, Z. & Bovik, A.C. (2002). A universal image quality index. *IEEE Signal Processing Letters*, 9

3-D MRI and DT-MRI Content-adaptive Finite Element Head Model Generation for Bioelectomagnetic Imaging

Tae-Seong Kim and Won Hee Lee
Kyung Hee University, Department of Biomedical Engineering
Republic of Korea

1. Introduction

One of the challenges of the 21st century is to understand the functions and mechanisms of the human brain. Although the complexity of deciphering how the brain works is so overwhelming, the electromagnetic phenomenon happening in the brain is one aspect we can study and investigate. In general, this phenomenon of electromagnetism is described as the electrical current produced by action potentials from neurons which are reflected as the changes in electrical potential and magnetic fields (Baillet et al., 2001). These electromagnetic fields of the brain are generally measured with electroencephalogrm (EEG) and magnetoencephalogram (MEG) that are actively used for bioelectromagnetic imaging of the human brain (a.k.a., inverse solutions of EEG and MEG).

In order to investigate the electromagnetic phenomenon of the brain, the human head is generally modelled as an electrically conducting medium and various numerical approaches are utilized such as boundary element method (He et al., 1987; Hamalainen & Sarvas, 1989; Meijs et al., 1989), finite difference method (Neilson et al., 2005; Hallez et al., 2008), and finite element method (Buchner et al., 1997; Marin et al., 1998; Kim et al., 2002; Lee et al., 2006; Wolters et al., 2006; Zhang et al., 2006; Wendel et al., 2008), to solve the bioelectromagnetic problems (a.k.a., forward solutions of EEG and MEG). Among these approaches, the finite element method (FEM) or analysis (FEA) is known as the most powerful and realistic method with increasing popularity due to (i) readily available computed tomography (CT) or magnetic resonance (MR) images where geometrical shape information can be derived, (ii) recent developments in imaging physical properties of biological tissue such as electrical (Kim et al., 2009) or thermal conductivity, which can be incorporated in to the FE models, (iii) numerical and analytical power that allow truly volumetric analysis, and (iv) much improved computing and graphic power of modern computers.

In applying FEA to the bioelectromagnetic problems, one critical and challenging requirement is the representation of the biological domain (in this case, the human head) as discrete meshes. Although there are some general packages available through which the mesh representation of simple objects is possible, their capability of generating adequate mesh models of biological organs, especially the human head, requires substantial efforts since (i) most mesh generators have some limitations of handling arbitrary geometry of

complex biological shapes, requiring simplification of complex boundaries, (ii) most mesh generation schemes use a mesh refinement technique to represent fine structures with much smaller elements. This tends to increase number of nodes and elements beyond the computational limit, thus demanding overwhelming computation time, (iii) most mesh generation techniques require careful supervision of users, and (iv) there is a lack of automatic mesh generation techniques for generating FE mesh models for individual heads. Therefore, there is a strong need for fully automatic mesh generation techniques.

In this chapter, we present two novel techniques that automatically generate FE meshes adaptive to the anatomical contents of MR images (we name it as *cMesh*) and adaptive to the contents of anisotropy measured through diffusion tensor magnetic resonance imaging (DT-MRI) (we name it as *wMesh*). The cMeshing technique generates the meshes according to the structural contents of MR images, offering advantages in automaticity and reduction of computational loads with one limitation: its coarse mesh representation of white matter (WM) regions, making it less suitable for the incorporation of the WM tissue anisotropy. The wMeshing technique overcomes this limitation by generating the meshes in the WM region according to the WM anisotropy derived from DT-MRIs. By combining these two techniques, one can generate high-resolution FE head models and optimally incorporate the anisotropic electrical conductivities within the FE head models.

This chapter introduces the cMesh and wMesh methodologies and their evaluations in their effectiveness by comparing the mesh characteristics including geometry, morphology, anisotropy adaptiveness, and the quality of anisotropic tensor mapping into the meshes to those of the conventional FE head models. The presented methodologies offer an automatic high-resolution FE head model generation scheme that is suitable for realistic, individual, and anisotropy-incorporated high-resolution bioelectromagnetic imaging.

2. Previous Approaches in Finite Element Head Modelling

Although the classical modelling of the head as a single or multiple spheres (thus called spherical head models) dates back much further than realistic boundary element and finite element head models, the early finite element head modelling was attempted by Yan et al. (1991). Then the later attempts are well summarized in a review paper by Voo et al. (1996). Medical image-based realistic finite element head modelling was introduced a year later by Awada et al. (1997) in 2-D and by Kim et al. (2002) in 3-D. Other than these works, numerous literatures have shown their own approaches of finite element head modelling. Lately, anisotropic properties of brain tissues including white matter and skull have been incorporated into the FE head models and their effects on the forward and inverse solutions have been investigated (Kim et al., 2003; Wolters et al., 2006). Recent studies focus on adaptive mesh modelling, high-resolution mesh generation, and influence of tissue anisotropies. More details can be found in (Lee et al., 2006, 2008; Wolters et al., 2006, 2007).

3. MRI Content-adaptive Finite Element Head Model Generation

The procedures of the content-adaptive finite element mesh (cMesh) generation are summarized as follows: namely, (i) MRI content-preserving anisotropic diffusion filtering for noise reduction and feature enhancement, (ii) structural and geometrical feature map generation from the filtered image, (iii) node sampling based on the spatial density of the

feature maps via a digital halftoning technique, and (iv) mesh generation. The cMesh generation depends on the performance of two key techniques: the quality of feature maps and the accuracy of content-adaptive node sampling. In this study, we focus on the former and its application to MR imagery to build more accurate and efficient cMesh head models for bioelectromagnetic imaging.

3.1 Gradient Vector Flow (GVF) Nonlinear Anisotropic Diffusion

To generate an effective and efficient cMesh head model, it is important to remove unnecessary properties of given images such as artifacts and noises. The content-preserving anisotropic diffusion offers pre-segmentation of sub-volumes to simplify the structures of the image and improvement of feature maps where mesh nodes are automatically sampled. In this study, the 3-D Gradient Vector Flow (GVF) anisotropic diffusion algorithm was used (Kim et al., 2003; Kim et al., 2004). The GVF nonlinear diffusion technique, which was successfully applied to regularize diffusion tensor MR images in a previous study (Kim et al., 2004), was proven to be much more robust in comparison to the conventional Structure tensor-based anisotropic diffusion algorithm (Weickert, 1997) and can be summarized as follows.

The GVF as a 3-D vector field can be defined as:

$$\mathbf{V}(i,j,k) = (\mathbf{u}(i,j,k), \mathbf{v}(i,j,k), \mathbf{w}(i,j,k)) . \tag{1}$$

The field can be obtained by minimizing the energy functional:

$$\begin{aligned} \varepsilon &= \iiint \gamma(\eta_{\mathbf{u}} + \eta_{\mathbf{v}} + \eta_{\mathbf{w}}) + |\nabla f|^2 |\mathbf{V} - \nabla f|^2 \, \partial x \partial y \partial z \\ \eta_{\mathbf{u}} &= \mathbf{u}_x^2 + \mathbf{u}_y^2 + \mathbf{u}_z^2 \\ \eta_{\mathbf{v}} &= \mathbf{v}_x^2 + \mathbf{v}_y^2 + \mathbf{v}_z^2 \\ \eta_{\mathbf{w}} &= \mathbf{w}_x^2 + \mathbf{w}_y^2 + \mathbf{w}_z^2 \end{aligned} \tag{2}$$

where f is an image edge map and γ is a noise control parameter.

For 3-D anisotropic smoothing, the Structure tensor $\mathbf{S}$ is formed with the components of $\mathbf{V}$

$$\mathbf{S} = \mathbf{V}(\mathbf{V})^T . \tag{3}$$

The 3-D anisotropic regularization is governed using the GVF diffusion tensor $\mathbf{D}_{\mathrm{GVF}}$ which is computed with eigen components of $\mathbf{S}$.

$$\partial J / \partial t = div[\mathbf{D}_{GVF} \nabla J] \tag{4}$$

where J is an image volume in 3-D. The regularization behavior of Eq. (4) is controlled with the eigenvalue analysis of the GVF Structure tensor (Ardizzone & Rirrone, 2003, Kim et al., 2003).

3.2 MRI Feature Map Generations

To generate better feature maps from the filtered images, tensor-driven feature extractors using Hessian tensor (Carmona & Zhong, 1998; Yang et al., 2003), Structure tensor (Abd-Elmoniem et al., 2002), and principal curvature methods such as Mean and Gaussian curvature (Gray, 1997; Yezzi, 1998) are utilized. The conventional feature maps proposed by Yang et al. (2003) showed the adequate procedures for the purpose of image representation that meshes are adaptive to the contents of an image where the extraction of image feature information from given image was performed using the Hessian tensor approach.

In the work of Yang et al. (2003), two approaches to generate the feature maps were proposed from the Hessian tensor of each pixel, **H**:

$$\mathbf{H} = \begin{bmatrix} I(i,j)_{xx} & I(i,j)_{xy} \\ I(i,j)_{yx} & I(i,j)_{yy} \end{bmatrix}, I_{xy} = I_{yx} \tag{5}$$

where I is an image, i and j are image indices, x and y indicate partial derivates in space. One feature map was derived from the maximum of the Hessian tensor components:

$$f_{\max}(i,j) = \max\{|I_{xx}(i,j)|, |I_{xy}(i,j)|, |I_{yy}(i,j)|\}. \tag{6}$$

Another proposed feature map was derived from the eigenvalues, μ's, of the tensor:

$$f_{H\max}(i,j) = \max\{|\mu_1(i,j)|, |\mu_2(i,j)|\}. \tag{7}$$

The two eigenvalues of the Hessian tensor matrix, denoted by μ_1 and μ_2 are given by

$$\mu_1 = \frac{1}{2}\left[(I_{xx} + I_{yy}) + \sqrt{(I_{xx} - I_{yy})^2 + 4I_{xy}^2}\right], \tag{8}$$

$$\mu_2 = \frac{1}{2}\left[(I_{xx} + I_{yy}) - \sqrt{(I_{xx} - I_{yy})^2 + 4I_{xy}^2}\right]. \tag{9}$$

The Hessian tensor approach extracts image feature information from the given MR image using the second-order directional derivatives, and its critical attribute is high sensitivity toward feature orientations. However it is known to be highly sensitive toward noise as well.

Currently, advanced differential geometry measures provide better options and choices in deriving feature maps with more effective and accurate properties. In this study, we derived advanced feature maps based on the Hessian and Structure tensor as alternative ways (Lee et al., 2006).

The Hessian tensor-driven feature maps are derived using the eigenvalues of the Hessian tensor in the following way:

$$f_{\mathbf{H}+}(i,j) = \sqrt{(\mu_1^{\mathbf{H}}(i,j) + \mu_2^{\mathbf{H}}(i,j))}, \tag{10}$$

$$f_{\mathbf{H}}(i,j) = \sqrt{\mu_1^{\mathbf{H}}(i,j)}\ , \tag{11}$$

$$f_{\mathbf{H}-}(i,j) = \sqrt{(\mu_1^{\mathbf{H}}(i,j) - \mu_2^{\mathbf{H}}(i,j))}\ , \tag{12}$$

where μ's are the positive eigenvalues of the tensor matrix.
Another approach is the use of the the Structure tensor due to robustness in detecting fundamental feature of objects. The Structure tensor **S** can be expressed as follows:

$$\mathbf{S} = \begin{bmatrix} I_x^2 & I_x I_y \\ I_y I_x & I_y^2 \end{bmatrix}. \tag{13}$$

We next derive the Structure tensor-driven feature maps with the eigenvalues of the Structure tensor as the same ways of the Hessian tensor:

$$f_{\mathbf{S}+}(i,j) = \sqrt{(\mu_1^{\mathbf{S}}(i,j) + \mu_2^{\mathbf{S}}(i,j))}\ , \tag{14}$$

$$f_{\mathbf{S}}(i,j) = \sqrt{\mu_1^{\mathbf{S}}(i,j)}\ , \tag{15}$$

$$f_{\mathbf{S}-}(i,j) = \sqrt{(\mu_1^{\mathbf{S}}(i,j) - \mu_2^{\mathbf{S}}(i,j))}\ . \tag{16}$$

The above feature map reflects the edges and corners of image structures for the plus sign. By taking the maximum eigenvalue, new feature map can be derived which is a natural extension of the scalar gradient viewed as the value of maximum variations. The other feature map represents the local coherence or anisotropy for the minus sign (Tschumperle & Deriche, 2002).
In addition, we generate new feature maps via the principal curvature. There are geometric meanings with respect to the eigenvalues and eigenvectors of the tensor matrix. The first eigenvector (corresponding eigenvalue represents the largest absolute value) is the direction of the greatest curvature. Conversely, the second eigenvector is the direction of least curvature. Also its eigenvalue has the smallest absolute value. The consistent eigenvalues are the respective amounts of these curvatures. The eigenvalues of tensor matrix with real values indicate principal curvatures, and are invariant under rotation.
The Mean curvature can be obtained from the Hessian tensor matrix (Gray, 1997; Yezzi, 1998). It is equal to the half of the trace of **H** which is invariant to the selection of x and y as well. The new feature map f_M using the Mean curvature can be expressed as follows:

$$f_M(i,j) = \frac{I_{xx}(1+I_y^2) - 2I_x I_y I_{xy} + I_{xy}(1+I_x^2)}{2(1+I_x^2+I_y^2)^{3/2}}\ . \tag{17}$$

From the Hessian tensor again, we also derive another feature map f_G using the Gaussian curvature as shown below:

$$f_G(i,j) = \frac{I_{xx}I_{yy} - I_{xy}^2}{(1 + I_x^2 + I_y^2)^2} . \tag{18}$$

3.3 Node Sampling via Digital Halftoning

In order to produce content-adaptive mesh nodes based on the spatial information of the feature map, we utilize the following popular digital halftoning algorithm. The Floyd-Steinberg error diffusion technique with the serpentine scanning is applied to create content-adaptive nodes in accordance with the spatial density of image feature maps (Floyd & Steinberg, 1975). This algorithm produces more nodes in the high frequency regions of the image. The sensitivity of feature map is controlled by regenerating a new feature map with the parameter, κ as shown below. In this way, the total number of content-adaptive nodes generated by the halftoning algorithm can be adjusted.

$$f'(i,j) = f(i,j)^{1/\kappa} \tag{19}$$

where f is a feature map and κ is a control parameter for the number of content-adaptive nodes.

3.4 FE Mesh Generation

Once cMesh nodes are generated from the procedures described above, FE mesh generation using triangular elements in 2-D and tetrahedral elements in 3-D is performed using the Delaunay tessellation algorithm (Watson, 1981).

3.5 Isotropic Electrical Conductivity in cMesh

In order to assign electrical properties to the tissues of the head, we segment the MR images into five sub-regions including white matter, gray matter, CSF, skull, and scalp. BrainSuite2 (Shattuck & Leahy, 2002) is used for the segmentation of the different tissues within the head. The first step is to extract the brain tissues from MR images other than the skull, scalp, and undesirable structures. Then, the brain images are classified into each tissue region including white mater, gray matter, and CSF using a maximum a posterior classifier (Shattuck & Leahy, 2002). The skull and scalp compartments are segmented using the skull and scalp extraction technique based on a combination of thresholding and morphological operations such as erosion and dilation (Dogdas et al., 2005).

The following isotropic electrical conductivity values according to each tissue type are used: white matter=0.14 S/m, gray matter=0.33 S/m, CSF=1.79 S/m, scalp=0.35 S/m, and skull=0.0132 S/m respectively (Kim et al., 2002; Wolters et al., 2006).

3.6 Analysis on the MRI Content-adaptive Meshes

3.6.1 Numerical Evaluation of cMeshes: Feature Maps and Mesh Quality

In order to investigate the effects of the feature maps on cMeshes, we used the following five indices as the goodness measures of content-adaptiveness: (i) correlation coefficient (CC) of the feature map to the original MRI, (ii) root mean squared error (RMSE), (iii) relative error (RE) between the original MRI and the reconstructed MRI based on the nodal MR intensity values (Lee et al., 2006), (iv) number of nodes, and (v) number of elements. For fair comparison of the content-adaptiveness of cMeshes, almost same number of meshes were generated by adjusting the mesh parameter κ as in Eq. (19). To test the content information of the non-uniformly placed nodes, the MR images were reconstructed using the MR spatial intensity values at the sampled nodes via the cubic interpolation method. Then the RMSE and RE values were calculated between the original and reconstructed MR images.

We next performed the numerical evaluations of cMesh quality, since the mesh quality highly affects computational analysis in terms of numerical accuracy on the solution on FEA. The evaluation of mesh quality is critical, since it provides some indications and insights of how appropriate a particular discretization is for the numerical accuracy on FEA. For example, as the shapes of elements become irregular (i.e, the angles of elements are highly distorted), the error of the discretization in the solutions of FEA is increased and as angles in an element become too small, the condition number of the element matrix is increased, thus the numerical solutions of FEA are less accurate. The geometric quality indicators were used for the investigation of cMesh quality as the mesh quality measures (Field, 2000). For a triangle element in 2-D, the mesh quality measure can be expressed as

$$q = \alpha \frac{A}{l_1^2 + l_2^2 + l_3^2} \tag{20}$$

where A represents the area of the triangle, and l_1, l_2, and l_3 are the edge lengths of the triangle element, and $\alpha = 4\sqrt{3}$ is a normalizing coefficient justifying the quality of an equilateral triangle to 1 (i.e., q=1, when $l_1 = l_2 = l_3$. If q>0.6, the triangle possesses acceptable mesh quality). The overall mesh quality was evaluated for triangle elements in terms of the arithmetic mean by

$$Q_a = \frac{1}{N} \sum_{i=1}^{N} q_i \tag{21}$$

where N indicates the number of elements.

Additionally, we counted the elements with the poor quality (i.e., q<0.6) as an indicator of the poor elements that affect the overall mesh quality. Certainly, other measures are available using other geometric quality indicators (Berzins, 1999).

Fig. 1 shows a set of results from 2-D cMesh generation obtained using the conventional techniques by Yang et al. (2003). Fig. 1(a) is a MR image, (b) conventional feature map obtained using $f_{\max}$, and (c) another suggested feature map using $f_{\mathrm{H}\max}$. Fig. 1(d) shows content-adaptive nodes from Fig. 1(c). Figs. 1(e) and (f) show content-adaptive meshes in 2-D from Figs. 1(b) and (c) respectively. There are 2327 nodes and 4562 triangular elements in

Fig. 1(e) and 2326 nodes and 4560 elements in Fig. 1(f). The triangle with different sizes indicates adaptive characteristics of mesh generation in accordance with the two different feature maps.

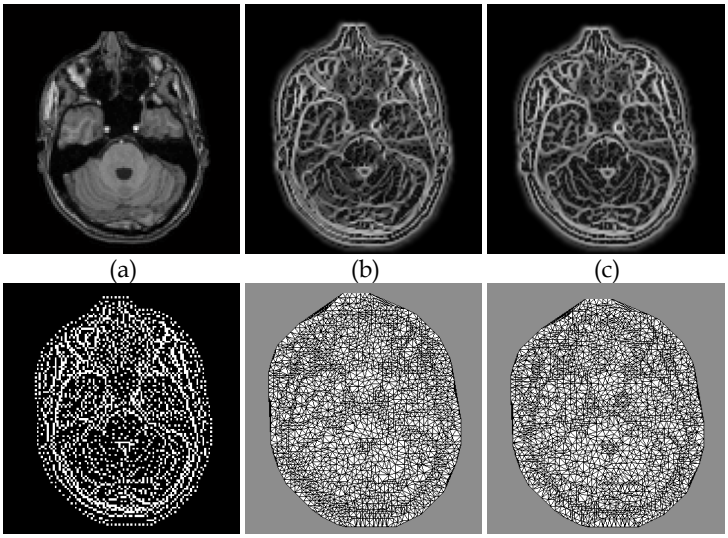


Fig. 1. Feature maps and cMeshes of a MR image: (a) a MR image, (b) feature map from (a) using f_{max}, (c) using f_{Hmax}, (d) content-adaptive nodes from (c), (e) cMeshes from (b) with 2327 nodes and 4562 elements, and (f) cMeshes from (c) with 2326 nodes and 4560 elements.

We also generated the cMeshes of the given MRI using the advanced feature maps. Figs. 2(a)-(c) display the feature maps obtained using f_{H+}, f_{H}, and f_{H-} derived from the Hessian approach. Their corresponding cMeshes are shown in Figs. 2 (d)-(f) respectively. There are 2326 nodes and 4560 elements in Fig. 2(d), 2324 nodes and 4556 elements in Fig. 2(e), and 2329 nodes and 4566 elements in Fig. 2(f). The high sensitivity of Hessian tensor to the structures of MRI is clearly visualized.

Fig. 3 shows a set of demonstrative results from the Structure tensor approaches. Figs. 3 (a)-(c) show the improved feature maps acquired using f_{S+}, f_{S}, and f_{S-} respectively. The corresponding cMeshs are shown in Figs. 3 (d)-(f). There are 2323 nodes and 4554 elements in Fig. 3(d), 2325 nodes and 4558 elements in Fig. 3(e), and 2323 nodes and 4554 elements in Fig. 3(f) respectively. Based on these results, it indicates that the Structure tensor-driven feature extractor yields optimal information on image features and their resultant cMeshes look most adaptive to the contents of the given MRI. That is larger elements are present in the homogeneous regions and smaller elements in the high frequency regions with reasonable numbers of nodes and elements. Content-adaptive nature is clearly visible in the contents of the given cMeshes.

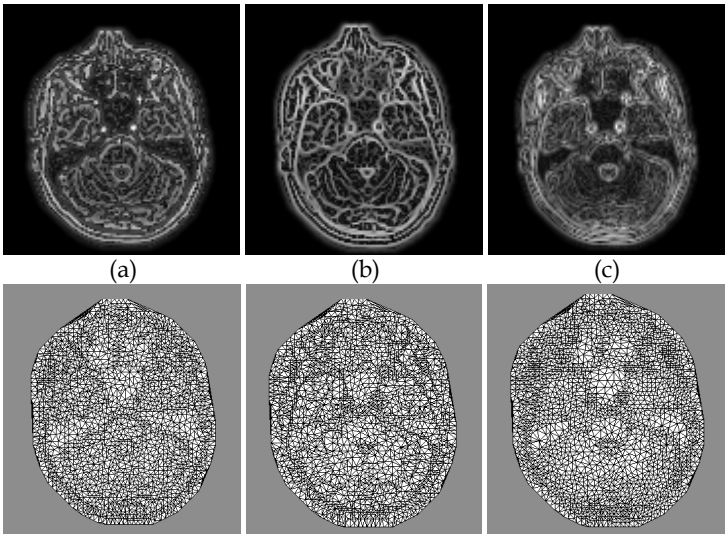


Fig. 2. Hessian tensor-derived feature maps and cMeshes: (a) feature map using f_{H+}, (b) using f_H, (c) using f_{H-}, (d) cMeshes from (a) with 2326 nodes and 4560 elements, (e) cMeshes from (b) with 2324 nodes and 4556 elements, (f) cMeshes from (c) with 2329 nodes and 4566 elements.

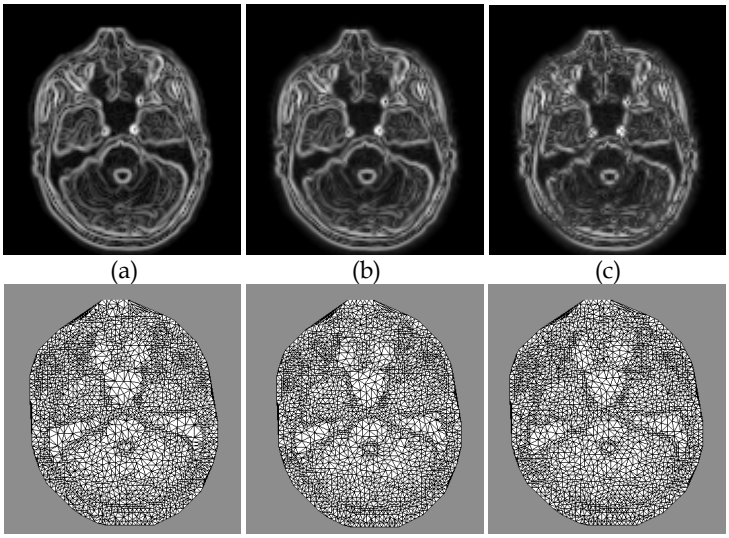


Fig. 3. Structure tensor-derived feature maps and cMeshes: (a) feature map using f_{S+}, (b) using f_S, (c) using f_{S-}, (d) cMeshes from (a) with 2323 nodes and 4554 elements, (e) cMeshes from (b) with 2325 nodes and 4558 elements, (f) cMeshes from (c) with 2323 nodes and 4554 elements.

In addition, by using the Mean and Gaussian curvature, the feature maps obtained using f_M and f_G are shown in Figs. 4(a) and (b) respectively. The resultant cMeshes are shown in Figs. 4(c) and (d). The characteristics of curvatures to the image features are clearly noticeable too.

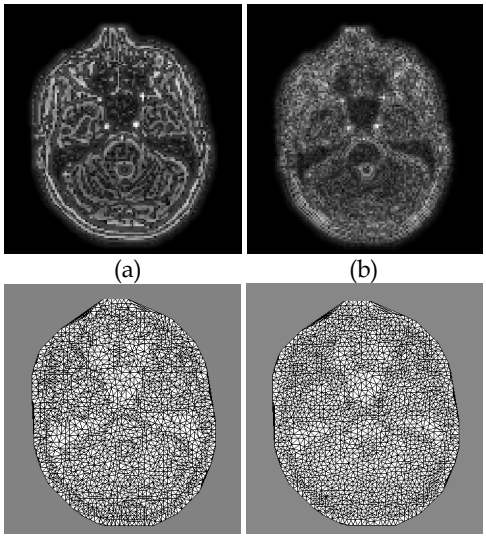


Fig. 4. Curvature-derived feature maps and cMeshes: (a) feature map using f_M, (b) using f_G, (c) cMeshes from (a) with 2326 nodes and 4560 elements, (d) cMeshes from (b) with 2325 nodes and 4558 elements.

The CC values in Table 1 show strong correlation between the Structure tensor-driven feature map and the original MRI, indicating the Structure-driven feature extractor generates much better content-adaptive features. Although the CC value of Structure tensor-driven approach is lower than the feature maps by f_{H+}, f_{H-}, f_M, and f_G, it produced much lower RMSE and RE values, indicating the reconstructed MRI is much closer to the original MRI. As for the cMesh quality, the result by f_G describes the highest value. Also, the Structure tensor approach show greatly acceptable values with much lower number of poor elements compared to other feature extractors, indicating the Structure tensor-driven approach will offer numerically accurate and efficient computational accuracy in FEA.

3.6.2 Numerical Evaluation of cMeshes: Regular Mesh vs. cMesh

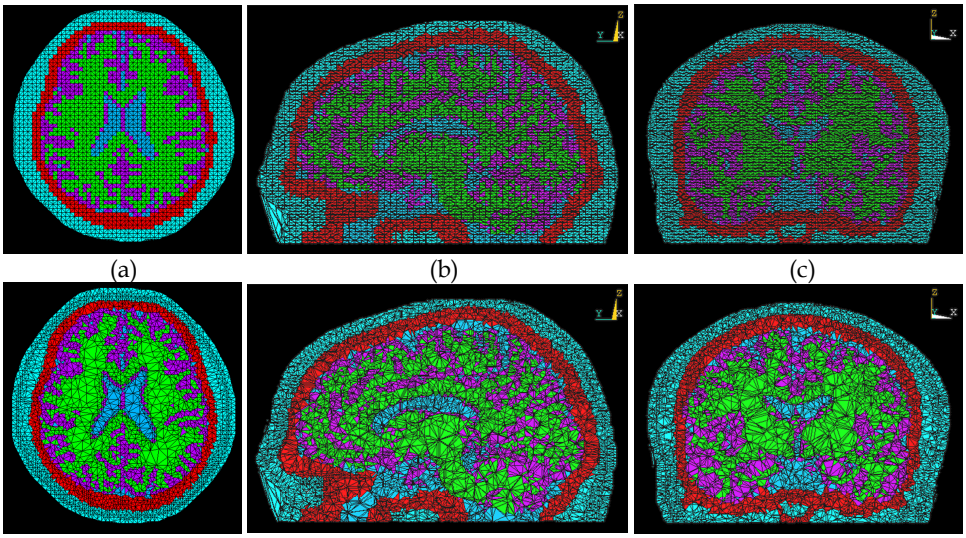
To evaluate numerical accuracy of the cMesh head model on FEA in 3-D against the conventional regular FE model commonly used in E/MEG forward or inverse problems, two 3-D cMesh models of the whole head (matrix size: 128×128×77, spatial resolution: 1×1×1 mm^3) differing in their mesh resolution were built using the Structure tensor-based (i.e., f_{S+}) cMesh generation technique as described earlier. For the reference model, the regular mesh head model was generated as the gold standard using fine and equidistant tetrahedral elements with inner-node spacing of 2 mm, since analytical solutions cannot be obtained for an arbitrary geometry of the real head.

The numerical quality of the cMesh head models were evaluated by comparing the scalp forward potentials computed from the cMesh models against those of the regular mesh model. To solve EEG forward problems governed by the Poisson's equation under the quasistatic approximation of the Maxwell's equation (Sarvas, 1987), the FE head models along with isotropic electrical conductivity information were imported into a software ANSYS (ANSYS, Inc., PA, USA). The forward potential solutions due to the identical current generator (Yan et al., 1991; Schimpf et al., 2002) were obtained using the preconditioned conjugate gradient solver of ANSYS. Then the scalp potential values from the cMesh head models were compared to those from the reference FE head model. As evaluation measures, both CC and RE were used along with the forward computation time (CT) as a numerical efficiency measure.

Fig. 5 shows a set of results from the 3-D regular and cMesh models of the whole head with isotropic electrical conductivities. In Figs. 5(a)-(c), there are 159,513 nodes and 945,881 tetrahedral elements in the regular FE head model. The cMesh model of the entire head with 109,628 nodes and 694,588 tetrahedral elements is given in Figs. 5(d)-(f). The mesh generation time for the 3-D regular and cMesh head models was 169.5 sec and 68.1 sec respectively on a PC with Pentium-IV CPU 3.0 GHz and 2GB RAM. In comparison to the regular mesh model in Figs. 5(a)-(c), the content-adaptive meshes are clearly visible according to MR structural information in Figs. 5(d)-(f). Various mesh sizes indicate the adaptive characteristics of meshes based on given MR anatomical contents as shown in Figs. 5(d)-(f).

Method	No. of Nodes	No. of Elements	MRI vs. Feature Map	MRI vs. Reconstructed MRI		cMesh Quality	No. of Poor Elements: q<0.6
			CC	RMSE	RE		
f_{max}	2327	4562	0.45	40.11	0.21	0.79 ± 0.16	503
f_{Hmax}	2326	4560	0.49	46.98	0.25	0.78 ± 0.16	617
f_{H+}	2326	4560	0.68	34.94	0.18	0.80 ± 0.15	418
f_{H}	2324	4556	0.50	46.38	0.24	0.78 ± 0.16	649
f_{H-}	2329	4566	0.62	34.94	0.18	0.82 ± 0.14	224
f_{S+}	2323	4554	0.60	31.96	0.17	0.81 ± 0.14	284
f_{S}	2325	4558	0.61	31.93	0.17	0.82 ± 0.14	243
f_{S-}	2323	4554	0.61	32.96	0.17	0.82 ± 0.14	254
f_{M}	2326	4560	0.68	34.94	0.18	0.80 ± 0.15	418
f_{G}	2325	4558	0.70	28.96	0.14	0.83 ± 0.13	166

Table 1. Numerical evaluations of the content-adaptiveness of cMeshes.



(a) (b) (c)

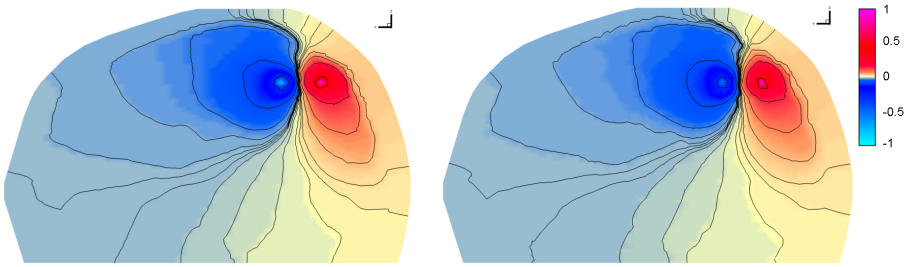
(d) (e) (f)

Fig. 5. Comparison of geometrical mesh morphology of the 3-D FE models of the whole head. Top row shows (a) a transaxial slice, (b) sagittal cutplane, and (c) coronal view from the regular mesh head model with 159,513 nodes and 945,881 tetrahedral elements through the five sub-regions segmented. Bottom row displays (d) a tranaxial slice, (e) sagittal cutplane, and (f) coronal view from the cMesh head model with 109,628 nodes and 694,588 elements. (cyan: scalp, red: skull, green: white matter, purple: gray matter, and deepskyblue: CSF).

Figs. 6(a) and (b) display the sagittal cutplanes of the 3-D forward potential maps from the regular FE (i.e., reference) and cMesh head model of the whole head respectively. The minor differences of the EEG electrical potential distribution between the regular vs. cMesh head models are directly noticeable in Figs. 6(a) and (b). In Table 2, the CC values show strong correlation of the scalp electrical potentials between the cMesh head models and reference model. The results from cMesh-2 show CC=0.999 and RE=0.037, indicating there is only minor difference in the scalp electrical potentials but significant gain in CT of 55% (5.47 to 3.02 min) with significantly reduced nodes and elements.

FE model	No. of Nodes	No. of Elements	CC	RE	CT (min)
Reference	159,513	945,881	1	0	1 (5.47)
cMesh-1	148,852	943,072	0.999	0.031	0.60 (3.28)
cMesh-2	109,628	694,588	0.999	0.037	0.55 (3.02)

Table 2. Numerical quality of the scalp electrical potentials in the cMesh head models.



(a) (b)

Fig. 6. Sagittal view of the 3-D forward potential maps from (a) the reference FE head model and (b) the cMesh-1 head model. The resultant EEG forward potentials are normalized by the maximum value of the EEG potential for isopotential visualization.

4. DT-MRI Content-adaptive Finite Element Head Model Generation

Fig. 7 describes the schematic steps of building wMesh head models along with the generation of the cMesh head model. The detailed technical steps are explained in the subsequent sections.

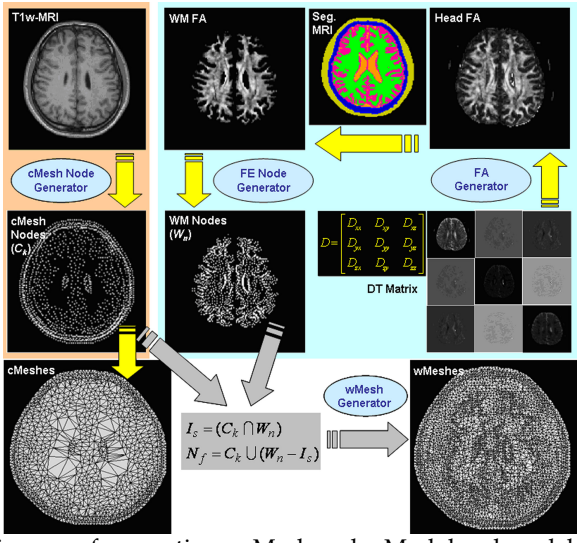


Fig. 7. Schematic diagram of generating a cMesh and wMesh head model.

4.1 DT-MRI Feature Map Generation

From DT-MRI data, the symmetric DT matrix is obtained: namely, the diffusion components along the x-y direction, the x-z direction, and the y-z direction (i.e., D_{xy}, D_{xz}, and D_{yz}) in addition to the traditional measurements of diffusivities along the x-, y-, and z-axes (i.e., D_{xx}, D_{yy}, and D_{zz}) (Bihan et al., 2001). The mathematical representation of the DT matrix is shown in Fig. 7.

For the wMesh head modeling, fractional anisotropy (FA) as an anisotropy feature map is used. The FA map is calculated using the eigenvalues of the DT matrix as follows:

$$FA = \frac{\sqrt{3}}{\sqrt{2}} \frac{\sqrt{(\lambda_1 - \lambda)^2 + (\lambda_2 - \lambda)^2 + (\lambda_3 - \lambda)^2}}{\sqrt{\lambda_1^2 + \lambda_2^2 + \lambda_3^2}}, \; 0 \le FA \le 1 \tag{22}$$

where λ_1, λ_2, and λ_3 are three eigenvalues and λ is the average of the eigenvalues.

The FA measures the ratio of the anisotropic part of the DT over the total magnitude of the tensor (Bihan et al., 2001). The minimum value of FA can occur only in a perfectly isotropic medium. The maximum value arises only when $\lambda_1 >> \lambda_2 = \lambda_3$. The FA is widely used to represent the anisotropy of the DT due to its robustness against to noise.

4.2 wMesh Generation

To build the wMesh head model, the first step is to co-register a set of T_1-weigthed MRIs to DT-MRIs using a voxel similarity-based affine registration technique (Maes et al., 1997). Then to generate the WM anisotropy-adaptive nodes, the head FA maps are derived from the measured DT matrix using Eq. (22). The WM FA maps are extracted from the head FA maps using the information of the WM regions segmented from the structural MRIs. To create the WM anisotropy-adaptive nodes based on the WM FA maps where the strong anisotropy is present, the node sampling is performed according to the spatial anisotropic density of the FA maps via the Floyd-Steinberg error diffusion algorithm technique (Floyd & Steinberg, 1975). Basically more nodes are created in the high anisotropic density regions of the FA maps.

In addition to the node generation in the WM regions based on the anisotropy feature maps, the cMesh nodes are generated from the T_1-weighted MRIs using our cMesh node generator as described in the previous sections. For the generation of the wMesh head models (see Fig. 7), the cMesh nodes C_k and WM nodes W_n are used which are expressed as:

$$C_k(x, y, z) = \{k \mid 1 \le k \le N\}, \tag{23}$$

$$W_n(x, y, z) = \{n \mid 1 \le n \le M\}, \tag{24}$$

where k and n are the nodal indices, x, y, and z the nodal coordinates in the Euclidean space, and N and M the total number of nodes of cMesh nodes C_k and WM nodes W_n respectively.

We find the intersectional node information (i.e., identical nodal positions, I_s) of C_k and W_n, using Eq. (25), since they share the same position of FE nodes which are overlapped in both the cMesh and WM node maps.

$$I_s(x, y, z) = (C_k \bigcap W_n), \tag{25}$$

$$I_s(x, y, z) = \{s \mid s \in (C_k \bigcap W_n)\}, \tag{26}$$

where s denotes the nodal indices intersected.

Then we compute the wMesh nodes N_f in the following way:

$$N_f(x,y,z) = C_k(x,y,z) \bigcup [W_n(x,y,z) - I_s(x,y,z)] . \quad (27)$$

The computed wMesh nodes N_f (i.e., the superfluous FE nodes I were removed) are used to generate the wMesh head model. The dense nodes in the WM regions are produced according to the WM anisotropic density over the cMesh nodes C_k. Once the wMesh nodes N_f are sampled from the procedures described above, the FE mesh generation using tetrahedral elements in 3-D is done via the Delaunay tessellation algorithm (Watson, 1981) to construct the wMesh head models. Fig. 7 shows the distinct mesh characteristics in the WM regions between the cMesh and wMesh head models.

4.3 Anisotropic Electrical Conductivity in wMesh

To set up the anisotropic electrical conductivity tensors in the WM tissue, we first hypothesize that the electrical conductivity tensors share the eigenvectors with the measured diffusion tensors according to the work of Basser et al. (2004). Then, we have adopted two different techniques of modeling WM anisotropy conductivity derived from the measured diffusion tensors: (i) a fixed anisotropic ratio in each WM voxel (Wolters et al., 2006) and (ii) a variable anisotropic ratio using a linear conductivity-to-diffusivity relationship in combination with a constraint on the magnitude of the electrical conductivity tensor (Hallez et al., 2008). Two different approaches of deriving the WM anisotropic conductivity tensors are briefly described as below.

To derive the WM anisotropic conductivity tensor with a fixed anisotropic ratio, the anisotropic conductivity tensor σ of the WM compartments is expressed as:

$$\sigma = \mathbf{S}\,\mathrm{diag}\left(\sigma_{long}, \sigma_{trans}, \sigma_{trans}\right)\mathbf{S}^{-1} \quad (28)$$

where $\mathbf{S}$ is the orthogonal matrix of unit length eigenvectors of the measured DT at the barycenter of the WM FEs. σ_{long} and σ_{trans} denote the eigenvalues parallel (longitudinal) and perpendicular (transverse) to the fiber directions, respectively, with $\sigma_{long} \geq \sigma_{trans}$.

Then we computed the longitudinal and transverse eigenvalues (i.e., anisotropic ratio of σ_{long} and σ_{trans}) using the volume constraint (Wolters et al., 2006) retaining the geometric mean of the eigenvalues. The volume of the conductivity tensor is calculated as follows:

$$\frac{4}{3}\pi\sigma_{iso}^3 = \frac{4}{3}\pi\sigma_{long}\sigma_{trans}^2 . \quad (29)$$

The anisotropic FE head models differing in the anisotropic ratio (i.e., 1:2, 1:5, 1:10, and 1:100) are generated using different conductivity tensor eigenvalues under the volume constraint algorithm, Eq. (29).

To compute the WM anisotropic conductivity tensors with the variable (or proportional) anisotropic ratios, a linear scaling approach of the diffusion tensor ellipsoids is used according to the self-consistent effective medium approach (EMA) (Sen et al., 1989; Tuch et

al., 1999; 2001). EMA states a linear relationship between the eigenvalues of the conductivity tensor σ and the eigenvalue of diffusion tensor d in the following way:

$$\sigma = \frac{\sigma_e}{d_e} d \tag{30}$$

where σ_e and d_e represent the extracellular conductivity and diffusivity respectively (Tuch et al., 2001). This approximated linear relationship assumes the intracellular conductivity to be negligible (Tuch et al., 2001; Haueisen et al., 2002). According to the proposition by Hallez et al. (2008), the scaling factor σ_e / d_e can be computed using the volume constraint in Eq. (29) as shown below.

The linear relationship between the conductivity tensor eigenvalues and diffusion tensor eigenvalues in the WM regions can be represented as

$$\frac{d_1}{\sigma_1} = \frac{d_2}{\sigma_2} = \frac{d_3}{\sigma_3} \tag{31}$$

where d_1, d_2, and d_3 are the eigenvalues of the diffusion tensor at each WM voxel. σ_1, σ_2, and σ_3 are the unknown eigenvalues of the electrical conductivity tensor at the corresponding voxel. Then the volume constraint algorithm as in Eq. (29) can be applied to compute the anisotropic electrical conductivities. The volume constraint equation can be rewritten as follows:

$$\frac{4}{3}\pi\sigma_{iso}^3 = \frac{4}{3}\pi\sigma_1\sigma_2\sigma_3 \tag{32}$$

where σ_1 is the eigenvalues to the largest eigenvector. σ_2 and σ_3 represent the eigenvalues to the perpendicular eigenvectors, respectively.

4.4 Analysis on DT-MRI Content-adaptive Meshes

4.4.1 Comparison of Anisotropy Adaptiveness and Anisotropy Tensor Mapping

To examine the effectiveness of the wMesh head model, we tested both anisotropy adaptiveness and the quality of anisotropic mapping into the meshes by comparing to the regular mesh and cMesh head models.

Fig. 8 shows a set of exemplary results from a regions of interest (ROI) to compare the anisotropy adaptiveness of the FE head models to the given mesh morphology. Fig. 8(a) shows a transaxial T_1-weighted MRI. The ROI, enclosing 38 × 38 voxels, is highlighted with a box in red on the T_1-weighted MRI. The enlarged ROI of the T_1-weighted MRI and its corresponding color-coded FA map derived from the DTs are given in Figs. 8(b) and (c) respectively. In Fig. 8(c), the projections of the principal tensor directions on the ROI color-coded FA map are visualized with white lines. Figs. 8(d)-(f) show the ROI regular meshes, cMeshes, and wMeshes respectively. In contrast to the regular meshes in Fig. 8(d), the anisotropy-adaptive characteristics of the wMeshes according to the WM anisotropy information is clearly noticeable in Fig. 8(f). Moreover, it appears that there is higher mesh

density in the WM regions where the degree of the anisotropy is strongly present. The results from wMeshes demonstrate that mapping the WM electrical anisotropy into the meshes could be performed more accurately. As mentioned previously, cMeshes in Fig. 8(e) show too coarse mesh characteristics in the WM tissues, which seem be unsuitable for the incorporation of the WM tensor anisotropy.

We next examined the quality of anisotropy mapping into the meshes which could be important since the correct representation of anisotropy affects the accuracy of FEA. Fig. 9 illustrates the projection of the DT ellipsoids overlaid on the transaxial slice of a T_2-weighted MRI. Fig. 9(a) displays the original DT ellipsoids in the WM tissues. In the corresponding WM regions, the DT ellipsoids at the barycenters of the WM elements from the wMeshes are shown in Fig. 9(b). The diameters in any directions of the DT ellipsoids reflect the diffusivities in their corresponding directions, and their major principle axes are oriented in the directions of maximum diffusivities. As observed in Fig. 9(d), the wMeshes are likely to provide a better way of reflecting the details of the directionality and magnitude of the anisotropic tensors due to the dense mesh features and anisotropy-adaptive characteristics in the WM regions. In other words, the wMesh head model better incorporate the WM anisotropic electrical conductivities and thereby the errors of the anisotropy modeling could be reduced.

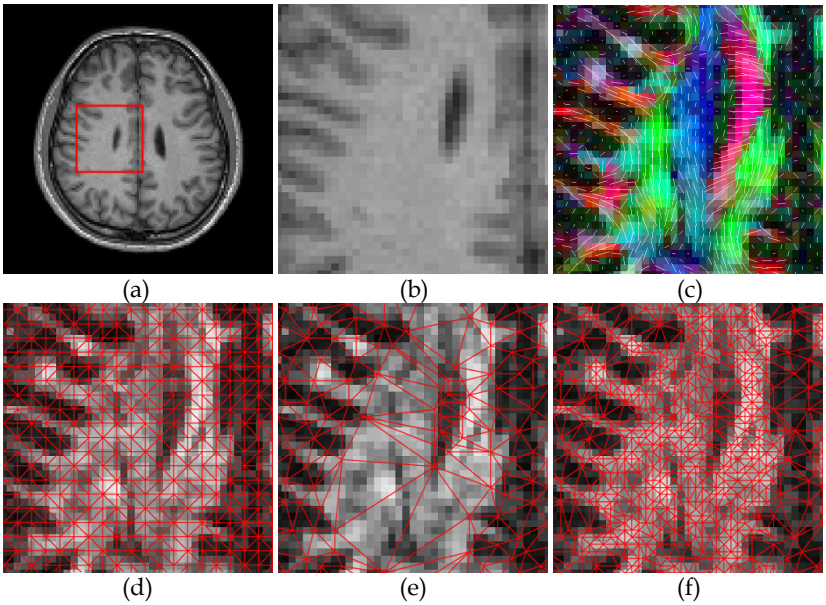
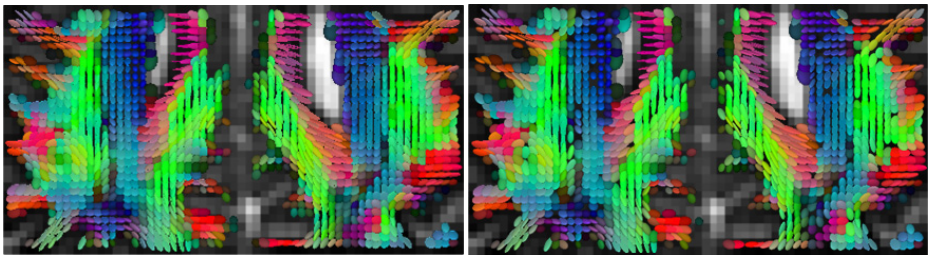


Fig. 8. Anisotropy adaptiveness of the FE head models. (a) a transaxial T_1-weighted MR slice including the red box which indicates the regions of interest (ROI), (b) a ROI (38 x 38 voxels) of the T_1-weighted MRI, (c) corresponding color-coded FA with the projections of the principal tensor directions shown as white lines, (d) regular meshes overlaid on the FA map, (e) cMeshes, and (f) wMeshes.



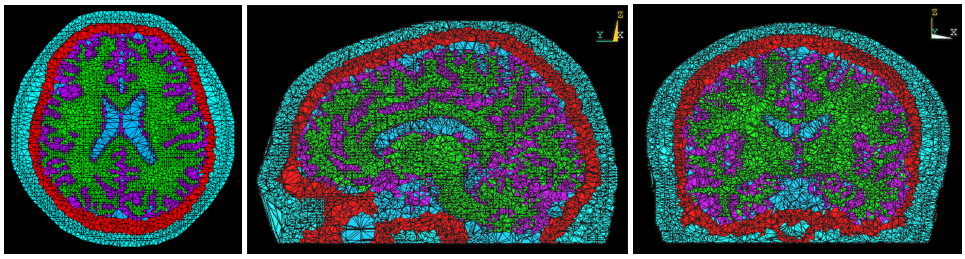
(a) (b)

Fig. 9. Mapping the DT ellipsoids of the WM regions onto a transaxial cut of the T_2-weighted MRI: (a) the original DT ellipsoids and (b) DT ellipsoids in the barycenters of the WM elements from the wMesh head model. The color indicates the orientation of the principal tensor eigenvector (red: mediolateral, green: anteroposterior, and blue: superoinferior direction).

4.4.2 Effect of Anisotropic Electrical Conductivity

To study the effects of the WM anisotropic electrical conductivity on the EEG forward solutions, we compared the EEG electrical potentials from the anisotropic wMesh head models against those of the isotropic models. To obtain the EEG forward potentials, we solved the Poisson's equation (Sarvas, 1987) due to the following current sources (Yan et al., 1991; Schimpf et al., 2002): as superficial sources, (i) an approximately tangentially oriented source (the posterior-anterior direction) and (ii) a radially oriented source (the inferior-superior direction) in the cortex; as a deep source (iii) an approximately radial source in the thalamus. Each dipole was placed in the isotropic gray matter regions with careful attention, since EEG fields are particularly sensitive to the conductivity changes of the brain tissue next to the dipole (Haueisen et al., 1997; Gencer & Acar, 2004).

Fig. 10 visualizes the wMesh model of the whole head through the five sub-regions segmented. There are 160,230 nodes and 1,009,440 tetrahedral elements in the wMesh head model. The fully automatic generation of the wMeshes took 80.3 sec on a PC with Pentium-IV CPU 3.0 GHz and 2GB RAM. Figs 10(a)-(c) display the transaxial, sagittal, and coronal view of the head model respectively. The wMeshes in Fig. 10 show dense and adaptive meshes in the WM regions generated based on the WM FA information. It is also seen that compared to the regular FE head model in Fig. 5(a)-(c), there are much smoother boundaries of the meshes at the skin, outer, and inner regions, thus possibly avoiding the stair-step approximation of curved boundaries (e.g., Wolters et al., 2007) and reducing EEG forward modeling errors. The WM anisotropy-adaptive meshing technique offers an optimal way of incorporating the WM anisotropic conductivity tensors into the meshes.



(a) (b) (c)

Fig. 10. Visualization of the 3-D wMesh model of the whole head with 160,230 nodes and 1,009,440 tetrahedral elements (color labeling as described in Fig. 5): (a) a transaxial slice, (b) sagittal cutplane, and (c) coronal view.

Fig. 11 displays the results of the EEG forward potential maps from the wMesh models of the whole head. According to the given source types, the resultant EEG forward distributions of the sagittal and coronal views from the isotropic wMesh models are visualized in Figs. 11(a)-(c) respectively. The EEG potential maps from the anisotropic wMesh head models at the 1:10 fixed anisotropic ratio are shown in Figs. 11(d) and (e). Fig. 11(f) shows the EEG potential distributions from the wMesh model at the anisotropic ratio of 1:100. Based on the observation in Fig. 11, the differences of the EEG electrical potential distributions between the isotropic vs. anisotropic wMesh models are directly noticeable through the altering directions and extension of the isopotential lines. In particular, the isopotentials in Fig. 11(f) show the greater effects of the WM anisotropic conductivities due to the strong anisotropy of 1:100.

To evaluate the numerical differences of the EEG forward solutions between the isotropic vs. anisotropic wMesh models, the scalp potential values were quantitatively compared using two similarity measures: relative difference measure (RDM) and magnification factor (MAG). Meijs et al., (1999) introduced these metrics to quantify the topography and magnitude errors. The quantitative results of the scalp electrical potentials according to different anisotropy settings are given in Table 3.

The results from the wMesh head model with the 1:10 anisotropic ratio using the tangential dipole show that the inclusion of the WM anisotropy resulted in the low RDM value of 0.037 and the MAG value of 0.959. On the other hand, a slightly larger influence (RDM=0.046 and MAG=0.910) was found in the wMesh model with the 1:10 anisotropic ratio for the radial dipole, thus indicating that the WM anisotropy led to the topography errors of the EEG and weakened the EEG fields. Moreover, the strong effects by the 1:100 WM anisotropy ratio were observed in the MAG value of 0.427, describing the WM anisotropy strongly weakened the EEG potential fields. The WM anisotropic conductivities around the deep source have a greater influence on the EEG forward solutions. In the case of the anisotropic models by the variable anisotropy setting, the results show smaller differences on the EEG forward solutions due to much lower variable anisotropic ratios of the WM anisotropic electrical conductivities.

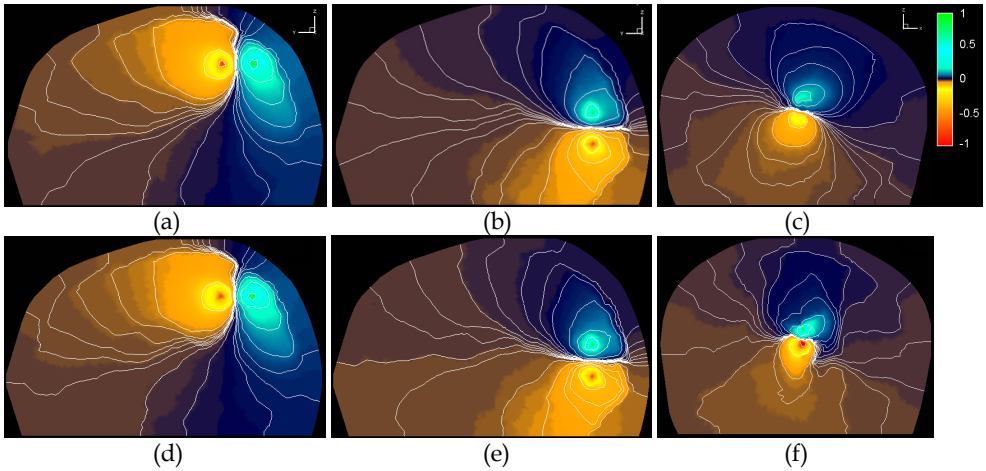


Fig. 11. EEG forward potential maps of the isotropic vs. anisotropic wMesh models of the whole head: Top row from the isotropic models, (a) the sagittal cutplane with a tangentially oriented dipole, (b) with a radially oriented dipole, and (c) coronal view with a deep source. Bottom row from the anisotropic models, (d) sagittal cutplane with a tangentially oriented dipole with the fixed anisotropic ratio of 1:10, (e) with a radially oriented dipole at the 1:10 fixed anisotropic ratio, and (f) coronal view with a deep source with the 1:100 fixed anisotropic ratio. The resultant EEG forward potentials are normalized by the maximum value of the EEG potential for isopotential visualization.

Anisotropy	Ratio	Tangential		Radial		Deep	
		RDM	MAG	RDM	MAG	RDM	MAG
Fixed	1:2	0.010	0.998	0.012	0.992	0.012	0.988
	1:5	0.024	0.983	0.030	0.957	0.053	0.924
	1:10	0.037	0.959	0.046	0.910	0.111	0.838
	1:100	0.080	0.790	0.153	0.640	0.540	0.427
Variable		0.022	0.992	0.022	0.986	0.045	0.982

Table. 3. Numerical differences of the scalp electical potentials between the isotropic vs. anisotorpic wMesh head models.

5. Conclusion

In this chapter, we have introduced how to generate MRI content–adaptive FE meshes (i.e., cMesh) and DT-MRI anisotropy–adaptive FE meshes (i.e., wMesh) of the human head in 3-D. These cMesh and wMesh generation methodologies are fully automatic with the pre-segmented boundary information of the sub-regions of the head (such as gray matter, white matter, CSF, skull, and scalp), DT information, and conductivity values of the segmented regions. Although the choice of using cMesh or wMesh depends on the aim of each FEA, the combination of these meshes should allow high-resolution FE modelling of the head. Also the presented technique should be extendable to other parts of the human body and their FEA of bioelectromagnetic phenomenon thereof.

6. Acknowledgement

This work was supported by a grant of Korea Health 21 R&D Project, Ministry of Health and Welfare, Republic of Korea (02-PJ3-PG6-EV07-0002). This work was also supported by the Korea Science and Engineering Foundation (KOSEF) grant funded by the Korea government (MEST) (2009-0075462).

7. References

Abd-Elmoniem, K. Z.; Youssef, A. M. & Kadah, Y. M. (2002). Real-time speckle reduction and coherence enhancement in ultrasound imaging via nonlinear anisotropic diffusion. *IEEE Trans. Biomed. Eng.*, Vol. 49, No. 9, 997-1014, 0018-9294

Ardizzone, E. & Rirrone, R. (2003). Automatic segmentation of MR images based on adaptive anisotropic filtering, Proceedings of *IEEE Int. Conf. Image Ana. Process.* (ICIAP'03), pp. 283-288, 0-7695-1948-2, Italy, Sept., 2003, IEEE

Awada, K. A.; Jackson, D. R.; Baumann, S. B.; Williams, J. T.; Wilton, D. R.; Baumann, S. B. & Papanicolaou, A. C. (1997). Computational aspects of finite element modeling in EEG source localization. *IEEE Trans. Biomed. Eng.*, Vol. 44, No. 8, 736-752, 0018-9294

Baillet, S.; Mosher, J. C. & Leahy, R. M. (2001). Electromagnetic brain mapping. *IEEE Sig. Process. Mag.,* Nov., 14-30, 1053-5888

Basser, P. J.; Mattiello, J. & Bihan, D. L. (1994). MR diffusion tensor spectroscopy and imaging. *Biophys. J.* Vol. 66, 259-67, 0006-3495

Berzins, M. (1999). Mesh quality: a function of geometry, error estimates or both?. *Eng. with Comp.*, Vol. 15, 236-247, 0177-0667

Bihan, D. L.; Mangin, J. F.; Poupon, C.; Clark, C. A.; Pappata, S.; Molko, N. & Chabriat, H. (2001). Diffusion tensor imaging: concepts and applications. *J. MRI,* Vol. 37, 534-546, 1053-1807

Buchner, H.; Knoll, G.; Fuchs, M.; Rienaker, A.; Beckmann, R.; Wagner, M.; Silny, J. & Pesch, J. (1997). Inverse localization of electric dipole current sources in finite element models of the human head. *Electroenceph. Clin. Neurophysiol.,* Vol. 102, 267-278, 1388-2457

Carmona, R. A. & Zhong, S. (1998). Adaptive smoothing respecting feature directions. *IEEE Trans. Image Process.*, Vol. 7, No. 3, 353-358, 1057-7149

Dogdas, B.; Shattuck, D. W. & Leahy, R. M. (2005). Segmentation of skull and scalp in 3-D human MRI using mathematical morphology. *Hum. Brain Mapping,* Vol. 26, 273-285, 1065-9471

Field, D. A. (2000). Qualitative measures for initial measures. *Int. J. Numer. Meth. Eng.,* Vol. 47, 887-906, 0029-5981

Floyd, R. & Steinberg, L. (1975). An adaptive algorithm for spatial gray scale. in SID *Int. Symp. Digest of Tech.* 36-37, 0003-966X

Gencer, N. G. & Acar, C. E. (2004). Sensitivity of EEG and MEG measurements to tissue conductivity. *Phys. Med. Biol.,* Vol. 49, 701-717, 0031-9155

Gray, A. (1997). The gaussian and mean curvatures and surfaces of constant gaussian curvature. §16.5 and Ch. 21 in *Modern Differential Geometry of Curves and Surfaces with Mathematica, 2nd ed.* Boca Raton, FL: CRC Press, 373-380 and 481-500, 1997

Hallez, H.; Vanrumste, B.; Hese, P. V.; Delputte, S. & Lemahiueu, I. (2008). Dipole estimation errors due to differences in modeling anisotropic conductivities in realistic head models for EEG source analysis. *Phys. Med. Biol.,* Vol. 53, 1877-1894, 0031-9155

Hamalainen, M. S. & Sarvas, J. (1989). Realistic conductivity geometry model of the human head for interpretation of neuromagnetic data. *IEEE Trans. Biomed. Eng.,* Vol. 36, No. 2, 165-171, 0018-9294

Haueisen, J.; Ramon, C.; Brauer, H. & Nowak, H. (1997). Influence of tissue resistivities on neuromagnetic fields and electric potentials studied with a finite element model of the head. *IEEE Trans. Biomed. Eng*., Vol. 44, No. 8, 727-735, 0018-9294

Haueisen, J.; Tuch, D. S.; Ramon, C.; Schimpf, P. H.; Wedeen, V. J.; George, J. S. & Belliveau, J. W. (2002). The influence of brain tissue anisotropy on human EEG and MEG. *NeuroImage,* Vol. 15, 159-66, 1053-8119

He, B.; Musha, T.; Okamoto, Y.; Homma, S.; Nakajima, Y. & Sato, T. (1987). Electric dipole tracing in the brain by means of the boundary element method and its accuracy. *IEEE Trans. Biomed. Eng.,* Vol. 34, No. 6, 406-414, 0018-9294

Katsavounidis, I. & Kuo, C-C. J. (1997). A multiscale error diffusion technique for digital halftoning. *IEEE Trans. Image Process.,* Vol. 6, No. 3, 483-490, 1057-7149

Kim, S.; Kim, T.-S.; Zhou, Y. & Singh, M. (2003). Influence of conductivity tensors on the scalp electrical potential: study with 2-D finite element models. *IEEE Trans. Nucl. Sci*. Vol. 50, No. 1, 133-138, 0018-9499

Kim, T.-S.; Zhou, Y.; Kim, S. & Singh, M. (2002). EEG distributed source imaging with a realistic finite-element head model. *IEEE Trans. Nucl. Sci*., Vol. 49, No. 3, 745-752, 0018-9499

Kim, T.-S.; Jeong, J.; Shin, D.; Huang, C.; Singh, M. & Marmarelis, V. Z. (2003). Sinogram enhancement for ultrasonic transmission tomography using coherence enhancing diffusion, *Proceedings of IEEE Int. Symposium on Ultrasonics,* pp. 1816-1819, 0-7803-7922-5, Hawaii, USA, Oct., 2004, IEEE

Kim, T.-S.; Kim, S.; Huang, D. & Singh, M. (2004). DT-MRI regularization using 3-D nonlinear gradient vector flow anisotropic diffusion, *Proceedings of Int. Conf. IEEE Eng. Med. Biol.,* pp. 1880-1883, 0-7803-8439-3, San Francisco, USA, Sep., 2004, IEEE

Kim, H. J.; Kim, Y. T.; Minhas, A. S.; Jeong, W. C.; Woo, E. J.; Seo, J. K. & Kwon, O. J. (2009). In vivo high-resolution conductivity imaging of human leg using MREIT: the first human experiment. *IEEE Trans. Med. Imag.,* Vol. 28, No. 1, 0278-0062

Lee, W. H.; Kim, T.-S.; Cho, M. H.; Ahn, Y. B. & Lee, S. Y. (2006). Methods and evaluations of MRI content-adaptive finite element mesh generation for bioelectromagnetic Problems. *Phys. Med. Biol*., Vol. 51, No. 23, 6173-6186, 0031-9155

Lee, W. H.; Seo, H. S.; Kim, S. H.; Cho, M. H.; Lee, S. Y. & Kim, T.-S. (2008). Influnce of white matter anisotropy on the effects of transcranial direct current stimulation: a finite element study, *Proceedings of Int. Conf. Biomed. Eng.,* pp. 460-464, 978-3-540-92840-9, Singapore, Dec., 2008, Springer Berlin Heidelberg.

Maes, F.; Collignon, A.; Vandermeulen, D.; Marchal, G.; Marchal, G. & Suetens, P. (1997). Multimodality image registration by maximization of mutual information. *IEEE Trans. Med. Imag.,* Vol. 16, 187-198, 0278-0062

Marin, G.; Guerin, C.; Baillet, S.; Garnero, L. & Meunier, G. (1998). Influence of skull anisotropy for the forward and inverse problem in EEG: simulation studies using FEM on realistic head models. *Hum. Brain Mapp.,* Vol. 6, 250-269, 1065-9471

Meijs, J. W. H.; Weier, O. W.; Peters, M. J. & Oosterom, A. V. (1989). On the numerical accuracy of the boundary element method. *IEEE Trans. Biomed. Eng.,* Vol. 36, No. 10, 1038-1049, 0018-9294

Neilson, L. A.; Kovalyov, M. & Koles, Z. J. (2005). A computationally efficient method for accurately solving the EEG forward problem in a finely discretized head model. *Clin. Neurophysiol.,* Vol. 116, 2302-2314, 1388-2457

Sarvas, J. (1987). Basic mathematical and electromagnetic concepts of the biomagnetic inverse problem. *Phys. Med. Biol.,* Vol. 32, 11-22, 0031-9155

Schimpf, P.; Ramon, C. & Haueisen, J. (2002). Dipole models for the EEG and MEG. *IEEE Trans. Biomed. Eng.*. Vol. 49, No. 5, 409-418, 0018-9294

Sen, A. K. & Torquato, S. (1989). Effective electrical conductivity of two-phase disordered anisotropic composite media. *Phys. Rev. B. Condens. Matter.,* Vol. 39, 4504-4515, 1098-0121

Shattuck, D. W. & Leahy, R. M. (2002). BrainSuite: an automated cortical surface identification tool. *Med. Image Anal.,* Vol. 8, 129-142, 1361-8415

Tschumperle, D. & Deriche, R. (2002). Diffusion PDEs on vector-valued images. *IEEE Sig. Proc. Mag.*, Sep., 16-25, 1053-5888

Tuch, D. S.; Wedeen, V. J.; Dale, A. M.; George, J. S. & Belliveau J. W. (1999). Conductivity mapping of biological tissue using the diffusion MRI. *Ann. N. Y. Acad. Sci.,* Vol. 888, 314-316, 0077-8923

Tuch, D. S.; Wedeen, V.; Dale, A.; George, J. & Belliveau, J. (2001). Conductivity tensor mapping of the human brain using diffusion tensor MRI. *Proc. Natl. Acad. Sci. USA,* Vol. 98, 11697-11701, 0027-8424

Voo, V.; Kumaresan, S.; Pintar, F. A.; Yoganandan, N. & Sances, A. (1996). Finite-element models of the human head. *Med. Biol. Eng. Comput.,* Vol. 34, No. 5, 375-381, 0140-0118

Watson, D. F. (1981). Computing the n-dimensional Delaunay tessellation with application to Voronoi polytypes. *The Comp. Jour.*, Vol. 24, No. 2, 167-172, 1460-2067

Weickert, J. (1997). A review of nonlinear diffusion filtering, In: *Scale-Space Theory in Computer Vision,* Romeny, B. ter Haar., Florack L, Koenderink J, Vierver M (Ed.) Vol. 1252, 3-28, Springer Berlin, 978-3-540-63167-5

Wendel, K.; Narra, N. G.; Hannula, M.; Kauppinen, P. & Malmivuo, J. (2008). The influence of CSF on EEG sensitivity distributions of multilayered head models. *IEEE Trans. Biomed. Eng.,* Vol. 55, No. 4, 1454-1456, 0018-9294

Wolters, C. H.; Anwander, A.; Tricoche, X.; Weinstein, D.; Koch, M. A. & MacLeod, R. S. (2006). Influence of tissue conductivity anisotropy on EEG/MEG field and return current computation in a realistic head model: a simulation and visualization study using high-resolution finite element modeling. *NeuroImage,* Vol. 30, 813-826, 1053-8119

Wolters, C. H.; Anwander, A.; Berti, G. & Hartmann, U. (2007). Geometry-adapted hexahedral meshes improve accuracy of finite-element-method-based EEG source analysis *IEEE Trans. Biomed., Eng.* Vol. 54, No. 8, 1446-1153, 0018-9294

Yan, Y.; Nunez, P. L. & Hart, R. T. (1991). Finite-element model of the human head: scalp potentials due to dipole sources. *Med. Biol. Eng. Comput.,* Vol. 29, 475-481, 0140-0118

Yang, Y.; Wernick, M. N. & Brankov, J. G. (2003). A fast approach for accurate content-adaptive mesh generation. *IEEE Trans. Image Process.*, Vol. 12, No. 8, 866-881, 1057-7149

Yezzi, A. (1998). Modified curvature motion for image smoothing and enhancement. *IEEE Trans. Image Process.*, Vol. 7, No. 3, 345-352, 1057-7149

Zhang, Y. C.; Ding, L.; van Drongelen, W.; Hecox, K.; Frim, D. M. & He, B. (2006). A cortical potential imaging study from simultaneous extra- and intracranial electrical recordings by means of the finite element method. *NeuroImage* Vol. 31, 1513-1524, 1053-8119

15

Denoising of Fluorescence Confocal Microscopy Images with Photobleaching compensation in a Bayesian framework

Isabel Rodrigues[1,3] and João Sanches[2]
Institute for Systems and Robotics[1], Instituto Superior Técnico[2], Instituto Superior de Engenharia de Lisboa[3]
Portugal

1. Introduction

Fluorescence confocal microscopy imaging is today one of the most important tools in biomedical research. In this modality the image intensity information is obtained from specific tagging proteins that fluoresce nanoseconds after the absorption of photons associated with a specific wavelength radiation. Additionally, the confocal technology rejects all the out-focus radiation, thereby allowing 3-D imaging almost without blur. Therefore, this technique is highly selective allowing the tracking of specific molecules, tagged with fluorescent dye, in living cells (J.W.Lichtman & J.A.Conchello, 2005). However, several difficulties associated with this technology such as multiplicative noise, *photobleaching* and photo-toxicity, affect the observed images. These undesirable effects become more serious when higher acquisition rates are needed to observe fast kinetic processes in living cells.

One of the main sources of these problems resides on the huge amount of amplification used to amplify the small amount of radiation captured by the microscope, required to observe the specimen. The amplification process, based on photo-multiplicative devices, generates images corrupted by a type of multiplicative noise with Poisson distribution, characteristic of low photon counting process (R. Willett, 2006). In fact the number of photons that are collected by the detector at each point in the image determines the signal quality at that point. The noise distorting these images is in general more limiting than the resolution of the confocal microscope.

The fluorophore is the component of the molecule responsible for its capability to fluoresce. The *photobleaching* effect consists on the irreversible destruction of the fluorescence of the fluorophores due to photochemical reactions induced by the incident radiation (J.Braga et al., 2004; Lippincott-Schwarz et al., 2003). Upon extended excitation all the fluorophores will eventually photobleach, which leads to a fading in the intensity of sequences of acquired images along the time. This effect prevents long exposure time experiments needed to analyse biologic processes with a long lasting kinetics (Lippincott-Schwarz et al., 2001).

The photochemical reactions associated with the *photobleaching* effect also produce free radicals toxic to the specimen. This photo-toxicity (J.W.Lichtman & J.A.Conchello, 2005) effect increases along with the power of the incident radiation.

Establishing the right amount of incident radiation is a key point in this microscope modality. On one hand, increasing the illumination increases the photon count which improves the quality of the signal, but on the other hand, this increasing of the incident radiation speeds up the *photobleaching* and photo-toxicity effects that increase the quality degradation of the acquired images and in the limit, may lead to a premature death of the cells.

Many algorithms that deal with this type of microscopy images are conceived under the assumption of the *additive white Gaussian noise* (AWGN) model. However, the *multiplicative white Poisson noise* (MWPN) model is more appropriated to describe the noise corrupting laser scanning fluorescence confocal microscope (LSFCM) images due to the photon-limited characteristics, whose main attribute is its dependence on the image intensity. In order to take advantage of all the knowledge on AWGN denoising, some authors, instead of using the Poisson statistics of the noisy observations, they prefer to modify it introducing variance stabilizing transformations, such as the Anscombe or the Fisz transforms (M.Fisz,1955; P.Fryźlewicz & G.Nason, 2001). However, even applying the Anscombe transform, the additive AWGN assumption is accurate only when each photon count is larger than thirty (R.Willett, 2006).

In the seventies, W.H. Richardson and L. Lucy in separate works developed a specific methodology for data following a Poisson distribution. The Richardson-Lucy (R-L) algorithm can be viewed as an Expectation-Maximization (EM) algorithm including a Poisson statistical noise model. This algorithm presents several weaknesses such as the amplification of the noise after a few iterations, in particular when the *signal to noise ratio* (SNR) is low, which is the case of LSFCM images.

More recently several works on denoising methods applied to photon-limited imaging have come up in the literature. Methods based on wavelet and other similar transforms were developed by several authors (K. Timmermann & R. Nowak, 1999), (P.Besbeas et al., 2004), (R.Willett & R. Nowak, 2004), among many others. In conjunction with the use of the Poisson statistics, in the Bayesian framework, several regularization schemes have been proposed. Dey et al. from INRIA have proposed diverse deconvolution/denoising methods in the Bayesian framework for confocal microscopy with *total variation* (TV) regularization (N. Dey et. al. 2004, 2006). The authors conceived a combination of the R-L algorithm with a regularizing TV based constraint, whose smoothing avoids oscillations in homogeneous regions while preserving edges. The TV regularization was also used in conjunction with a multilevel algorithm for Poisson noise removal (Chan & Chen, 2007). Adaptive window approaches have been conceived for Poisson noise reduction and morphology preserving in Confocal Microscopy (C.Kervrann & A.Trubuil, 2004). Non-parametric regression methods have been developed in (J.Boulanger et al., 2008) for the denoising of sequences of fluorescence microscopy volumetric images (3-D+t). In this case the authors adopted a variance stabilizing procedure with a generalized Ascombe transform to combine Poisson and Gaussian noise models and proposed an adaptive patch-based framework able to preserve space-time discontinuities and simultaneously to reduce the noise level of the sequences. Other approach was proposed by Dupé (Dupé, et al., 2008) where a deconvolution algorithm uses a fast proximal backward-forward spitting iteration which

minimizes an energy function whose data fidelity term accounts for Poisson noise and an L_1 non-smooth sparsity regularization term acts upon the coefficients of a dictionary of transforms such as wavelets and curvelets.

Here a denoising algorithm for Poisson data that explicitly takes into account the *photobleaching* effect is presented, under the assumption that among all the complex mechanisms associated to overlapping phenomena that can cause the fading of the intensity in fluorescence microscopy, the photochemical one is the most relevant.

The main goal of the proposed algorithm is to estimate the time and space varying morphology of the cell nucleus and simultaneously the intensity decay rate due to *photobleaching* of fluorescence microscopy images of human cells.

The intensity decrease along the time is modelled by a decaying exponential with a constant rate. The algorithm is formulated in the Bayesian framework as an optimization task where a convex energy function is minimized.
Maximum a posteriori (MAP) estimation criterion is employed since it has been successfully used in other modalities, specially for image restoration purposes.
In general the denoising process is an ill-posed and an ill-conditioned problem (Vogel, 1998) requiring some sort of regularization. In the Bayesian framework the regularization effect is achieved by using *a prior* distribution functions that are jointly maximized with the distribution functions associated with the observation model describing the noise generation process.
Given the characteristics of these images, the local markovianity of the nucleus morphology seams to be a reasonable assumption. Thus, according to the Hammersley-Clifford theorem (J.Besag, 1986), a Gibbs distribution with appropriate potentials can be considered as the ***a prior*** knowledge about the cell nucleus morphology.
Several potentials have been proposed in the literature (T.Hebert & R.Leahy, 1989) and among them, one of the most popular of these functions is the quadratic, mainly for the sake of mathematic simplicity. However this function over-smoothes the solution. Since it is assumed that the morphology of the cell consists of sets of homogeneous regions separated by well defined boundaries, an alternative is the use of *edge preserving priors* such as *total variation* (TV) based potential functions that have been applied with success in several problems (L. I. Rudin et al., 1992; J. Bardsley & A. Luttman, 2006; N. Dey et al., 2004).
Very recently a new type of norms, called log-*Euclidean* norms, was proposed in (V. Arsigny et al., 2006). The interaction between neighbouring pixels that regularizes the solution imposed by the potential functions using this type of norms is based on the ratio of their intensities and not on its difference. This new approach is particularly suitable to be used in this case due to the positive nature of the optimization task associated with the denoising process of the LSFCM images. The advantage of this type of norms is more perceivable in small intensity regions where differences between neighbours are small while their ratios may exhibit relevant values. The penalization cost obtained with difference based priors may not be enough to remove the noise in these small intensity regions while the penalization costs induced by the ratio based priors may be strong enough to do it.

In this paper these log-Euclidean norms are jointly used with the total variation based priors to improve the performance of the denoising algorithm in the small intensity regions and simultaneously preserve the transitions across the entire image due to the TV approach.
Synthetic data were generated with a low level of SNR and Monte Carlo experiments were carried on with these data in order to evaluate the performance of the algorithm.
Real data of a HeLa immortal cell nucleus (D.Jackson, 1998), acquired by a laser scanning fluorescence confocal microscope (LSFCM), are used to illustrate the application of the algorithm.

2. Problem Formulation

Each sequence of $M \times N$ fluorescence microscopy images under analysis, Y, corresponds to L observations of a cell nucleus acquired along the time. Data can be represented by a 3D tensor, $Y = \{y_{i,j,t}\}$, with $0 \le i, j, t \le N - 1, M - 1, L - 1$. Each pixel, $y_{i,j,t}$, is corrupted by Poisson noise and the time intensity decrease due to the *photobleaching* effect is modelled by a decaying exponential whose rate, denoted by λ, is assumed to be constant in time and in space.
The goal of the algorithm described here is the estimation of human cells morphology along the time as well as the intensity decay rate, λ, associated with the *photobleaching* effect, from the noisy sequence Y, usually exhibiting a low *signal to noise* ratio (SNR).

The proposed method consists of an iterative algorithm performed in two-steps. In the first step the intensity decay rate coefficient, λ, is estimated jointly with a crude time invariant basic morphology version of the cell.
In the second step a more realistic time and space varying version of the cell nucleus morphology is estimated by using the intensity decay rate coefficient, λ, obtained in the previous step.
The overall estimation process needs to be decomposed in these two steps in order to decouple the sources of intensity changes which are the *photobleaching* effect, estimated in the first step, and the real cell morphology changes in time and space, estimated in the second step.

2.1 Step one

Let us focus now on the first step of the algorithm where the rate of decay due to the *photobleaching* is estimated.
Let X be an approximated noiseless version of the noisy data, Y, where

$$x_{i,j,t} = g_{i,j}\ e^{-\lambda t} \tag{1}$$

$\mathbf{G} = \{g_{i,j}\}$, with $0 \le i, j \le N - 1, M - 1$, represents a time invariant version of the cell morphology and $e^{-\lambda t}$ represents the time intensity decay term that models the *photobleaching* effect. By adopting this model all the time variability of the intensity in the images is caught by the exponential term in order to accurately estimate the rate of decay due to the *photobleaching*.

A Bayesian approach using the *maximum a posteriori* (MAP) criterion is adopted to estimate **G** and λ. The problem may be formulated as the following energy optimization task

$$\left(\hat{\mathbf{G}},\hat{\lambda}\right)=\underset{\mathbf{G},\lambda}{\operatorname{argmin}}\ \mathbf{E}\left(\mathbf{G},\lambda,\mathbf{Y}\right) \tag{2}$$

where the energy function $\mathbf{E}(\mathbf{G},\lambda,\mathbf{Y})=\mathbf{E}_{\mathbf{Y}}(\mathbf{G},\lambda,\mathbf{Y})+\mathbf{E}_{\mathbf{G}}(\mathbf{G})$ is the sum of two terms, a data fidelity term, $\mathbf{E}_{\mathbf{Y}}(\mathbf{G},\lambda,\mathbf{Y})$, and a prior term, $\mathbf{E}_{\mathbf{G}}(\mathbf{G})$, needed do regularize the solution. The *a priori* information for λ is merely its overall constancy. The first term of this sum pushes the solution towards the observations according to the type of noise corrupting the images and the *a priori* energy term penalizes the solution in agreement with some previous knowledge about **G** (T. K. Moon & W. C. Stirling, 2000).

Assuming the independence of the observations the *data fidelity term*, which is the anti-logarithm of the *likelihood* function, is

$$\mathbf{E}_{\mathbf{Y}}(\mathbf{G},\lambda,\mathbf{Y})=-\log\left[\prod_{i,j,t=0}^{N-1,M-1,L-1}p\left(y_{i,j,t}\,|\,g_{i,j},\lambda\right)\right] \tag{3}$$

where $p\left(y_{i,j,t}\,|\,g_{i,j},\lambda\right)=\dfrac{\left(g_{i,j}e^{-\lambda t}\right)^{y_{i,j,t}}}{y_{i,j,t}!}e^{-g_{i,j}e^{-\lambda t}}$ is the Poisson distribution, yielding

$$\mathbf{E}_{\mathbf{Y}}(\mathbf{G},\lambda,\mathbf{Y})=\sum_{i,j,t}\left[g_{i,j}e^{-\lambda t}-y_{i,j,t}\log\left(g_{i,j}e^{-\lambda t}\right)\right]+C \tag{4}$$

The prior term regularizes the solution and helps to remove the noise. By assuming **G** as a *Markov random field* (MRF), $p(\mathbf{G})$ can be written as a Gibbs distribution, $p(\mathbf{G})=\dfrac{1}{Z}e^{-\sum_{c\in C}V(g_c)}$, where Z is the partition function and $V(.)$ are the *clique potentials* (S.Geman & D.Geman, 1984). The sum of all *clique potentials*, the negative of the exponential argument function, is called the Gibbs energy, $\mathbf{E}_{\mathbf{G}}(\mathbf{G})$. In order to preserve the edges of the cell morphology *log-total variation* (log-TV) potentials are used in the regularization term. These potential functions have shown to be appropriated to deal with this type of optimization problems in R_{+}^{N} (V. Arsigny et al., 2006).

The regularization based on quadratic potentials is often used because they simplify the mathematical formulation of the estimation problem. However, they over-smooth the solution, leading to significant loss of morphological details. On the contrary, the log-TV prior is more efficient to attenuate small differences among neighbouring nodes due to the noise, but it penalizes less the large amplitude differences due to the transitions. Additionally, this prior is able to penalize differences between neighbouring pixels when their amplitude is very small. This does not happen with quadratic priors that are based on differences between pixels, g_i-g_{iv}, and not on amplitude ratios, g_i/g_{iv}, on which the log-TV prior is based.

The log-TV Gibbs energy function is defined as follows

$$E_{\mathbf{G}}(\mathbf{G}) = \alpha \sum_{i,j,t} \sqrt{\log^2\left(\frac{g_{i,j}}{g_{i-1,j}}\right) + \log^2\left(\frac{g_{i,j}}{g_{i,j-1}}\right)} \tag{5}$$

and therefore the overall energy function to be minimized is

$$\mathbf{E}(\mathbf{G}, \lambda, \mathbf{Y}) = \sum_{i,j,t} \left[g_{i,j} e^{-\lambda t} - y_{i,j,t} \log\left(g_{i,j} e^{-\lambda t}\right)\right] + \alpha L \sum_{i,j} \sqrt{\log^2\left(\frac{g_{i,j}}{g_{i-1,j}}\right) + \log^2\left(\frac{g_{i,j}}{g_{i,j-1}}\right)} \tag{6}$$

where α is a tuning parameter used to control the regularization strength that is kept constant in this step.

The minimization of the energy function (6) with respect to $g_{i,j}$ leads to a non-convex problem (Stephen Boyd & Lieven Vandenberghe, 2004) since it involves non-convex functions (e.g. $\sqrt{\log^2(x/a) + \log^2(x/b)}$). However, performing an appropriate change of variable, $s_{i,j} = \log(g_{i,j})$, it is possible to turn it into convex. Due to the monotonicity of the logarithmic function, the minimizers of both energy functions $\mathbf{E}(\mathbf{G}, \lambda, \mathbf{Y})$ and $\mathbf{E}(\mathbf{S}, \lambda, \mathbf{Y})$ are related by $\mathbf{S}^* = \log(\mathbf{G}^*)$.

The new objective function for the first step of this model is then

$$\mathbf{E}(\mathbf{S}, \lambda, \mathbf{Y}) = \sum_{i,j,t} \left[e^{s_{i,j} - \lambda t} - y_{i,j,t}\left(s_{i,j} - \lambda t\right)\right] + \alpha L \sum_{i,j} \sqrt{\left(s_{i,j} - s_{i-1,j}\right)^2 + \left(s_{i,j} - s_{i,j-1}\right)^2} \tag{7}$$

The minimization of this equation is accomplished by finding its stationary points, performing iteratively its optimization in S with respect to each component $s_{i,j}$, one at a time, considering all other components in each iteration as constants.

Let us explicitly represent the terms involving a given node $s_{i,j}$ in the energy function (7)

$$\begin{aligned} \mathbf{E}(\mathbf{S}, \lambda, \mathbf{Y}) = & \sum_{t} \left[e^{s_{i,j} - \lambda t} - y_{i,j,t} s_{i,j}\right] + \\ & \alpha L \left[\sqrt{\left(s_{i,j} - s_{i-1,j}\right)^2 + \left(s_{i,j} - s_{i,j-1}\right)^2} + \sqrt{\left(s_{i+1,j} - s_{i,j}\right)^2 + \left(s_{i+1,j} - s_{i+1,j-1}\right)^2} + \right. \\ & \left. \sqrt{\left(s_{i,j+1} - s_{i-1,j+1}\right)^2 + \left(s_{i,j+1} - s_{i,j}\right)^2} \right] + C \end{aligned} \tag{8}$$

where C is a term that does not depend on $s_{i,j}$. To cope with the difficulty introduced by the non-quadratic terms, a Reweighted Least Squares based method is used (B.Wohlberg & P.Rodriguez, 2007). The minimizer of the convex energy function (8), s^*, is also the minimizer of the following energy function with quadratic terms

$$\begin{aligned}\mathbf{E}(\mathbf{S},\lambda,\mathbf{Y}) = &\sum_t\left[e^{s_{i,j}-\lambda t} - y_{i,j,t}s_{i,j}\right] + \\ &\alpha L\left[w\left(s^*_{i,j}\right)\left[\left(s_{i,j}-s_{i-1,j}\right)^2+\left(s_{i,j}-s_{i,j-1}\right)^2\right]+\right. \\ &w\left(s^*_{i+1,j}\right)\left[\left(s_{i+1,j}-s_{i,j}\right)^2+\left(s_{i+1,j}-s_{i+1,j-1}\right)^2\right]+ \\ &\left.w\left(s^*_{i,j+1}\right)\left[\left(s_{i,j+1}-s_{i-1,j+1}\right)^2+\left(s_{i,j+1}-s_{i,j}\right)^2\right]\right]+C\end{aligned} \tag{9}$$

where

$$w\left(s^*_{i,j}\right)=\frac{1}{\sqrt{\left(s^*_{i,j}-s^*_{i-1,j}\right)^2+\left(s^*_{i,j}-s^*_{i,j-1}\right)^2}} \tag{10}$$

Since the weights $w\left(s^*_{i,j}\right)$, $w\left(s^*_{i+1,j}\right)$ and $w\left(s^*_{i,j+1}\right)$ depend on the unknown minimizer $s^*_{i,j}$, an iterative procedure is used, where in the k^{th} iteration, the estimated value $s^{(k-1)}_{i,j}$, computed in the previous iterations, is used instead of $s^*_{i,j}$. For sake of simplicity let us denote the weights $w\left(s^{(k-1)}_{i,j}\right)$, $w\left(s^{(k-1)}_{i+1,j}\right)$ and $w\left(s^{(k-1)}_{i,j+1}\right)$ just by w, w_c and w_d respectively. The minimization of (9) with respect to $s_{i,j}$ is performed by finding its stationary point,

$$\sum_t\left(e^{s_{i,j}-\lambda t} - y_{i,j,t}\right)+h_{i,j}=0 \tag{11}$$

where

$$h_{i,j}=2\alpha L\left[(2w+w_c+w_d)s_{i,j}-w\left(s_{i-1,j}+s_{i,j-1}\right)-w_c s_{i+1,j}-w_d s_{i,j+1}\right] \tag{12}$$

The minimizer of (7) with respect to λ can be obtained through the computation of its stationary point, which is accomplished by solving the equation

$$\sum_{i,j,t}\left[-te^{s_{i,j}-\lambda t}+ty_{i,j,t}\right]=0 \tag{13}$$

Using the Newton's method the solutions of (11) and (13) can be iteratively obtained by

$$s^{(k+1)}_{i,j}=s^{(k)}_{i,j}-\frac{\sum_t\left(e^{s^{(k)}_{i,j}-\lambda t}-y_{i,j,t}\right)+h_{i,j}}{\sum_t\left(e^{s^{(k)}_{i,j}-\lambda t}\right)+2\alpha L(2w+w_c+w_d)} \tag{14}$$

$$\lambda^{(k+1)} = \lambda^{(k)} - \frac{\sum_{i,j,t}\left(-te^{s_{i,j}^{(k)}-\hat{\lambda}t} + ty_{i,j,t}\right)}{\sum_{i,j,t}\left(t^2 e^{s_{i,j}^{(k)}-\hat{\lambda}t}\right)} \tag{15}$$

The stopping criterion is based on the norm of the error of λ between consecutive iterations and on the number of iterations. The norm of the error of $s_{i,j}$ was also computed but only for control purposes, since it acts as an auxiliary variable to estimate λ.

The estimated parameter $\hat{\lambda}$ is used in the next step as a constant, under the assumption that the intensity decay due the *photobleaching* effect was totally caught in this step.

2.2 Step two

The ultimate goal of the second step of the proposed algorithm is to estimate the time and space varying cell nucleus morphology, denoted by $\mathbf{F} = \{f_{i,j,t}\}$, where the intensity decay rate due the *photobleaching* is characterized by the parameter λ estimated in the previous step. Each point of the noiseless image sequence, $\mathbf{X} = \{x_{i,j,t}\}$ to be estimated is defined in this step as

$$x_{i,j,t} = f_{i,j,t}\ e^{-\hat{\lambda}t} \tag{16}$$

The estimation of the parameters $f_{i,j,t}$, performed in a Bayesian framework by using the maximum *a posteriori* (MAP) criterion, may be formulated as the following optimization task

$$\hat{\mathbf{F}} = \arg\min_{\mathbf{F}}\ \mathbf{E}\left(\mathbf{F}, \hat{\lambda}, \mathbf{Y}\right) \tag{17}$$

where the energy function $\mathbf{E}(\mathbf{F}, \hat{\lambda}, \mathbf{Y}) = \mathbf{E}_{\mathbf{Y}}(\mathbf{F}, \hat{\lambda}, \mathbf{Y}) + \mathbf{E}_{\mathbf{F}}(\mathbf{F})$, as before, is the sum of two terms, $\mathbf{E}_{\mathbf{Y}}(\mathbf{F}, \hat{\lambda}, \mathbf{Y})$, the data fidelity term and $\mathbf{E}_{\mathbf{F}}(\mathbf{F})$, the energy associated to the *a priori* distribution for $\mathbf{F}$.

To preserve the edges of the cell morphology, log-TV and L_1 (L_1 norm) potential functions are used in space and in time respectively. The regularization is performed simultaneously in the image space and in time using different prior parameters which means that this denoising iterative algorithm involves an anisotropic 3-D filtering process that is able to accomplish different smoothing effects in the space and in the time dimensions.

The energy function related to the a priori distribution of **F** is given by

$$E_{\mathbf{F}}(\mathbf{F}) = \alpha \sum_{i,j,t} \sqrt{\log^2\left(\frac{f_{i,j,t}}{f_{i-1,j,t}}\right) + \log^2\left(\frac{f_{i,j,t}}{f_{i,j-1,t}}\right)} + \beta \sum_{i,j,t} \left|\log\left(\frac{f_{i,j,t}}{f_{i,j,t-1}}\right)\right| \tag{18}$$

Therefore the overall problem consists on minimizing the following function

$$\begin{aligned}\mathbf{E}(\mathbf{F},\hat{\lambda},\mathbf{Y}) = &\sum_{i,j,t}\left[f_{i,j,t}e^{-\hat{\lambda}t} - y_{i,j,t}\log\left(f_{i,j,t}e^{-\hat{\lambda}t}\right)\right]\\ &+\alpha\sum_{i,j,t}\sqrt{\left(\log\left(f_{i,j,t}\right)-\log\left(f_{i-1,j,t}\right)\right)^2+\left(\log\left(f_{i,j,t}\right)-\log\left(f_{i,j-1,t}\right)\right)^2}\\ &+\beta\sum_{i,j,t}\left|\log\left(f_{i,j,t}\right)-\log\left(f_{i,j,t-1}\right)\right|\end{aligned} \tag{19}$$

where α and β are tuning parameters to control the strength of the regularization in space and in time respectively. The parameter α is adaptive and β is constant. The standard deviation of the logarithm of the morphology, computed for each image, seems to perform an important role in adapting the strength of the regularization in the space domain. Thus, for both synthetic and real data, $\alpha = \alpha_0 \times \mathrm{std}(\log(f_{i,j,t}))$, where α_0 is a constant, is used.

As before, in the previous step, the energy function (19) with respect to $f_{i,j,t}$ is non convex. Once again, to make it convex, the following change of variable is performed: $z_{i,j,t} = \log(f_{i,j,t})$. Due to the monotonicity of this function, the minimizer of $\mathbf{E}(\mathbf{F},\hat{\lambda},\mathbf{Y})$ is related to the one of $\mathbf{E}(\mathbf{Z},\hat{\lambda},\mathbf{Y})$ by $\mathbf{Z}^* = \log(\mathbf{F}^*)$, where the log function of tensor $\mathbf{F}$ is taken component-wise.

The objective function to be minimized with respect to the unknowns $z_{i,j,t}$ in this second step is

$$\begin{aligned}\mathbf{E}(\mathbf{Z},\hat{\lambda},\mathbf{Y}) = &\sum_{i,j,t}\left[e^{z_{i,j,t}-\hat{\lambda}t} - y_{i,j,t}\left(z_{i,j,t}-\hat{\lambda}t\right)\right]+\\ &\alpha L\sum_{i,j,t}\sqrt{\left(z_{i,j,t}-z_{i-1,j,t}\right)^2+\left(z_{i,j,t}-z_{i,j-1,t}\right)^2}+\\ &\beta L\sum_{i,j,t}\left|z_{i,j,t}-z_{i,j,t-1}\right|\end{aligned} \tag{20}$$

The estimation of $\mathbf{Z}$ is performed by using the ICM (Iterated Conditional Modes) method (J.Besag, 1986) where (20) is minimized with respect to each unknown $z_{i,j,t}$ at a time, keeping all other unknowns constant.

As before, let us consider explicitly the terms involving a given node $z_{i,j,t}$ in the energy equation

$$\begin{aligned}E(\tilde{Z},\lambda,Y) = &\sum_t\left[e^{z_{i,j,t}-\hat{\lambda}t}+y_{i,j,t}z_{i,j,t}\right]+\alpha L\sqrt{\left(z_{i,j,t}-z_{i-1,j,t}\right)^2+\left(z_{i,j,t}-z_{i,j-1,t}\right)^2}+\\ &\sqrt{\left(z_{i+1,j,t}-z_{i,j,t}\right)^2+\left(z_{i+1,j,t}-z_{i+1,j-1,t}\right)^2}+\\ &\sqrt{\left(z_{i,j+1,t}-z_{i-1,j+1,t}\right)^2+\left(z_{i,j+1,t}-z_{i,j,t}\right)^2}+\\ &\beta\left[\left|z_{i,j,t}-z_{i,j,t-1}\right|+\left|z_{i,j,t+1}-z_{i,j,t}\right|\right]+C\end{aligned} \tag{21}$$

where C is a term that does not depend on $z_{i,j,t}$. The optimization of (21) is performed by using the Reweighted Least Squares method, as before in the first step, to cope with the non quadratic prior terms. The minimizer of the convex energy function (21), $\mathbf{Z}^*$,is also the minimizer of the following energy function with quadratic terms

$$\begin{aligned}\mathbf{E}(\mathbf{Z},\lambda,\mathbf{Y}) = \sum_t \Big[e^{z_{i,j,t}-\hat{\lambda}t} - y_{i,j,t}z_{i,j,t} \Big] + \\ \alpha\Big[w\left(z^*_{i,j,t}\right)\Big[\left(z_{i,j,t}-z_{i-1,j,t}\right)^2+\left(z_{i,j,t}-z_{i,j-1,t}\right)^2\Big] + \\ w\left(z^*_{i+1,j,t}\right)\Big[\left(z_{i+1,j,t}-z_{i,j,t}\right)^2+\left(z_{i+1,j,t}-z_{i+1,j-1,t}\right)^2\Big] + \\ w\left(z^*_{i,j+1,t}\right)\Big[\left(z_{i,j+1,t}-z_{i-1,j+1,t}\right)^2+\left(z_{i,j+1,t}-z_{i,j,t}\right)^2\Big]\Big] + \\ \beta\Big[v(z_{i,j+1,t})\left(z_{i,j,t}-z_{i,j,t-1}\right)^2 + v(z_{i,j,t+1})\left(z_{i,j,t+1}-z_{i,j,t}\right)^2\Big] + C\end{aligned} \tag{22}$$

where

$$w\left(z^*_{i,j,t}\right) = \frac{1}{\sqrt{\left(z^*_{i,j,t}-z^*_{i-1,j,t}\right)^2+\left(z^*_{i,j,t}-z^*_{i,j-1,t}\right)^2}}\ , \tag{23}$$

and

$$v\left(z^*_{i,j,t}\right) = \frac{1}{\left|z^*_{i,j,t}-z^*_{i,j,t-1}\right|} \tag{24}$$

Since the weights $w\left(z^*_{i,j,t}\right)$, $w\left(z^*_{i+1,j,t}\right)$, $w\left(z^*_{i,j+1,t}\right)$, $v\left(z^*_{i,j,t}\right)$ and $v\left(z^*_{i,j,t+1}\right)$ depend on the unknown minimizer $\mathbf{Z}^*$, the same iterative procedure used in the first step is adopted here, where the estimation of $\mathbf{Z}^*$ at the previous $(k-1)^{th}$ iteration, $\mathbf{Z}^{(k-1)}$, is used. Let us denote these weights by w, w_c, w_d, v_a and v_c respectively.

The minimization of (22) with respect to $z_{i,j,t}$ is obtained by finding its stationary point,

$$\frac{\partial E(\mathbf{Z},\lambda,\mathbf{Y})}{\partial z_{i,j}} = e^{z_{i,j}-\hat{\lambda}t} - y_{i,j,t} + h_{i,j,t} = 0 \tag{25}$$

where

$$\begin{aligned}h_{i,j,t} = 2\left(2\alpha w + \alpha w_c + \alpha w_d + \beta v_a + \beta v_c\right)z_{i,j,t} - 2\alpha w\left(z_{i-1,j,t}+z_{i,j-1,t}\right) - 2\alpha w_c z_{i+1,j,t} - \\ 2\alpha w_d z_{i,j+1,t} - 2\beta\left(v_a z_{i,j,t-1} + v_c z_{i,j,t+1}\right)\end{aligned} \tag{26}$$

Using the Newton's method the solution of (25) can be obtained iteratively by

$$z^{(k+1)}_{i,j,t} = z^{(k)}_{i,j,t} - \frac{e^{z_{i,j,t}-\hat{\lambda}t} - y_{i,j,t} + h_{i,j,t}}{e^{z_{i,j,t}-\hat{\lambda}t} + 2\alpha\left(2w + w_c + w_d\right) + 2\beta\left(v_a + v_c\right)} \tag{27}$$

3. Experimental Results

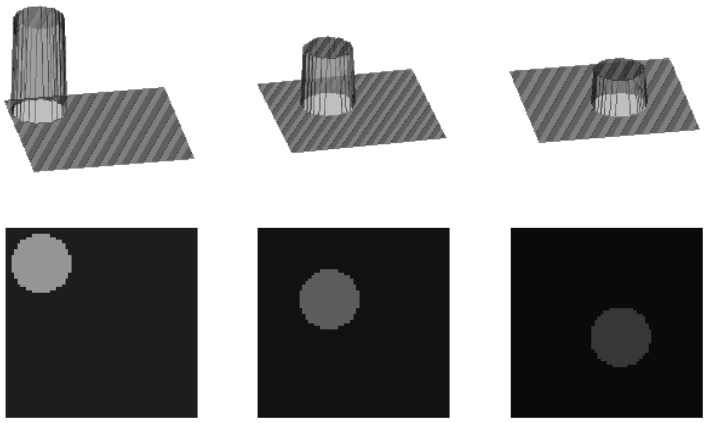
In this section experimental results with synthetic and real data are presented.

3.1 Synthetic data

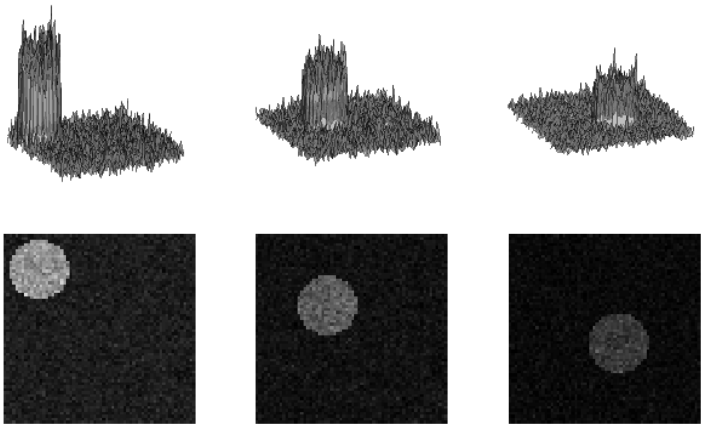
The synthetic data is formed by a set of 64 images of a white circle moving from the left top corner of the first image down to the right, along the diagonals of 64 squares, 64×64 pixels, with constant velocity, over a grey background. This set of images was conceived having in mind the real situation where the morphology of a cell nucleus changes along the image with time. For the first image of the sequence (t=0), the intensities of the circle and of the background were set equal to 50 and 10 respectively. The final sequence of synthetic data was then generated by applying an exponential decay in time (t=0,…,63) with rate equal to λ=0.025 image^{-1} , followed by corruption with Poisson noise. This rate of decay can be considered realistic under the hypothesis of an acquisition rate of 10s, which means λ=0.0025s^{-1} .

Fig. 1 shows images for three time instants of the synthetic sequence. The images on the first double row (a) belong to the original sequence, before being corrupted by Poisson noise. The same images corrupted with Poisson noise are shown in (b). The third double row (c), with the results of the reconstruction according to eq. 16 of the second step of the algorithm, show the ability of this methodology for removing noise, although providing good preservation of the edges of the moving circle.

(a)



(b)



(c)

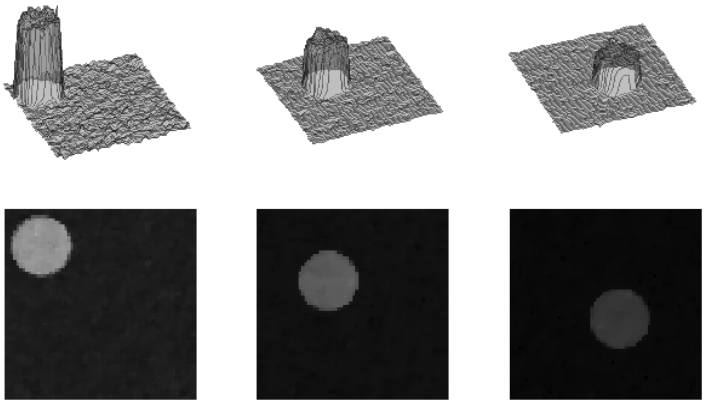


Fig. 1. (a),(b),(c) Three time instants (1, 20, 40) of the true, noisy and reconstructed synthetic sequences and respective mesh representations.

The mesh representations of the estimated morphology for three different time instants of the sequence show the ability of the algorithm to recover the true morphology whose shape is a constant height cylinder and whose behaviour in time is to slide down along the diagonal of a 64×64 pixels square. Both the position and the height of the cylinder are correctly estimated for the complete sequence.

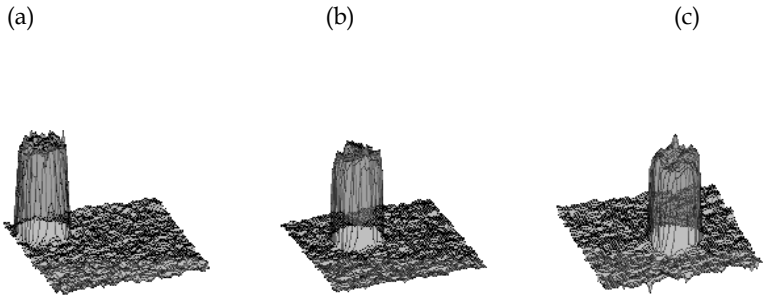


Fig. 2. Mesh representation of the estimated morphology $\hat{f}_{i,j,t}$ for images 1 (a), 20 (b) and 40 (c).

An example of the denoising and recovery capabilities of the presented methodology is shown in Fig. 3. where the true (a), the noisy (b) and the estimated (c) morphologies corresponding to image 48 of the sequence are displayed. Subplot (d) shows profiles taken along the diagonal of (a), (b) and (c). In spite of the degradation of the image, it is noticeable how well the algorithm manages to recover the morphology.

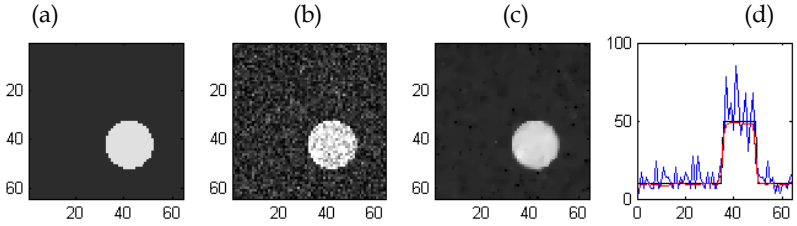


Fig. 3. (d) Morphology profiles, image 48, (a) True (black line in (d)), (b) Noisy (blue line in (d)), (c) Estimated (red line in (d)).

The root mean square error (RMSE) of the estimated morphology of the complete sequence was computed for each of the 300 iteration of the second step and its plot can be seen in Fig. 4. During the first 20 iterations the RMSE drops very quickly and keeps that trend till the end of the iterative procedure, which suggests effective convergence of the algorithm.

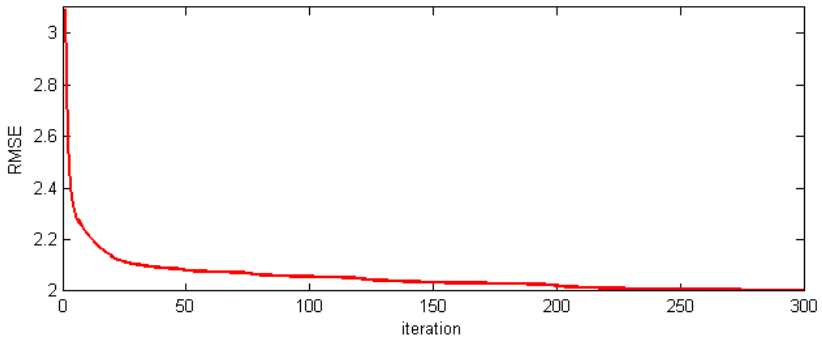


Fig. 4. Root mean square error (RMSE) of the estimated morphology of the complete sequence, in every iteration.

In order to evaluate the quality of the presented algorithm, the *signal to noise ratio* (SNR), the *mean square error* (MSE) and the *Csiszáér I-divergence* (I-div) were adopted. The literature is not very conclusive on what concerns to the choice of the *figure of merit* more suitable to evaluate the quality of an algorithm that deals with Poisson multiplicative noise.
Some authors use the SNR although there is strong evidence that it gives a more efficient quality evaluation in the Gaussian denoising situations than in the Poissonian ones.
As in section 2, let $X = \{x_{i,j,t}\}$ and $\hat{X} = \{\hat{x}_{i,j,t}\}$ with $0 \le i, j,t \le N-1, M-1, L-1$, be respectively the noiseless and the estimated sequences of images. The SNR of image t of the estimated sequence can be defined as:

$$SNR(t) = 10\log_{10}\left(\frac{\sum_{i,j}\left(x_{i,j,t}\right)^2}{\sum_{i,j}\left(x_{i,j,t}-\hat{x}_{i,j,t}\right)^2}\right) \quad 28$$

The MSE is extensively used with the purpose of evaluating the quality of the denoising algorithm, independently of the noise statistics and is defined as:

$$MSE(t) = \frac{1}{MN}\sum_{i,j}\left(x_{i,j,t}-\hat{x}_{i,j,t}\right)^2 \quad 29$$

According to (N. Dey et al., 2004), to quantify the quality of the denoising procedure in the presence of non-negativity constraints, which is the case of the Poisson denoising, the Csiszáér I-divergence (Csiszáér, 1991) is the best choice.
The I-Divergence between the t^{th} image of the original (noiseless) sequence X and the t^{th} image of the restored sequence $\hat{X}$ is given by:

$$I\_div_{X,\hat{X}}(t) = \sum_{i,j}\left[x_{i,j,t}\log\left(\frac{x_{i,j,t}}{\hat{x}_{i,j,t}}\right) - \left(x_{i,j,t}-\hat{x}_{i,j,t}\right)\right] \quad 30$$

The I-Divergence can be interpreted as a quantifier of the difference between the true image and the estimated one. Ideally, a perfect denoising should end with an I-div equal to zero.

A Monte Carlo experiment with 500 runs, based on sequences similar to the described above, was carried out. For each run, the rate of decay λ was estimated in the first step and used to estimate the morphology $f_{i,j,t}$ in the second step. The final reconstruction is obtained by $\hat{x}_{i,j,t} = \hat{f}_{i,j,t} e^{-\hat{\lambda} t}$.

The SNR, the MSE and the I-div were computed for every image in each of the 500 runs and the means and standard deviations of the estimated lambda, $\hat{\lambda}^{(run)}$, of the SNR of the reconstruction, $SNR_{\hat{X}}^{(run)}$, of the MSE of the morphology, $MSE_{\hat{F}}^{(run)}$, of the MSE of the reconstruction, $MSE_{\hat{X}}^{(run)}$, of the I-div of the morphology, $I-div_{\hat{F}}^{(run)}$ and of the I-div of the reconstruction $I-div_{\hat{X}}^{(run)}$, were computed.

The mean of the estimated rate of decay is 0.025061, with a mean square error (MSE) of 0.00011126, which is very close to the original one (0.025) .

Fig. 5. (a) shows the mean of the SNR for each image of the noisy sequences used in the Monte Carlo experiment (black line) and the mean of the SNR of the respective reconstruction. As can be noticed, the SNR improvement is $\approx 10dB$ and is almost constant throughout the sequence of images.

The mean of the MSE of the reconstruction is plotted in Fig. 5. (b) strengthen the evidence of the ability of the presented algorithm to restore this type of sequences.

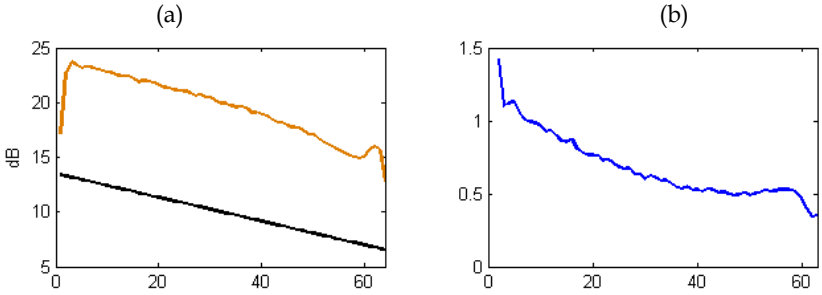


Fig. 5 (a) Mean of the SNR over the 500 runs computed from the noisy sequence (black line) and from the reconstructed sequence $\hat{X}$ (red line), (b) Mean of the MSE of the reconstructed sequence.

In the present situation the mean of the I-div of the reconstructed images (Fig. 6.) is not zero as it would be in an ideal case, but it is well bellow the one obtained with the noisy sequences.

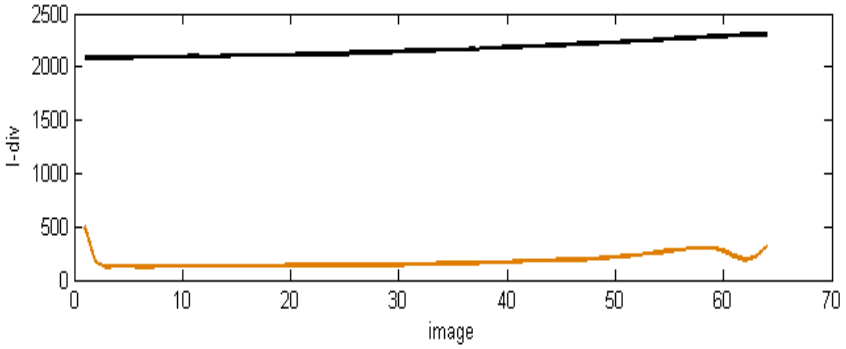


Fig. 6. I-div mean over 500 runs from the noisy sequence y (black line) from the reconstructed sequence $\hat{X}$ (red line).

3.2 Real data

Three sets of real CLSFM images of cell nucleus, identified as 2G100, 7GREEN_FRAP and BDM_FLIP, were analyzed.

The sequence 2G100 consists of 100 CLSFM images of a Hela cell nucleus, acquired at a rate of 23s, in normal laboratory conditions, using a continuous, low intensity laser illumination. During the acquisition of the 2G100 sequence, no additional techniques such as FRAP (Fluorescence Recovery After Photobleaching) or FLIP (Fluorescence Loss In Photobleaching) were employed. The aim is the observation of a cell nucleus where certain particles are tagged with fluorescent proteins, for quite a long time, in order to acquire data where the *photobleaching* effect occurs without the interference of important diffusion and transport phenomena.

Three images, 1, 20, 45 of this sequence, corresponding to the time instants 0s, 460s and 1035s after the beginning of the acquisition process, are displayed in Fig. 7. (a), (b) and (c). The appearance of these images is noisy, with an SNR decreasing very quickly with the time.

Using the previously described methodology, the rate of decay due to the *photobleaching*, λ, and the cell nucleus morphology, $F_{i,j,t} = \{f_{i,j,t}\}$, were estimated. The achieved value for the rate of decay was $\hat{\lambda} = 3.9988 \times 10^{-4}\,s^{-1}$.

Fig. 8. (a), (b) and (c) show images of the reconstructed sequence for the same time instants as in Fig. 7, where a considerable reduction of noise can be observed while their morphological details are preserved.

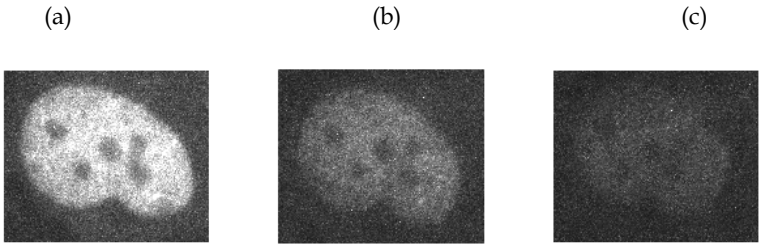


Fig. 7. Noisy images 1 (a), 20 (b), 45 (c) from the real data set 2G100.

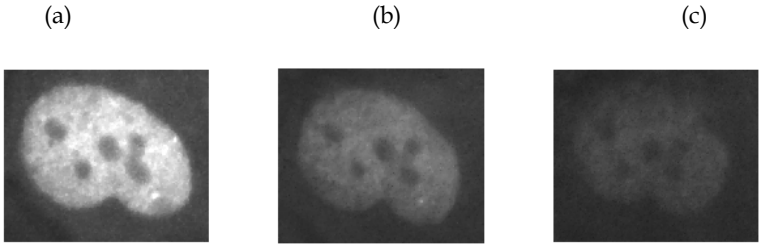


Fig. 8. Images 1 (a), 20 (b), 45 (c) from reconstructed sequence (2G100).

Three images of the estimated morphology can be seen in Fig. 9. (a), (b) and (c). It is noticeable the substantial improvement in the quality of the details of the cell nucleus structure. In particular, the comparison between the images displayed in Fig.7.c) and Fig.9.c) reveals the ability of the algorithm to recover information from original images where almost no information is available.

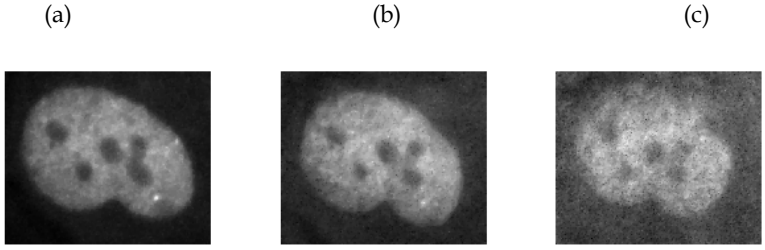


Fig. 9. Estimated morphology corresponding to images 1 (a), 20 (b), 45 (c) of the data set 2G100.

The plot of the horizontal and vertical profiles along the dashed lines in image (Fig. 10. (a), (b) and (c)) reinforces this idea. The intensity valleys become more perceptible and identifiable. The noise undergoes an effective reduction and the edges are well defined and preserved.

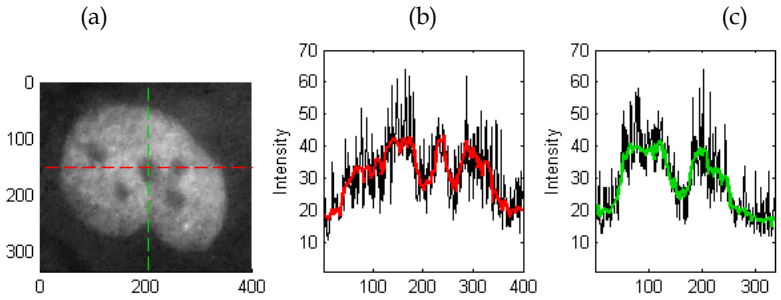


Fig. 10. (a) Reconstructed image 20 of the sequence 2G100. The red dash line corresponds to the horizontal profile in (b). The green dashed line stands for the vertical profile in (c).

The means and standard deviations (std) of the images of the noisy and reconstructed sequences were computed and the results are presented in the plot of Fig. 11. (a) and (b). As can be observed in (a), the mean of the images is well preserved in the reconstruction; in addition, the pattern of the std becomes smoother.

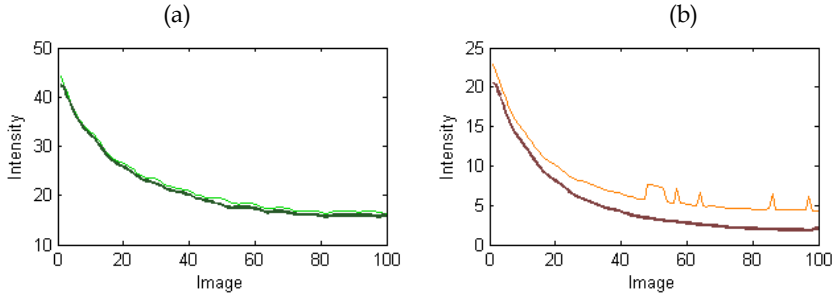


Fig. 11. (a) Mean of each image of the sequence 2G100. Dark and light green lines correspond to the reconstructed and to the noisy sequences respectively. (b) Standard deviation of each image of the sequence 2G100. Dark and light red lines correspond to the reconstructed and to the noisy sequences respectively.

In order to test the behaviour of the proposed algorithm in more complex situations where diffusion and transport phenomena are present, the sequences 7GREEN_FRAP and BDM_FLIP are used.

Sequence 7GREEN_FRAP consists of 108 images acquired at a rate of 9.4s *per* image, using the FRAP (Fluorescence Recovery After Photobleaching) technique (R. Underwood, 2007). In this technique a pre-defined region in the cell is illuminated with a high intensity focused laser beam, during a small period of time, to force the irreversible *photobleaching* of the fluorescence tagged molecules present within that region. The movements of the bleached molecules out of the bleached region and of the surrounding unbleached molecules into the bleached area lead to a recovery of fluorescence into the bleached area. The acquisition

process consists on monitoring this recovery over time, at low laser power, to prevent further bleaching. In this situation the fluorescence recovery superimposes the global *photobleaching* due to the low intensity illumination of the cell nucleus, which explains the slight increase of the intensity with the time.
As expected, the estimated rate of decay is negligible. Eventual small intensity fluctuations that may be observed in the images are due to the manual fine tuning required to keep the target points in focus during the experiment.

In Fig. 12. (a), (b) and (c) and Fig. 13. (a), (b) and (c), images corresponding to the time instants t=9.4s, 460.6s and 1005.8s of the raw data and of the reconstruction are shown. Also the corresponding estimated morphology is presented in Fig. 14. (a), (b) and (c). These images denote a substantial improvement in the details of the nucleus structure.

(a) (b) (c)

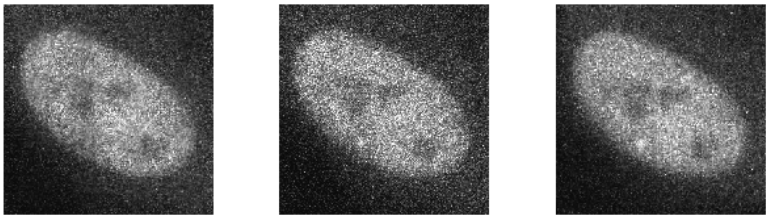


Fig. 12. (a) Noisy images 2 (a), 50 (b), 108 (c), real data set 7GREEN_FRAP.

(a) (b) (c)

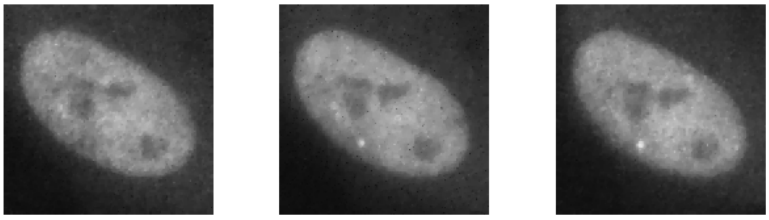


Fig. 13. Reconstructed images 2 (a), 50 (b), 108 (c), sequence 7GREEN_FRAP.

(a) (b) (c)

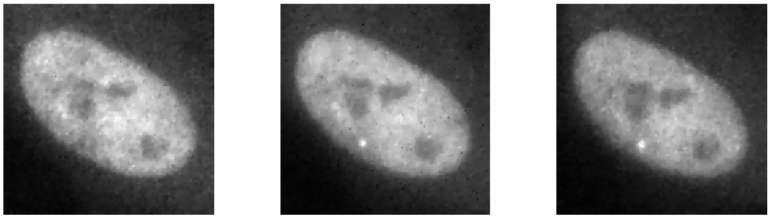


Fig. 14. (a) Morphology 2 (a), 50 (b), 108 (c), sequence 7GREEN_FRAP.

The profiles in Fig. 15. (a), (b) and (c) confirm this assertion. In this figure two images and respective profiles are displayed to illustrate the increase of the intensity with time.

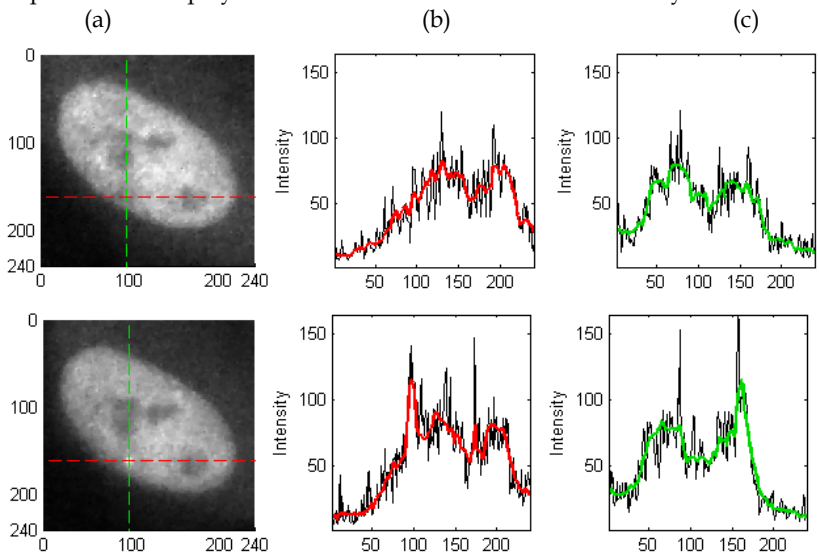


Fig. 15. (a) Reconstructed image 2 (first row) and 108 (second row) of the sequence 7GREEN_FRAP. The red dash line corresponds to the horizontal profile in (b). The green dashed line stands for the vertical profile in (c).

Plots of the intensity means of the noisy and of the reconstructed sequences show the constancy characteristic of an almost zero rate of decay. In addition, the reconstruction mean (Fig.16. (a)), is well preserved while the std (Fig. 16. (b)) is smoother than the corresponding to the noisy sequence.

From all these assertions it is possible to assume that the presented methodology is appropriate for denoising and for morphology estimation when the FRAP technique is used.

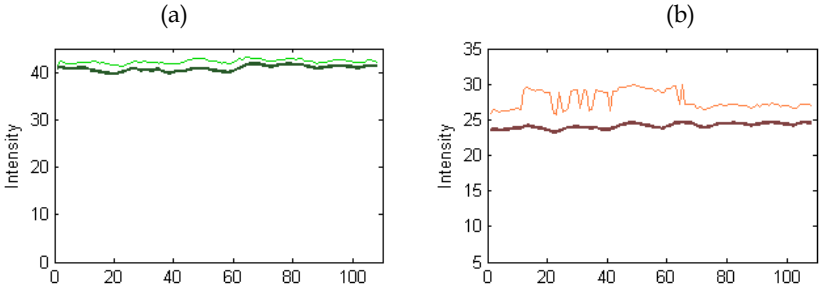


Fig. 16. (a) Mean of each image of the sequence 7GREEN_FRAP. Dark and light green lines correspond to the reconstructed and to the noisy sequences respectively. (b) Standard deviation of each image of the sequence 7GREEN_FRAP. Dark and light red lines correspond to the reconstructed and to the noisy sequences respectively.

The BDM_FLIP sequence consists of 175 images acquired at a rate of 7.4s and is the result of applying the FLIP (Fluorescence Loss In Photobleaching) technique (R.Underwood, 2007)to the Hela cell nucleus. This technique makes use of the *photobleaching* effect as a perturbing agent of the distribution of fluorescent molecules in the cell nucleus. In a FLIP experiment, during a certain time interval, a small defined region within the nucleus, expressing fluorescently tagged proteins, is illuminated with repetitive bleach pulses of a high intensity focused laser beam, in order to force the occurrence of the *photobleaching* effect. The surrounding area is then monitored for a decrease in the level of fluorescence. Any fraction of the cell nucleus connected to the area being bleached will gradually fade owing to the movement of bleached molecules into the bleached region. The resulting information from the experiment can then be used to determine the kinetic properties, including diffusion coefficients, mobile fraction and transport rate of the fluorescently labeled molecules.
Three images from the sequence (t=29.6s, t=362.6s and t=843.6s) and respective reconstruction are displayed in Figs. 17. (a), (b) and (c) and Figs. 18. (a), (b) and (c). It is easy to perceive that the intensity decrease is quite fast which is very inconvenient when the acquisition times are long.

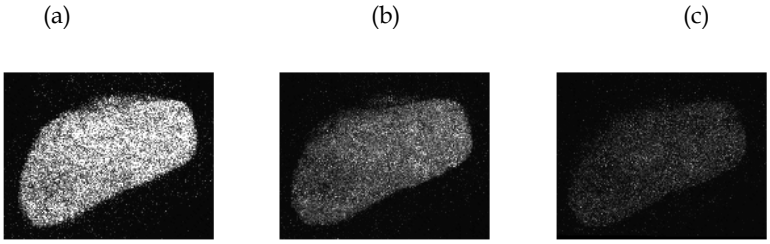


Fig. 17. (a) Noisy images 5 (a), 50 (b), 115 (c), real data set BDM_FLIP.

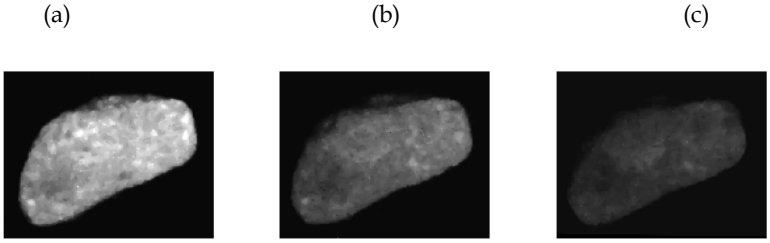


Fig. 18. (a) Reconstructed images 5 (a), 50 (b), 115 (c), sequence BDM_FLIP.

The estimated value of the rate of decay of the intensity is $\hat{\lambda} = 1.5 \times 10^{-3} \, s^{-1}$. This value is obviously larger than the one obtained for the sequence 2G100, due to the use of the FLIP technique that reinforces the decrease of the intensity.
The estimated morphology is shown in Fig. 19. (a), (b) and (c) and the improvement in the details of the nucleus structure are noticeable.

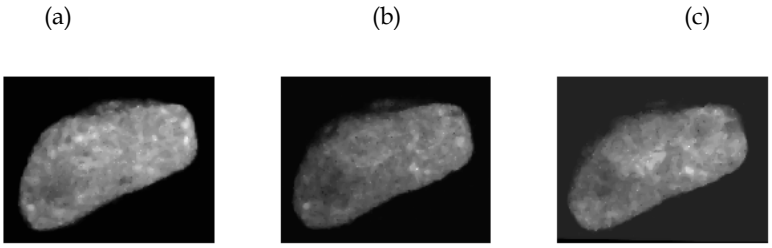


Fig. 19. (a) Morphology images 5 (a), 50 (b), 115 (c), BDM_FLIP.

The profiles in Fig. 20. (a), (b) and (c) were chosen to show the bleached region of the nucleus. This region can be identified in the profiles as a depression of the intensity.

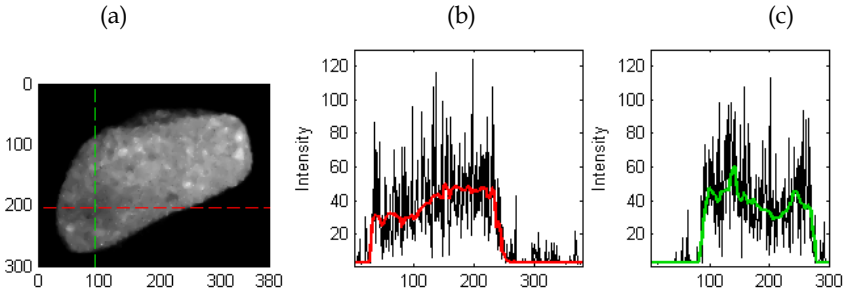


Fig. 20. (a) Reconstructed image 10 of the sequence BDM_FLIP. The red dash line corresponds to the horizontal profile in (b). The green dashed line stands for the vertical profile in (c).

The mean of the reconstruction (Fig. 21. (a)) is preserved and the std (Fig. 21. (b)) is smoothed.

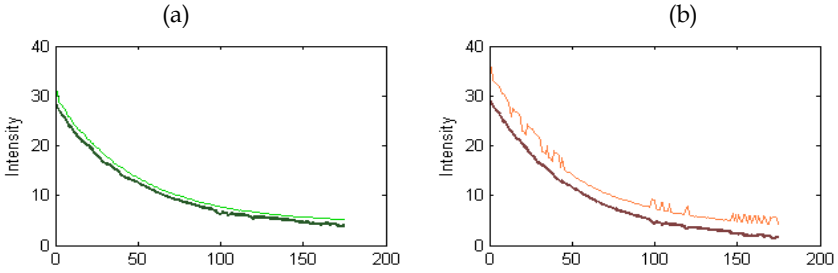


Fig. 21. (a) Mean of each image of the sequence BDM_FLIP. Dark and light green lines correspond to the reconstructed and to the noisy sequences respectively. (b) Standard deviation of each image of the sequence BDM_FLIP. Dark and light red lines correspond to the reconstructed and to the noisy sequences respectively.

4. Conclusion

In this chapter a new denoising algorithm for fluorescence microscopy (FM) imaging, to deal with the Poissonian multiplicative noise corrupting the images, is proposed. Furthermore, image sequences acquired along the time present an intensity decreasing due to the permanent fluorophore loss of its ability to fluoresce. This effect is caused by photochemical activity induced by the incident light that favours reactions of the fluorescent protein with other surrounding molecules. This intensity decrease prevents long time acquisition experiments because of the lack of morphological details in the last images of the sequence, making the biological information recovery a difficult task.

The proposed algorithm is designed in the Bayesian framework, as an optimization task, with the maximum *a posteriori* (MAP) criterion. The observation model takes into account the *photobleaching* effect in order to recover as much as possible all the information present in the sequence, even from the last images exhibiting a very low *signal to noise ratio* (SNR). This is possible because the algorithm establishes temporal correlation between consecutive images within the sequence.

The algorithm is developed iteratively in two steps. The intensity decay rate due to the *photobleaching* is estimated in the first step assuming a time invariant cell morphology. The goal in this step is to capture exclusively the intensity time variation caused by the *photobleaching* in the exponential term. In the second step the time and space varying cell morphology is estimated after compensating the *photobleaching* effect with the decay rate estimated in the first step.

The energy functions used in the optimization problem are designed to be convex and their minimizers are computed by using the Newton's algorithm and a reweighted least squares based method. This approach guarantees a continuous convergence towards the global minimum. Furthermore, the prior distributions needed to regularize the solutions use edge preserving potential functions. The associated Gibbs energy function is based on log-Euclidean total variation functions in time and in space, appropriated to this positively constrained optimization problem.

Monte Carlo tests with synthetic data were used to assess the performance of the algorithm. These tests have shown the ability of the algorithm to strongly reduce the Poisson multiplicative noise, to estimate both the decay rate due to the *photobleaching* and the underlying morphology and, simultaneously, to preserve the transitions.

Tests with real data from *laser scanning fluorescence confocal microscopy* were also performed where it is shown the effectiveness of the algorithm to cope with this type of noise and low SNR, even when the FRAP and FLIP techniques are used. Images and profiles extracted from the original and processed images are displayed for visualization and comparison purposes.

5. Acknowledgment

The authors thank Dr. José Rino and Profª Maria do Carmo Fonseca, from the Instituto de Medicina Molecular, Faculdade de Medicina da Universidade de Lisboa, for providing biological technical support and the real data used in this paper.
This work was supported by Fundação para a Ciência e a Tecnologia (ISR/IST plurianual funding) through the POS Conhecimento Program which includes FEDER funds.

6. References

Arsigny, V.; Fillard, P.; Pennec X. & Ayache N. (2006). Log-Euclidean metrics for fast and simple calculus on diffusion tensors, *Magnetic Resonance in Medicine*, vol. 56, no. 2, pp. 411–421, August 2006.

Bardsley, J. & Luttman, A. (2006). Total variation-penalized Poisson likelihood estimation for ill-posed problems, Depart. Math Sci., Univ. Montana Missoula, Tech. Rep. 8, 2006.

Besag, J. (1986). On the statistical analysis of dirty pictures, *J. R. Statist. Soc. B*, vol. 48, no. 3, pp. 259–302, 1986.

Besbeas, P.; Italia De Feis & Sapatinas T. (2004). A comparative simulation study of wavelet shrinkage estimators for Poisson counts., *Int. Stat. Rev.*, vol. 72, no. 2, pp. 209–237, 2004.

Boulanger, J.; Sibarita, J.-B.; Kervrann, C.; Bouthemy P. (2008). Non-Parametric regression for patch-based fluorescence microscopy image sequence denoising, *Proc. IEEE Int. Symp. on Biomedical Imaging, ISBI'08*, , pp. 748-751, Paris, France, May, 2008.

Boyd, S. & Vandenberghe, L. (2004). *Convex Optimization*. Cambridge University Press.

Braga, J.; Desterro, J.M. & Carmo-Fonseca; M. (2004). Intracellular macromolecular mobility measured by fluorescence recovery after photobleaching with confocal laser scanning microscopes. *Mol Biol Cell.*, 15:4749-60.

Csiszáér, I. (1991)., Why least squares and maximum entropy? an axiomatic approach to inference for linear inverse problems, *The Annals of Statistics*, vol. 19, no. 4, pp. 2032–2066, 1991.

Dey, N., Blanc- Féraud, L.; Zimmer, C.; Kam, Z.; Olivo-Marin, J.-C. & Zerubia J. (2004). A deconvolution method for confocal microscopy with total variation regularization, *IEEE International Symposium on Biomedical Imaging: Nano to Macro, 2004*, vol. 2, pp. 1223–1226, April 2004.

Dey, N.; Blanc- Féraud, L.; Zimmer, C.; Roux, P.; Kam, Z.; Olivo-Marin, J.-C. & Zerubia, J. (2006). "Richardson–Lucy Algorithm With Total Variation Regularization for 3D Confocal Microscope Deconvolution", Microcopy Research and Technique, 69:260-266, 2006.

Dupé, F.X.; Fadili, M.J.; Starck, J-L. (2008). Deconvolution of confocal microscopy images using proximal iteration and sparse representations, *ISBI 2008*, Paris, France, 2008.

Fisz, M. (1955). The limiting distribution of a function of two independent random variables and its statistical application, *Colloquium Mathematicum*,, vol. 3, pp. 138 – 146, 1955.

Fryźlewicz, P. & Nason, G. (2001)., Poisson intensity estimation using wavelets and the Fisz transformation, Department of Mathematics, University of Bristol, United Kingdom, Tech. Rep. 01/10, 2001.

Geman S. & Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI, no. 6, pp. 721–741, Nov. 1984.

Hebert, T. & Leahy, R. (1989). A generalized EM algorithm for 3-D Bayesian reconstruction from Poisson data using Gibbs priors, *IEEE Transactions on Medical Imaging*, vol. 8, no. 12, pp. 194 – 202, 1989.

Jackson, D.; Iborra, F.; Manders, E. & Cook P. (1998). Numbers and organization of RNA polymerases, nascent transcripts, and transcription units in HeLa nuclei. *Mol. Biol. Cell*, vol. 9, pp. 1523–1536, 1998.

Kervrann, C. & Trubuil, A. (2004). An adaptive window approach for Poisson noise reduction and structure preserving in confocal microscopy, *International Symposium on Biomedical Imaging, ISBI'04*, Arlington, VA, April 2004.

Lichtman, J. W. & Conchello, J. A. (2005). Fluorescence microscopy. *Nat Methods*, vol. 2, no. 12, pp. 910–9, 2005.

Lippincott-Schwartz, J.;Altan-Bonnet, N. & Patterson, G. H. (2003). Photobleaching and photoactivation: following protein dynamics in living cells. *Nat Cell Biol*, vol. Suppl,S7-14 September 2003. [Online]. Available: http://view.ncbi.nlm.nih.gov/pubmed/14562845.

Lippincott-Schwartz, J.; Snapp, E. & Kenworthy, A. (2001). Studying protein dynamics in living cells. *Nat Rev Mol Cell Biol*. 2:444-56.

Moon, T. K. & Stirling, W. C. (2000). *Mathematical methods and algorithms for signal processing*. Prentice-Hall.

Rudin, S. L. & Fatemi, E. (1992). Nonlinear total variation based noise removal algorithms, *Physica D*, vol. 60, pp. 259–268, 1992.

Timmermann, K.E. & Nowak, R.D.(1999). Multiscale modeling and estimation of Poisson processes with application to photon-limited imaging, *IEEE Transactions on Information Theory*, vol. 45, no. 3, pp. 846–862, Apr 1999.

Underwood, R. (2007). Frap and flip, photobleaching technique to reveal cell dynamics, 2007. [Online]. Available: http://www.nikoninstruments.com/images/stories/litpdfs/nikon_note_3nn06_8_07_lr.pdf.

Vogel, C. & Oman, M. (1998). Fast, robust total variation-based reconstruction of noisy, blurred images, 1998. [Online]. Available: citeseer.ist.psu.edu/vogel98fast.html.

Willett, R. (2006). "Statistical analysis of photon-limited astronomical signals and images, *Statistical Challenges in Modern Astronomy IV (SCMA IV)*, 2006. [Online]. Available: http://www.ee.duke.edu/ willett/papers/WillettSCMA2006.pdf.

Willett, R.M. & Nowak, R.D. (2004). Fast multiresolution photon-limited image reconstruction, *IEEE International Symposium on Biomedical Imaging: Nano to Macro, 2004.*, pp. 1192–1195 Vol. 2, 15-18 April 2004.

Wohlberg, B. & Rodríguez, P. (2007). An iteratively reweighted norm algorithm for minimization of total variation functionals, *IEEE Signal Processing Letters*, vol. 14, no. 12, pp. 948 – 951, 2007.

16

Advantages of virtual reality technology in rehabilitation of people with neuromuscular disorders

Imre CIKAJLO and Zlatko MATJAČIĆ
Institute for rehabilitation, Republic of Slovenia, Slovenia

1. Introduction

Computer generated virtual environments (VE) offer capability to provide real time feedback to the user during the practice. The user can see the effect/consequence of his/her action and change the strategy/strength to achieve the desired goal. Such goal can be presented in virtual reality (VR) environment, which is a powerful tool and can efficiently replace tasks from the real life (Rose et al, 2000). Besides augmented feedback, the VR provide an option to practice motor control learning within environment to the people with neuromuscular disorders. Practice, which is essential in motor learning, can be facilitated by designing attractive VR environments, rather fun tasks or games.

The motor control training and muscle strengthening is often performed within clinical environment with predefined tasks and real objects. In contrary recently developed technologies enable development of similar tasks in a virtual world, using computer graphics. The tasks created within the virtual reality (VR) environment offer a capability to provide real time feedback to the subject during practice. Subjects can see the effect of his/her action immediately after the intervention and change the strategy to achieve the directed goal (Holden & Dyar, 2002). Such goal presented in the VR environment may have similar rehabilitation effect as the tasks applied in the real world (Rose et al, 2000). Besides augmented feedback, the VR can also enable selective motor control learning to the people with disabilities (Holden, 2005). The motor learning is defined as a sensory perceptual mechanism by which a new motor skill is learned. An essential component of the motor learning is feedback, which could be in the real world provided by the physiotherapist or in the virtual world provided by visual scenery. In contrary to the task performed in the real environment the virtual environment can be easily changed, adapted to the subject's needs, speed and level and keep all the modifications of eventually complex virtual scenarios within controllable and reproducible limits. Successful applications have been shown in stroke subjects (Sisto et al., 2002, Yang et al., 2008).

In the proposed chapter two various approaches are presented; a selective motor control training using position controlled single-joint rehabilitation robot (Cikajlo, 2008) and virtual reality balance training (VRBT) via telerehabilitation service (Smart Home iRiS). Both approaches apply VR technology to implement rehabilitation tasks into the clinically proven

rehabilitation devices. In the selective motor control training 10 neurologically intact individuals and a cerebral palsy (CP) child participated to demonstrate the novel concept. The results showed a significant progress in single joint torque control. The motor control training may also have significant impact on CP child's gait pattern, one of the major rehabilitation issues in CP children treatment. The dynamic balance training within virtual environment offers various tasks that require from the subject, standing in the dynamic balance frame, anterior-posterior and medio-lateral movements. In the case study a stroke subject demonstrated comparable clinical outcomes to conventional therapy. Additionally the telerehabilitation approach enables the subject to practice balance at home while the physiotherapist can supervise and provide instructions remotely.

2. Rehabilitation issues

Rehabilitation of neurologically impaired subjects tends to enhance motor control skill, which is claimed to be a direct result of repeatable practice (Sisto et al, 2002). The repeatable practice, particularly the frequency and duration of the practice, are often limited in rehabilitation. Physical therapy treatment in the rehabilitation center is often limited to three times per week in the acute settings and also the individual treatments can not last more than an hour due to fatigue. But to maximize the recovery an intensive treatment should be considered as the intensive rehabilitation may lead to better functional outcomes. Besides intensity, also repeated and targeted actions in therapy play an important role in the outcome enhancement. The goal is to achieve independence in a short time, but that is the limitation in the existing health care systems. One of the possible solutions in the future might be a telemedicine – a home based rehabilitation (Deutsch et al, 2007).
The improvement of functional outcomes could be achieved only with a certain level of subject's motivation and further enhanced with the variety of repeatable in targeted actions in the environment where the events can be added in the controllable manner. But it is nearly impossible to add unexpected events in the real life experiment in a safe and repeatable manner. Therefore the application of virtual environment may contribute to the solution of the problem.

3. Virtual reality technology in rehabilitation

Virtual environment (VE) is immersion of a person or an object in a computer generated environment such that the person experiences stereovision, correct perspective for all objects regardless of the motion, and objects in the environment move according to the subject motion (Kenyon et al, 2004). In order to assure these characteristics, applications of certain technologies are required (Burdea & Coiffet, 2003). These technologies apply several sensor types to generate artificial sensory information to allow individuals to experience and interact with the environment. But, the interaction and experience must be realistic enough, therefore the computer must generate new images fast enough to assure movement adequate to real-time responses. Simulation of the 3D environment can be presented through the immersion or nonimmersion scenario (Sisto et al, 2002). The immersion scenario is referred to the use of a head-up mounted display (HMD), equipped with 3D tilt sensor that tracks all the changes and sends them to the virtual environment. Hereby the important issue is the synchronization between the head movement and presented image. In the case

when the image update is to slow and not synchronized with the head movement, the subject may experience dizziness, nausea or other inconveniences, which may cause sickness. The state of the art, multiperson, room size, high resolution 3D video and audio system is the CAVE™ (University of Illinois, Chicago). The nonimmersion VE can be displayed on computer monitor or projected to the large screen in front of the subject. These displays have been preferred to use in clinical studies due to the relatively low price and no reports on cyber sickness (Holden, 2005).

The concept of rehabilitation is based on repetition of movement, providing feedback and motivation. Repetition is important for local and central nervous system, resulting in motor learning and cortical changes. Besides repetition of the movements during rehabilitation the tasks must be linked to some successfully accomplished assignments, i.e. target oriented approach. The subject practices movements during the rehabilitation treatment with medical professionals' assistance, but only can keep up with the extensive practice, when being well motivated. Therefore the motivation factor plays an important role in rehabilitation practice (Holden et al, 2002). And here the virtual reality (VR) comes in handy, providing visual feedback, repetitive practice and motivation. The VR technology offers enormous variations of objects, orientations, creative environments, augmented feedback. Besides that all the parameters of the VE like objects size, position, velocity, movement's speed, augmented support, etc can be varied under supervised repeatable conditions.

It has been reported on successful use of VE training in hemiplegia, where the upper extremity reaching task applied 16 times for 1 to 2 hours resulted in clinical and functional motor improvement (Jack et al, 2001). The authors also pointed out that VR offered motivation and fun for practice. Some authors investigated the effect of VE on visual function in rehabilitation which may be also important in the recovery of motor function, because the visual system is needed to guide movements (Sisto et all, 2002).

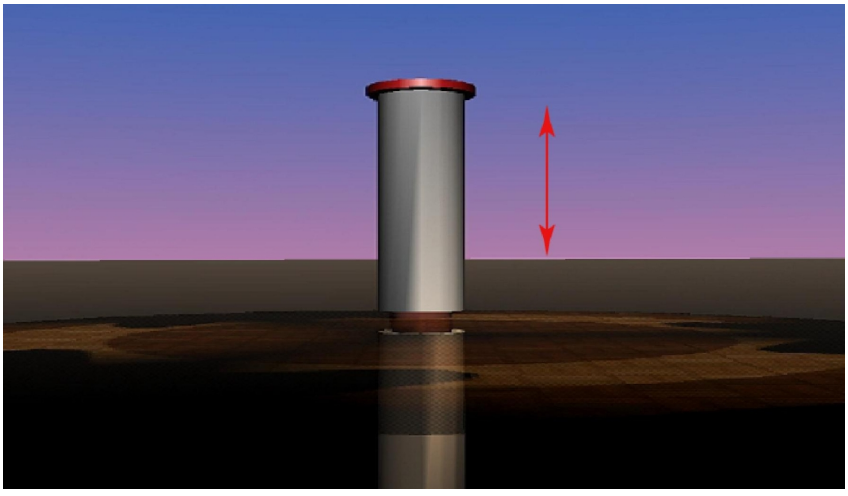


Fig. 1. The virtual reality environment for maximal torque assessment. The subject was asked to push the top button as high as possible by generating a knee joint torque.

Some examples of VE for motivation in maximal dynamometer based joint torque assessment (Fig. 1), selective motor control training (Fig. 2) and VRBT (Fig. 3) are presented.

Their common attributes of the presented VE in the nonimmersion scenario running on the commercially available inexpensive personal computer with 3D graphic adapter. In cerebral palsy children the VE with a growing cylinder with top red button (Fig. 1), associated with the amount of the generated joint torque presented a considerable issue and motivated the child to play the game. The goal was to push the top red button higher and higher by generating joint torque in isometric and isokinetic mode (Dvir, 2004) and simultaneously assess muscle electromyography to identify muscle power and selective motor control. The selective motor control training VE (Fig. 2) applied trajectory controlled dynamometer to impose joint movement, while the task required from the subject to generate adequate joint torque to move the object (bee) flying in the virtual office and hitting targets (flowers) in the VE. When the object touched the target, the target disappeared and at the end of the game the number of hits scored. Besides the difficulty level was pre-defined by the level of generated joint torque tolerance.
The VE for VRBT was designed as a game of virtual walk. The task required from the subject to "walk" by tilting the dynamic standing frame (Cikajlo & Matjačić, 2009; Medica Medizintechnik, Germany) forward and "turn" by tilting the frame in frontal plane. The task (Fig. 3) was divided in three difficulty levels, each comprising additional obstacles on the way.

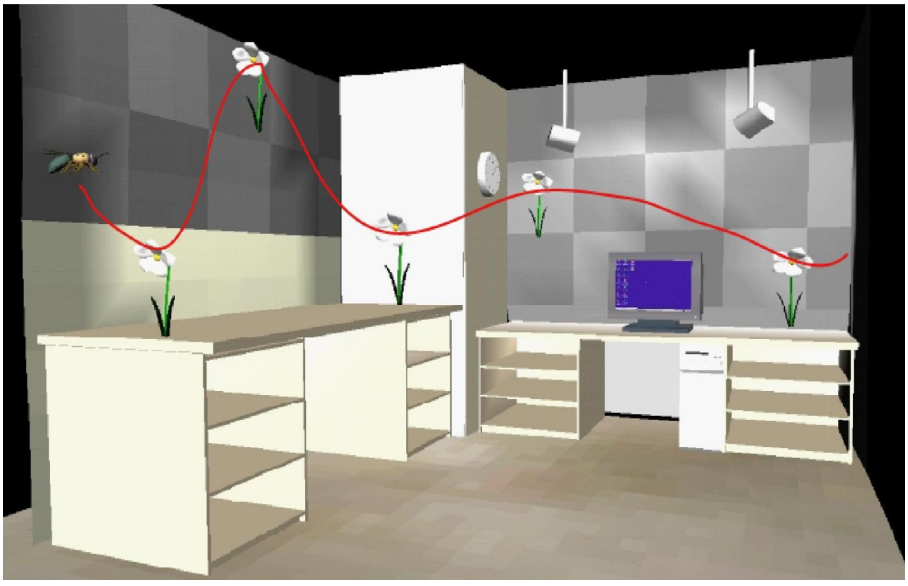


Fig. 2. The virtual reality environment for motor control learning; by generating adequate joint torque the subject can move the object (bee). The extension knee torque causes upward movement and the flexion torque a downward movement of the object, while the object moves in transversal plane synchronized (time) with the dynamometer.

In the first and the easiest level (level 1) there were no obstacles on the tree lined path, where the subject passed by the woman on the left and at the dolphin statue turned left and continue to the buffet, where the obstacles were two tables and chairs. Afterwards the

subject made a turn around the tables and returned back to the dolphin statue and continued his way passing by the security guard to the entrance of the building. When the subject entered through the door, the task restarted from the begging. At the second level (level 2) four benches were added as obstacles which the subject needed to avoid. And the third level (level 3) added three additional cans and two pools near the dolphin statue (Fig 7, Fig. 12).
VE in rehabilitation is believed to positively influence on practice, because of the ability to make tasks easier, less dangerous, more customized, fun and easy to learn due to the feedback provided. Some studies have examined these advantages and provide experimental evident that motor learning in the VE may be superior (Holden, 2005).

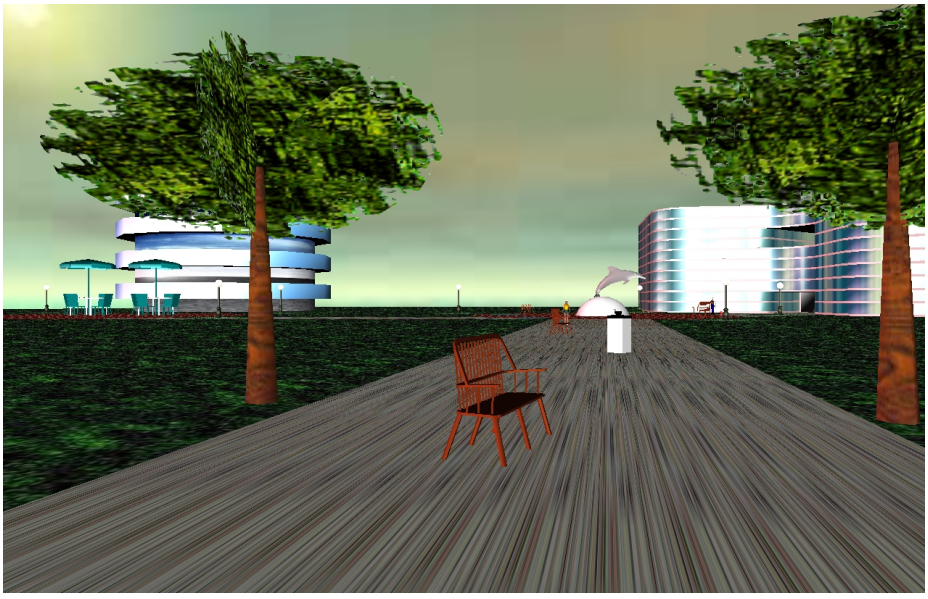


Fig. 3. Virtual reality technology was used to implement the dynamic balance training task, where the subject "walked" through the scenery by tilting the apparatus. The task runs in web-explorer, which made it easy to implement it in the telerehabilitation service.

4. Telerehabilitation

The low cost personal computer based VE equipment is nowadays inexpensive and available to wide range of users and will eventually allow VE based rehabilitation in locations other than the rehabilitation center hospital, e.g. a patient's home. Patients who experienced any disabling event such as stroke will most likely return to their home after the rehabilitation treatment in the rehabilitation hospital. Nowadays the treatment is becoming more and more expensive and also effective, therefore we may expect that the patient would return to his/her home sooner than in the past. But the majority will still need to continue with the rehabilitation through the outpatient service or through the home care. For those patient who do not own a vehicle or lack of public transportation or live in the remote areas

the outpatient service can present a barrier. On the other hand the insurance companies would most likely prefer a service that could be performed on patient's home, the patient would be given instructions over the communication media and if possible posses a device that could evaluate his functional motor capabilities. All this would significantly reduce the number of outpatient visits and in some cases reduce the length of the hospital treatment.
The remote rehabilitation or telerehabilitation service can be provided through the broadband internet connection between the medical centre and the user's home environment or local medical institution. The presented approach has been demonstrated in the modern smart home iRis (www.ir-rs.si/en/smarthome) equipped with rehabilitation aids integrated into the smart home's technologies. The application runs on the personal computer connected to the living room 42" LCD, which could be remotely operated, even from the electric wheelchair. The non-immersion scenario VE (Fig. 3) was designed to run in web-explorer on the LCD and enabled the subject to practice balance in the home environment. The supervising physiotherapist or other medical professional in the medical centre could monitor remotely the VE and the task being performed in real-time. The currently available technology can enable the VE monitoring over the internet using the web browser (Internet Explorer, MicroSoft with blaxxun plug-in). Moreover, the videoconference session with the rehabilitation centre was on his disposal anytime in order to provide additional tasks and important advices to the subject during the dynamic balance training. The videoconference was established by using a free-ware Skype (Skype Technologies S.A., Luxembourg, EU). Full-duplex voice and video enabled the physiotherapist to advise the subject to correct the posture during the therapy. Data of each training session were recorded and presented important information on subject's activity or inactivity and an objective information on subject's balance capabilities. On the basis of that information the subject is invited to the outpatient visit to the rehabilitation centre, where other clinical tests can be carried out.

5. Methods

In the chapter two various approaches of VR applications are presented:

- VR based motor control training using dynamometer
- VR based dynamic balance training with telerehabilitation service

Both used nonimmersion VR scenarios presented on the LCD. The first approach with the dynamometer demonstrated the power of visual motivation and the possible application of selective motor control training based on gait kinematic parameters. The second approach demonstrated the use of VR to motivate the stroke subject for balance training. Besides the task was supervised remotely through the internet and the subject was given instructions via the videoconference session.

5.1 VR based motor control training using dynamometer

5.1.1 Equipment

The proposed system (Fig. 4) consists of Biodex System 3 dynamometer (Biodex Medical Systems, New York, USA) with additional analog output and optional serial interface, personal computer with multifunction DAQ board (NI-PCI- 6259, National Instruments,

Austin, Texas, USA) servicing data assessment, Biodex control, graphical user interface (programmed in Matlab, The MathWorks, Inc., Natick, MA, USA) and VR environment (blaxxun Contact VRML plug-in). Muscle activity was measured by electromyography (EMG), MyoSystem 2000 (Noraxon U.S.A. Inc., Arizona, USA).

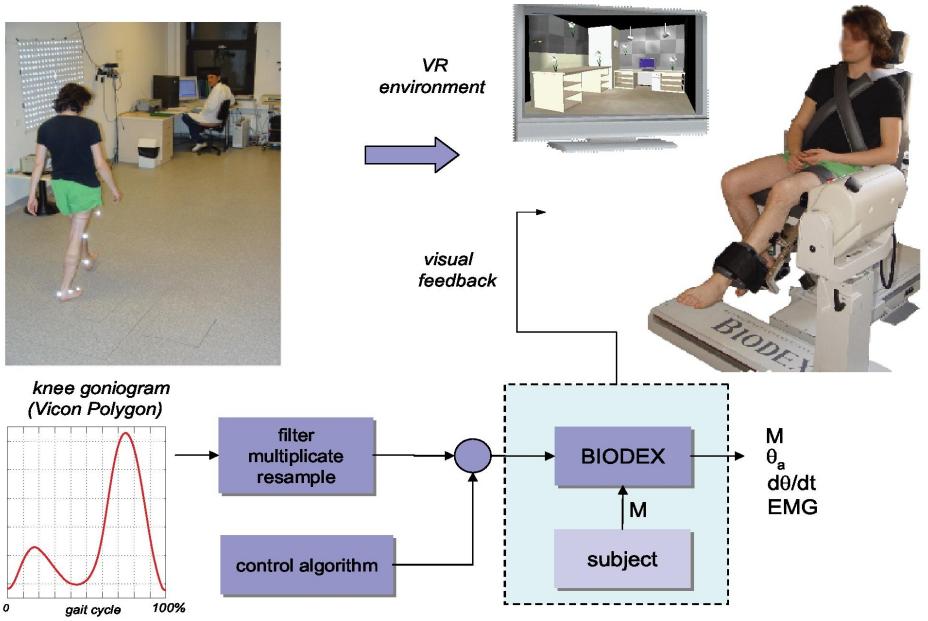


Fig. 4. The goniogram controlled dynamometer performs pre-programmed joint movement while the subject moves the object in Virtual scenery by generating joint moment. The goniogram from gait analysis software is applied in position control of the dynamometer. The subject's task is to hit all the targets (flowers) by providing joint moment (adequate to avatar's vertical position) while the avatar (bee) is synchronized with the goniogram.

5.1.2 Dynamometer control

The control of a dynamometer, which is in general intended for isokinetic or isometric dynamometry (Dvir, 2004), has been modified to the level that enabled position control. The joint goniograms (joint angle per gait cycle), assessed during participating subject's gait analysis, have been averaged and normalized to a single gait cycle (GC). The obtained GC goniogram was resampled and presented the desired trajectory in position control of the dynamometer (in this case can be considered as a 1-dof robot). Thus the angular velocity of the dynamometer in controlled mode was determined by the selected goniogram and sample time/GC time. Simultaneously the subject was asked to generate the adequate joint torque in order to accomplish the required task. The joint torque was measured by Dynamometer's built-in torque sensor and calibrated on-line to compensate for gravity (Herzog, 1988). The information on joint torque was introduced into the VR environment, where the moment value caused an adequate movement of the object (Fig. 4).

Virtual reality environments have been designed to provide feedback information to the user/subject. In the first stage of development two specific VR tasks were designed:

- Environment for maximal joint torque (Fig. 1) assessment with specific visual object that stimulated the subject to put his maximal effort to accomplish the task. The subject was asked to push the top button as high as possible by generating a knee joint torque;
- Environment for motor control learning (Fig. 2) where the specific task required from the subject to generate specific joint torque and control the object in the VR environment to accomplish the task - by generating adequate joint torque the subject can move the object (bee). The extension knee torque causes upward movement and the flexion torque a downward movement of the object, while the object moves in transversal plane synchronized (time) with the dynamometer.

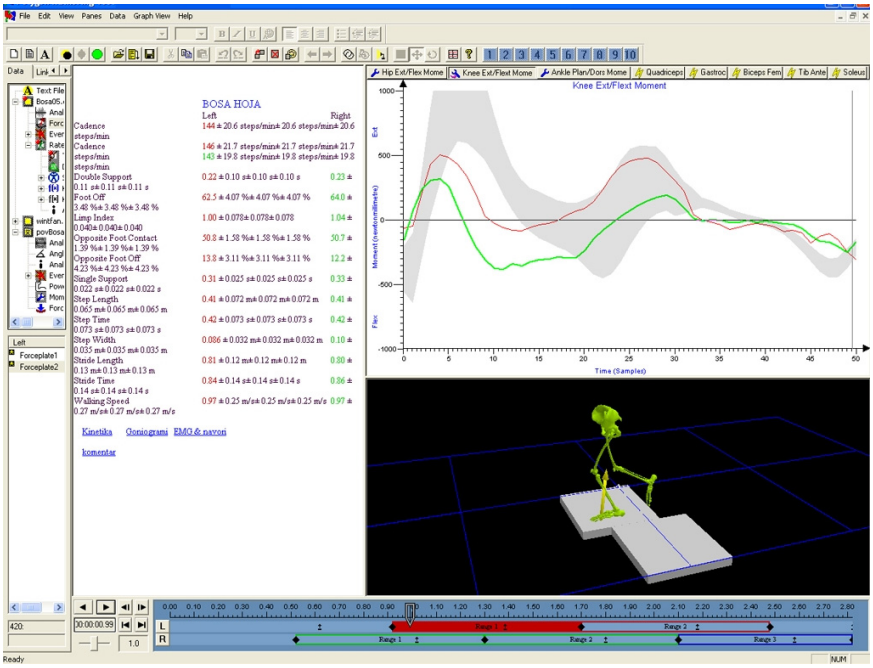


Fig. 5. Knee torque per gait cycle assessed and analyzed with commercial gait analysis software was exported to file and uploaded into developed program for dynamometer control.

The VR task required from the user to push the cylinder and the button as high as possible (Fig. 1) or hit as many targets as possible (Fig. 2). The VR object (bee) moved in the environment in transversal plane over time, synchronized with the joint goniogram, while the object's vertical movement was related to the generated joint torque value. The joint extension torque provoked the object's upwards movement and the flexion torque moved the VR object downward. The targets were positioned in the VR world in a way that the generated joint torque trajectory (Fig. 5) would equal to the moment reference value [8] during the gait, if the subject succeeds to hit all the targets. The target was considered hit when the generated torque was within the selected limits. The selected limits defined influenced on the task execution difficulty and thus defined the difficulty levels. For the

level 1 the torque variation could be within 30% (±15%), for the level 2 20% (±10%) and for the level 3 10% (±5%) of the accurate torque, required to hit the target.

5.1.3 Subjects

10 neurologically intact volunteers (27.2 SD 3.5 years, 71.8 SD 8.1 kg, 173.4 SD 5.5 cm) participated in the virtual environment task and protocol development and preliminary evaluation. The volunteers had no muscular-skeletal impairment or any disease that would affect motor control capabilities. Besides, a 12-year old boy with cerebral palsy (CP II, spastic diparesis, 149 cm height, weight 48kg, functional electrical stimulation 1-ch peroneal nerve 1 hour/day, crouch gait, poor selective motor control), a patient of the Rehabilitation hospital, voluntarily participated in the case study due to his good cognitive capabilities. The child had no prior experience with VR.

The methodology was approved by local ethics committee and the subjects/the subject's parents/ gave informed consent.

5.1.4 Protocol

The subjects were seated on the Biodex reclining chair. Stabilization of the subjects was achieved by placing velcro straps across the chest, around the waist, just above the right knee and just above the right ankle, which secured the right lower leg to the input shaft of the dynamometer. The dynamometer lever was set to 45^{o} flexion and the dynamometer operated in goniogram position control (Cikajlo, 2008) or isometric mode (Dvir, 2004). In addition, we visually aligned the estimated transverse rotational axis of the dynamometer. The surface EMG electrodes were placed on the right lower extremity's quadriceps muscle. The entire test procedure was repeated for the subjects over 3 consecutive days at approximately the same time of the day to limit the extent of possible diurnal variation. Each day the VR task for maximal joint torque assessment (Fig. 1) was performed; 3 times for knee flexion and 3 times for knee extension. The protocol proceeded with the VR task for selective motor control training (Fig. 2). During the task execution the surface EMG of quadriceps muscles were assessed. On the first day a slow (GC = 15s) task speed was selected for each difficulty level, on the second day the task execution speed was increased for 50% (GC = 10s) and on the third consecutive day for additional 100% (GC =5s), in total 3 times faster than the task execution speed at the beginning.

5.2 VR based dynamic balance training with telerehabilitation service

5.2.1 Balance training

Balance training was based on dynamic balance training standing frame made of steel base construction placed on four wheels, which when unlocked enable the apparatus mobility. The later is important in clinical environment where the rehabilitation aids have no dedicated space. The standing frame is made of aluminum and fixed to the base with passive controllable spring defining the stiffness of the two degrees of freedom (2 DOF) standing frame. The stiffness of the frame is set up according to the individual's requirements. On the top of the standing frame a wooden table with safety lock for holding the subject at the level of pelvis was mounted.

The subject was standing in vertical position in the balance trainer with his hands placed on the wooden table in front of him and secured with safety lock from behind at the level of

pelvis, preventing to fall backward (Fig. 6). The standing frame can tilt in sagital and frontal plane for ±15O. The tilt of the balance standing frame (BalanceTrainer, Medica Medizintechnik, Germany) was measured by commercially available three-axis tilt sensor (Xsens Technologies, Enschede, The Netherlands).
The task for dynamic balance training was based on subject's movement, i.e. weight transfer in sagital and frontal plane, resulting in BalanceTrainer tilt. The tilt, assessed by sensor mounted on BalanceTrainer frame, resulted in the immediate action in the designed virtual environment (Fig. 3).

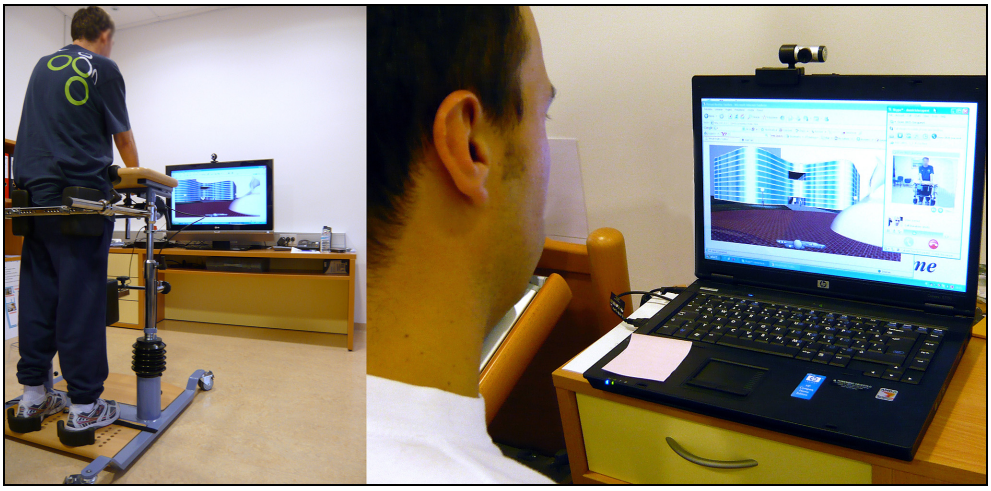


Fig. 6. Telerehabilitation in practice. A physiotherapist (right photo) is supervising and assisting the subject remotely during dynamic balance training (left photo) in Smarthome (Smart Home iRiS) The BalanceTrainer apparatus was equipped with tilt sensor and detected the subject's movement. Adequately the virtual environment changed and the subject was required to avoid obstacles and to reach the goal – the door in the building.

5.2.2 Subject

In the pilot study a 47 year old subject (male, 180 cm, 80kg, intracerebral hemorrhage a month before the therapy) with right side affected and a slight motor dysphasia and dysarthria. The subject has been cardiovascular compensated (Julu et al, 2003), without additional diseases and medications that may affect the balance. His cognitive functions were preserved, allowing him to follow the given instructions.
The methodology was approved by local ethics committee and the subjects gave informed consent.

5.2.3 VR supported telerehabilitation protocol

The testing of the VE based telerehabilitation balance training took place in the SmartHome Iris (Smart Home iRiS), a well equipped demonstrating apartment for people with special needs. The subject was standing in the BalanceTrainer (Fig. 6 left), secured with safety rod at the pelvis level, and placed his hands at the front table. In front of the subject was a LCD screen with a multimedia camera with built-in microphone on the top. An engineer who

was responsible for the subject in the SmartHome Iris established a videoconference call with the physiotherapist who was located in the remote room. The physiotherapist could see the subject and provided him instructions how to correct his posture while the subject was performing the VE based task (Fig. 7).
The subject performed the therapy five times a week, each time for 17 to 20 minutes. On the first week the subject used level 1 of the VR task for balance training, on the second week, when the physiotherapist estimated the subject's progress, the training started with level 2 and on the third and fourth week with level 3.
The subject balance capabilities were also assessed with clinical instrument (Berg Balance Scale – BBS, Berg et al., 1992) at the beginning of the therapy and after four weeks. Besides BBS, the standing alternatively on the healthy and the affected lower extremity, the »timed up and go« test and the 10-m rapid walk test were also performed.
Between the VE task sessions the subject took part in other standard neurotherapeutic programs (cognitive rehabilitation, treadmill training, gait…, etc), but no balance training.

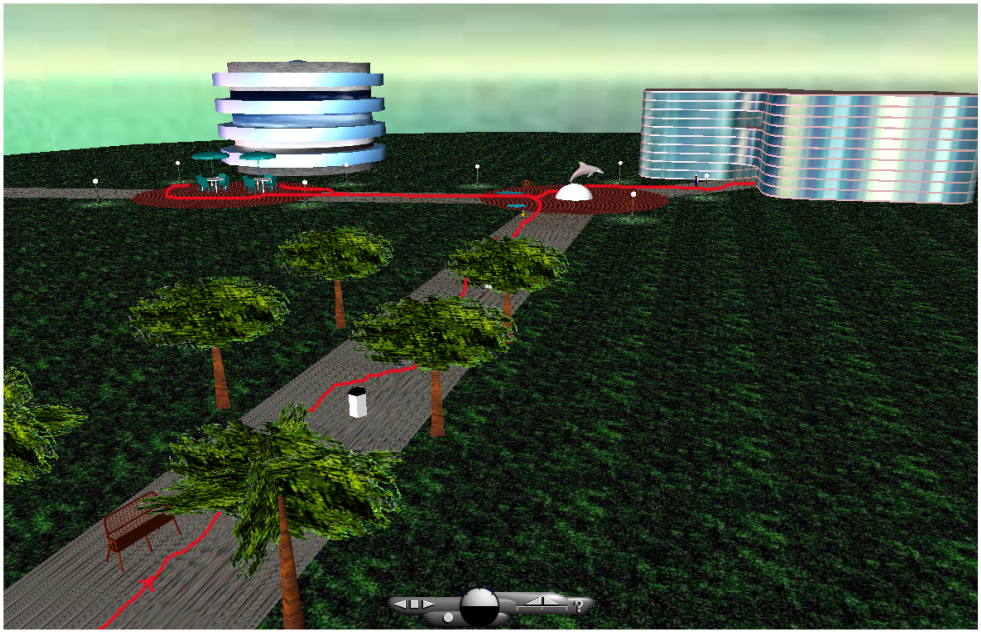


Fig. 7. Telerehabilitation task in VE (perspective view). The subject moved in the VE by tilting the standing frame. The designed task's path (red line) required from the subject to transfer the weight from left to right side and vice-versa and in this manner perform voluntary balance control training.

6. Results

6.1 VR motor control training using dynamometer

The knee joint torque assessed in neurologically intact volunteers when performing the maximal moment VR task extended between 111.3 (SD 23.2) Nm and 122 (SD 29.5) Nm (Table 1).

The obtained joint torques were normalized and averaged on GC and contrasted to the knee (Fig. 8) normative (Winter, 2005) torque (reference values), which was generated in neurologically intact subjects during gait (shaded area). In the upper Fig. 8 the triangles presents the targets (flowers) from the VR environment. There were no significant changes in joint torques when the task difficulty level was changed, but the generated torque was noticeably delayed with increased velocity (target 2 in Fig. 8). There were almost no changes in EMG responses in neurologically intact subjects (Fig. 9). The expected knee extensor muscles were activated simultaneously with the generated joint torque. The Fig. 9 below shows the actual knee joint angle during VR task execution normalized on GC.

MaxTorque	Variable	Day		
		1	2	3
Knee / Nm (SD)	extension	122.6 (29.5)	112.9 (22.2)	111.3 (23.2)
Knee goniogram - control	Difficulty level	Velocity		
		1	2	3
hits / %	level 1*	94.0 (12.9)	94.9 (10.1)	95.4 (12.8)
	level 2*	93.7 (14.1)	91.4 (14.9)	88.5 (15.9)
	level 3*	64.0 (32.1)	61.1 (27.4)	54.3 (28.0)

*paired T-test (statistically significant - $p < 0.05$)

Table 1. Numerical values of generated maximal joint torque for each day and percentage of hits for each difficulty level and velocity in neurologically intact volunteers.

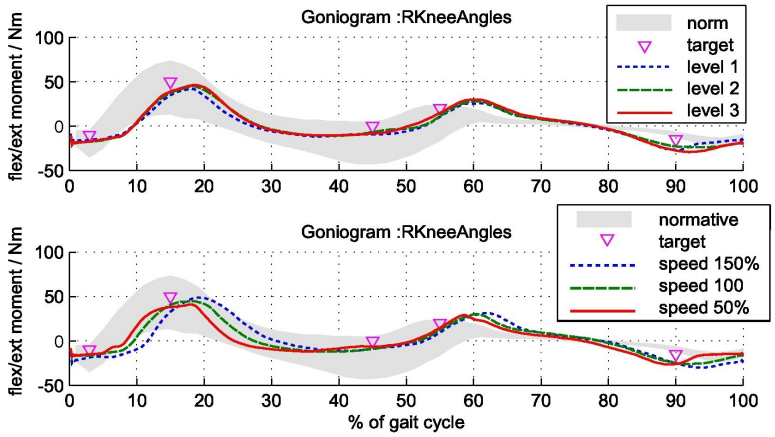


Fig. 8. The knee torque generated by the neurologically intact subjects with contrasted reference values (shaded area) taken from the gait knee joint moment (Winter, 2005). There were no significant changes in the torque when the task difficulty level was changed, but the torque applied was noticeably delayed with increased velocity (target 2) in neurologically intact volunteers.

The overall hit score (Table 1) over all targets demonstrates that the increasing velocity had minor effect in comparison with the difficulty level, what may be the consequence of the fact that the subjects have mastered the VR task. For the easiest difficulty task level (1) a change in hit score with increasing velocity was minor, slightly larger change could be noticed for the task level 2 and poor hit score with high standard deviation was recorded in level 3 (the most difficult), where the success rate was less than 65%.

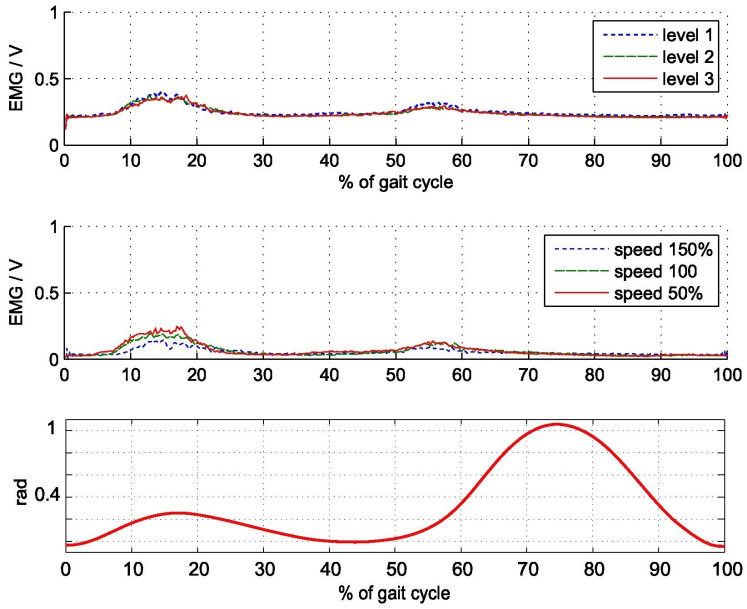


Fig. 9. There were almost no changes in EMG responses in neurologically intact volunteers with increasing difficulty level, but only expected latency which was related to the delayed torque generation. Bellow: The actual position of the dynamometer corresponded to gait knee joint goniogram.

The results of the max torque task (Table 1, Fig. 1) in the CP child show a day-to-day increase in average maximal knee extension joint torque generation in isometric conditions: day 1 – 12.3 Nm (SD 1.1), day 2 – 14.1 Nm (SD 2.3) and day 3 – 17.7 Nm (SD 1.2). The average for knee flexion joint torque: day 1 – 15.3 Nm (SD 2.1), day 2 – 24.1 Nm (SD 3.1) and day 3 – 23.5 Nm (SD 2.1). The normalized time course of the generated joint torque in the CP child showed high oscillations on the first day of the training session VR target tracking task (Fig. 2). The subject was also trying to correct his action by rapid flexion-extension movement, especially when trying to hit the 2nd target. The action resulted in oscillations at the moment where action was required, i.e. between 10 and 30% of the GC. The GC was 15s and treated as very slow action. On the second day of training (Fig. 10 middle panel) the subject showed more precise control of generated joint torque and improved tracking resulted in hitting the target nr. 2. The task execution speed (GC = 10s) was also increased. More precise joint torque control has been demonstrated through the complete task. The

lower panel of Fig. 10 shows that the oscillations caused by the quick knee extension/flexion torque have disappeared and the subject managed to control the joint torque as requested by the task. The bold line in Fig. 10 shows the mean value joint torque assessed each day for specific task execution speed and task difficulty level 1.

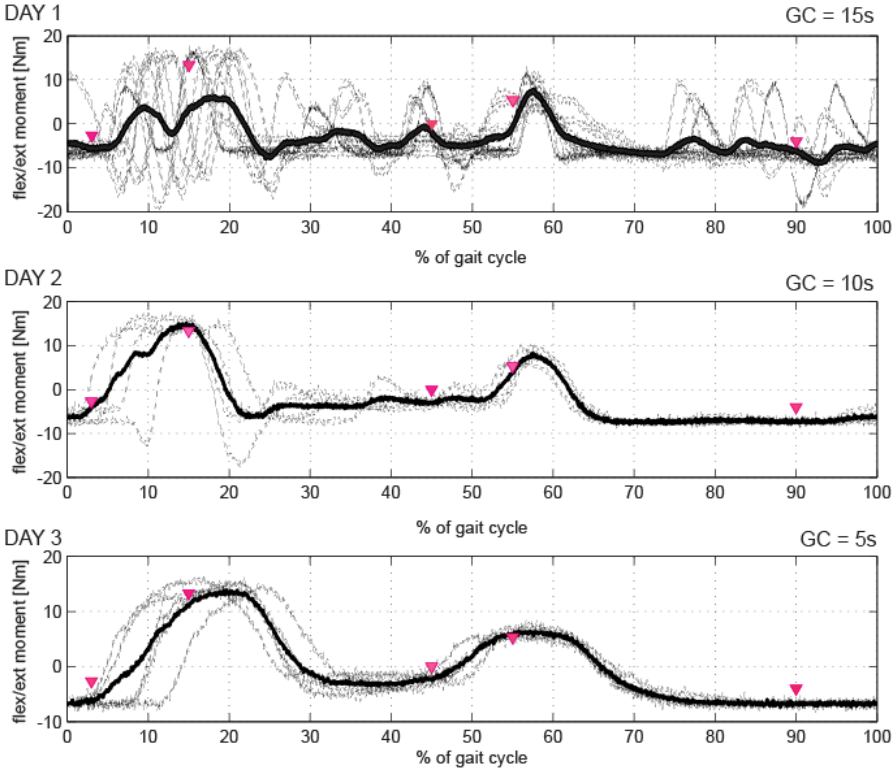


Fig. 10. The assessed knee joint torque in CP child shows oscillations in the range (e.g. target nr 2) where high torque was required on the first day of training. The knee joint torque control has evidently improved as the subject was capable of target tracking almost without any oscillations at the end of the third day of training. Targets are marked with a triangle.

The overall hit statistics (all levels, all targets) also demonstrated an increase in mean hit score (Fig. 11) in the CP child. Results in the lower right Fig. 11 demonstrate the gradual increase of each target hits over all task difficulty levels; target 1 (76 ->86%), target 2 (41 ->50%), target 3 (79 -> 88%), target 4 (67 -> 88%) and target 5 (85 -> 98%). The only decrease (90.5 -> 88.0%) took place at the target 3 from day 2 to day 3, but was considered insignificant ($p > 0.05$). The target 2, which required the highest peak torque generation, also demonstrated statistically significant (41.2 -> 50.0%, $p < 0.05$) target hits increase between the first and the last session.

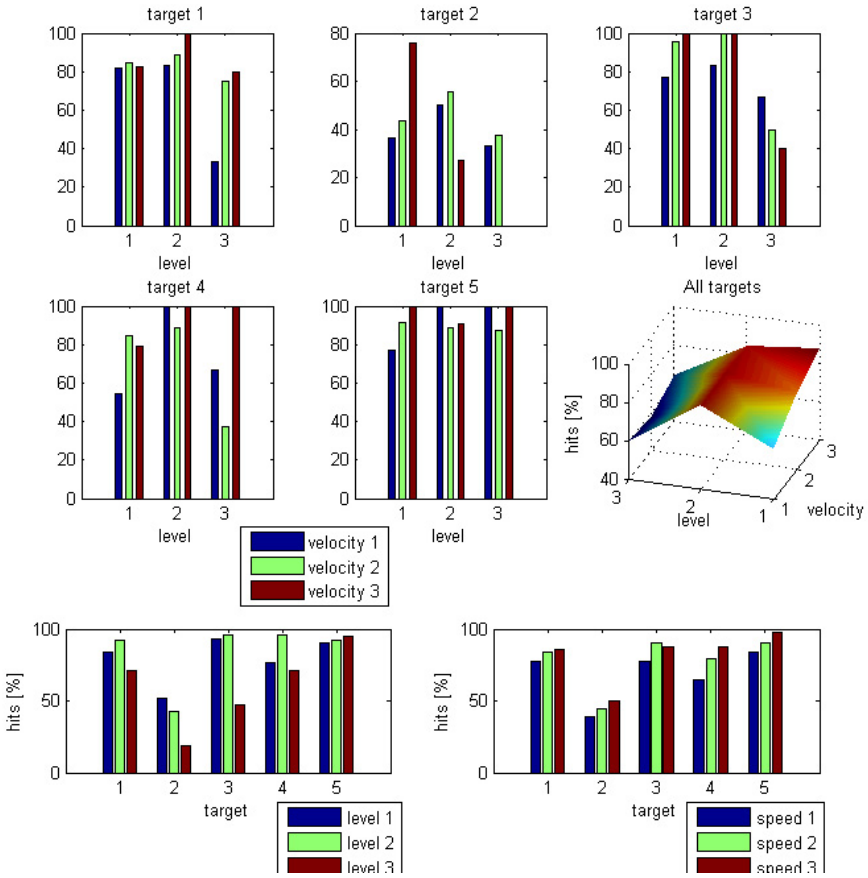


Fig. 11. The mean target hit rate in CP child for each target in VR target tracking (Fig. 2) for each day/speed with standard deviation. The bar graphs show increase in overall target hit rate in three consecutive training days. The VR target hits rate has increased for all targets (except for target nr 3 from day/speed 2 to day/speed 3) in three consecutive training days regardless of the task difficulty level (lower right graph).

6.2 VR dynamic balance training

Fig. 13. shows the mean track times, achieved at the beginning of the therapy and a week later (start end) for each difficulty level. In the first week the subject made an enormous progress, reducing the mean track time from 48s to 42s ($p < 0.05$). On the second week the task was a step more difficult with obstacles on the way, but the score was still encouraging, from 56s to 52s ($p = 0.031$). The level 3 with even more obstacles, which also lengthened the path, also demonstrated progress from 53s to 51s ($p = 0.0157$).

The score of each training track was displayed in seconds together with the object hits (penalty time 5s) and total score:

51s 3 hits (+15s) 66s

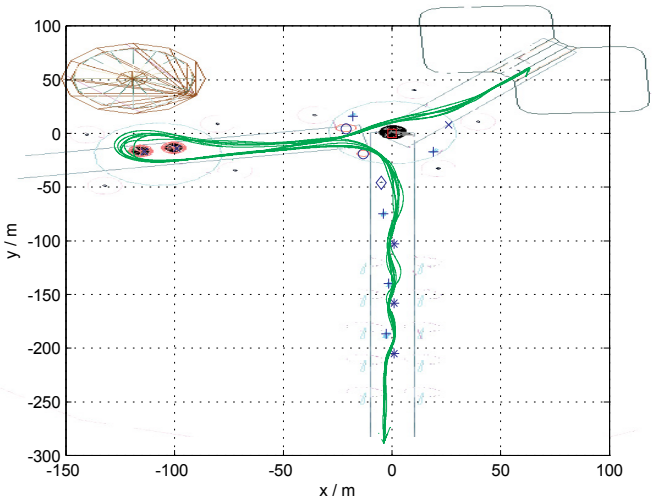


Fig. 12. The top view of the VE task objects (* cans, + benches, O pools, x hydrant,…) with training tracks appended. The VR task started at the lower end of the figure and finished at the upper right corner with entering the building. It is obvious that the subject has hit the tables at the buffet twice, crossed the pool and hit the 2nd and 3rd can.

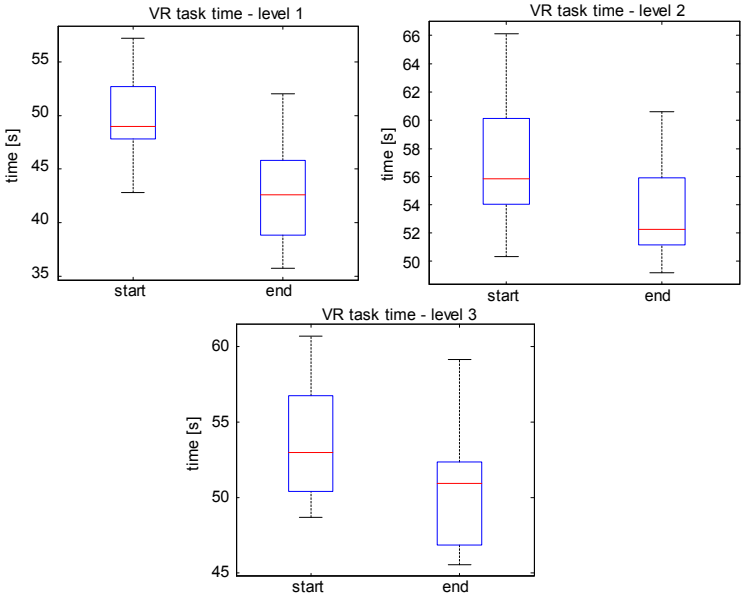


Fig. 13. The upper left graph: Time to complete the task was shortened for 15% in the first week of therapy ($p < 0.05$). The upper right graph: Time to complete the task (level 2) at the end of the second week of therapy was shortened for 8% comparing with results from the beginning of the second week ($p = 0.031$). Bellow: At the end of the third week the average time to complete the task (task level 3) was shortened for 4% ($p = 0.0157$).

For the clinical purposes the therapy has been also evaluated with clinical instruments; BBS, Standing on a single extremity and Timed Up & Go test and 10m walk. The BBS has increased from 50 up to 54 after the treatment. The results of the "Standing on healthy extremity test" improved from 41.10s to 46.82s and on the "affected extremity" improved from 4.87s to 13.63s. Similar improvements were found in "Timed Up & Go test", (from 14.93s to 11.47s) and in "10m walk", where the subject managed to beat the distance in 11.20s, a way faster than before the treatment (12.57s).

7. Discussion

The presented examples demonstrate the use of virtual reality technology in rehabilitation to motivate the subject during the target oriented task and therefore contribute to the augmented rehabilitation efficacy. Both examples enable task difficulty control and task velocity adjustment. The first example introduces a novel approach in selective motor control training. The VR tasks in this example enable and motivate the subjects to fully cooperate with the task which may enhance the rehabilitation outcomes and on the other hand allow creating unique training environments where task constraints can be modified and supervised. Besides, the single muscle group tasks enable identification of some pathological gait causes, which could not be identified only with computer gait analysis (e.g. weak muscle, selective motor control...). The results demonstrated that the day-to-day training in VR environment with integrated real-time feedback may improve the subjects' perceptive and predictive capabilities which may have impact on selective motor control; neurologically intact individuals managed to follow the task difficulty change as they learned the task very quickly. In this case the increased task velocity led to the generated joint torque and adequate muscles' EMG delay. The goal of the particular VR task was to generate joint torque similar to the reference assessed in normal gait (Winter, 2005). The mean generated joint torque was within the standard deviation of the gait normative. Only the curve shape has changed in relation to the task execution velocity. The peak torque at the VR target nr 2 (Fig. 2) was delayed due to the required rapid and strong action to achieve the required torque in order to hit the target. Here the CP child, who had major problems with selective motor control, especially in hip and knee joints, managed to generate the adequate knee joint torque only after a three days of intensive training. We assume that such training may also have impact on gait pattern change in terms of improved selective muscle control during stance.

The second example shows a right-side affected subject who had major problems with balancing and weight transfer, especially when loading his affected extremity, experiencing a VRBT. The subject sometimes left the desired trajectory in VE and hit some obstacles. But as the subject was highly motivated, he was improving his performance on a daily basis. The subject managed to accomplish the task faster, more accurate, even when the physiotherapist increased the difficulty level. The outcomes demonstrated that the subject achieved practically the same track time at the end of therapy with level 3 as at the beginning with level 1, where no obstacles were present. This may lead to the conclusion that the subject has completely mastered the VR task. Besides the objective engineering indicators also the clinical instruments revealed improvement of the subject's clinical status; improved weight balance – standing on the affected extremity, faster in timed up&go and 10m walk test and increased BBS. Especially the 10m walk might be very relevant when

considering that it is argued that timing of gait over 10 meters is a valid reliable measure that is currently underused (Wade et al, 1987).
Both presented examples are pilot studies or proof of concept within limitations, but may demonstrate a successful application in clinical/home environment. Besides fun and motivation, which were the key of successful selective motor control training in CP child, it has been demonstrated that hospitalized stroke subject may continue the rehabilitation process at home. Such service may cut the cost on the long term and patients from rural communities and even urban settings with poor access to transportation and therefore limited access to outpatient service could also benefit from telerehabilitation service (Rosen, 1999).

8. Conclusion

After all, "Why actually use the virtual reality in rehabilitation?" might be a legitimate question and is entitled to explanation. Creating appropriate virtual environments for rehabilitation is rather a difficult task, requiring proficiency in subjects' cognitive capabilities, motor control, etc… and is presumably related to high costs. Therefore it would be more convenient to use real environments. But as shown in the chapter the VEs enable creating various tasks, changing parameters, adding, subtracting objects and other modifications that may influence on enhanced learning. Important for motor learning practice is that the VE enables the use of augmented feedback (Holden & Dyar, 2002) about the performance. Another important fact is that the VE can be very simple in the early stage of rehabilitation, constraining the subject to focus only on the key elements of the task – targeted tasks. The VE could be customized for different therapeutic approaches and provide options to guide the subject through the VE and help him to correct errors. One might consider that errors in real environment may result in physical injury or damage of auxiliary objects. Therefore one must consider some of the important advantages that VE offer over real environments; safety, smaller space and less equipment requirements, repeatability (Mirelman et al, 2009) faster change of task requirements or task adaptation and finally lower cost, especially when the patients can practice on their own without presence of the medical professional or even at home – telerehabilitation.

9. References

Bryanton, C.; Bossé, J.; Brien, M.; McLean, J.; McCormick, A. & Sveistrup, H. (2006). Feasibility, motivation, and selective motor control: virtual reality compared to conventional home exercise in children with cerebral palsy, *Cyberpsychol Behav*, Vol 9, No. 2, 123-128.

Burdea, G & Coiffet, P. (2003). *Virtual Reality Technology*, Second Edition with CD-ROM, Wiley, ISBN 0471360899, New Jersey

Berg, K.O.; Wood-Dauphinee, S.L.; Williams, J.I. & Maki, B. (1992). Measuring balance in the eldery: validation of an instrument. *Can. J. Public. Health*. Vol 25, 7-11,

Cikajlo, I. (2008). Integration of Virtual Reality based task into controlled dynamometry to enhance motor rehabilitation, *Proceedings of IEEE Virtual Rehabilitation, 2008*, pp 157-162, ISBN: 978-1-4244-2701-7, Vancouver, Canada, August 2008, Omnipress USA,

Cikajlo, I. & Matjačić, Z. (2009). Directionally specific objective postural response assessment tool for treatment evaluation in stroke patients. *IEEE Trans Neural Syst Rehabil Eng* Vol. 17, No. 1, 91-100.

Dvir, Z. (2004). *Isokinetics: Muscle Testing, Interpretation and Clinical Applications*, 2rd ed., Churchill Livingstone (Elsevier Limited), ISBN 0-443-07199-3, UK, 2004.

Deutsch, J.E.; Mirelman (2007), A. Virtual reality-based approaches to enable walking for people poststroke, *Top Stroke Rehabil.* Vol. 14, 45-53. Review.

Deutsch, J.E.; Lewis, J.A. & Burdea, G. (2007). Technical and patient performance using a virtual reality-integrated telerehabilitation system: preliminary findings. *IEEE Trans Neural Syst Rehabil Eng* Vol. 15, No. 1, 30-34.

Edmans, J.A.; Gladman, J.R.F.; Cobb, S.; Sunderland, A.; Pridmore, T.; Hilton, D. & Walker, M.F. (2006). Validity of a Virtual Environment for Stroke Rehabilitation, *Stroke*. Vol. 37, 2770-2775,

Hercog, W. (1988). The Relation Between the Resultant Moments at a Joint and the Moments measured by an Isokinetic dynamometer, *J Biomech*, Vol. 21, No. 1, 5-12.

Holden, M.K. & Dyar, T. (2002). Virtual environment training: A new tool for neurorehabilitation. *Neurology Report*, Vol. 26, No. 2, 62-71,

Holden, M.K. (2005). Virtual Environments for Motor Rehabilitation: Review, *CyberPsychology & Behavior*, Vol. 8, No. 3, 187-211,

Holden, M.K.; Dyar, T.A. & Dayan-Cimadoro, K. (2007). Telerehabilitation using a virtual environment improves upper extremity function in patients with stroke. *IEEE Trans Neural Syst Rehabil Eng*., Vol. 15, No. 1, 36-42.

Julu, P. O. O.; Cooper, V. L. ; Hansen, S. & Hainsworth, R. (2003).Cardiovascular regulation in the period preceding vasovagal syncope in conscious humans. *J Physiol,* Vol. 549, No. 1, 299-311, DOI: 10.1113/jphysiol.2002.036715

Kenyon, R.V.; Leigh, J. & Keshner, E.A. (2004). Considerations for the future development of virtual technology as a rehabilitation tool. *J NeuroEng Rehab*,Vol. 1, 1-13.

Mirelman, A; Bonato, P & Deutsch J.E. (2009), Effects of training with a robot-virtual reality system compared with a robot alone on the gait of individuals after stroke, *Stroke*; Vol. 40, 169-174.

Nyberg, L; Lundin-Olsson, L; Sondell, B.;Backman, A.; Holmlund, K.;Eriksson, S.; Stenvall, M.; Rosendahl, E.; Maxhall, M.; Bucht, G. (2006), Using a virtual reality system to study balance and walking in a virtual outdoor environment: a pilot study., *Cyberpsychol Behav*. Vol. 9, 388-395,

Rose, F.D.; Attree, E.A. & Brooks, B.M. (2000). Training in virtual environments: transfer to real world tasks and equivalence to real task training. *Ergonomics*, Vol. 43, 494-511.

Rosen, M. (1999). Telerehabilitation. *NeuroRehabilitation*, Vol. 12, 11-26.

Sisto, S.A.; Forrest, G.F. & Glendinning, D. (2002). Virtual reality applications for motor rehabilitation after stroke, *Top Stroke Rehabil*, No. 8, 11-23,

Yang, Y.R.; Tsai, M.P.; Chuang, T.Y.; Sung, W.H. & Wang, R.Y. (2008). Virtual reality-based training improves community ambulation in individuals with stroke: a randomized controlled trial, *Gait Posture*. Vol. 28, No.2, 201-206,

Winter, D. (2005). *Biomechanics and motor control of human movement*, 3rd, ed., John Wiley & Sons, Hoboken, New Jersey; 2005.

Wade, D. T.; Wood, V.A.; Heller, A.; Maggs, J.; Langton Hewer, R.(1987). Walking after stroke: measurement and recovery over the first 3 months. *Scand. J. Rehabil. Med.* Vol 19, 25-30

blaxxun Contact VRML plug-in - http://www.blaxxun.com, blaxxun interactive, Inc., The Virtual Worlds Company, 550 Bryant Street, Suite 770 San Francisco, California 94103 USA

Smart Home iRiS, Institute for rehabilitation, Republic of Slovenia, Available: http://www.ir-rs.si/en/smarthome

Virtual Reality ToolboxTM, The MathWorks, Inc., 3 Apple Hill Drive Natick, MA 01760-2098 USA

A prototype device to measure and supervise urine output of critical patients

A. Otero[1], B. Panigrahi[1], F. Palacios[2], T. Akinfiev[3], and R. Fernández[3]
[1]*Department of Information and Communications Systems Engineering. University San Pablo CEU. 28668 Madrid, SPAIN. Phone: +34 91 372 4040, fax: +34 91 372 4824*
e-mail: abraham.otero@usc.es
[2]*Critical Care Unit. Universitary Hospital of Getafe. Getafe, 28905 Madrid, SPAIN*
e-mail: palaciosfra@gva.es
[3]*Automatic Control Department, IAI/CSIC-Industrial Automation Institute, Spanish Council for Scientific Research, La Poveda 28500 Arganda del Rey, Madrid, SPAIN*
e-mail: teodor@iai.csic.es,roemi@iai.csic.es

1. Introduction

Patient monitoring history can be considered to have started in 1887, when the British Augustus D. Waller made the first electrocardiogram (ECG) recording on a human being (Waller, 1887). The first commercial monitoring device was invented by the Nobel Prize winner, Willem Einthoven, who in 1903 embarked on negotiations with the Cambridge Scientific Instruments Company to commercialise his "string galvanometer" for recording electrocardiograms (Einthoven, 1903). Since then, commercial monitoring devices capable of recording and supervising the status of many other physiological parameters have been developed: heart rate, respiratory rate, systolic, diastolic and mean blood pressures, blood levels of oxygen saturation, brain waves, intracranial pressure, partial pressure of expired oxygen, nitrogen and carbon dioxide -just to mention a few.

However, at present there is a very relevant physiological parameter that is still measured and supervised manually by critical care unit staff: urine output. This parameter is the best indicator of the state of the patient's kidneys. If a kidney is producing an adequate amount of urine it means that it is well perfused and oxygenated. Otherwise, it indicates that the patient is suffering from some pathology. When the urine output of a patient is too low the patient is said to have oliguria. If the patient does not produce urine at all, then he/she is said to have anuria. Sometimes, the urine output can be too high; in these cases the patient is said to have polyuria.

Urine output is also essential for calculating the patient's water balance; and is used in multiple therapy protocols to asses the reaction of the patient to the treatment. Two of the more prominent clinical algorithms where urine output plays a central role are the resuscitation of septic shock patients (Rivers, 2001), and the resuscitation and early management of burn patients (Mitra, 2006). In the latter patients, where endovenous resuscitation is required, a Foley catheter is placed very early in the treatment process in

order to monitor urine output. While the end points of the resuscitation are debatable, hourly urine output is a well-established parameter in the fluid management of these patients, as well as one of the most reliable assessments of the patient's state and evolution. In the case of the septic shock patient resuscitation, achieving a certain minimum value for the urine output itself is a therapeutic target.

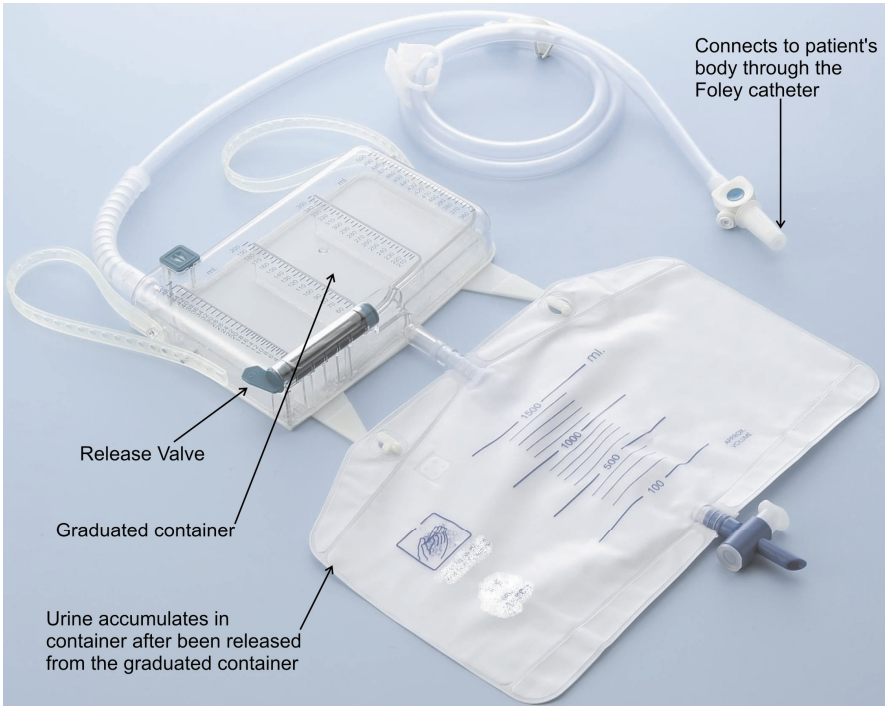


Fig. 1. Commercial urine meter

At present, critical patients' urine is collected in a graduated container (see Fig. 1). Every hour, the nursing staff manually records the reading of the container of every patient, and operates a valve which releases the urine into a larger container. In critical care units, this procedure must be performed 24 times a day, 365 days a year. As any repetitive and monotonous task, this one is prone to errors. A device capable of automatically measuring the patients' urine output, and supervising the attainment of the established therapeutic goals, would release the healthcare staff from a considerable amount of work, and would permit measurements to be carried out more frequently.

This paper presents a device capable of recording and supervising the urine output of critically ill patients. Section 2 describes this device. Section 2.1 presents a sensor specifically designed for this purpose upon which the device is based. Section 2.2 describes the operation of a microcontroller that takes the sensor readings and sends them via Bluetooth to a PC. The microcontroller also controls the operation of an actuator that releases the urine contained in the sensor into a larger container. The actuator is presented in section 2.3. Through a program installed in a PC, the healthcare staff can inspect the production of urine

during the patient's stay in the critical care unit, and set therapeutic goals for this parameter. These goals are automatically supervised; if they are not met, the PC triggers an auditory warning. Section 2.4 describes this software. The device has been successfully validated (using water) by operating it continuously for four consecutive days. The results of this validation are shown in Section 3. Section 4 discusses the results of this work, and Section 5 contains a series of conclusions and lines of future extension.

2. Device design

One of our design goals was to create a device that could be competitive in cost with the manual urine meters currently employed. Due to the higher complexity of our system, that will require sensors, actuators and microcontrollers, it will not be feasible to reach a price similar to the commercial urine meters -around five euros. However, it may be feasible to design a device whose disposable parts -all those which come into contact with the urine- have a low cost, although the price of non-disposable parts is higher. The latter parts can be amortized over a longer period of time; hence its economic impact on the overall price of the device can be negligible.
This goal, along with that of obtaining a robust and simple to operate device, were the main factors that guided the design process.

2.1 The sensor

In the design of the sensor used to measure the urine output a large number of constraints must be taken into account. On the one hand, any component of the device that is in contact, or may enter in contact, with the urine of the patient cannot be reused for different patients. Furthermore, the current commercial devices are usually changed approximately every week for hygienic reasons. Therefore, any component of the device that is or may be in contact with the urine must be easy to dispose of, and should have a low price. Being in contact with urine also means that the component is in indirect contact with the patient's body through the Foley catheter -the flexible tube that is passed through the urethra into the bladder during urinary catheterization to drain urine. Therefore, the component has to be sterilized before it is used. Sterilization of sophisticated sensors that have a lot of parts, possibly encapsulated in some type of covering, can be complicated.
The urine usually is acidic, although sometimes it can be slightly basic –its pH varies between 4.5 and 8. It contains Uric acid (between 25-75 mg/l), Urea (between 15-34g/24h), Sodium (between 130-260 mEq/24h), Potassium (less than 90 mEq/24h), Chlorine (between 110-250 mEq/24h), and Copper (less than 30 mcg/24h), among other components (Brunzel, 2004). Thus, it can be quite corrosive, especially for metals.
Finally, the device must allow the system to provide feedback on the fulfillment of the therapeutic goals for the urine output at least every hour, in order to be able to provide similar information to that produced manually by the nursing staff. In some cases, these therapeutic goals may be as low as 0.5 milliliters of urine per hour per kilogram of the patient's weight. For a 60 kg patient, this means that the sensor must be capable of providing reliable measurements of at least 30 ml of urine, although it would be desirable to be able to measure smaller quantities.

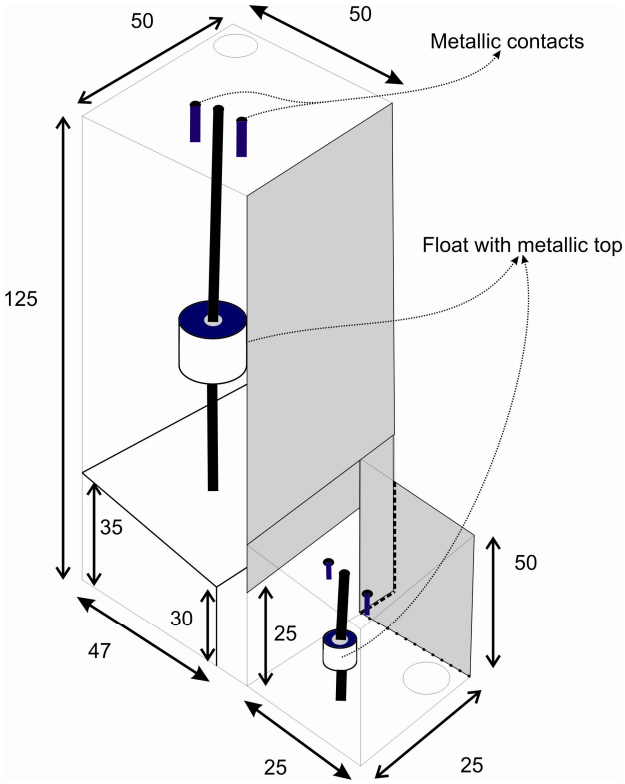


Fig. 2. Three-dimensional view of the sensor. All sizes are shown in millimeters.

Many options were considered in the construction of the sensor. The use of ultrasound sensors was studied, a solution that has been proposed before to measure the height of a column of liquid (USP 3693445, 1972). However, this type of sensor is not practical to measure volumes as small as those under consideration here. Its low accuracy -approximately 1 cm- would require a fairly narrow container to counteract the inaccuracies of the sensor measurements - a circular container could have a maximum radius of 2.5 cm. to permit the identification of oliguria or anuria for a 60 kg patient in not more than one hour. Such a narrow container would cause the ultrasound beam -which usually has an opening angle of 23 degrees (Everett, 1995)- to bounce several times against the container walls before reaching the liquid. The echoes of the collisions can trigger a sensor reading before the ultrasound beam arrives at the liquid.

A laser would provide more accuracy in the readings, but it would considerably increase the final cost of the device, making it unable to compete with the current manual measuring devices. Commercial flowmeters also were considered, but these devices are designed to measure a continuous flow of liquid, while urine output is usually a small erratic drip. Furthermore, the sensor would have to be disposed of because it would be in direct contact with the urine. This would make it cost prohibitive.

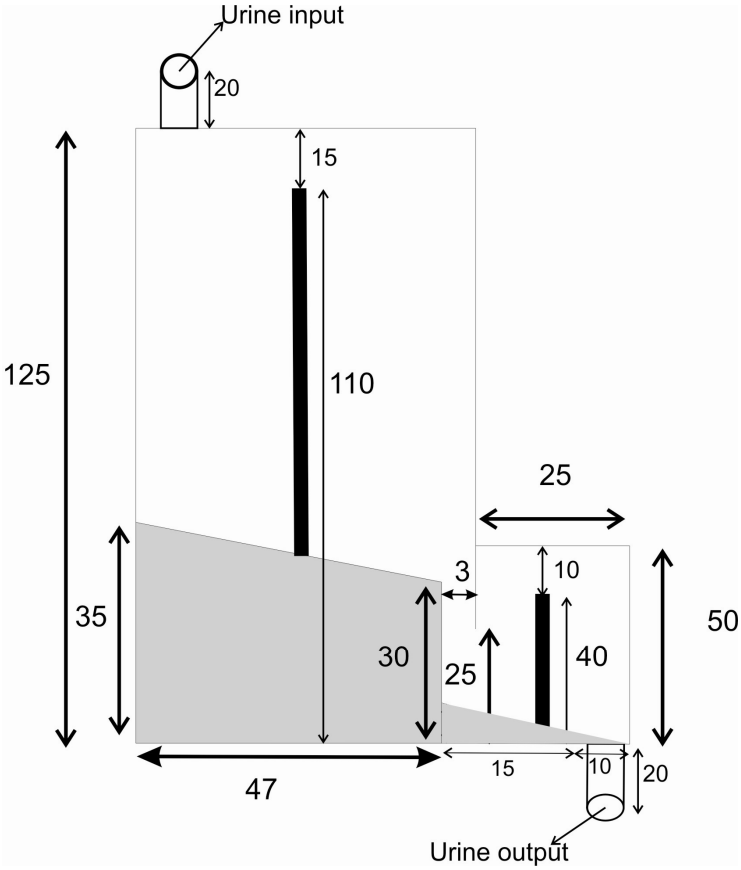


Fig. 3. Lateral view of the sensor. All sizes are shown in millimeters. Floats and the metal contacts have been excluded from the drawing. Gray areas correspond to solid regions.

We finally opted for building a sensor specifically designed for this problem. The sensor is based on a float which moves vertically along a pole and has a metallic surface on top. Urine makes the float rise along the pole until the metallic surface reaches two metal contacts placed at the top of the container, indicating that the container is full (see Fig. 2). This simple design has a flaw: the container must have a volume small enough to monitor the urine output of a patient who may have oliguria. This may be as small as 30 ml per hour. Thus, the container must have a maximum volume of 30 ml to be able to identify the oliguria state in less than an hour for an adult patient.

However, if the patient were producing normal amounts of urine or, even worse, if he/she has polyuria -these patients may produce up to 1 litre of urine per hour- a container of such a low volume would require its content to be released up to several hundred times per day. Our intention is to incorporate an actuator that automatically releases the urine to a container of larger volume, similar to the way the nursing staff manually releases the urine into a larger container every hour. The fact of carrying out several hundred valve operations per day can cause a considerable stress both in the valve and on the actuator. The result

would be an increase in price of an actuator and valve capable of supporting several thousand openings during its lifetime without breaking the sterility of all tubes and containers that come into contact with urine and, therefore, are in indirect contact with the patient's body.

To resolve this problem, we designed a sensor with two different volume containers, each one of them equipped with a float similar to that we described above (see Figs. 2 and 3). The smaller container has a volume of 15 ml, permitting an oliguria warning to be triggered for a 60 kg patient in half an hour. This is half of the time that would be required to detect this problem from the measures taken manually by the nursing staff. For a patient of average weight, who does not have oliguria or anuria, the container should be filled in less than 20 minutes.

The purpose of the smaller container is to provide an early warning of low urine output. If the patient has oliguria, a precise and continuous monitoring of the urine output is needed. Thus, if the small container does not fill in the expected time, when it gets full it will be emptied, and we shall measure again the time required to fill it. Since the patient is producing a very small amount of urine, the small container will be filled between 10 to 30 times a day; therefore, the actuator will not be subjected to significant stress.

The sum of the volumes of both containers is approximately 180 milliliters. If the patient produces normal amounts of urine, or if he/she has polyuria, when the small container gets full, its content will not be released, and the larger container will start to fill. Therefore, for a patient who produces 5 liters per day, the actuator will be triggered about 25 times a day.

In order to operate correctly, this sensor must be built in such a way that the urine does not begin to accumulate in the large container until it has completely filled the small one. This was achieved by making the bottom of the large container solid (see the grey area of Fig. 3) and inclined towards the small container. The bottom of the small container is also inclined towards an exit tube placed on the right part of the bottom of the small container (see Fig. 3), to facilitate the emptying of both containers.

The fact that the small container is at a lower level than the big one can cause the metallic contacts of the small container, and the top metal surface of the container's float, to be under the urine when the large container begins to fill. As we have already indicated, the urine can be quite corrosive for metals. The first sensor design showed that, after 48-72 hours of immersion in a saline solution with similar properties to urine, the metal oxide accumulated on the metal parts -which were made of copper and aluminum- caused the sensor to malfunction, since it was not always able to detect the filling of the small container.

To address this problem, the small container was designed in such way that when it gets full the top has an empty space 15 mm in height. This space is enough to prevent both the metal contacts and the top of the float from coming into contact with the urine.

The sensor described here was designed using the software Pro/Engineer Wildfire, and was built using a rapid prototyping printer -Prodigy Plus- manufactured by Stratasys (Stratasys, 2009). This printer builds components through the deposition of successive layers of molten material and is accurate to a few tenths of a millimeter in the measures of the components it builds. The entire structure of the sensor was built with this printer; only the floats and the metal contacts had to be added to complete it.

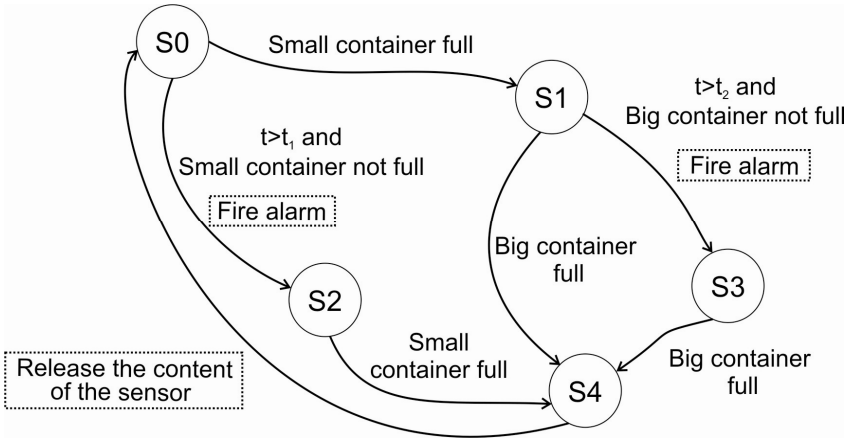


Fig. 4. Finite state machine describing the operation of the micro controller. The conditions of the transitions are shown on each arc, and the actions taking place in each transition are indicated by text surrounded by a dotted line.

2.2 Actuator design

One of our goals is to provide more frequent measurements of urine output. This has led us to design a container-sensor with low volume -15 ml for the small container, and 180 for the overall sensor. The container currently used to take manual measurements usually has volume of 500 ml. The use of a smaller volume container makes it necessary to release its content more frequently.

Another of our goals is to alleviate the workload of healthcare staff by automating as much as possible all the tasks related to monitoring and supervising the urine output. Thus, we must provide an automatic mechanism to release the content of the sensor to a higher volume container, avoiding the need of healthcare staff to manually perform this operation several times each hour. Furthermore, the release of the content of the sensor must be performed exactly in the moment the container gets full, because once the container is filled the urine produced will not be registered until it is emptied. Therefore, the delays that might occur if the nursing staff had to manually release the urine -for example, because at the releasing time some emergency that requires the nursing staff attention has occurred in the critical care unit- would introduce errors in the measurements.

Several options were considered for an actuator for the urine release: a shaded pole motor, a stepper motor moving a screw, magnetism, etc. Finally, we opted for using linear solenoids because of their low cost, and because they are easy to control using a microcontroller. One solenoid is used for opening the valve, and a second one for closing it. However, adapting the design of the device proposed here to use any other actuator capable of opening and closing a valve to release the urine is trivial.

2.3 The microcontroller

The microcontroller we used in our prototype was an Atmel AT89S52, a 50 cents-a-unit, low-power, high-performance CMOS 8-bit microcontroller with 8K bytes of in-system

programmable Flash memory, 256 bytes of RAM, 32 I/O lines, two data pointers, three 16-bit timers, a full duplex serial port, on-chip oscillator, and clock circuitry.
As soon as the device is turned on, the microcontroller starts measuring the elapsed time. One of the three 16-bit timers is used for this purpose. The timer gets incremented once every instruction cycle. Since there are only two bytes devoted to the value of this timer, the maximum value it may have is 65,535. One instruction cycle in the AT89S52 has a duration of 1.085 microseconds. Three additional 8 bits registers are used to store the number of times that the timer overflows. The maximum time that can be measured is ((65535 x 1.085 /10E6)x255x255x255)/3600 hours, which is equal to 327.5 hours, close to two weeks. Thus, as long as the patient produces at least 15 ml of urine every two weeks, this timer will be enough.

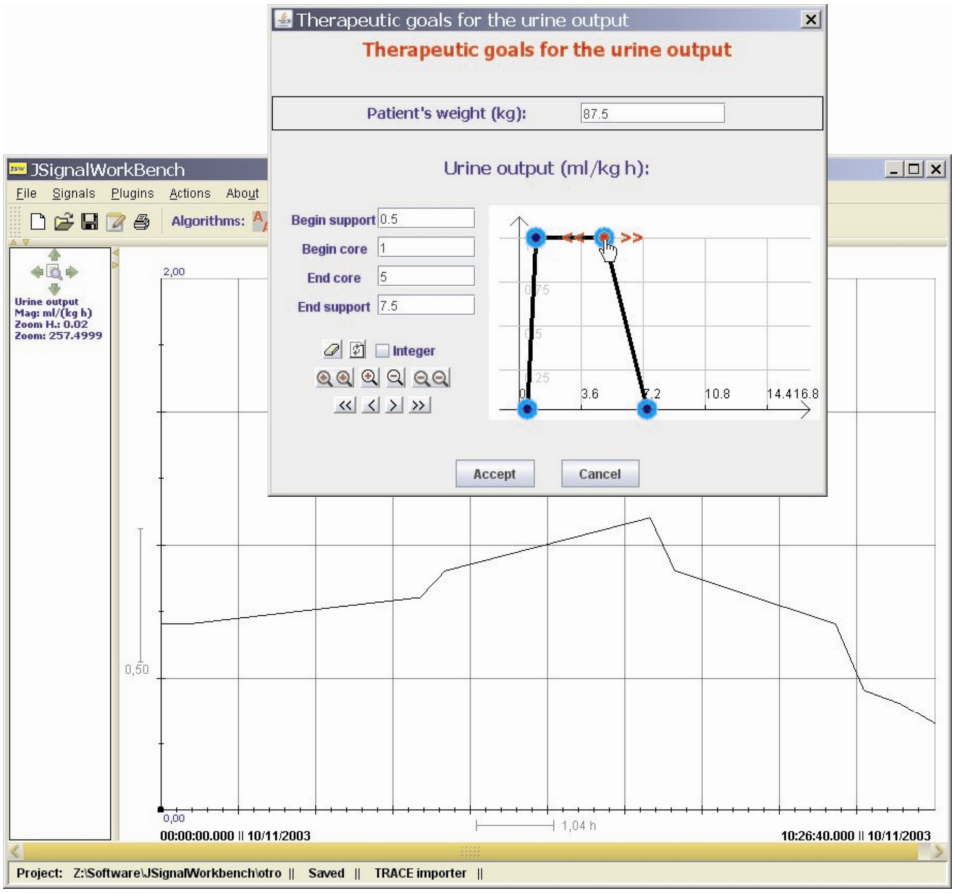


Fig. 5. Main screen of the Java application and the window that allows healthcare staff to set the therapeutic goals for urine output.

The physician sets the therapeutic goals for urine output on a computer using a program built for this purpose. From the therapeutic goals and the weight of the patient, this

program calculates the maximum time that should be required for filling the small container, if urine output is within the therapeutic goals, and the maximum time required for the large container. We shall call these times t1 and t2, respectively. These times are sent to the microcontroller using the serial port. When the microcontroller receives this information, it replies sending back the time at which the monitoring started. This permits the synchronizatiom of the microcontroller and the program installed on the PC.
Then, the microcontroller waits for the signal that indicates that the small container is full. If this signal arrives before t1 seconds, therapeutic goals are being met and it is not necessary a very precise monitoring of urine output. Therefore, the valve shall not be opened, and the larger container starts to fill. If after t1 seconds the signal indicating that the small container is full has not arrived, the microcontroller turns on an LED indicating that the therapeutic goals are not being met. In this case, a detailed monitoring of urine output is required. Thus, when the small container is full the actuator will be activated and the timer will be resested in order to measure, again, the filling time of the small container. If the small container has been filled within the expected time, but the large container is not filled before t2 seconds, the alarm LED is also turn on. In either case, when the big container is full, the actuator will be activated.
Figure 4 shows a finite state machine detailing the operation of the microcontroller. The conditions that trigger the transitions are displayed as text on the arches that represent the possible transitions. t represents the time elapsed since the device was turn on, or since the last time the content of the sensor was released. Some transitions involve performing some action, such as a firing an alarm or open a valve. These actions are displayed as text drawn over the arches of the transitions and surrounded by a dotted line. All state transitions are communicated to the program installed in the PC via the serial port. This allows the PC program to calculate urine production at any given moment, and to find out whether the patient has anuria or oliguria.
In Figure 4, S0 represents the initial state; S1 corresponds to the filling of the small container before t1 seconds; S2 to the filling of the large container before t2 seconds, S3 to the non-filling of the small container before t1 seconds; S4 to the non-filling of the large container before t2 seconds; and S5 corresponds to the activation of the solenoid and the release of the urine. When the micro controller reaches the S5 state it immediately transitions to S0; S5 is a fictitious state which has been added to the diagram for sake of clarity.

2.4 Client software

We have developed a Java application that receives the readings of the urine output from the microcontroller through the RS232 port. To avoid the need for wiring, a serial-port-to-Bluetooth adapter has been used both in the PC and in the micro controller. The program allows the healthcare staff to inspect a graph showing the patient's urine output in milliliters per hour (see Figure 5). It also has a screen showing a set of statistics, like the hourly and daily urine output throughout the stay of the patient in the critical care unit.
Using this program, the healthcare staff can set therapeutic goals for the urine output. These therapeutic goals are represented with the aid of the Fuzzy Set Theory, a tool which has proved its value for handling and representing experience-based heuristic knowledge, such as that commonly used in the medical domain (Barro & Marin, 2002).
We shall introduce some basic concepts of Fuzzy Set Theory on which our solution is based. Given as discourse universe $\Re$, we define the concept of fuzzy value C by means of a

possibility distribution π_C defined over $\Re$. Given a precise value $\upsilon \in \Re, \pi_C(\upsilon) \in [0,1]$ represents the possibility of C being precisely υ. A fuzzy number is a normal and convex fuzzy value. A fuzzy value C is normal if and only if $\exists \upsilon \in \Re, \pi_C(\upsilon) = 1$. C is said to be convex if and only if $\forall \upsilon, \upsilon', \upsilon'' \in \Re, \upsilon' \in [\upsilon, \upsilon''], \pi_C(\upsilon') \geq \min\{\pi_C(\upsilon), \pi_C(\upsilon'')\}$.

We obtain a fuzzy number C from a flexible constraint given by a possibility distribution π_C, which defines a mapping from $\Re$ to the real interval [0,1]. A fuzzy constraint can be induced by an item of information such as "x has a low value", where "low value" will be represented by π_C. Given a precise number $\upsilon \in \Re$, $\pi_{C=\text{low}}(\upsilon) \in [0,1]$ represents the possibility of C being precisely υ; i.e., the degree with which υ fulfills the constraint induced by "low value".

Normality and convexity properties are satisfied by representing π_C, for example, by means of a trapezoidal representation. In this way, $B = (\alpha, \beta, \gamma, \delta), \alpha \leq \beta \leq \gamma \leq \delta$, where $[\beta, \gamma]$ represents the core, $core(\upsilon) = \{\upsilon \in \Re, \pi_C(\upsilon) = 1\}$, and $]\beta, \delta[$ represents the support, $supp(\upsilon) = \{\upsilon \in \Re, \pi_C(\upsilon) > 0\}$ (see Figure 5).

We shall represent the therapeutic goals for urine output by a trapezoidal possibility distribution. The minimum and maximum values aceptable for the urine output are the beginning and end of the support of the distribution, respectively. If urine output is lower (or higher) than this value the program produces an audible warning until it is turn off by the healthcare staff. The purpose of this warning is to identify oliguria and polyuria states. The beginning and end of the core of the trapezoidal possibility distribution are the limits of the interval within which ideal values of the urine output lay in.

The semantics of this possibility distribution is "adequate urine output"; therefore the degree of membership of a value of urine output to the possibility distribution indicates the adequacy of the patient's urine output regarding to the therapeutic goals. If the degree of membership is zero, either the urine output is less than the minimum acceptable value -the patient has oliguria or anuria-, or greater than the maximum acceptable -the patient has polyuria. If the degree is 1, the urine output is within the range of ideal values. The closer the degree is to 1, the closer the patient's urine output to the ideal value it is, and the closer the value to zero it is, the closer the patient is to oliguria or polyuria.

This information is represented by a color code that is used when drawing the graphs of urine output of the patient: a membership of 0 is associated with the color red; a value which lies in $]0,0.2]$ is associated with purple; $]0.2,0.4]$ with pink; $]0.4,0.6]$ with orange; $]0.6,0.8]$ with yellow; $]0.8,1[$ with blue; and green with the total membership. In this way the graph of diuresis provides an instantaneous visual feedback on the patient state (Otero *et al.*; 2008).

Figure 5 shows the screen that permits the definition of the therapeutic goals. The trapezoidal possibility distribution that represents "adequate urine output" is obtained as the fuzzy product of the weight of the patient by the possibility distribution representing the number of $ml/(kg \bullet h)$ of urine output. This fuzzy product is defined as

$W \otimes (\alpha, \beta, \gamma, \delta) \coloneqq (W \bullet \alpha, W \bullet \beta, W \bullet \gamma, W \bullet \delta)$ (Kaufmann & Gupta, 1984), where W is the patient's weight and $(\alpha, \beta, \gamma, \delta)$ represents therapeutic goals for the urine output.

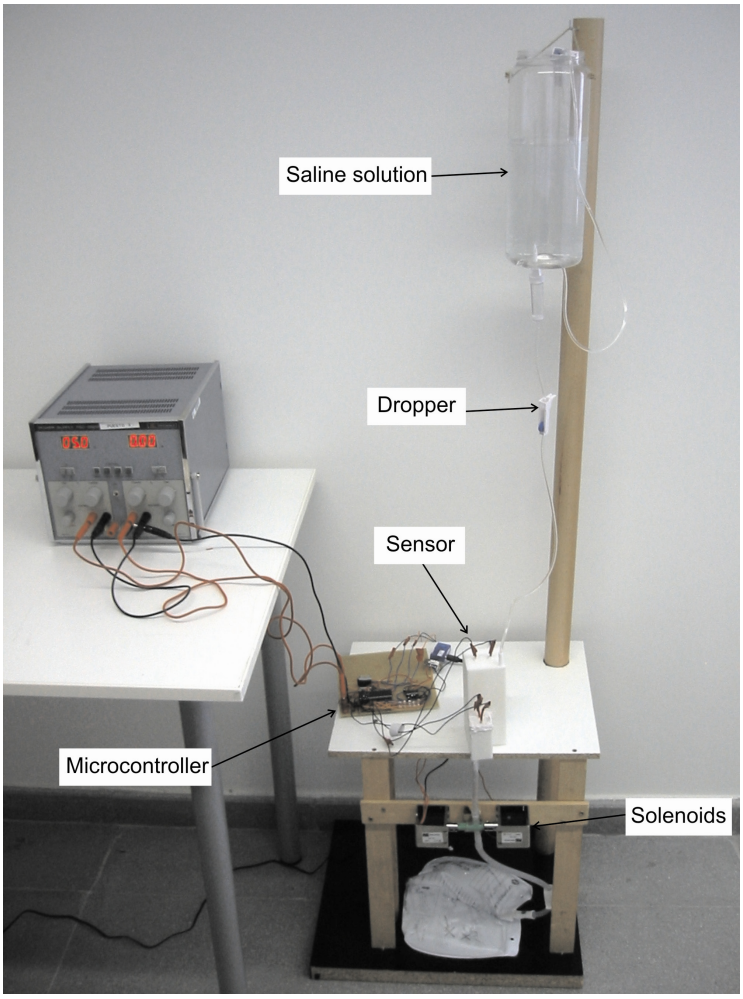


Fig. 6. Picture of the urine meter prototype.

3. Device validation

Once the device was built, a series of tests to verify its proper operation were performed. In the test a saline solution with similar properties to urine was used. This liquid was stored in a container placed at a higher level than the sensor, and a dropper was used to regulate the flow of fluid from this container to the sensor (see Figure 6).

The first tests failed for various minor mechanical problems. For example, the valve that releases the urine did not closed properly after the first opening, which produced a small dripping. Once this mechanical glitches were solved, a number of tests which forced the device to operate in all the states shown in the finite state machine of Figure 4 were performed. Several tests in which the amount of fluid sent to the sensor was carefully measured were also conducted, and then we verified that the production of urine provided by the software installed on the PC was accurate.
Finally, a stress test in which the system worked continuously for four days was conducted. After the four days, we check the state of all parts of the device, mainly the opening and closing valve –it still was able to completely seal the exit of liquid when it was closed; and the state of the floats and the metal contacts of the sensor –they showed mild corrosion that did not affect their operation. This suggests that the design of the sensor did effectively prevent the metal parts of the small container from being submerged in liquid.

4. Discussion

The tests we have performed show that our prototype device permits the automatic measuring of urine output with reliabily and accuracy over a period of several days, and the generation of alarms when urine output deviates from the established therapeutic goals. To the best of our knowledge, no commercial or research device with these characteristics has been built until now.
Currently, in the critical care units, patients' urine output is registered manually by the nursing staff. Every hour a nurse must record the measurement of urine output from each of the patients in the unit, and must operate the valve that releases the urine from a graduated container to a larger container. This has to be repeated 24 times a day, 365 times a year. Every one or two days is also necessary to change the container where all the urine accumulates. We estimate that the amount of time required for the tasks associated with the manual monitoring of urine for every 10 patients admitted in a critical care unit is equivalent to the annual full-time dedication of a nurse. Therefore, total or partial automation of the tasks related to this activity can lead to considerable economic savings for the institutions that provide the care services.
The total cost of our prototype device was about 300 €. The most expensive parts were the serial-port-to-Bluetooth adapter, and the solenoids. We estimate that the cost of the disposable parts, if produced industrially, is less than 10 €; thus, the final cost of the device is comparable to the cost of the devices that are used to manually measure urine output. However, the workload that the device can potentially save, and therefore the savings in the cost of the health care service, is considerable.
On the other hand, all monotonous and repetitive tasks, as the one under consideration here, are prone to error. Therein lays the importance of automating them as much as possible. To provide a computer system to record and supervise continuously the urine output of all patients admitted to a critical care unit, and alert the healthcare staff on any deviation from the therapeutic goals, can potentially prevent human errors in this task.
Moreover, the monitoring interval currently employed in the urine output -one hour- tries to reach a compromise between avoiding risk states for the patient and not placing an excessive burden on the nursing staff. An automatic system, as the one described here, permits more frequent measurements to be carried out; thus it enables the identification of

deviations from normality at earlier stages. For example, the device described in this chapter, for a patient of about 80 kg, can generate a warning indicating oliguria in approximately 15 minutes.

5. Conclusion

We have built a device capable of automatic monitoring and supervising the urine output of critical care unit patients. The device is based on a sensor specially designed for this task. The sensor identifies the filling of two containers, one of 15 ml of capacity and another of 180 ml. A microcontroller processes the sensor output and, based on the filling time of the containers, it supervises if the therapeutic goals are being achieved. If the small container is filled on time, the patient is producing an acceptable amount of urine and it is not necessary a very precise monitoring of urine output; thus the microcontroller allows the larger volume container to start filling. If not, when the smaller container gets full, the microcontroller activates an actuator that releases its content. This enables a more accurate monitoring of urine output to be carried out, while limiting the stress suffered by the device's actuators. Whenever the big container is full, the actuator releases the content of both containers.
Therapeutic goals are established for each patient by the healthcare staff using a PC program. The program allows the healthcare staff to indicate the weight of the patient, the minimum and maximum acceptable values, and the ideal values for the urine output. The program also displays the patient urine output throughout his/her stay in the critical care unit, and alerts the healthcare staff of any deviation that occurs with respect to the acceptable values for the urine output. Using a colour code, it also indicates the deviation of the patient state from the ideal state.
The cost of non-disposable parts of the device is similar to the cost of current manual urinemeters. However, this device can offset a considerable amount of workload from the nursing staff, translating in savings in the cost of the healthcare service rendered to the patient. This device can also help avoiding typical errors of monotonous and repetitive tasks, such as the one we have at hand, by automating the monitoring of therapeutic goals for the production of urine and warning when a deviation occurs.
Our future work is oriented towards the construction of a device similar to the one presented in this chapter, but sterilized, so it can be used in a pilot tests in the intensive care unit of the University Hospital of Getafe. A device of this type may also be the basis for carrying out a series of clinical studies based on a more continuous and accurate monitoring of urine output throughout the stay of a patient in the critical care unit.

6. Acknowledgements

We want to thank Javier F. Sarria Paz and Manuel A. Armada for their support in building the containers of the sensor used in our device. We also would like to acknowledge the aid for the work presented in this paper of the Spanish MEC and the European FEDER under the grant TIN2006-15460-C04-02, and of the University San Pablo CEU under the grant TIN2006-15460-C04-02 and USP 04/07. T. Akinfiev acknowledges the financial support received from CSIC under the project "New actuators with high efficiency and control algorithms for automation and robotics". R. Fernández acknowledges the financial support received from Ministry of Science and Innovation under JAEDoc Programme.

7. References

Barro, S. & Marin, R. (2002). A call for a stronger role for fuzzy logic in medicine. *Fuzzy Logic in Medicine* (pp. 1-17). Springer-Verlag, 3-7908-1429-6.

Einthoven, W. (1903). Die galvanometrische registrirung des menschlichen elektrokardiogramms, zugleich eine beurteilung der anwendung des capillarelektrometers in der physiologie. *Archiv für die gesammte Physiologie des Menschen und der Thiere*, 99, 472-480.

Everett, H. R. (1995). Sensors for Mobile Robots. Theory and Application. *AK Peters*, 1-56881-048-2.

Kaufmann, A. & Gupta, M.M. (1984) Introduction to Fuzzy Arithmetic. *Van Nostrand Reinhold Company Inc.*, 978-0442008994.

Liquid Level Measurement Device (1972). United States Patent 3693445.

Otero A. & Félix, P. & Zamarrón C. (2008). Visual knowledge-based metaphors to support the analysis of polysomnographic recordings. *Advances in Soft Computing*, 49, 10–20, 1867-5662.

Mitra, B. & Fitzgerald, M. & Cameron, P. & Cleland, H. (2006). Fluid resuscitation in major burns. ANZ Journal of Surgery. 76, 35-38.

Nancy, B. (2004). Fundamentals of Urine & Body Fluid Analysis. Saunders, 978-0-7216-0178-6.

Rivers, E. & Nguyen, B. & Havstad, S. & Ressler, J. & Muzzin, A. & Knoblich, B. & Peterson, E. & Tomlanovich, M. (2001). Early Goal-Directed Therapy in the Treatment of Severe Sepsis and Septic Shock. *New England Journal of Medicine*. 345, 1368-1377.

Stratasys (2009). Rapid prototyping printer, http://www.stratasys.com/.

Waller, A. (1887). A demonstration on man of electromotive changes accompanying the heart's beat. *Journal of Physiology*, 8 (5), 229-234.

Zadeh, L.A. (1975). The concept of a linguistic variable and its application to approximate reasoning. *Information Science*, 8, 199–249, 1741-6485.

18

Wideband Technology for Medical Detection and Monitoring

Mehmet Rasit Yuce, Tharaka N. Dissanayake and Ho Chee Keong
The University of Newcastle
Australia

1. Introduction

Biomedical devices benefit from the recent and rapid growth of wireless technology for measuring physiological signals for both implantable and wearable systems. The use of wireless in healthcare systems provides great mobility and increases comfort level of patients by freeing them from hospital equipments. However, it is important to select a proper and safe wireless band for medical data transmission as it is crucial for the patient 's safety. Historically narrow band wireless technologies have been used in medical monitoring widely. This chapter introduces the use of wideband technology for future medical monitoring systems. The important parameters of the wideband technology are its low power transmitter design, low-interference effect in medical environment and high data rate capability. The chapter describes low power implementation of ultra-wideband (UWB) and applies the technology to some of emerging biomedical applications.

Wideband technology can find applications in biomedical monitoring especially for neural recording and multi-channel continuous signal monitoring such as Electroencephalography (EEG), Electrocardiogram (ECG) and Electromyography (EMG). The potential benefits are the low-power transmitter to increase the battery life, high data rate to increase the resolution and performance, and less interference effect on the other wireless system in medical centers. Wideband technology is particularly suitable for telemetry systems requiring high data rate transmission such as wireless endoscope and multi-channel continuous biological signal monitoring.

The design of a UWB wireless chip has been difficult for chip designers due to the difficulty in the demodulation of narrow pulses. Generally a UWB receiver circuit has demonstrated a power consumption higher than that of narrow band systems. One way to eliminate the high power consumption of an ultra wideband receiver is to use a transmitter only method for medical monitoring and detection. This chapter will address and discuss implementation of such a system. The system includes only a transmitter to send data from body to an external device for monitoring and recording. The transmitter in the wideband telemetry system is attached to or implanted in the body. Meanwhile an off-the-shelf receiver system is placed at a location between 0.5-10 meters away to detect the transmitted signal.

Methods and design techniques to use ultra wide band (UWB) technology for wearable and implantable physiological monitoring systems will be presented. The chapter covers two main applications of wideband technology: In the first part of the chapter, a band limited impulse radio and antennas based on the UWB system will be discussed to investigate the implementation of wideband signals for implantable medical devices. An example of a wireless endoscope device using an impulse based wideband radio system will be described in detail. In the second part of the chapter, a medical monitoring for wearable systems using wideband technology will be explained with implementation details.

2. Wideband Technology for Medical Telemetry

Electrocardiogram (ECG) and temperature recording have been used for more than 50 years in medical diagnosis to understand the biological activities (Mackay, 1961; Lefcourt, 1986). Electronics for biological signals (i.e. Bioelectronics) is designed taking into account very specific requirements for a a given wireless telemetry application. Wireless technologies developed in commercial domain cannot be used directly in medical implants or wearable systems because of two reasons: (1) they have been optimized for general use and are thus complex; (2) the device size exceeds the required size limitation of the current implant technology. Furthermore on-body medical monitoring devices also require miniaturization and emission constrains so that they can be wearable.

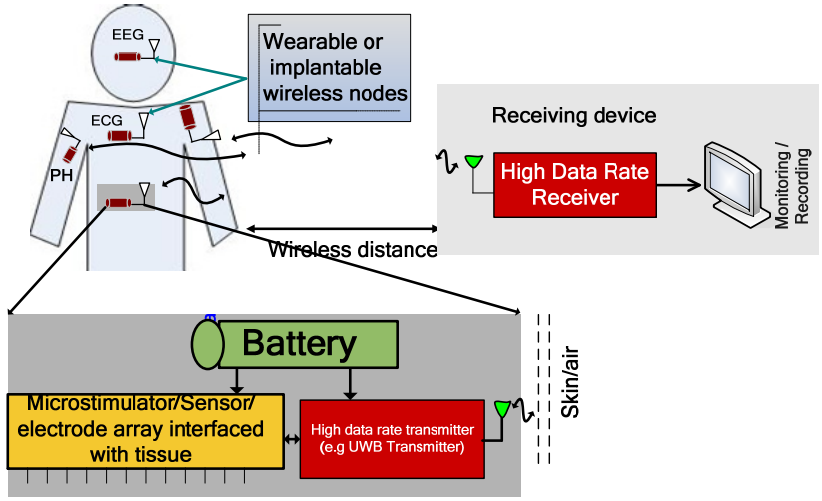


Fig. 1. A modern medical monitoring and detection system.

Fig. 1 shows a medical monitoring system that will most likely be implemented widely in medical centers in the near future. A monitoring/detection device in a telemetry system consists of sensors/electrodes that detect biological signals from the body, a battery and a transmitter to transmit signals from inside or on-body to a remote receiving device for monitoring and diagnosis. The term monitoring is used in the medical telemetry for two different functionalities: monitoring of the body signals which are used for medical diagnosis and monitoring the status of the prosthetics devices such as a pacemaker, or a

cochlear implant. A monitoring device is required almost in all telemetry systems in order to send the medical status to a remote system for better treatments.

Most of implantable systems such as retina prostheses and cochlear implants have a very short transmission distance for wireless telemetry. An inductive link is constructed between the external and the implanted devices with a distance of few centimeters. In addition, implantable devices dedicate a very small size to the wireless telemetry section of an implant. Consequently a wireless system with a very simple communication scheme such as binary ASK (amplitude shift keying), FSK (frequency shift keying) and PSK (phase shift keying) modulators and demodulators can be employed (Liu et al, 2005). The transmission frequency used is usually lower than 20 MHz. The frequency 13.56 MHz in ISM (Industry, Scientific and Medical) band is usually the most common for such inductive link telemetry systems and is also used for RFID (Radio Frequency Identification) applications. A low transmission frequency usually less than 20 MHz is utilized mainly because of the simplicity in the design and to avoid the use of power hungry blocks such as mixer, oscillators. When many of such implantable systems are used in the same environment or for the same patient, such simple communication systems will face the problems of interference. Thus a more advanced wireless technology will be required in the future to accommodate better radio links for medical implants.
For some of advanced medical implants such as pacemaker, implantable cardioverter defibrillators and electronic pill (i.e. wireless endoscope), a much longer range is required for the wireless telemetry. Moving to higher frequencies is the only way to increase the range and to dedicate enough spectrums for a reliable communication in the future. However, at higher frequencies a wireless transceiver requires the use of RF blocks such as voltage-controlled oscillators (VCOs), mixer and phase-locked loops (PLLs) to down convert (or up convert for the transmitter case) the frequencies in order to process using integrated circuit technology. These blocks are constructed using inductors and capacitors on chip or off-chip which increases the physical size of the wireless chip.

Medical implants have physical limits for the electronics of the wireless telemetry and cannot afford to accommodate such blocks. Thus, in order to alleviate some of the issues mentioned above, US Federal Communication Commission (FCC) and some of international authorities have allocated a new band at 402-405 MHz with 300 KHz channels to enable the wireless communication of such implantable devices to deliver high level of comfort, mobility and better patient care (Tekin, 2008; Bradley, 2006). With the advance of radio frequency IC (RFIC) technology, this frequency band promises high-level of integration (compared to inductive link designs) which results in miniaturization and low power consumption.
Medical telemetry can be categorized into two groups: high data rate and low data rate systems. Multi channel recording (i.e. neural recording systems) for implantable system and multi-channel continuous signals such as EMG, ECG and EEG necessitate a high data rate communication. As an example, scientists aim to achieve the recording of more than 100 channels in order to simultaneously record brain functions; a data rate more than 20 Mbps is required (Yuce et al, 2007, Chae et al, 2008). A similar figure is also useful for a wireless endoscope implant to obtain higher resolution pictures and images. The MICS has channels with 300 KHz width and thus cannot provide such data rates.

	Frequency band	Bandwidth (or data rate)	Trans. Power
WLANs (802.11b/g)	2.4 GHz	>11 Mbps	250 mW
IEEE 802. 15.1 (Bluetooth)	2.4 GHz	1 MHz, 1 Mbps	4 dBm,20 dBm
IEEE 802.15.4 (ZigBee)	2.4 GHz	250 kbps	0 dBm
WMTS	608-614, 1395-1400, 1429-1432 MHz	6 MHz	≥10 dBm and< 1.8dB
MICS	402-405 MHz	3 MHz	- 16 dBm
UWB	0-960 MHz 3.1 -10.6 GHz	800 kbps, 27.24 Mbps	-41dBm
Inductive link	<20 MHz, 13.56 MHz ISM	Usually around kbps	> 0dB

Table 1. Existing wireless systems.

Although the advances in RFIC technology for wireless communication technologies have been significant in the commercial domain, these technologies are not directly transferable to medical applications due to the differing power, size, and safety related radiation requirements of medical devices. Most popular wireless technologies are shown in Table 1. These wireless systems target a wide range of applications. The existing advanced wireless systems such as ZigBee (IEEE 802.15.4), WLANs, and Bluetooth (IEEE 802.15.1) operate at 2.4 GHz ISM band and may suffer from the strong interference from each other when they are located in the same environment (Shin et al., 2007). Some of the commonly used wireless platforms (either wireless chip or complete board) from Crossbow[1], Texas Instrument[2] and Zarlink [3] are shown in Table 2. This table summarizes the properties of wireless boards or chips used in most of Today's low-power applications.

As can been seen, two chips that meet the requirement of a medical device the most are zarlink's new MICS chip and Crossbow's Mica2DOT device. Although the physical dimension of the Zarlink's device is for the chip package area, a small board can easily be designed, similar to that of Mica2DoT shown in Fig. 2 as the chip requires only few external components. Zarlink provides one of the lowest-power wireless chips available today. The low-power achieved by reducing the supply voltage to value as low as 1.2 V while most of the others operate with a voltage 1.8V and higher. When comparing the available technologies with current state of the art UWB technologies, the unique properties of UWB that outperforms others are high data rate and its transmitter power and physical size which can be extremely small. Based on the key requirements from medical monitoring, UWB could be the best choice in term of power consumption, scalability and size when a transmitter only approach is followed (Yuce et al., 2007).

[1] http://www.xbow.com.
[2] http://www.ti.com/.
[3] http://www.zarlink.com/

Model	Company	Frequency	Data Rate	RF Power	Physical Dimension	Current	
						Tx	Rx
UWB	Ref1 ** Ref2***	3.1 – 10.6 GHz	20 Mbps	-41 dBm	Very small	1.6mW 2mW	16 mA*
Mica2 (MPR400)	Crossbow[1]	868/916 MHz	38.4 kbps	-24 -+5 dBm	58 x 32 x 7 18 grams (board)	27 mA	10 mA
MicAz	Crossbow[1]	2.4 GHz	250 kbps	-24 -0 dBm	58 x 32 x 7 18grams (board)	17.4 mA	19.7 mA
Mica2DOT	Crossbow[1]	433 MHz	38.4 kbps	-20-+10 dBm	25X6 mm^2 3 gram (board)	25 mA	8 mA
CC1010	TI[2]	300 to 1000 MHz	76.8 kbps	-20-+10 dBm	12X12 mm^2 (chip)	26.6 mA	11.9 mA
CC2400	TI[2]	2.4 GHz	1 Mbps	-25-0 dBm	7.1X7.1 mm^2 (chip)	19 mA	23 mA
MICS	Zarlink[3] (ZL70250)	402-405 MHz, 433 MHz ISM	800 kbps	< 0 dBm	7X7 mm^2 (chip)	5 mA, continuous TX / RX	

Table 2. Comparison of hardware designs for wireless systems
*Reference (Ryckaert et al., 2007), **(Chae et al., 2008), ***(Ryckaert et al., 2005).

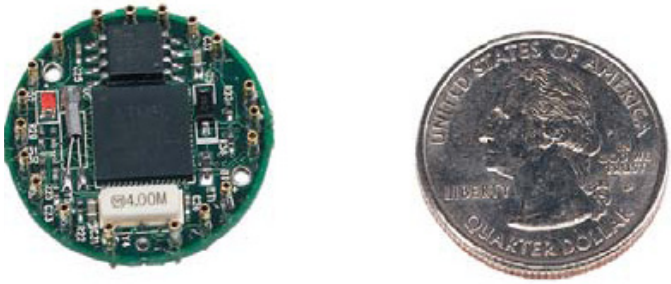


Fig. 2. A Mica2DOT board (taken from www.xbow.com).

This chapter investigates the use of wideband technology (UWB) technology for medical monitoring devices. UWB is one of the most recent wireless technologies using narrow pulses to carry data (Arslan et al., 2006). The major drawback of a UWB system is the high power consumption of the UWB receiver as indicated in Table 2. Power consumption at UWB's receiver is very high, usually higher than the narrow band wireless communications. Although recent developments have shown some promising results (Ryckaert et al.,2007), a low-power complete wireless UWB transceiver still consumes more than traditional narrowband devices. It is important to note that the transmitter part of UWB consumes very little power as it is very easy to generate pulses in a circuit (Ryckaert et al., 2005). Thus if

there can be a trade off between the transmitter and receiver, medical monitoring will be one of the most attractive applications for UWB communication (Yuce et al., 2007). UWB has two major advantages, high data rate and low power consumption of the transmitter. High data rate property can be incorporated with medical sensors requiring multi channel continuous monitoring from a single patient (Especially multi channel EEG signals are used in some healthcare applications for better diagnosis) (Ho & Yuce, 2008). It may also be possible in future wideband technology to be integrated together with narrow band medical system easily for a reliable wireless communication as well as to cover different environments in medical centers. High data rate transmission, as commonly known, is not the only unique property of an UWB based telemetry system. The advantages of a wide band technology can be summarized as follows:

- **Low power transmitter design**. A transmitter circuit can be designed with few components and may consume extremely low power when designed with an integrated circuit technology. Thus if a medical monitoring device can be designed based on a transmitter only approach, significant power and size reduction can be achieved.
- Dedicated band to wideband technology is **very large, ranges from 3.1 to 10 GHz**. This range can be divided into different bands where each band can be used for one medical device. This way a multi-access scheme can easily be arranged in the case more than one medical device is used in the same environment.
- The wide band technology has **less interference effect** on other wireless devices since the regulated transmitter power is very low (i.e. -41.dB).
- **High data rate capability**
- **Miniaturized antenna** design at high frequencies.

2.1 Transmitter Design for Wideband Communication

The basis of any UWB transmitter is a narrow rectangular pulse train and some form of filtering to meet the spectrum mask requirement (Arslan et al., 2006). Wide-band communication can be broadly classified into Impulse UWB (I-UWB) and Carrier based UWB (MC-UWB). I-UWB requires fewer components and has the advantage of simple and low power transmitter designs. MC-UWB normally divides the bandwidth into channels of 500MHz; it performs better in avoiding interference from narrowband systems and useful for multi-access communications. UWB conveys information using very narrow pulses typically in the range of a few hundreds of picoseconds. For an I-UWB system, the desired spectrum shape is achieved through pulse shaping; hence no carrier is required. The Gaussian pulse and its derivatives are the most commonly used pulse shapes for UWB analysis (Marchaland et al., 2005). However, practical implementation of a higher order Gaussian pulse is difficult (Hyunseok, 2003). Therefore, most practical transmitters use a monocycle pulse, which is a first order derivative of the Gaussian pulse. As the monocycle pulse does not meet the FCC's spectrum requirements, some form of filtering is necessary.

The narrow UWB pulses can be obtained by using completely digital or mixed digital–analog techniques. The use of analog blocks such as mixer and VCO in a UWB system is less attractive for low power, low cost medical applications. Among all methods using the delay-and-AND gate or delay-and-XOR gate is the least complex way in CMOS integrated circuit

technology. Basically any type of pulse generator with a pulse shaping circuits will provide an I-UWB based transmitter. However, a transmitter design that controls the pulse width using a bandpass filter as well as controlling the energy from the frequency domain side lobes of the narrow rectangular pulse is found to be very suitable for low power medical applications. A general scheme for pulse generations is given in Fig. 3. The delay unit can be realized using digital gates such as inverters, analog differential delay cells, flip-flops and controllable capacitors (Yuce et al., 2007; Ho & Yuce, 2008; Marchaland et al., 2005). In Fig. 4, a time diagram is depicted to show how a narrow band pulse can be generated from data bits. The data signal $s(t)$ and the delayed replica $s_d(t)$ are passed through XOR gate or an AND gate to obtain a UWB narrow pulse $x(t)$ (e.g. $x(t)=S(t).S(t-\tau)$).

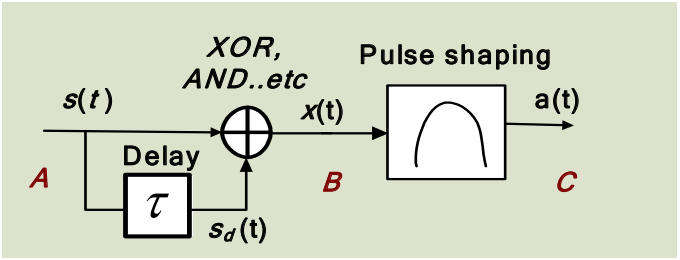


Fig. 3. Narrow pulse generator for UWB communication

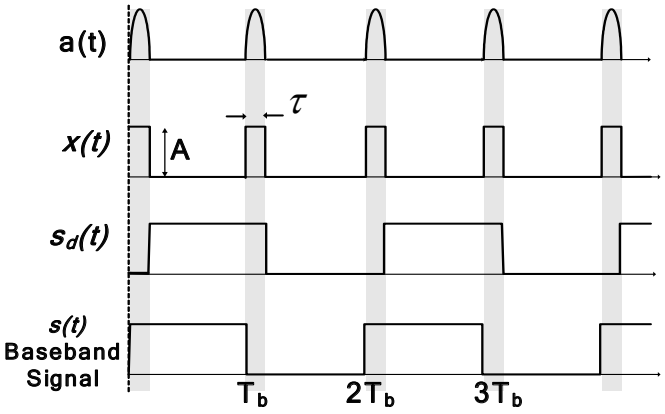


Fig. 4. Timing diagram for UWB pulse generation

2.2 Analysis of Band Limited Narrow Pulses for Medical Telemetry

In this section UWB pulse generation schemes in terms of time and frequency domain characteristics will be explained. The term "band limited " is used here because a band limited UWB signal has advantages by eliminating interference, if any, since the allocated band is very wide. It also presents the opportunity to divide the UWB spectrum into different bands when more than one device is used in the same environment a frequency hoping (different band allocations) can be applied. Assuming a UWB signal using a coding scheme such as Manchester non-return zero (NRZ) or Pulse Position Modulation (PPM), the

Fourier series expansion in terms of a rectangular pulse train can be represented by (Yuce et al., 2007)

$$x(t)=\frac{A\tau}{T_b}+\frac{2A\tau}{T_b}\sum_{k=1}^{\infty}\frac{\sin(\pi k\tau/T_b)}{\pi k\tau/T_b}\cos(k\omega t) \tag{1}$$

where A is the pulse amplitude, T_b is the bit period, τ is the pulse width obtained from the delay unit in Fig.4, ω is $2\pi/T_b$. In order to satisfy the FCC's spectrum requirements, the signal must be band limited, which can be given by (2).

$$x(t)=\frac{2A\tau}{T_b}\sum_{k=n1}^{n2}\frac{\sin(\pi k\tau/T_b)}{\pi k\tau/T_b}\cos(k\omega t) \tag{2}$$

where n_1 is ω_1/ω, n_2 is ω_2/ω, ω_1 and ω_2 are the lower and upper cutoff frequencies of the bandpass filter, respectively. Since the UWB is a power limited system, maximizing the transmission power within the spectrum mask is an important design consideration. Detailed analysis of (2) shows the impact of the various pulse parameters on the transmission power. Although this analysis is illustrated using a rectangular pulse train, the fundamental principle can be applied to all types of pulse shapes, including those used in carrier based systems.

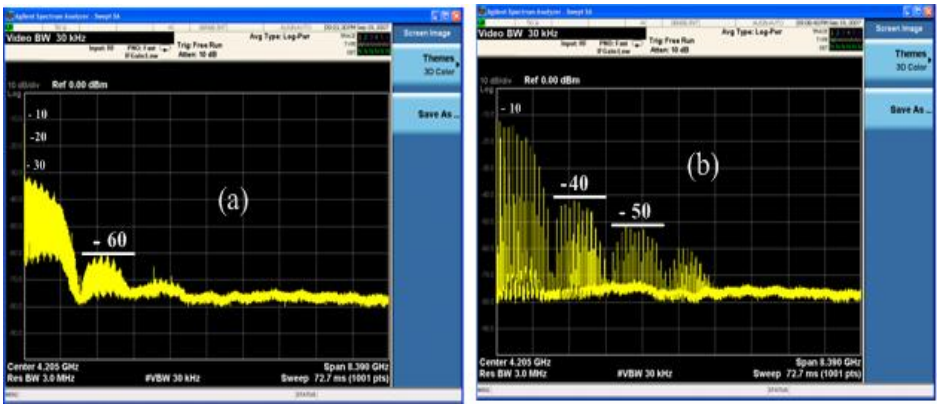


Fig. 5. Spectrums of 10MHz (a) and 100 MHz (b) pulse trains.

The key to maximize the transmitted power is to have the maximum number of approximately equal amplitude spectral lines within a given bandwidth. The maximum number of spectral lines depends on the data rate. Lower data rates contain more spectral lines but are lower in amplitude. For example, for a system with a 2 GHz bandwidth spanning from 3.1 GHz to 5.1 GHz and a data rate of 100 Mbps, the maximum possible number of spectral lines is 21. The amplitude of spectral lines is multiplied by a factor of

$2A/\pi k$ where k ranges from 30 to 51. While a system with a data rate of 10 Mbps contains 200 spectral lines (k ranges from 300 to 500). Fig. 5 shows the spectrum plots of a 10MHz signal and a 100 MHz signal with 1 ns pulse width. From the plots, it is evident that the envelope is determined by the 1ns pulse width. The amplitude of the frequency components depends on the data rate. Although amplitudes of the spectrum lines for the high data rate are higher, it does not mean that the spectrum contains higher energy in a specified bandwidth. Both signals will have the same amount of energy as the low data rate has more spectral lines.

The relationship between the pulse width (τ), the pulse period (T_b) and the number of spectral lines within a 2GHz bandwidth is shown in Fig. 6. The smallest number of spectral lines occurs when τ = 50% duty cycle where null occurs at every even integer multiple of k. To ensure a maximum number of spectral spikes in a given bandwidth, two conditions have to be satisfied. The center frequency of the desired frequency band, $(f_2-f_1)/2$, must align with $0.5/\tau$, and in addition $1/\tau$ must be greater than $f_2 - f_1$. When both conditions are met, a plot as shown in Fig. 7-(a) can be obtained. If $1/\tau$ is less than the bandwidth of the bandpass filter, the null occurs inside the band of interest as shown in Fig. 7-(b).

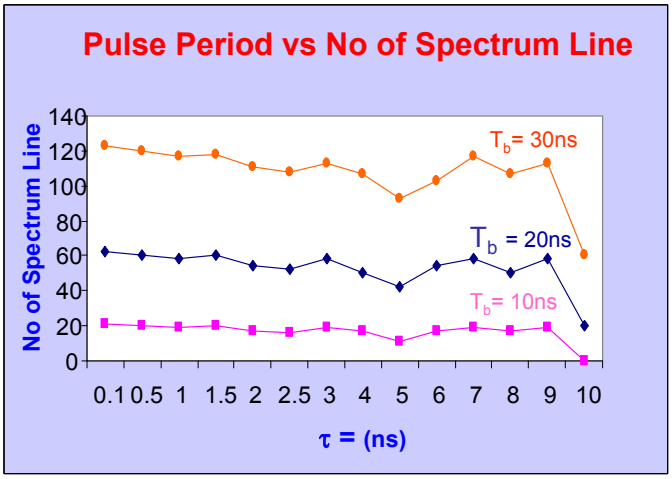


Fig. 6. Number of spectral lines vs. pulse width.

Another important factor that determines the transmission power is the relationship between the filter bandwidth (BPF) and the pulse width. Wideband filters with narrowband interference cancellation capability normally come in the form of FIR filters, which require high computing power and are power hungry. The practical implementation of wideband low-noise amplifier (LNA) and antennas is also hard to achieve. Therefore it is necessary to have a good balance between pulse widths, the filter bandwidth and the signal strength. If both τ and the filter bandwidth are large, the signal to noise ratio (SNR) will be low. A general guideline for selecting these two parameters is to maintain $1/\tau > BW$. The output of a bandpass filter resembles a high order Gaussian pulse. Fig. 8-(a) shows the output pulse from a filter bandwidth of 7.5 GHz and Fig. 8-(b) with a bandwidth of 1GHz. The wider the

bandwidth, the narrower the output pulse can be obtained. Thus the width of the output pulse can be determined by the bandwidth of the bandpass filter. Sometimes this feature could be useful in the demodulation of the UWB modulated signals.

As explained earlier, the width of the input pulse determines the energy level. It is thus important to adjust the BPF such that a null will not appear within the spectrum and also the transmitted signal energy is optimized with respect to parameters explained in Fig. 6&7. It should also be noted that the distance between two pulses as shown in Fig. 8 ensures their distinction. A larger bandwidth has a finer time resolution and is thus easier to distinguish the adjacent pulses during the demodulation process at the receiver side.

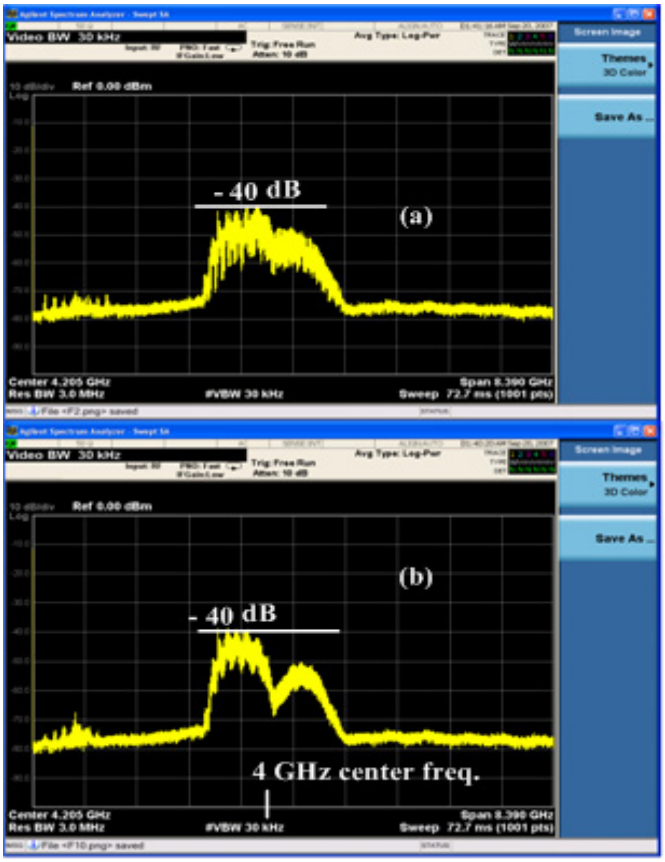


Fig. 7. Band limited pulse spectrums using 1 GHz BPF.

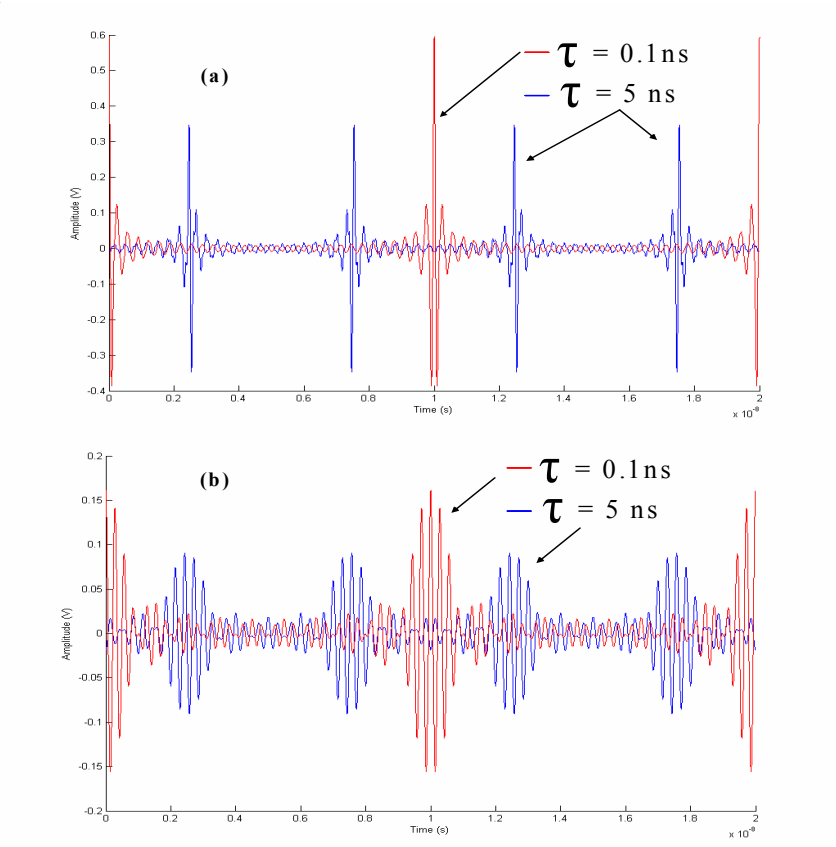


Fig. 8. The effect of bandwitdh of BPF on UWB pulse.

3. Wide Band Telemetry for Implantable Systems

Many efforts are being undertaken by scientists to understand the functions of inner organs that may be useful to diagnose and treat patients better. The early implantable[4] devices were constructed with simple electronic structures to make the device small enough so that it can be inserted in the body. A basic transmitter connected to a sensor has been used to send the signal from inside the body to external devices for tracking physiological parameters of organs (Mackay, 1961). Today's technology is able to miniaturize complex sensor devices e.g camera in order to insert inside the body so that it can travel and capture pictures for medical monitoring and detection. Prosthetic devices like Retina and Cochlear implants require a distance of few centimeters for wireless data transmission while the new implantable systems like wireless endoscope and drug delivery implants necessitate a range of 0.5-2 meters as the devices are inserted deeply inside as well as to allow patients free movement in the medical centers. In order to extend current capability of an electronic pill

[4] The term implantable refers to devices that are inserted or ingested into the body.

technology, scientists work on the development of a high capacity radio system for better resolution as well as small enough to be swallowable or implantable in the human body.

This section will discuss the research activities for the development of a high resolution wireless endoscope device that uses wideband technology to improve medical detection and treatment. To study the feasibilty of UWB signal tranmission within a human body, a band limited UWB prototytpe system explained earlier has been tested in a laboratory environtment. Integration of antenna with UWB transmitter electronics has been considered in a capsule shaped structure. In this section the implementation details and measurement results interms of time signals and frequency spectrums at different stages of the UWB prototype system will be presented, with capsule-shaped antennas at both the transmitter and receiver ends.

3.1 UWB Technology for Wireless Endoscope

The wireless capsule endoscope is a recent implanted system that requires the integration of more complex systems on the same platform when compared to conventional implantable systems. Wireless endoscope (i.e. electronic pill) is an alternative to fiber based endoscope used in diagnosing diseases related to gastrointestinal tract, which is often inconvenient to the patient. Furthermore, capsule endoscope can reach areas such as small intestine and deliver real time images wirelessly to an external console (Meng et al., 2004). Fig. 9 shows an example of wireless endoscope used in a medical monitoring system. The device travels through the digestive system to collect image data and transfers them to a nearby computer for display with a distance 1 meter or more. A high resolution video based capsule endoscope produces a large amount of data, which should be delivered over a high capacity wireless link.

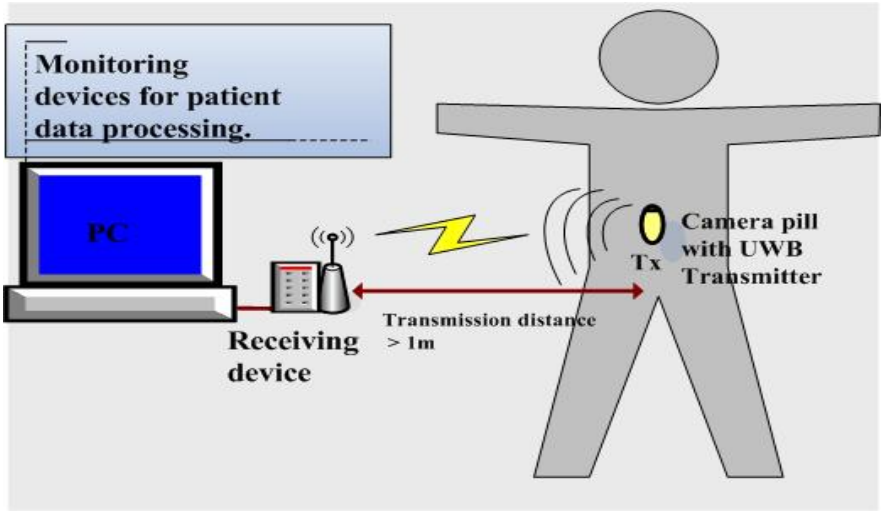


Fig. 9. A wireless endoscope monitoring system.

Since its early development, wireless endoscope designs have been based on narrow band transmission and thus have limited number of camera pixels (Nagumo, 1962; Meron, 2000). These systems are bulky due to large electronic components and batteries used. Current attempts in wireless endoscope systems or in other implantable systems have been limited to low frequency transmission. The low frequency transmission is easy to design and is found attractive due to its high efficiency. However a low frequency link requires large electronic components such as capacitors and inductors, which makes it difficult to realize a complete integrated system. Thus a high frequency link is required for better resolution and a miniaturized system. A miniaturized integrated system with a short-medium range wireless capability will make a significant impact in the performance of an electronic pill.

For low-power, high data rate and short-medium range applications of this kind, UWB (Ultra-wideband) communication is an ideal physical layer solution, which can achieve a data rate equal or higher than 100 Mbps(Kim et al., 2007; Lee et al., 2005). There have been ongoing advancements in UWB communication for short range applications (WIMEDIA, online 2009); however they cannot directly be applied to implantable telemetry because of different design and optimizations required due to stringent physical requirements and biological safety.

In recent designs (Xie *et al*., 2006; Park et al., 2002), narrow band wireless systems have been used for bi-directional telemetry systems, with limited data rate capability. The use of wideband technology for medical implants should overcome unique challenges associated with the high frequency implementation. To address these challenges, preliminary work is presented in this section to show a complete working UWB prototype with a capsule-shaped antenna specifically designed for an electronic pill technology. The selected band is 3.5-4.5GHz (i.e. band-limited), which avoids narrowband systems operating in the ISM (Industry, Scientific and Medical) bands. Frequencies beyond 5GHz are avoided since tissue imposes strong attenuation in that range. The initial design has been tested in a laboratory environment demonstrating that an impulse based UWB system is an attractive design for wireless endoscope monitoring.

Current wireless endoscope device commercially available by "Given Imaging" is used to diagnose disorders such as Crohn's disease, Celiac disease, benign and cancerous tumors, ulcerative colitis, gastrointestinal reflux disease (GERD), and Barrett's esophagus. (Givenimaging, online, 2009). The pill uses the Zarlink's RF chip for wireless transmission (Zarlink, online, 2009). The chip uses the MICS (Medical Implant Communication Service) band that is also allowed for unlicensed use and implantable devices such as cardiac pacemakers, hearing aids, and neurostimulators (Bradley, 2006). However, the allowable channel bandwidth for this band is only 300 kHz. It is difficult to assign enough data rate for the high quality image and video data at the moment for a real time data transfer and monitoring. It is quite obvious that there is an immediate need for higher-bandwidth data transmission for electronic pill to provide a real time data video and image monitoring that could facilitate a better diagnosis by medical professionals. This wideband technology can also be used for other medical sensing devices and prostheses such as implantable drug delivery and cochlear implants, and many other applications across the range.

3.2 Implementation and Testing

The feasibilty of UWB signal tranmission within a human body is shown in this section. A band limited UWB prototytpe system described earlier has been tested in a laboratory environtment for wireless endocsope monitoring systems. In this section the implementation details and measurement results interms of time signals and frequency spectrums at different stages of the UWB prototype system are presented, with capsule-shaped antennas at both the transmitter and receiver end. Main challenges associated with the design of microelectronics for implantable electronics are miniaturization, antenna design and saving the battery life. The microsystems will contain four main blocks, battery/power management circuitry, camera/sensors, transmitter (UWB transmitter) and antenna design. Integration of antenna with UWB transmitter electronics should be considered in a capsule shaped structure, ideally size000. Since miniaturization is important, different design approaches can be followed. As an example, each block on a separate board layer and then integrate them on top of each other as shown in Fig. 10 is a good approach to follow for a better miniaturization. In a different design shown in Fig. 10-(a) antenna can be designed such that it can easily be inserted on top of the transmitter layer. In Fig. 10-(b), the capsule shape is divided into two regions where antenna will be designed to be placed in upper-half whereas the remaining electronic units could be placed in the lower-half. Placing electronic units on one side of antenna is another possibility, Fig. 10-(c). There are commercially available mini cameras that can easily be integrated in electronic pill technology (STMicroelectronics. Online. (2009)). Small miniature rechargeable battery technologies are also being developed (smallbattery, 2009; buybionicear (http://www.buybionicear.ca/), 2009). These batteries have a dimension around 5 mm and can easily be integrated in a capsule shape structure shown in Fig 10.

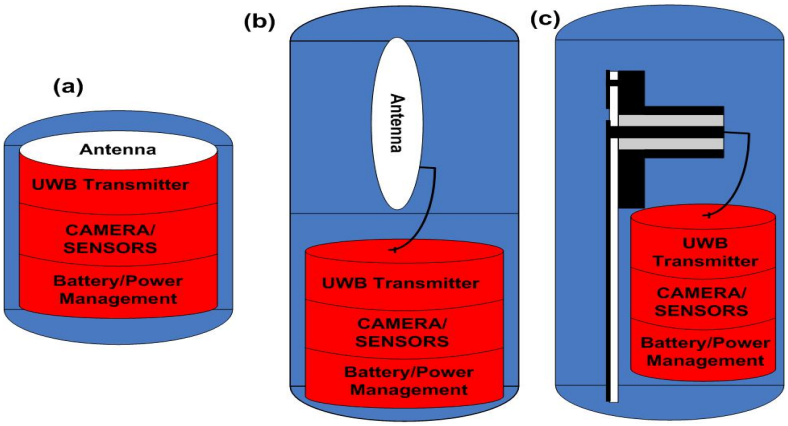


Fig. 10. Possible physical shapes for future implantable electronic pills.

The antennas that have been previously reported for endoscope applications operate in a lower frequency band (Kwak et al., 2005) A low-cost, printed, capsule-shaped UWB antenna has been designed for the targeted application (Dissanayake et al., 2009). The printed antenna presented herein demonstrates good matching in the frequency band of 3.5-4.5GHz and the radiation performance has been evaluated experimentally using a low-power I-

UWB transmitter/receiver prototype to show that it is suitable for the implantable wireless endoscope monitoring. The antenna matching has been optimized using CST microwave studio commercial electromagnetic simulation software. Proposed antenna is printed on a 0.5mm thick RO4003 capsule-shaped, low loss, dielectric substrate ($\varepsilon_r = 3.38$). It can easily fit inside a size-13 capsule (Capsule, 2000) , ingestible by large mammals. Overall length and width of the antenna is 28.7mm and 14mm, respectively. It is primarily a planar dipole, which has been optimized using simulations and printed on one side of the substrate together with a Grounded-CPW (Coplanar Wave Guide) feed as shown in Fig. 11.

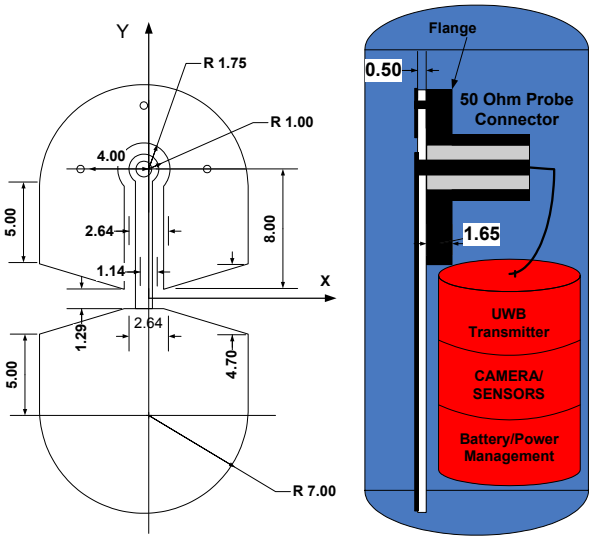


Fig. 11. A wireless endoscope monitoring system with antenna dimensions.

Grounded-CPW has characteristic impedance of 50 Ohms and the ground plane on the opposite side of the substrate is intended to support other electronics as shown in Fig. 11. This avoids performance degradation upon integration with other electronics, batteries and connectors. A panel mount SMA connector is used in place of these electronics for testing. Flange of the connector acts as a ground plane to the CPW. The circular pad in one end of the grounded-CPW facilitates broadband coaxial-to-CPW transition (Kamei et al., 2007).

The feed line has an effective dielectric constant of 2.62 at 3.5 GHz (lower end of the matched band). Therefore, the guided wavelength at that frequency is approximately 53mm, which is less than that of a CPW. The overall antenna length, 28.7mm, is close to half the guided wavelength, which is typical for a dipole. Hence the additional ground plane, which also is a part of the feed line, has contributed to the miniaturization of the antenna. As a result, largest dimension of the proposed antenna is only 0.3 times the free space wavelength at 3.5 GHz, 40% less compared to half of free space wavelength. On top of this dielectric loading of the antenna may be employed to achieve further antenna miniaturization. Three symmetrically placed vias ensure electrical connection between the patch on one side of the substrate and the flange of the connector on the other side. The

radius of each via is 0.75mm. Parametric studies have shown that the distance to the vias from the center of the coaxial feed affects the input impedance of the antenna. Note that the patch, flange and each via form shorted transmission line resonators. At certain lengths, the resonant frequency of the standing waves created by via reflections can be between 3.5 and 4.5 GHz, resulting an in-band notch, which is not desirable. Thus we have selected 4mm as the optimum distance.

Two antenna prototypes have been fabricated using conventional printed circuit board design techniques. This makes the antenna low cost. Reflection coefficients of both antennas have been measured using E5071B vector network analyzer from Agilent. Measured results and simulated S11 values from CST Microwave Studio are shown in Fig. 12. There is a good agreement between measured and theoretical S11 results. Antennas have greater than 10dB return loss from 3.4-4.6 GHz. Simulations suggests that the proposed antenna has radiation patterns (not shown) similar to that of a dipole antenna. Theoretical gain at 4 GHz is 2.23dBi. It allows about -45dBm/Hz output power of the UWB transmitter under the regulations in free space. Higher transmitter power or antenna gain is possible for in-body transmission as we shall discuss shortly.

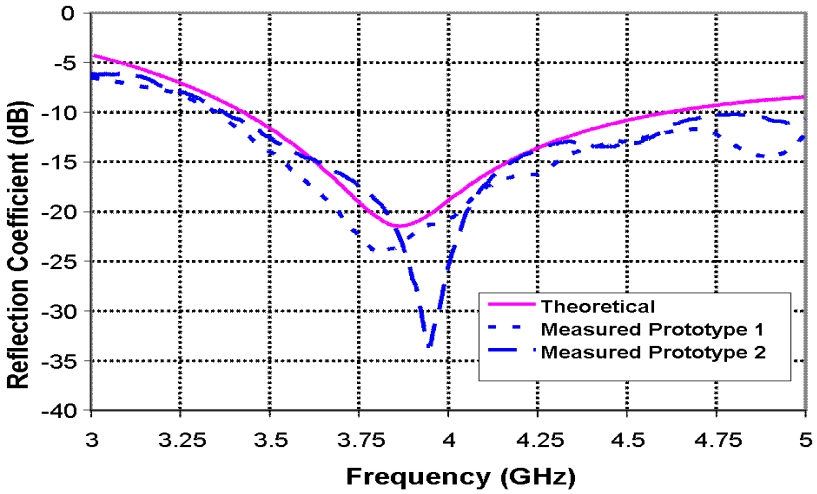


Fig. 12. Theoretical and measured reflection coefficients of the UWB antenna.

3.3 Experiments for Tissue Penetration

Our objective is to demonstrate the designed antenna and UWB prototype is capable of supporting a low-power UWB communication, which will be ultimately used to form an in-body-to-air link, without FCC violating regulations. The setup used in the experiment is shown in Fig. 13. The diameter of the plastic container is 75mm. The network analyzer (VNA) used is calibrated for full range. Salt reduced Corned Beef Silverside has been used as meat. One antenna is fixed at the bottom of the container, while the other is flushed into meat during the measurement. Both antennas were coated with clear rubber coating from Chemsearch™, to prevent any contact with meat or fluids.

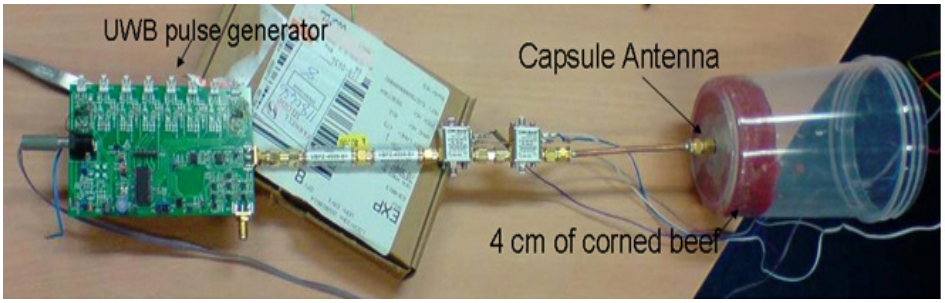


Fig. 13. Experimental setup of a UWB transmitter with capsule shaped antenna loaded with tissue material.

The coating did not have any effect on the antennas' characteristics. Antennas were held parallel so that coupling through meat is in bore sight. Prior to each measurement, jacket of aluminum foil covered the outer surface of the container to minimize outside coupling paths between the antennas. Measured S21 using the VNA is shown in Fig. 14. Coupling between antennas in the same laboratory environment and instrument calibration, for both through the meat and free space, are shown for comparison. There is about 20-30 dB attenuation through meat within 3-5GHz band for every 2 cm. This attenuation is not entirely due to absorption by meat. The antenna mismatch due to presence of meat also contributes to this.

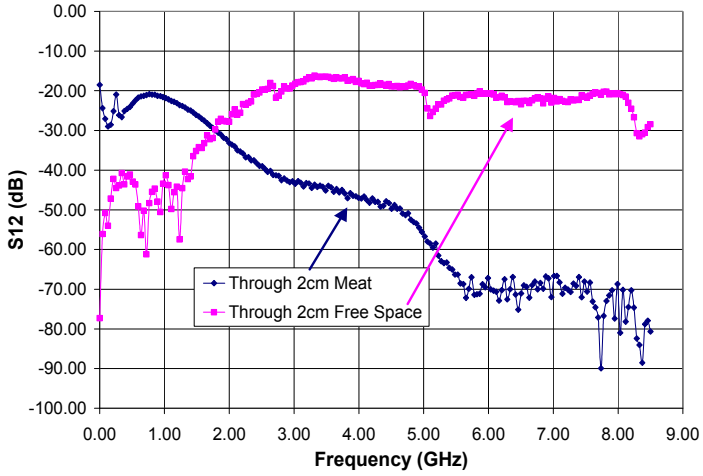


Fig. 14. Antenna coupling through meat (s21 measurement).

For a UWB transmitter, the regulation requires the signal output to be -41 dBm/Hz and lower with 0dBi antenna gain (Arslan et al., 2006). To make the UWB transmission feasible for implantable devices, higher transmitted signal levels can be used at the implanted transmitter side. The UWB signal power is arranged such that when the signal is radiated through the skin, the power level should meet the FCC mask. Fig. 15 shows acceptable transmitted power levels of the implanted transmitter for different penetration depths,

approximately based on the results of our experiment. At 2cm, we can allow for as much as 20 dBm of transmitted power, which would ultimately meet regulated spectral density requirements after penetration through tissue. Thus considering the strong attenuation through body tissue, the transmitter power level can be adjusted from -20 dBm to 20 dBm in the system, without violating power levels of FCC regulation. Of course, the power levels should not reach above regulated in-body tissue absorption levels. A special case of electronic pills is that the device travels in the body, it does not stay in the same area (unlike the stationed implants), and thus increasing power levels will not increase the heat much at the tissue of a certain body part.

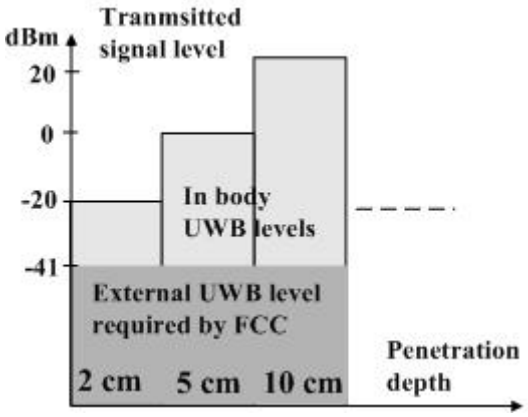


Fig. 15. Power levels of transmitted UWB signal in body.

3.4 Testing and Measurements

In the I-UWB setup, pulses have been generated based on an all digital approach described in section 2.2. Fig. 16 shows the UWB prototype with transmitter and receiver with waveforms shown explicitly. Short pulses are generated according to the on-off keying (OOK) modulated signal. At the transmitter, the pulse generator unit produces a rectangular-shaped pulse with 1ns width, as shown in Fig 16 (a). The spectrum of the rectangular pulse extends over an unlimited frequency band. Thus a Band Pass Filter (BPF) centered at 4 GHz with 1 GHz bandwidth is used to constrain the signal power under the FCC emission mask (i.e. a band limited UWB system). The energy of the side lobes is maximized within the bandwidth of the bandpass filter as discussed in Section 2.2. The filtered pulses are fed into our custom made UWB antenna. The UWB signal has shown good performance in the frequency band of 3.5- 4.5 GHz. It has also shown its ability to form a 0.6 m UWB link across the laboratory both in free-space and when loaded with meat emulating an implant once a high gain antenna is used at the receiver instead of one shown in Fig. 16-(b).

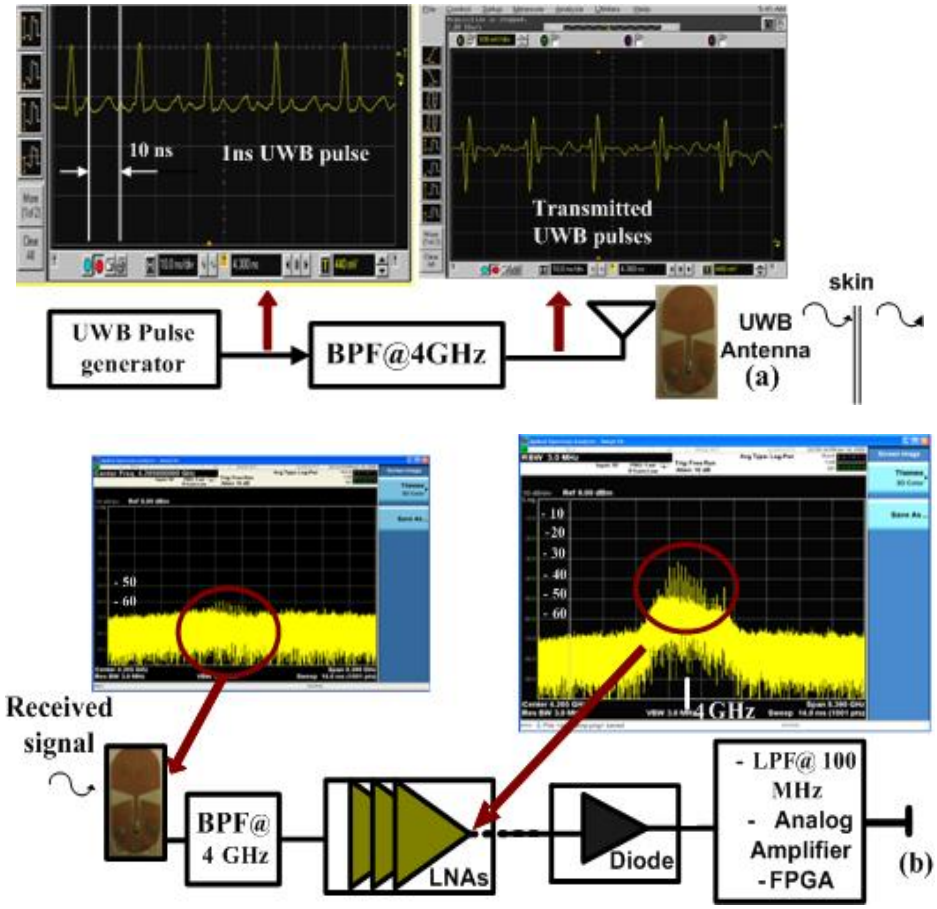


Fig. 16. A ultra wideband (UWB) wireless telemetry prototype and measurement results,(a) transmitter with 1 ns UWB pulse, and (b) receiver with spectrums at the output of antenna and after RF amplifications.

Despite the simplicity of the transmitter design, several limitations arise when designing a practical UWB receiver. A major challenge faced by an UWB receiver is its capability to demodulate the narrow pulses. A coherent receiver requires a very high speed ADC (Analog-to-Digital Converter) with a large analog input bandwidth. Secondly, it is hard to achieve precise synchronization, which is critical for the reliable operation of coherent receiver. In this experiment, a non-coherent energy detector method is used to demodulate the received signal.

There are different receiver architectures that can easily be constructed using high performance off-shelf RF components. Usually a mixer is used to down convert the high frequencies to low frequencies (Ryckaert et al., 2007). Herein a diode is used due to simplification in the successive blocks (See Fig. 16 (b)). The received signal is passed through a BPF, whose center frequency is 4 GHz, to eliminate possible interference from the

frequencies of Wireless Local Area Network (WLAN) standards (for example 2.4 GHz and 5 GHz). The signal is then amplified by the Low Noise Amplifier (LNA). A diode and a Low Pass Filter (LPF) down converts the UWB signal and the baseband data is finally recovered by the FGPA.

At the receiver end, the main component is the diode detector. When small input signals below -20dBm are applied to the diode, it translates the high frequency components to their equivalent low frequency counterparts due to its nonlinear characteristics. Measurement results, shown in Fig. 16(b) are spectrum plots at the outputs of the receive antenna and the low-noise amplifiers. The transmitted narrow UWB pulses are recovered at the output of the diode. The 50 MHz data stream is obtained at the FPGA after the demodulation process. The time domain signals before and after the FPGA are shown in Fig. 17. The recovered signal is a 50 Mbps pulse obtained from pulses with width of 1ns.

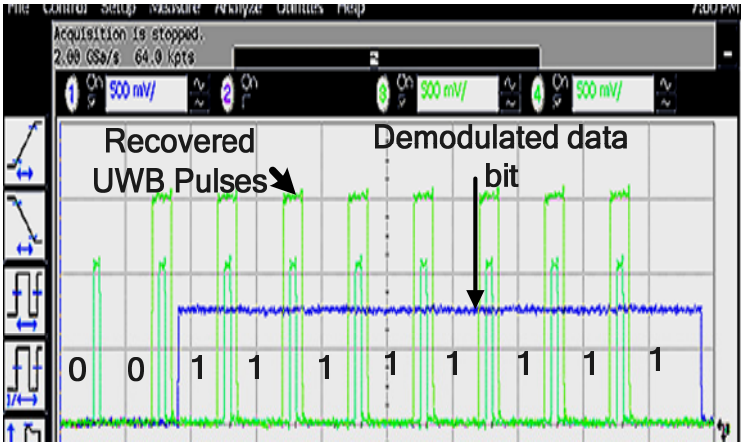


Fig. 17. Received and demodulated UWB signals.

4. Wearable Medical Monitoring System

Deployment of wireless technology for wearable medical monitoring has improved patient's quality of life and efficiency of medical staff. Several wireless technologies based on Bluetooth, ZigBee, and WLAN are available for sensor network applications (given in Table 1); however they are not optimized for medical sensor networks and lack interoperability. Therefore, there is a need for standardization to provide an optimized solution for medical monitoring systems. A group (IEEE802.15.6) was formed in November 2007 to undertake this task (WBAN standard, online, 2009). Low data rate UWB is one of the potential candidates under consideration, to overcome the bandwidth limitations of current narrowband system, and to improve the power consumption and size. In this part of the chapter, a multi-channel wearable physiological signals monitoring system using ultra wideband technology will be described.

4.1 Continuous Sign Monitoring Using UWB

An ultra wideband based low data rate recording system for monitoring multiple continuous electrocardiogram (ECG) and electroencephalogram (EEG) signals have been designed, and tested to show the feasibility of low data rate UWB in a medical monitoring systems. There has been a wide spread use of wireless monitoring systems both in hospital and home environments. Ambulatory ECG monitoring, EEG monitoring in emergency departments, respiratory rate, SPO2 and blood pressure are now performed wirelessly (WBAN standard, 2009; Ho & Yuce, 2007). The various wireless technologies adopted for medical application are shown in Table 1. Low data rate UWB is suitable for vital signs monitoring system as its transmission power is lower than those of WLAN, Bluetooth and Zigbee (See Table 1), and is less likely to affect human tissue and cause interference to other medical equipments. Furthermore, it is able to transmit higher data rates, which makes it suitable for real time continuous monitoring of multiple channels. Currently, the task group for Wireless Body Area Network (IEEE802.15.6) is considering the low data rate UWB transmission as one of the wireless technologies for the wireless devices operating in or around human body. Herein, a multiple channel monitoring system is designed and tested to show the suitability of low data rate UWB transmission for non-invasive medical monitoring applications. An 8-channel UWB recording system developed to monitor multiple ECG and EEG signals is presented in Fig. 18. Commercial off-the-shelf digital gates have been used for designing this UWB prototype system.

The system is designed to operate with a center frequency of 4 GHz and a pulse width of 1 ns, which is equivalent to 1 GHz bandwidth. An UWB transmitter is assembled using commercial off-the-shelf components for transmission of physiological signals from an on-body sensor node (Fig. 19). The UWB pulses are generated in a way to occupy the spectrum efficiently and thus to optimize the wireless transmission The transmitter as shown in Fig. 19 generates and transmits multiple pulses per bit. A clock in the transmitter is used for this

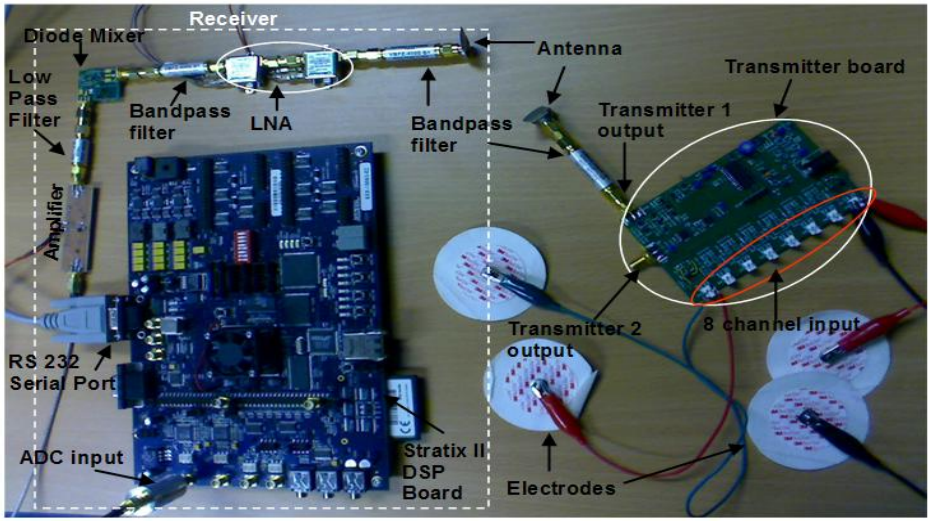


Fig. 18. Photograph of complete UWB prototype for physiological signal monitoring.

purposes and thus the number of pulses per bit can easily be adjusted. Sending more pulses per bit increases the power level at the transmitted band at 4 GHz. All the blocks (off-te-shelf components) in the transmitter consume a micro watt range power except the delay unit used to obtain very short pulses and the amplifier at the output used to arrange the output signal power for longer distances. These blocks can be designed with the recent low power integrted circuit technolgies that can easily lead to low power consumption. During the wireless transmission the ECG signal is digitised using a 10 bit-ADC in the microcontroller and the data is arranged based on the UART format in the sensor node. Each 10 bits data output from the ADC is transmitted with one start bit before the start of a byte and one stop bit at the end, which forms a periodic sequence that is used in the demodulation at the receiver.

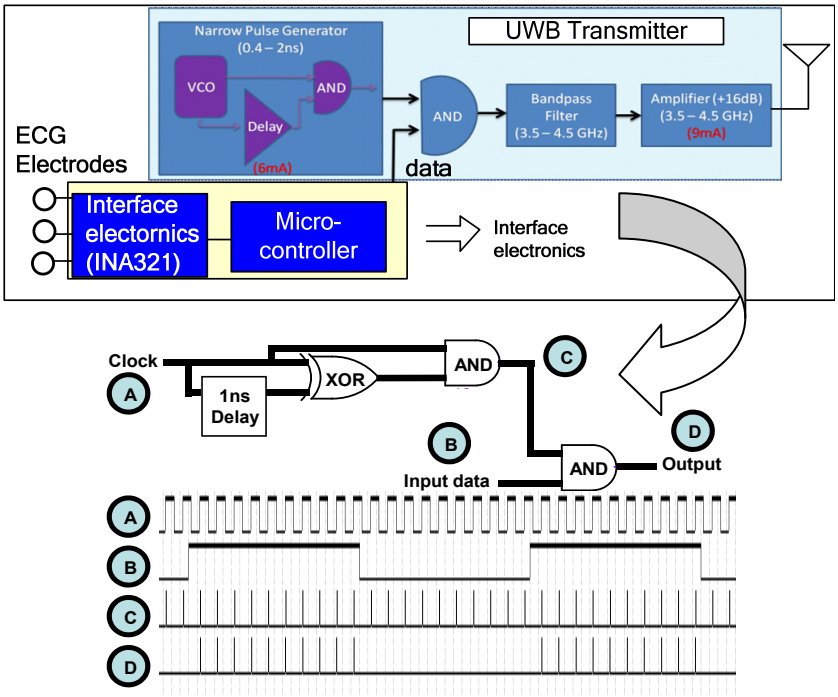


Fig. 19. ECG sensor nodes and UWB transmitter block diagram using off shelf components.

The non-coherent receiver and a field programmable gate array (FPGA) explained in the previous section is used to demodulate the data. The signals are monitored at the computer (PC) via the serial port based on UART format. Using the UWB prototype, multichannel ECG monitoring has been successfully performed showing the feasibility of low data rate UWB transmission for medical monitoring applications. Front ends for both the high data rate electronic pill system (section 3.1.) and low data rate UWB based wearable sensor system receiver for on body sensors are similar. However different data demodulation approaches are applied for the data recovery. Since here the UWB transmitter sends multiple pulses per bit to increase the processing gain, the receiver is designed to sample at

a rate much higher than the data rate. The information in the bit is determined, only after performing several samples; this increases the reliability of the system.

The ECG data is obtained from the body using the instrumental amplifier (INA321) from Texas Instruments. The ECG signals are transmitted and received wireless using the UWB pulses. The result is displayed using MATLAB in Fig. 20 on the remote computer. The signal is corrupted by the 50 Hz noise as can been seen in the waveform obtained from the oscilloscope before transmitting (Fig. 20-(a)), after receiver and monitoring in MATLAB in time (Fig. 20-(b)) and the frequency domain (c). The signal is passed through a 50 Hz digital notch filter designed using a MTLAB program. The 50 Hz noise is successfully removed and the ECG signal recovered. Removing the 50 Hz noise at the PC instead of the receiver helps to reduce the complexity and the programming power required at the receiver. The whole measurement has been carried out in our lab where there were other wireless standards (e.g WiFi) and equipments operating. The ECG signal has successfully been monitoring without any error.

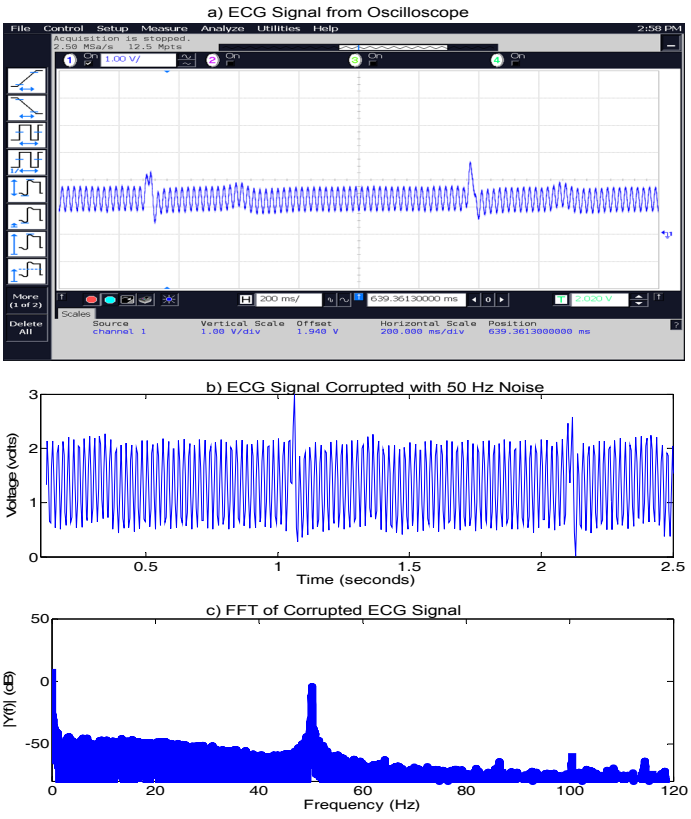


Fig. 20. Monitored ECG waveforms with 50 Hz noise

Alternatively, another program written using Visual Basic is developped to decode the data; it performs filtering as well as helps to displays the received multiple channel signals on the

screen. Parity bit check is performed on the received data to ensure all data received correctly. Once the received data is decoded, it is formatted back into a 10 bit word and separated based on the information embedded in the channel bits. Digital filtering is also performed on the received signal to remove the 50 Hz noise, which comes from the power supply. The ECG signal in Fig. 21 is successfully monitored in our lab environment with other wireless devices operating. The graphical user interface (GUI) program can display any eight channels by changing the button "channel selection" shown in the window.

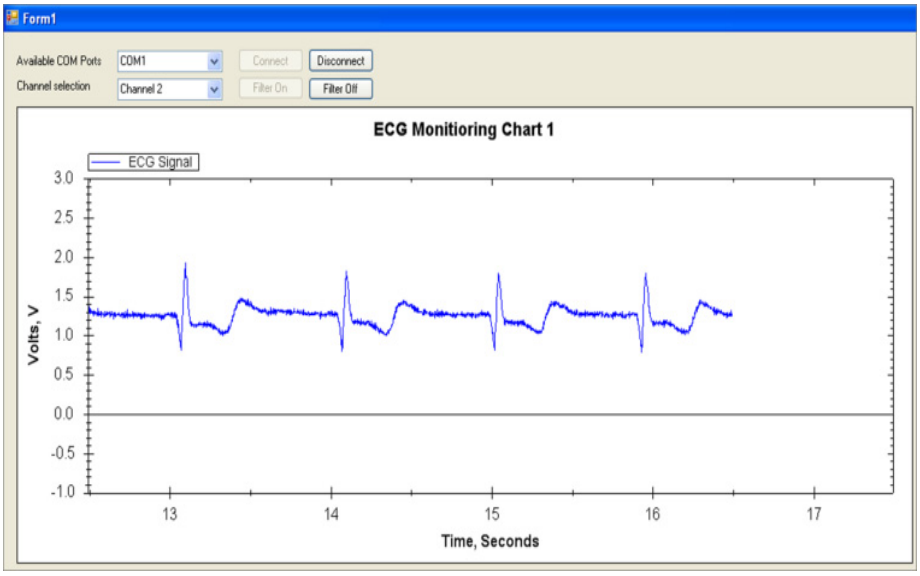


Fig. 21. Multi-channel ECG Signal detection via UWB wireless communication

5. Summary

This chapter has addressed the use of wideband signals in medical telemetry systems for monitoring and detection. The demonstrated UWB techniques provide an attractive means for UWB signal transmission for in-body and on-body medical applications. A band limited UWB telemetry system and antennas have been explained extensively to show the feasibility of UWB signals for implantable and wearable medical devices. The design of UWB transmitters are explained and analyzed to show its suitability for both high data rate and low data rate biomedical applications. Although the UWB system has higher penetration loss in an implantable environment compared to the conventional narrow band telemetry systems, a power level higher than the UWB spectrum mask can be used since it is a requirement for the external wireless environment. Thus an implanted UWB transmiiter should have the abilty to generate higher transmission power levels to eliminate the effect of strong attenuation due to tissue absorbtion. It should be noted that there will be a trade-off between the transmitted power levels and the desired communication range. A multiple channel EEG/ECG monitoring system using low data rate UWB technology has also been given in this chapter. The UWB receiver in the prototype is able to receive and recover

sucessfully the UWB modulated ECG/EEG signals. The real time signals are displayed on PC for non-invasive medical monitoring. Wideband technology can be targeted and utilized in medical applications for its low power transmitter feature and less interference effect. When a transmitter only approached is used, the transmitter design's complexity can be traded off with that of the receiver as the receiver will be located outside and its power consumption and size are not crucial.

6. References

Arslan, H.; Chen, Z. N. & Di Benedetto, M-G. (2006). *Ultra Wideband Wireless Communication,* Wiley-Interscience, ISBN: 978-0-471-71521-4, October 13, 2006, USA.

Bradley, P. D. (2006). An ultra low power, high performance medical implant communication system (MICS) transceiver for implantable devices, *Proceedings of the IEEE Biomedical Circuits and Systems Conference (BioCAS '06)*, pp. 158-16, , ISBN: 978-1-4244-0436-0, November -December 2006, IEEE, London, UK.

BUYBIONICEAR. http://www.buybionicear.ca/, 2009.

Capsule. "Capsule Size Chart," Fairfield, NJ, USA: Torpac Inc., 2000

Chae, M.; Liu, W. & Yang, Z. & Chen, T. & Kim, J. & Sivaprakasam, M &Yuce, M. (2008). A 128-channel 6mW Wireless Neural Recording IC with On-the-fly Spike Sorting and UWB Transmitter, *IEEE International Solid-State Circuits Conference (ISSCC'08),* pp. 146-603, 978-1-4244-2010-0, February 2008, IEEE, San Francisco, USA.

Dissanayake, T.; Yuce, M. R. & Ho C. K. (2009). Design and evaluation of a compact antenna for implant-to-air UWB communication. *IEEE Antennas and Wireless Propagation Letters,* vol. 8, Page(s):153 - 156, 2009, ISSN: 1536-1225.

Givenimaging, http://www.givenimaging.com/ , 2009

Ho, C. K. & Yuce M. R. (2008). Low Data Rate Ultra Wideband ECG Monitoring System, *Proceedings of IEEE Engineering in Medicine and Biology Society Conference* (*IEEE EMBC08*), pp. 3413-3416, ISBN: 978-1-4244-1814-5, August 2008,Vencouver, Canada.

Hyunseok, K.; Dongwon, P. & Youngjoong, J. (2003). Design of CMOS Scholtz's monocycle pulse generator, *IEEE Conference on Ultra Wideband Systems and Technologies*, pp. 81-85, ISBN: 0-7803-8187-4 , 16-19 November 2003, Virginia, USA.

Kamei, T.; et al. (2007). Wide-Band Coaxial-to-Coplanar Transition. *IEICE Transactions in Electronics,* vol. E90-C, no. 10, pp. 2030-2036, 2007, ISSN: 0913-5685

Kim, C.; Lehmann, T. & Nooshabadi, S. & Nervat, I. (2007). An ultra-wideband transceiver architecture for wireless endoscopes, *International Symp. Commun. and Information Tech.(ISCIT 2007)*, pp. 1252-1257, ISBN: 978-1-4244-0976-1, October 2007, Nice France

Kwak, S. I.; Chang, K. &Yoon, Y. J. Ultra-wide band Spiral Shaped Small Antenna for the Biomedical Telemetry, *Proceedings* of *Asia Pacific Microwave Conference*, 2005, vol 1, pp. 4, ISBN: 0-7803-9433-X, December 2005, China.

Lefcourt, AM.; Bitman, J. & Wood, D. L. & Stroud, B. (1986). Radiotelemetry system for continuously monitoring temperature in cows. *Journal of Dairy Science,* Vol. 69,(1986) page numbers (237-242).

Lee, C. Y. & Toumazou, **C.** (2005). Ultra-low power UWB for real time biomedical wireless sensing, *Proceedings of IEEE International Symposium on Circuits and Systems*, pp. 57 - 60 , ISBN: 0-7803-8834-8, May 2005, Kobe Japan

Liu, W.; et al. Implantable biomimetic microelectronic systems design. *IEEE Eng. Med. Biol. Mag.*, vol. 24, pp. 66, Sep.–Oct. 2005, SSN: 0739-5175.

Mackay, R.S. & Jacobson, B. (1961). Radio telemetering from within the human body. *Science* vol. 134, October 1961, pp. 1196-1202.

Marchaland, D.; Baudoin, G. & Tinella, C. & Belot, D. (2005). System concepts dedicated to UWB transmitter, *The European Conference on Wireless Technology*, pp. 141-144, ISBN: 2-9600551-1-X, october 2005

Meng, M. Q. H.; et al. (2004). Wireless Robotic Capsule Endoscopy: State-of-the Art and Challenges, *Proceedings of the 5th World Congress on intelligent Control and Automation*, vol. 6, pp. 5561-5565 ISBN: 0-7803-8273-0, 2004

Meron, G. (2000). The development of the swallowable video capsule (M2A), *Gastrointestinal Endoscopy*, vol. 6, pp. 817-8199, 2000

Nagumo, J.; et al. (1962). Echo capsule for medical use. *IRE Transaction on Bio-medical Electronics*, vol. 9, pp. 195-199, 1962 , ISSN: 0096-1884

Park, H. J. et al. (2002). Design of bi-directional and multi-channel miniaturized telemetry module for wireless endoscopy, in *Proc. 2nd Int. IEEE-EMBS Conf. Microtechnologies in Medicine and Biology*, 2002, pp. 273-276, ISBN: 0-7803-7480-0, Madison, WI, USA.

Ryckaert, J.; et al. (2007). A CMOS Ultra-Wideband Receiver for Low Data-Rate Communication. *IEEE J. of solid state circuits*, vol. 42, pp. 2515-2527, Nov. 2007 ISSN: 0018-9200.

Ryckaert, J.; et al. (2005). Ultra-wide band transmitter for low-power wireless body area networks: Design and evaluation. *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 52, pp. 2515, Dec. 2005, ISSN: 1549-8328

Smallbattery. http://www.smallbattery.company.org.uk/hearing_aid_batteries.htm, 2009

Shin, S. Y.; Park, H. S. & Kwon, W. H. (2007). Mutual interference analysis of IEEE 802.15.4 and IEEE 802.11b. *Computer Networks,* Vol. 51 **,** August 2007, pp. 3338-3353, ISSN: 1389-1286

STMicroelectronics.(2009).http://www.st.com/stonline/products/literature/bd/14370.pdf,

Tekin, A.; Yuce, M. R. & Liu, W. (2008). Integrated VCOs for Medical Implant Transceivers. *VLSI Design,* vol. 2008, January 2008, ISSN:1065-514X

WBAN standard; (2009). http://www.ieee802.org/15/pub/TG6.html, (WBAN standard group) 2009.

WIMEDIA, http://www.wimedia.org/en/index.asp, 2009.

Xie, X.; et al. (2006). A low-power digital IC design inside the wireless endoscope capsule. *IEEE. J. Solid State Circuits*, vol. 41, pp. 2390-2400, Nov. 2006, ISSN: 0018-9200

Yuce, M. R.; et al. (2007). A wideband telemetry unit for multi-channel neural recording systems, *IEEE International Conference on Ultra-Wideband (ICUWB)*, pp. 612-617, ISBN: 978-1-4244-0521-3 September 2007, Singapore.

ZARLINK; http://www.zarlink.com/zarlink/hs/4889.htm, 2009.

19

"Hybrid-PLEMO", Rehabilitation system for upper limbs with Active / Passive Force Feedback mode

Takehito Kikuchi and Junji Furusho
Osaka University
Japan

1. Introduction

The aging society and physical deterioration of the aged people have become a serious problem in many countries. Moreover, there are many patients of ataxia: paralysis caused by brain stroke, or asynergia. Early detection of the functional deterioration and sufficient rehabilitative trainings are necessary for such patients.

In general, therapists make rehabilitative programs based on inspections and measurements for each patient. However, it is difficult to adopt appropriate rehabilitation programs for all patients, because the evaluation method is based on experiences of each therapist. Nowadays, Evidence Based Medicine (EBM) is strongly required in the field of rehabilitation (Miyai et al., 2002). Therefore, the rehabilitation systems using robotics technologies or virtual reality technologies are expected to quantify the effect of rehabilitative trainings. Furthermore, robot system can enhance motivation of patients by creating new and unique training methods that have not existed yet.

Until now, some kinds of rehabilitation systems for upper limbs have been reported and developed (Krebs et al., 2000; Loureiro & Harwin, 2007, Lum et al., 2004; Zhang et al., 2007; Perry et al., 2007; Wolbracht et al., 2006; Nef et al., 2007). Almost all rehabilitation robots have utilized electric motors or other actuators. Such actuator-based (active type) systems have great advantages in rehabilitative activities, for example, those systems can perform assistive forces, spring-like reactions and so on. But from a view point of safety, we have room to consider utilizations of brake-based (passive type) rehabilitation systems.

Munir S., et al. (Munir S., et al., 1999) have developed passive haptic devices. In their system, conventional powder brakes were used as haptic generators. Grossly speaking, the response time of the powder brake is more than 50ms and it causes lack in quality of force feedbacks. To solve this problem, we have developed several types of haptic devices for upper limbs rehabilitation with ER fluid (Electrorheological fluid) brakes (Kikuchi T., et al., 2007). Thanks to the quick response of the ER fluids, these systems presented high quality haptics.

However, the effects and roles of active / passive force feedback for rehabilitative trainings have not been clarified yet. In this study, we have developed an active / passive switchable rehabilitation system for upper limbs (Hybrid-PLEMO), and planed to address its

effectiveness. In this chapter, we will explain a basic structure, properties and results of functional tests on the Hybrid-PLEMO.

2. Reaching function of brain-injured patient and its rehabilitation

Motor palsy is a decrease in physical capabilities of a voluntary movement. It appears clinically as a muscular weakness. Motor palsy is recognized as abnormal posture, movements, and abnormal motion patterns in the rehabilitation medicine a scapular girdle, a shoulder joint, an elbow joint, a wrist joint, and fingers cannot be moved separately. For severely impaired stroke survivors, such abnormal coordination is characterized with enforced co-activations between shoulder adductors and elbow extensors (extensor synergy) as well as between shoulder abductors and elbow flexors (flexor synergy) (Brunnstrom S., 1970). These synergy patterns gradually decrease depending on recovery of paresis with adequate rehabilitative trainings.

Upper extremity is mainly used for operations of objects; reaching, grasping and releasing. A normal reaching action takes great amount of efforts to adequately adjust a combination of motions of a shoulder, an elbow, a wrist joint and fingers. In many cases, the normal reaching is a very difficult task for the patients with ataxia because of their synergy movements.

In the rehabilitation to the paretic upper extremities, an improvement of the reaching function is one of the most important objectives. It is thought that stroke patients with the synergy pattern can improve their performances of upper extremities by acquiring the movement free from synergy patterns (Brunnstrom S., 1970). It is reported that, 30 to 66 percent of stroke patients do not use their upper extremity functions in daily life (Johanna H., et al., 1999). Two factors are related to this fact. Firstly, a lot of stroke patients tend to do almost all of ADL (Activities of Daily Living) with compensations of a normal side limb and they rarely use a paretic side limb, which is called “learned-non-use” (Wolf SL, Et al., 1989). Secondly, once their brains are damaged, excitements of the non-damaged side increase (Liepert J., et al., 2000) and it results in excessive weakening of the function of the damaged side.

Plautz et al. (Plautz EJ., et al., 2000) studied on the brain recovery using a squirrel monkey and its damaged-brain model. In their research, it is clarified that re-composition of a cerebral cortex is promoted by not only using the hand but also by advanced operation training with a motor learning. This indicates that re-composition of cerebral cortex can be facilitated by an advanced accurate operation task such as drawing tracks accurately. Moreover this can bring about good effects to improvements of stroke patient's upper extremity functions.

3. Development history

In our previous researches, clutch-type actuators with functional fluids have been adopted for torque control of rehabilitation systems. A conceptual diagram of ER fluid clutch actuator (ER actuator) is shown in Figure 1. Basic concept for safety with this clutch-type actuator was reported (Furusho J. & Kikuchi T., 2007). Then its applications for “EMUL” system, 3-D rehabilitation system for upper limbs (Furusho J., et al., 2005), and “Robotherapist”, 6-DOF rehabilitation system for upper limbs (Furusho J., et al., 2006) were also reported. These actuator-based (active type) machines have great advantages of

variation, accuracy and other performance of haptic forces. However, due to the usage of many actuators, these systems have disadvantages of cost, space and usability.

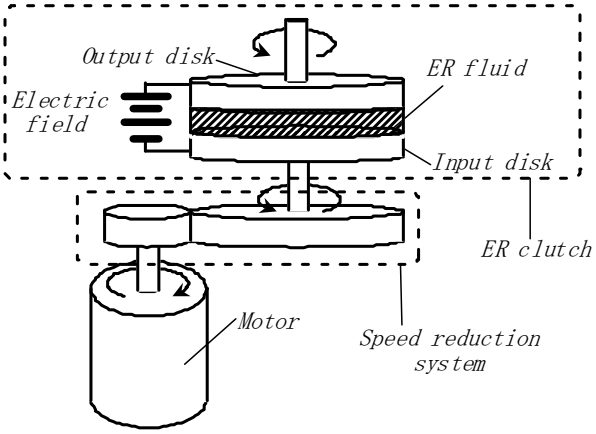


Fig. 1. ER-clutch-type actuation system for safety

In late years, we developed PLEMO systems with another concept for safety (Kikuchi T., et al., 2007). We have developed the PLEMO systems with demand of downsizing, low-cost, good usability and more advanced safety. The PLEMO systems have only 2-DOF force feedback function on a working plane for downsizing and cost-cutting, but the working plane can be adjusted its inclined angle. We named this system "Quasi-3-DOF Rehabilitation System for Upper Limbs" (Figure 2). For another feature of PLEMOs, its haptic control is conducted by only brakes with ER fluid (ER brake). These systems are safer than any other rehabilitation systems with actuators. The features of active / passive force feedback are compared in Table 1. As shown in this table, active type (actuator-based) machines have a great advantage on applicability for users. On the other hand, passive type (brake-based) machines have merits of safety, cost and size. The PLEMO systems are now under the clinical tests (Ozawa T., et al., 2009) (Figure 3).

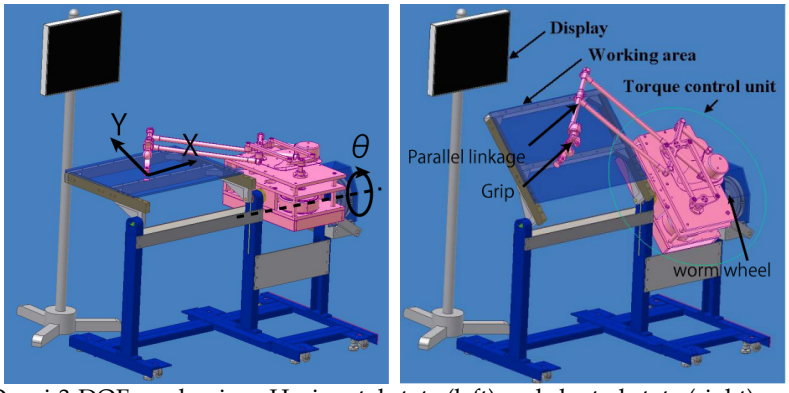


Fig. 2. Quasi-3-DOF mechanism; Horizontal state (left) and slanted state (right)

Feedback mode	Active force feedback	Passive force feedback
Force feedback	Actuator	Brake
Subject	Every subjects	Patient who can move his arm voluntarily
Safeness	Less safer than passive force feedback	Safe in mechanism
Cost	Expensive	Less expensive than active force feedback

Table 1. Comparison between active / passive force feedbacks in rehabilitation system

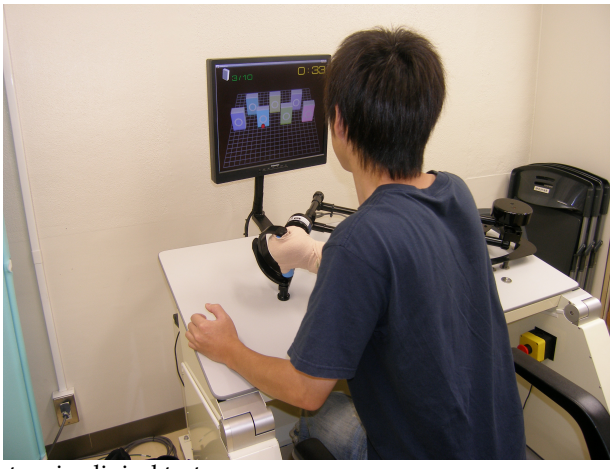


Fig. 3. PLEMO system in clinical tests

Table 1 shows comparisons in only engineering factors. However, it has not been cleared how active / passive forces effect to the upper limbs function in rehabilitation. We need a haptic device that provides active / passive haptic forces on the same environment to discuss this question. Then, we decided to develop the active / passive switchable haptic device for upper limbs rehabilitation; Hybrid-PLEMO (Kikuchi T., et al., 2008), mentioned in following sections.

4. Core technology

4.1 ER Fluid

ER fluid is one of the functional fluids of which rheological properties can be changed by applying electrical fields (Winslow W.M., 1949). In this paper, a particle-dispersed-type ER fluid is used. The characteristics of the fluid are shown in Figure 4. As shown in this figure, its shear stress depends on the application of electric field from 0.0kV/mm to 2.0kV/mm and does not depend on shear rate. The time constant of the viscosity change is several millseconds, and the response is stable and repeatable. Thanks to these characteristics, we can build up clutch / brake devices utilizing the ER fluid.

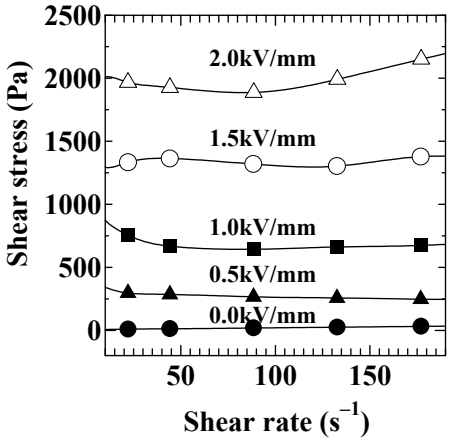


Fig. 4. Flow curve of ER fluid (Particle-dispersed type)

4.2 Basic structure of ER Actuator & brake

Figure 5 shows a basic structure of a cylindrical-type ER brake. It consists of a fixed cylinder and a rotating cylinder with the ER fluid between them. These cylinders also play the role of a pair of electrodes. The rotating cylinder is fixed on the output shaft and driven by external forces through this shaft. When a voltage is applied between the pair of cylinders, the electric field is generated within the ER fluid, and then the viscosity of the fluid increases. This increase of viscosity generates the braking torque and reduces the rotational speed.

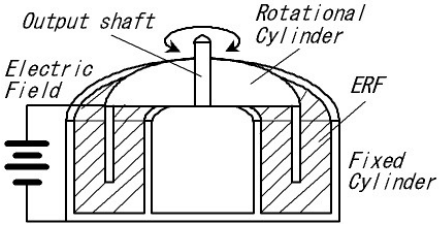


Fig. 5. Basic structure of ER Brake

With the same configuration and rotation of the fixed-side of the ER brake, we can compose ER actuator (see Figure 1) (Furusho J. & Sakaguchi M., 1999). In the configuration of the ER actuator, a conventional motor generates driving torque from input part of the ER clutch. Additionally, output torque of the ER actuator is controlled with the ER clutch separated from motor rotation. By restricting the rotational speed of the motor, we can easily keep safe state. This system has good controllability of torque, low inertia and high safety, which is suitable for human-machine coexisting systems, for example haptic displays or rehabilitation systems.

4.3 Double-Output ER Fluid Clutch / Brake device

Figure 6 shows an appearance and a cross section of the double-output and multilayered-disk-type ER fluid clutch/brake device developed in this study. This device has two groups of multilayered disks (input disks / output disks) in its package. Stator disks (input disks) are fixed on the casing for each group. However, when the casing is rotated by an electric motor, these disks are rotated simultaneously and the device works as a clutch. When the casing is locked, input disks are also locked and the device works as a brake. Two groups of output disks are connected to the inner shaft and the outer shaft, respectively. The particle type ER fluid is filled between each disk and we can control 2 output torques independently.

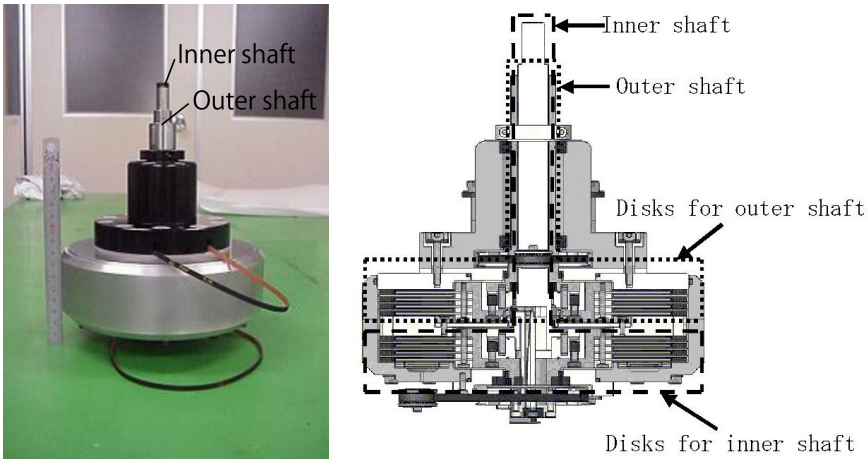


Fig. 6. Double-output ER fluid clutch / brake (left: Appearance, right: Sectional view)

Specification of the device is shown in table 2. Figure 7 shows output torque of this device. We can control transmission (or braking) torque by application of the electric field between rotor / stator disks accurately and rapidly.

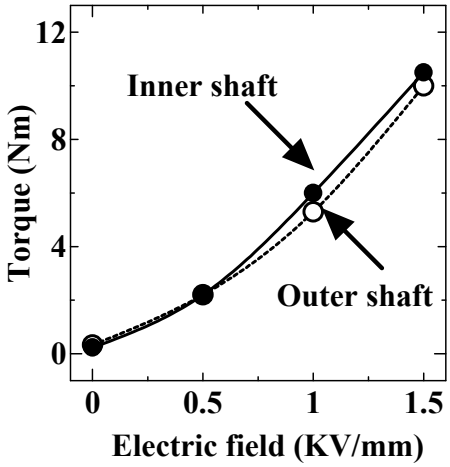


Fig. 7. Output torque of double-output ER fluid clutch

Total diameter	192 mm
Total heigth (include output shaft)	225 mm
Num. of disks (for inner / outer parts each)	4 (ER fluid layer: 8)
Diameter of disk	155 mm
Thickness of disk	1 mm
Disk gap	1 mm

Table 2. Specification of double-output ER fluid clutch

5. Basic structure and property of Hybrid-PLEMO

5.1 Concept

The PLEMO has 2 controllable DOFs on a working plane and 1 passive DOF of the inclined angle of the working plane as shown in Figure 2. We defined this working space as a "Quasi-3-DOF Working Space". An operator grasps a handle on the end-effecter of its arm, watches visual information on a display and plays application software as rehabilitative trainings and evaluation tests.

In a previous report (Kikuchi T., et al., 2007), we used only ER brakes for its torque control. Therefore, its haptic control was passive. In a new type of haptic device developed in this research, we use ER actuators for its haptic control with the quasi-3-DOF mechanism mentioned above. At the same time, we adopt a switchable mechanism between active / passive modes by releasing / fixing rotation of input parts of the clutches. We named this new haptic devices "Hybrid-PLEMO". Figure 8 shows the Hybrid-PLEMO, and table 3 shows specifications of it.

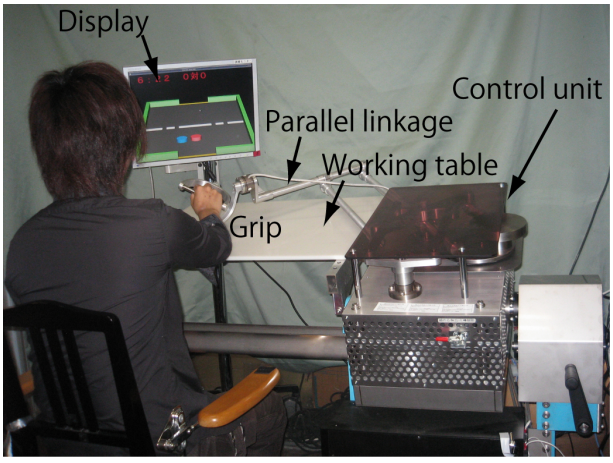


Fig. 8. Hybrid-PLEMO

Size	W0.6m x D0.5m x H1.0m
Motion region	W0.6m x D0.5m Adjustable angle of the inclination is -30~90deg
Maximum force	4kfg at end-effector
Num. of double-output ER clutch	2
Power of motor	40W

Table 3. Specification of Hybrid-PLEMO

5.2 Force control mechanism

Haptic force on the end-effector of the Hybrid PLEMO is controlled by a torque control unit with ER actuators mentioned above. In Figure 1, the motor is rotated by simply constant voltage (without feedback control) in order to assure high safety of the clutch-type actuator. Therefore, the rotation direction of the ER actuator is basically one way. We need two actuators for CW (clockwise) direction and CCW (counterclockwise) direction for one controllable DOF.

To realize two controllable DOFs of the Hybrid-PLEMO, we utilized two sets of double-output ER fluid clutch/brake devices described above. The one is rotated in CW direction. The other is rotated in CCW direction. Driving parts of the ER actuators are shown in Figure 9. As shown in this figure, both CW and CCW direction are generated by gears and one way rotation of a DC servo-motor. Each CW and CCW rotation is transmitted by belt-pulley system to the "ER Device1" and the "ER Device2". Additionally, when the motor is locked by a disk brake built in this system, each clutch works as a brake.

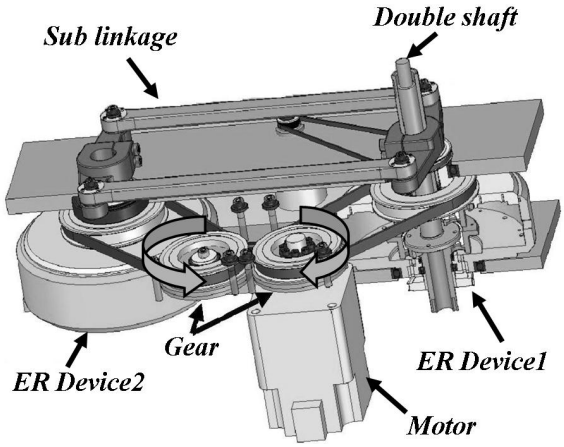


Fig. 9. Driving parts of ER actuators

A parallel linkage mechanism of the Hybrid-PLEMO is shown in Figure 10. the "ER Device1" and the "ER Device2" have a pair of two controllable shafts, which are a pair of an outer shaft and an inner shaft. Two outer shafts with opposite rotations are connected with the "Sub Link1". In same manner, two inner shafts are connected with the "Sub Link2". By

using sub-links, we can realize two controllable DOFs for haptic control. These two DOFs are converted to orthogonal two directions of the end-effector by using main parallel linkage which consists of the "Link1" and the "Link2".

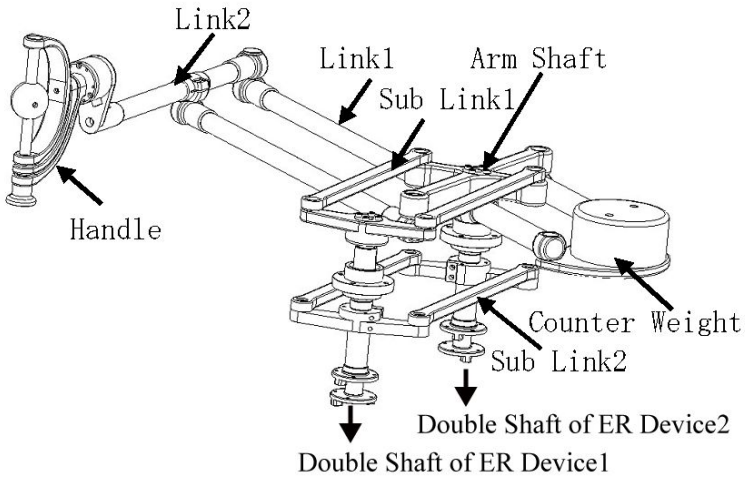


Fig. 10. Parallel linkage system for Hybrid PLEMO

5.3 Control system

Figure 11 shows the control system for the Hybrid-PLEMO. Absolute encoders (FA Coder, TS566N320, Tamagawa Seiki Inc., Japan, resolution: 17bits) measure the rotational angle of ER actuators or brakes. We can calculate the position and the velocity of the handle depending on each angle. A Digital / Input/ Output (DIO) board (PCI-2154C, Interface Inc., Japan) loads this information to a controller (personal computer). The handle includes a force sensor (OPFT-220N, Minebea Co. Ltd., Japan), and operating force is measured by this sensor. A potentiometer (CP-2F, Midori Precision Inc., Japan) measures the inclination of the worktable and the angle is loaded by an Analog/Digital (A/D) converter board (PCI-3165, Interface Inc., Japan, resolution: 16bits). The brake torque of the ER brake is controlled by applied voltage from high voltage amplifiers (High voltage amplifier, MAX-ELECTRONICS, Co. Ltd., Japan). A Digital/Analog (D/A) converter board (PCI-3338, Interface Inc., Japan, resolution: 12bits) outputs the reference signal to the amplifiers.

A controller is a personal computer (DOS/V), and an operating system (OS) is Vine Linux 3.2 and ART-Linux (kernel 2.4.31) as a real-time OS. Open-GL and Glut3.7 are used for the graphic libraries. A graphic process and a control process are executed by one PC. Multi-process programming is used to realize it. The control process is repeated every 1 ms exactly.

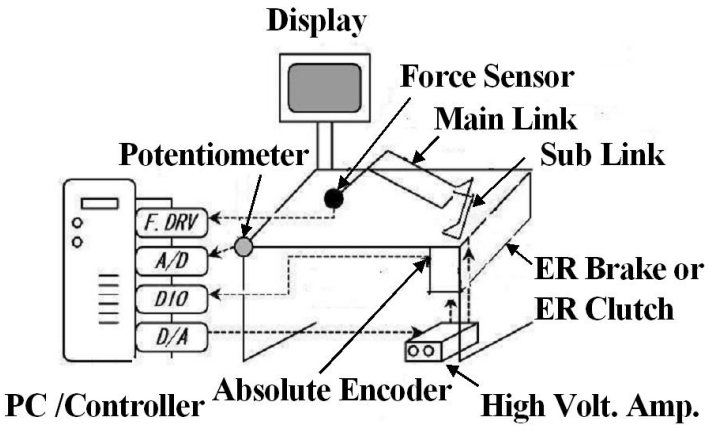


Fig. 11. Control System for Hybrid PLEMO

5.4 Experimental result for constant resistive force

To evaluate the performance of the haptic force of the Hybrid-PLEMO, constant resistive forces with active / passive haptic controls were measured. Figure 12 shows experimental results. A solid line represents force with the active haptic control, and a dashed line represents force with the passive haptic control. In these experiments, an operator manipulated a handle forward from an initial position. Initially, resistive forces were zero, but when the hand entered a friction area, resistive forces of constant 5N were instantly presented. As shown in this figure, haptic forces with both active/passive haptic methods were accurately controlled.

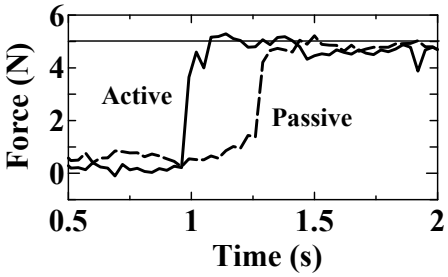


Fig. 12. Constant resistive force by passive / active mode

6. Reaching with active / passive force guidance

6.1 Motivation

Reaching motion is one of the most important tasks for rehabilitative trainings. In physical therapy (Voss D.E., et al., 1985) or occupational therapy (Trombly C.A., 1983), the reaching training is done without pulling patient's hand in a correct direction, but with giving resistance against the correct direction. This resistance force activates patient's intention to move their own arm by themselves.

In this section, we suggest two types of reaching programs with the Hybrid PLEMO system. These two programs are based on the reaching training with force guidance or resistance mentioned above. However, the methods of force feedback are different each other. The one utilizes active force feedback mode, and the other utilizes passive mode.

6.2 Method

For lab-level tests, examinees were healthy students (22~30 years old) in this section. Figure 13 shows a view displayed during this test. Initially, an examinee sets a handle position (red circle) on the starting position by manipulating the handle of the system. After that, subjects move their arm toward "Y" direction shown in Figure 13 along a target trajectory. However, a correct trajectory is not displayed in order to evaluate simply the effect of force information. Subjects have to find the correct trajectory from 5 choices (numbered from "0" to "4") based on only force information of the Hybrid-PLEMO. The aim of this program is to operate the handle, search target trajectories with only force information and trace it.

Fig. 13. View of the reaching program

Methods of force display in the active / passive modes are shown in Figure 14. In these figures, broken lines show the target trajectories. In active mode, the PLEMO system generates outgoing-vector forces of 5N from the target trajectory. On the other hand, in passive mode, the system generates a distribution of resistant forces (0N or 3N or 5N). The nearer to the target the hand position is, the stronger the resistance is. The start position was same (X = 0cm) in every experiments. The target trajectory was changed in a random manner.

In this experiments, "X" position of each trajectory (Num.0~ Num.4) is set as follows; X = -20 cm (Num.0), -10 cm (Num.1), 0 cm (Num.2), 10 cm (Num.3), 20 cm (Num.4).

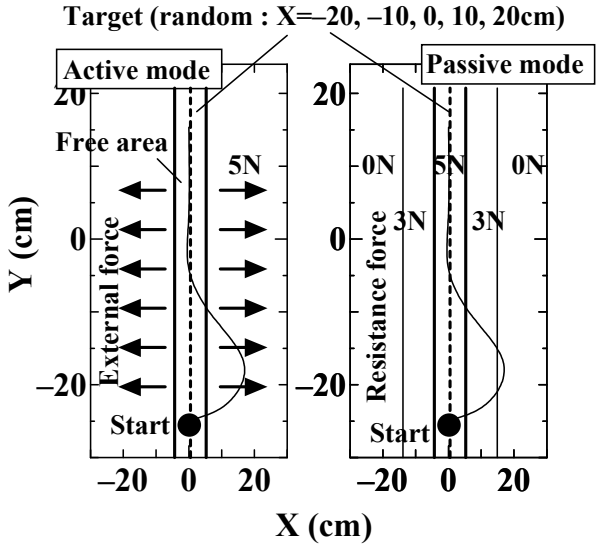


Fig. 14. Force feedback method for reaching program (Left: active mode, right: passive mode)

6.3 Result

The target trajectories were decided in the random manner. Figure 15 shows a set of experimental results for the same target position (Num.3; X = 10cm) with active/passive mode each. Broken lines show the target trajectories, and black dots show the starting positions of the handle. As shown in the left side of Figure 15, the operator can recognize the target position smoothly with active-type force guidance. On the other hand, as shown in the right side of Figure 15, it took more time to recognize the target position with passive-type force guidance than active mode.

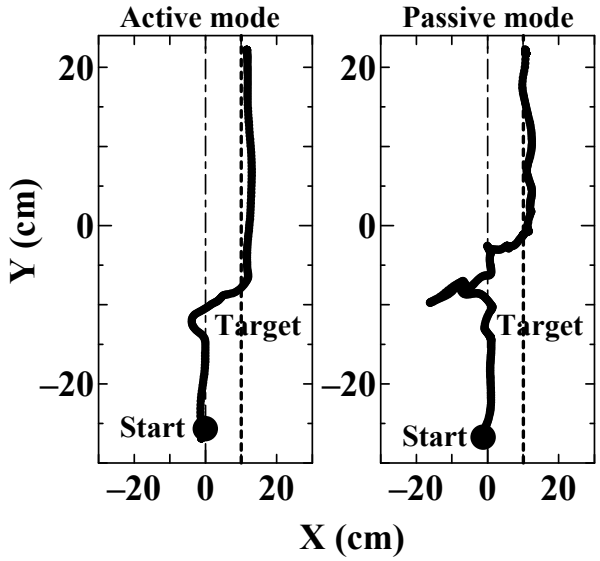


Fig. 15. Experimental Results (Left: active mode, right: passive mode)

6.4 Discussion

In order to evaluate the delay for searching target trajectories, we categorized a reaching path by two phases; a "search-phase" and a "reach-phase" shown in Figure 16. In this section, we defined the "search-phase" as a period from the beginning to the condition which meets both of following two equations,

$$\dot{Y} \equiv Vy > 10\,[cm/s], \tag{1}$$

and,

$$|\dot{X}| \equiv |Vx| < 10\,[cm/s]. \tag{2}$$

Therefore, above two equations mean conditions for the beginning of the "reach-phase". Equation (1) was defined to detect subject's intention for reaching. Equation (2) was defined to detect subject's intention for end of searching.

Experiments were conducted by 50 times each for the active / passive modes. We picked out 10 data whose target number was "1" from each data of the active / passive modes as shown in Table 4. The time of the "search-phase" in the active mode was longer than that in the passive mode.

The reason of this delay is thought that, in the passive mode, the operator needs more time to understand the distribution of force field with his own motion, and recognize correct direction toward the target.

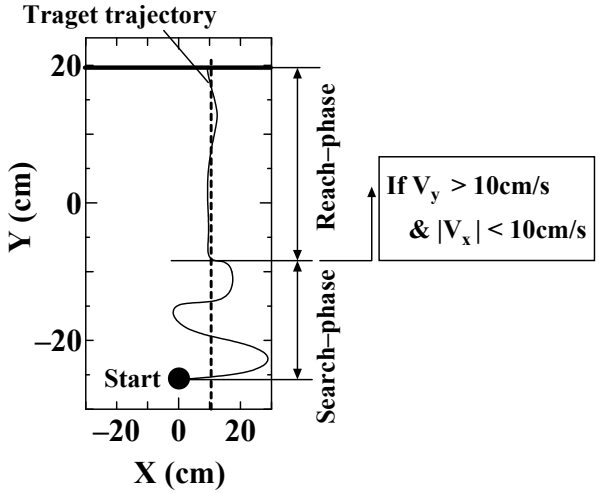


Fig. 16. Experimental Results (Left: active mode, right: passive mode)

	Search-Phase[s]	Reach-Phase[s]	Total time [s]
Active 1	2.059	0.562	2.621
Active 2	1.709	0.680	2.389
Active 3	1.639	0.789	2.428
Active 4	1.699	0.776	2.475
Active 5	1.429	0.952	2.381
Active 6	1.457	0.906	2.363
Active 7	1.309	0.988	2.297
Active 8	1.229	0.827	2.056
Active 9	1.269	0.757	2.026
Active 10	1.139	0.874	2.013
Ave.	**1.491**	**0.811**	**2.305**
SD	**0.282**	**0.128**	**0.207**
Passive 1	1.739	1.537	3.276
Passive 2	2.479	1.509	3.988
Passive 3	2.379	1.018	3.397
Passive 4	2.089	1.284	3.373
Passive 5	1.739	1.480	3.219
Passive 6	2.018	1.322	3.340
Passive 7	1.839	0.994	2.833
Passive 8	2.079	1.131	3.210
Passive 9	2.119	1.229	3.348
Passive 10	1.329	2.278	3.607
Ave.	**1.981**	**1.378**	**3.359**
SD	**0.335**	**0.370**	**0.295**

Table 4. Duration of "search-phase" and "reach-phase" in active / passive force feedback mode

7. Conclusion & Future works

This chapter described the development and evaluation of the "Hybrid-PLEMO", which is one of the rehabilitation systems for upper limbs with the quasi-3-DOF mechanism and the switchable mechanism between the active/passive modes. In the last part of this chapter, we compared the effect of the active / passive forces on the reaching tests for a healthy person. The duration of "search-phase" in the "passive mode" was longer than that in the "active mode".

In the future works, we will clarify the effects and roles of the active / passive force feedback in rehabilitative trainings with this system. It will be possible to measure the effect of rehabilitation with EEG or NIRS for the same subjects under the same environments by using this system.

8. Acknowledgments

This work was financially supported by a JAPAN Grant-in-Aid for Scientific Research. We thank Dr. Akio Inoue, who is the president of ER tec Co. Ltd. and contributed to the developments of all ER fluid devices.

9. References

Brunnstrom S. (1970). *Movement therapy in hemiplegia*, Harper & Row, New York

Furusho J. & Kikuchi T. (2007). A 3-D Rehabilitation System for Upper Limbs "EMUL", and a 6-DOF Rehabilitation System "Robotherapist", and Other Rehabilitation System with High Safety, *Rehabilitation Robotics* (Edited by Sashi S Kommu), Chapter 8, I-Tech Education and Publishing, pp.115-136

Furusho J., Koyanagi K., et al. (2005). Development of a 3-D Rehabilitation System for Upper Limbs Using ER Actuators in a NEDO Project, International Journal of Modern Physics B, Vol. 19, No. 7-9, pp.1591-1597

Furusho J. & Sakaguchi M. (1999) New actuators using ER fluid and their applications to force display devices in virtual reality and medical treatments, *International journal of Modern Physics B*, vol.13, no.14, 15 & 16, pp.2151-2159

Furusho J., Shichi N., et al. (2006). Development of a 6-DOF Force Display System with High Safety and its Application to rehabilitation, *Proceedings of the 2006 IEEE International Conference on Mechanism and Automation*, pp. 962-967

Johanna H, van der Lee JH, Wagenaar RC, Lankhorst GJ, Vogelaar TW, Deville WL, Bouter LM. (1999). Forced use of the upper extremity in chronic stroke patients: results from a single-blind randomized clinical trial, *Stroke*, Vol.30, No.11, pp.2369-2375

Kikuchi T., Jin Y., Fukushima K., Akai H. & Furusho J. (2008). "Hybrid-PLEMO", Rehabilitation system for upper limbs with Active / Passive Force Feedback mode, *Proceedings of the 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp.1973-1976

Kikuchi T., Xinghao H., et al. (2007). Quasi-3-DOF Rehabilitation System for Upper Limbs: Its Force-Feedback Mechanism and Software for Rehabilitation, *Proceedings of IEEE International Conference on Rehabilitation Robotics 2007*, pp.24-27

Krebs H.I., Volpe B.T., et al. (2000). Increasing productivity and quality of care: Robot-aided neuron rehabilitation, *Journal of Rehabilitation Research and Development*, Vol.37, No.6, pp.639-652

Liepert J, Storch P, Fritsch A, Weiller C (2000). Motor cortex disinhibition in acute stroke, *Clin Neurophysiol.*, Vol.111, No.4, pp.671-676

Loureiro R.C.V. & Harwin W.S. (2007) Reach & Grasp Therapy: Design and Control of a 9-DOF Robotic Neuro-rehabilitation System, *Proceedigs on the IEEE 10th International Conference on Rehabilitation Robotics*, pp. 757-763

Lum P.S., Burgar C.G. & Shor P.C. (2004). Evidence for improved muscle activation patterns after retraining of reaching movements with the MIME robotic system in subjects with post-stroke hemiparesis, *Proceedings on the IEEE Transactions on Neural Systems and Rehabilitation Engineering*, Vol.12, pp.184-194

Miyai I., Yagura H., et al. (2002). Premotor Cortex Is Involved in Restoration of Gait in Stroke, *Annals of Neurology*, Vol.52, No.2, pp.188-194

Munir S., Tognetti L. & Book W.J. (1999). Experimental Evaluation of a New Braking System for Use in Passive Haptic Displays, *Proceedings of the American Control Conference*, pp.4456-4460

Nef T., Mihelj M., Kiefer G., Perndl C., Muller R. & Reiner R. (2007). ARMin - Exoskeleton for arm therapy in stroke patients, *Proceedings on the IEEE 10th International Conference on Rehabilitation Robotics*, pp. 68-74

Ozawa T., Kikuchi T., et al. (2009). Study on Clinical Evaluation for Stroke Patients with "PLEMO", Rehabilitation System for Upper Limbs, *Proceedings of IEEE International Conference on Rehabilitation Robotics 2009*, in press

Perry J.C., Rosen J. & Burns S. (2007). Upper-Limb Powered Exoskeleton Design, *IEEE Transactions on Mechatronics*, vol.12, pp.408-417

Plautz EJ, Milliken GW, Nudo RJ. (2000). Effects of repetitive motor training on movement representations in adult squirrel monkeys: role of use versus learning, *Neurobiol Learn Mem*, Vol.74, No.1, pp.27-55

Trombly C.A. (1983), *Occupational Therapy for Physical Dysfunction (Second Edition)*.

Voss D.E., Ionta M. K. & Myers B. J. (1985). *Proprioceptive Neuromuscular Facilitation (Third edition)*.

Winslow, W.M. (1949). *Journal of Applied Physics*, Vol.20, pp.1137-1140

Wolbrecht E.T., Leavitt J., Reinkensmeyer D.J. & Bobrow J.E. (2006). Control of a pneumatic orthosis for upper extremity stroke rahabilitation, *Proceedings on the IEEE Engineering in Medicine and Biology Conference*, pp. 2687-2693

Wolf SL, Lecraw DE, Barton LA, Jann BB (1989). Forced use of hemiplegic upper extremities to reverse the effect of learned non-use among chronic stroke and head-injured patients, *Exp Neurol*, Vol.104, No.2, pp.125-132

Zhang L.Q., Park H.S. & Ren Y. (2007). Developing An Intelligent Robotic Arm for Stroke Rehabilitation, *Proceedings of the IEEE 10th International Conference on Rehabilitation Robotics*, pp. 984-994

20

Fractional-Order Models for the Input Impedance of the Respiratory System

Clara Ionescu[1], Robin De Keyser[1], Kristine Desager[2] and Eric Derom[3]
[1]*Ghent University, Department of Electrical energy, Systems and Automation, Technologiepark 913, Gent 9052, Belgium*
[2]*University Hospital Antwerp, Department of Pulmonary Medicine, Campus Drie Eiken, D.T.428, Universiteitsplein 1, 2610 Wilrijk, Belgium*
[3]*Ghent University Hospital, Department of Respiratory Medicine, De Pintelaan 185, Gent 9000, Belgium*

1. Introduction

Thanks to the technological advances, complex mathematical tools have been enabled for general use and applications in the field of biomedical engineering. Moving from fiddler's paradise of integer order models for biological systems, the modern scientist has vast possibilities to explore new horizons in the novel generalized order modeling concepts (Losa *et al*, 2005).

One of these novel concepts in modeling and identification is that of fractals, self-similarity in geometrical structures (Mandelbrot, 1983). Although originally applied in mathematics and chemistry, the signal processing community introduced the concept of fractional order modelling in several areas (Eke *et al.*, 2002; Losa *et al*, 2005; Jesus *et al*, 2008). A perfect example of fractal structure is that of the lungs (Mandelbrot, 1983; Weibel, 2005) and of the circulatory system (Gabrys *et al*, 2004). Regarding the lungs, there exist observations to support the claim that dependence exists between the viscoelasticity and the air-flow properties in the presence of airway mucus with disease. It is also agreed that fractional orders appear intrinsically in viscoelastic materials (i.e. soft lung tissue, soft arterial tissue, polymers) (Suki *et al.,* 1994; Adolfsson *et al.,* 2005; Craiem & Armentano, 2007). When characterization of the respiratory tree is envisaged, the mechanical properties are captured by measuring the input impedance, which gives insight upon airway and tissue resistance and compliance (Oostveen *et al*, 2003; Ionescu & De Keyser, 2008; Ionescu *et al*, 2009b).

This chapter will discuss the use of fractional order (FO) models for characterizing the input impedance of the human respiratory system in relation to its fractal structure. Given a brief summary of our previously gathered insight in the intrinsic properties of human respiratory tree, we introduce here several competitive models for characterizing the total input impedance. A comparison with the well-inherited fractional order model from the specialized literature and recently published hot-stone articles, will situate our results within the overall research on this challenging topic.

2. Materials and Methods

2.1 Subjects

The first group evaluated in this study consists of volunteers without a history of respiratory disease, whose lung function tests were performed in our laboratory and Table 1 presents their biometric parameters, whereas a detailed analysis on their respiratory impedance parameters can be found in (Ionescu *et al*., 2009a).

HEALTHY	male (n=15)	female (n=8)
Age (yrs)	23±0.7	23±1.3
Height (m)	1.76±0.062	1.68±0.032
Weight (kg)	73±5.1	63±2.8

Table 1. Biometric parameters of the healthy subjects; values are presented as mean±SD (SD: standard deviation).

A second group consists of patients under observation at the "Leon Danielo" Hospital in Cluj-Napoca, Romania, pulmonary division, diagnosed with COPD (Chronic Obstructive Pulmonary Disease). The latter group of patients consisted of former coal miners from the Petrosani area in Romania. Table 2 presents the corresponding biometric and spirometric parameters of the COPD group (Ionescu & De Keyser, 2009c).

COPD	male (n=21)
Age (yrs)	58±9
Height (m)	1.81±0.08
Weight (kg)	76±4.8
VC % pred	86±6.7
FEV_1 % pred	43±9

Table 2. Biometric and spirometric parameters of the COPD patients. Values are presented as mean±SD; % pred: predicted according to the asymptomatic males of the present study; VC: vital capacity; FEV_1: forced expiratory volume in one second.

2.2 Measurement Procedure

In this study, the Forced Oscillations Technique (FOT) non-invasive lung function test was applied (Oostveen *et al*., 2003). Air-pressure variations P, with respect to the atmospheric pressure and corresponding air-flow Q during the FOT lung function test can be measured either at the mouth of the patient, either endotracheal, either at body surface (Northrop, 2002). If the impedance is measured at the mouth of the patient, then it is called *input impedance*. In the case when the measurements are done across the body surface, this is then called *transfer impedance*. Using electrical analogy, where the P corresponds to voltage and Q corresponds to current, the respiratory impedance Z_r can be defined as their spectral (frequency domain) ratio relationship (Daröczy & Hantos, 1982). The present study is restricted to measurements of input respiratory impedance, that is, P and Q are measured at the mouth of the patient with reference to the atmospheric pressure (Oostveen *et al*., 2003).

Typically, the resulting impedance is a frequency dependent complex representation of mechanical properties and defines a real part R_{rs} – called *resistance* – and an imaginary part X_{rs} – called *reactance*. The real part describes the dissipative mechanical properties, whereas the imaginary part is related to the energy storage capacity and determined by both elastic and inertive properties.

The subject is connected to the typical setup from figure 1 via a mouthpiece, suitably designed to avoid flow leakage at the mouth and dental resistance artefact. The oscillatory flow U(t) in most recent FOT devices is generated by a loudspeaker (LS) connected to a chamber (Birch *et al.*, 2001). The LS is driven by a power amplifier fed with the oscillating signal generated by a computer. The movement of the LS cone generates a pressure oscillation inside the chamber, which is applied to the patient's respiratory system by means of a tube connecting the LS chamber and the bacterial filter (bf). A side opening (BT) of the main tubing allows the patient to decrease dead space re-breathing. Ideally, this bias tube will exhibit high impedance at the excitation frequencies to avoid the loss of power from the LS pressure chamber. It is advisory that during the measurements, the patient should wear a nose clip and keep the cheeks firmly supported to reduce the artefact of upper airway shunt. Pressure and flow are measured at the mouthpiece, respectively by means of **i)** a pressure transducer (PT) and **ii)** a pneumotachograph (PN) plus a differential pressure transducer (PT). The FOT pressure signal should be kept within a range of a peak-to-peak size of 0.1-0.3 kPa, in order to ensure optimality, patient comfort and stay within a narrow range in order to assume linearity (Desager *et al.*, 1997). Averaged measurements from 3-5 technically acceptable tests should be taken into consideration for further signal processing. Typical recorded signals are depicted in figure 1-B.

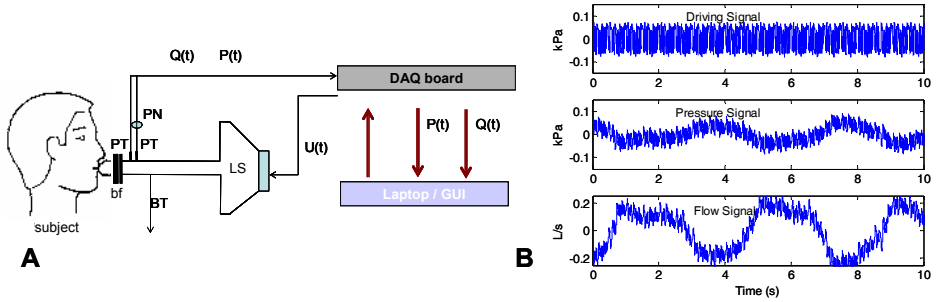


Fig. 1. (**A**) A schematic overview of the FOT measurement setup. (**B**) Typical measured signals from one subject: oscillatory driving pressure; trans-respiratory pressure and air-flow. The breathing of the patient (low frequency) can be observed superimposed on the multisine signals. See text for symbol explanations.

Several alternative setups for FOT have been developed during the last decades, such as the head plethysmograph (Govaerts *et al*, 1994), infants (Desager *et al*, 1991), or replacing the pneumotachograph by a known tube impedance (Franken *et al*, 1981).

The most common periodic excitation is the sinusoidal excitation; this signal is of great interest when frequency-domain identification is targeted, providing unbiased estimates. Besides, it allows direct interpretation of the mechanical load and poses a high signal-to-

noise ratio. The drawback is that only one point in the frequency domain is excited, so the information is not sufficient to assess the mechanical properties of the lungs. In order to avoid this drawback, *multi-sine waves* are applied to excite the system over the desired range of frequencies, in one experiment. The limitation, however, is that the amplitude (power) of the signal at a specific frequency decreases, due to the fact that the overall power spectrum of the multi-sine must be kept within the linearity, safety and comfort range for the patient. This constraint leads to limitations in the peak-to-peak amplitude of the FOT signal, i.e. between 0.1-0.3kPa (Oostveen *et al*., 2003; Van De Woestijne *et al*., 1994). In this study, we consider applying multi-sine oscillations in the frequency range: 4-48Hz. An overview of other type of signals is given in (Oostveen *et al*., 2003; Ionescu *et al*., 2009a).

3. Theoretical Background

The following sections provide the theoretical principles and experimental results underpinning the proposed fractional-order models.

3.1 Respiratory Input Impedance

One of the most common non-parametric representations of the *input impedance* Z_r is obtained assuming no correlation between the breathing and the excitation signal. This is done in the conditions that the breathing and the oscillations at the mouth of the patient are superimposed (Daröczi & Hantos, 1982), thus breathing can be considered as noise in this identification task. Apart from the errors introduced by the linear assumptions, the spectral representation of Z_r is a fast, simple and fairly reliable evaluation. The algorithm for estimating Z_r can be summarized starting from:

$$P(s) = Z_r(s)Q(s) + U_r(s) \tag{1}$$

where s denotes the Laplace operator. If the excitation signal is designed to be uncorrelated with the breathing of the patient and correlation analysis applied to the measured signals, one can estimate the respiratory impedance as:

$$\widehat{Z_r}(j\omega) = \frac{S_{PU_g}(j\omega)}{S_{QU_g}(j\omega)} \tag{2}$$

where $S_{ij}(j\omega)$ denotes the cross-correlation spectra between the various input-output signals, ω is the angular frequency and $j = (-1)^{1/2}$ and $U_g(t)$ denotes the excitation signal generated by the loudspeaker (multisine). From the point of view of the forced oscillatory experiment, the signal components of respiratory origin, (U_r) have to be regarded as pure noise for the identification task. However, to fulfill this condition it is necessary that:

i) the test signal U_g is designed such that it is not correlated with the normal respiratory breathing signal U_r and
ii) the conversion from voltage to pressure oscillation follows a linear relationship.

By definition, the modulus $|Z_r|$ is a measure of the total mechanical load of the respiratory system at the respective oscillation frequencies. The phase of respiratory impedance Φ_r is defined as the phase lag between $P(t)$ and $Q(t)$ and it is computed as the ratio between the time lag and the oscillation period T: $\Phi_r = 360 \times \Delta t / T$. The frequency where Φ_r =0 is called the *resonance frequency* and it depends on the balance between the different kind of mechanical properties (elastic, inertial). This then allows for differentiating between healthy and pathologic cases, since the resonance frequency changes significantly from typically 8Hz for a healthy adult to 14Hz for a patient with mild airway obstruction and from 20Hz onward in cases of severe obstruction.

According to (Pasker *et al.*, 1997), the real (*Rrs*) and imaginary (*Xrs*) parts of the impedance can be predicted from their biometric data as given below:

Female

Rrs0f=-0.4300*h+0.00165*w-0.00070*a+0.9312 (RSD=0.0619)
Rrs1f=0.01176*h-0.000106*w-0.000045*a-0.00817 (RSD=0.00256)
Xrs0f=0.2487*h-0.001700*w-0.00053*a-0.2158 (RSD=0.0406)

Male

Rrs0m=-0.2454*h+0.001564*w-0.00055*a+0.5919 (RSD=0.0493)
Rrs1m=0.01176*h-0.000106*w-0.000045*a-0.00817 (RSD=0.00197)
Xrs0m=0.2487*h-0.001700*w-0.00053*a-0.2158 (RSD=0.0306)

where *a* denotes age in *yrs*, *h* denotes height in *m*, *w* denotes weight in *kg* and *RSD* is the residual standard deviation. The *0f* and *1f* coefficients are related to the *E* and *D* coefficients respectively, resulting from fitting the polynomial given by:

$$R_{rs} \ (\text{or } X_{rs}) = Df + E \tag{3}$$

to the real or imaginary data sets, with *D, E* identified constants and *f* the oscillatory frequency, in this case between 4-48Hz. Confidence intervals of the identified values for 95% were calculated from RSDx1,96.

With the real (*Rrs*) and imaginary (*Xrs*) parts of the impedance from (2), parametric identification can be performed and the model parameters estimated using a nonlinear least squares optimization algorithm, by use of the MatLab® function *lsqnonlin*. The optimization algorithm is a subspace trust region method and is based on the interior-reflective Newton method described in (Coleman *et al*, 1996). In this application, the lower bounds were set to 0 (model parameters cannot have negative values) and no upper bounds. The optimization stopped either when a high number of iterations reached 100*nr. of variables, or a termination tolerance value of 1*e*-8. In all cases we obtained a correlation coefficient between data and model estimates above 80%.

Along with the corresponding model estimates, the error on the real and imaginary part respectively and the total error between the real patient impedance and the model estimated impedance are calculated according to the formulae:

$$E_R = \sqrt{\frac{1}{N}\sum_{1}^{N}(r-\hat{r})^2}; \quad E_X = \sqrt{\frac{1}{N}\sum_{1}^{N}(x-\hat{x})^2}; \qquad E_{Total} = \sqrt{E_R^2 + E_X^2} \tag{4}$$

with *r* denoting values in the real part of the impedance, x denoting values in the imaginary part of the impedance and *N* the total number of data samples (*N*=23).

3.2 Data Validation and Statistical Analysis

Since the healthy groups consisted of volunteers with no record of respiratory pathology, a validation was done using the reference values given in the previous section. All patients were within the 95% confidence intervals (Ionescu *et al*. 2009a). An one-way analysis of variance (*t* student test) was used to compare model parameters among the two groups: healthy and COPD. Results were considered significant at $p \le 0.05$. Further on, model parameters from the separate groups were evaluated using boxplots. The boxplot is typically a box and whisker plot for each column of the matrix, whereas here the columns are respectively the parameters for healthy group and for COPD group. The box has lines at the lower quartile, median, and upper quartile values. The whiskers are lines extending from each end of the box to show the extent of the rest of the data. Outliers are data with values beyond the ends of the whiskers (see, e.g. figure 5).

3.3 The Origins of Fractional-Order Models

The concept of fractional order (or non-integer order) systems refers to those dynamical systems whose parameters contain arbitrary order derivatives and/or integrals. The FO derivatives and integrals are tools of the Fractional Calculus theory (Podlubny, 1999). The dynamical systems whose model can be represented in a natural way by FO parameters, exhibit specific features:

- viscoelasticity;
- diffusion;
- fractal structure.

Viscoelasticity has been shown to be the origin of the appearance of FO models in polymers (from the Greek: *poly* - many and *meros* -- parts) (Adolfsson *et al*, 2005) and other resembling biological tissues (Suki *et al*, 1994). Diffusion phenomena have been intensively studied in the field of chemistry and dielectrics (Oustaloup, 1995; Machado & Jesus, 2004) and only recently in biology (Losa *et al*, 2005). Finally, it has been shown that the fractal structure leads to a FO model in some geometrical structures and electrical networks (Oustaloup, 1995; Ramus *et al*, 2002). Although viscoelastic and diffusive properties were intensively investigated in the respiratory system, the fractal structure was surprisingly ignored. Probably one of the reasons is that the respiratory system is not symmetric, thus failing to satisfy one of the conditions for being a typical fractal structure. Nonetheless, some degree of recurrence has been recognized in the airway generation models (Mandelbrot, 1983; Weibel, 2005), but simulation models including this specific property are not yet available.

The literature reports the existence of a FO model based on viscoelastic assumptions (Hantos *et al*, 1992). The model provides an expression for the input impedance (measured at the mouth of the patient) as:

$$Z_r(s) = \frac{P(s)}{Q(s)} = R + Ls + \frac{1}{Cs^{\beta}} \tag{5}$$

with P - pressure in kPa; Q - flow in l/s; Z_r - the impedance; R - airway resistance kPa-s/l; L - inductance kPa-s^2/l; C - capacitance in l/kPa; $0 \leq \beta \leq 1$ a fractional order and s the Laplace operator. This model, although broadly used by researchers and providing valid parameter values in several groups of patients, is unable to characterize increasing resistance with frequency at frequencies higher than 30Hz (Ionescu & De Keyser, 2008). As a result of the limitation, researchers have performed and reported studies in which the parameter values are meaningless from a physiological standpoint, i.e. negative (Hantos *et al*, 1982; Suki *et al*, 1997).

In the same context of characterizing viscoelasticity, (Suki *et al,* 1994) established possible scenarios for the origin of viscoelastic behaviour in the lung parenchyma. The authors apply the model from (5) in the form:

$$Z_r(s) = \frac{1}{Cs^{\beta}} \tag{6}$$

in which the real part denotes elastance and the imaginary part the viscance of the tissue. Similarly to (5), this model was also referred as the *constant-phase model* because the phase is independent with frequency, implying a frequency-independent mechanical efficiency (Oustaloup, 1995).

In his paper, Suki *et al.* (1994) recognizes five classes of systems admitting power-law relaxation or constant-phase impedance:

- **Class 1**: *systems with nonlinear constitutive equations*; a nonlinear differential equation may have a At^{-n} solution to a step input. Indeed, lung tissue behaves nonlinearly, but this is not the primary mechanism for having constant-phase behaviour, since the forced oscillations are applied with small amplitude to the mouth of the patient to ensure linearity.
- **Class 2**: *systems in which the coefficients of the constitutive differential equations are time-varying*; the linear dependence of the pressure-volume curves in logarithmic time scale does not support this assumption.
- **Class 3**: *systems in which there is a continuous distribution of time constants that are solutions to integral equations*. By aid of Kelvin bodies and an appropriate distribution function of their time constants, a linear model has been able to capture the hysteresis loop of the lungs, capturing the relaxation function decreasing linearly with the logarithm of time. This is a class of systems which may be successful in acknowledging the origin of the constant-phase behaviour, but there is no micro--structural basis.
- **Class 4**: *complex dynamic systems exhibiting self-similar properties (fractals)*. This class is based on the fact that the scale-invariant behaviour is ubiquitous in nature and the stress relaxation is the result of the rich dynamic interactions of tissue strips independent of their individual properties. Although interesting, this theory does not give a straightforward explanation for the appearance of constant-phase behaviour.

- **Class 5**: *systems with input-output relationships including fractional order equations*; borrowed from fractional calculus theory, several tools were used to describe viscoelasticity by means of fractional order differential equations (Suki *et al*, 1997; Craiem & Armentano, 2007).

Referring to the specific application of respiratory mechanics, Classes 3-5 are most likely to characterize the properties of lung parenchyma. The work presented in this chapter deals primarily with concepts from Class 4, but addresses also several items from Class 5.
Hitherto, the research community focused on the aspect of viscoelasticity in soft biological tissues. The other property of the lungs which can be related to fractional-order equations is diffusion and some papers discuss this aspect (Losa *et al*, 2005).
Surprisingly, the plain fractal-like geometry of the airways has been completely ignored throughout the decades. Perhaps one of the reasons for this lack of interest from the research community is that the lungs are not perfectly symmetric and even more, the quasi-symmetry disappears completely with disease. While most biologic processes could be described by models based on power law behaviour and quantified by a single characteristic parameter (one fractional order), the necessity arises to introduce multi-fractal models. The study presented in this chapter will address both the single- and multi-fractal models for the respiratory impedance.

3.4 Some Concepts from Fractional Calculus

The fractional calculus is a generalization of integration and derivation to non-integer (fractional) order operators: D_t^n. Several definitions of this operator are available; see e.g. (Podlubny, 1999). All of them generalize the standard differential--integral operator in two main groups:

(a) they become the standard differential--integral operator of any order when *n* is an integer;
(b) the Laplace transform of the operator D_t^n is s^n (provided zero initial conditions), and hence the frequency characteristic of this operator is $(j\omega)^n$.

The Laplace transform for integral and derivative order *n* are, respectively:

$$\begin{aligned} L\left\{D_t^{-n} f(t)\right\} &= s^{-n} F(s) \\ L\left\{D_t^{n} f(t)\right\} &= s^{n} F(s) \end{aligned} \tag{7}$$

where $F(s) = L\left\{f(t)\right\}$ and *s* is the Laplace variable. The Fourier transform can be obtained by replacing *s* with *jω* in the Laplace transform and the equivalent frequency-domain expressions are:

$$\begin{aligned} \frac{1}{(j\omega)^n} &= \frac{1}{\omega^n}\left(\cos\frac{\pi}{2} + j\sin\frac{\pi}{2}\right)^{-n} = \frac{1}{\omega^n}\left(\cos\frac{n\pi}{2} - j\sin\frac{n\pi}{2}\right) \\ (j\omega)^n &= \omega^n\left(\cos\frac{\pi}{2} + j\sin\frac{\pi}{2}\right)^{n} = \omega^n\left(\cos\frac{n\pi}{2} + j\sin\frac{n\pi}{2}\right) \end{aligned} \tag{8}$$

Thus, the modulus and the argument of the FO terms are given by:

$$\begin{aligned} |M|_{dB} &= 20\log\left|(j\omega)^{\mp n}\right| = \mp 20n\log|\omega| \\ \angle|_{rad} &= \arg\left((j\omega)^{\mp n}\right) = \mp n\frac{\pi}{2} \end{aligned} \tag{9}$$

resulting in:

- a Nyquist contour of a beeline with a slope $\mp n\frac{\pi}{2}$ anticlockwise rotation of the real axis in the complex plane around the origin according to variation of the FO value n;
- Magnitude (dB *vs* log-frequency): straight line with a slope of $\mp 20n$ passing through 0dB for $\omega = 1$ (see figure 2);
- Phase (rad *vs* log-frequency): horizontal line, thus independent with frequency, with value $\mp n\frac{\pi}{2}$ (see figure 2).

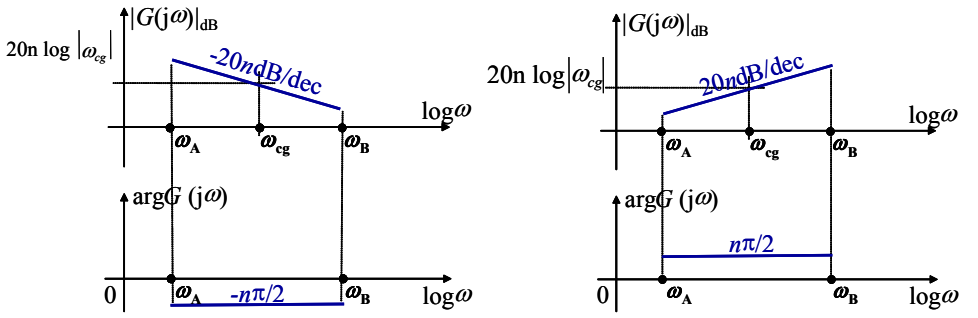


Fig. 2. Sketch representation of the FO integral and derivate operators in frequency domain, by means of the Bode plots (Magnitude, Phase)

4. Proposed Fractional-Order Models for Evaluation

Several attempts have been made to obtain an electrical equivalent of the respiratory tree (Farre *et al*, 1989; Diong *et al*, 2007). However, the models reported hitherto are an approximation of the tree rather than a precise formulation, and do not preserve the intrinsic geometry. Thanks to technological advances, information on airway radius, length and thickness is available. Our previous work provides a simulator for the dichotomous airway structure and validates the appearance of the fractional-order behaviour in the impedance, arising intrinsically from the fractal structure of the lungs (Ionescu *et al*, 2009d; Muntean *et al*, 2009).

The FO terms in the impedance models arise either from intrinsic viscoelastic (at low frequencies), either intrinsic fractal structure of the airway tree. Viscoelastic properties and FO models have been already presented and validated for the respiratory tree in (Suki *et al*, 1994; Hantos *et al*, 1992). Furthermore, based on electrical analogy to ladder networks and corresponding theoretical background (Oustaloup, 1995), we have shown that the FO properties expressed in (9) arise from the fractal structure of the airway tree (Ionescu *et al*, 2009d). A similar analysis was conducted for other systems, e.g. the arterial tree (Gabrys *et al*, 2004) and the stemming leaf-structure (Ionescu & Machado, *in pres*s).
It is interesting to compare the models existing in literature with some similar candidate models, on the same range of frequencies 4-48Hz, which is commonly evaluated in clinical trials (Oostveen *et al*., 2003). We therefore propose the following FO models, in order of complexity. The first model, from here-on referred to as FO1, is based on (5):

$$Z_{FO1}(s) = R_1 + \frac{1}{C_1 s^{\beta_1}} \tag{10}$$

with R_1 the resistance, C_1 the capacitance and $0 \leq \beta \leq 1$. This model was initially used at frequencies<5Hz, whereas the effect of the inductance is negligible. Therefore, evaluating such model at 4-48Hz frequency interval, one may expect low performance results.
The second model proposed here, referred to as FO2, is in fact (5), by adding the inductance term:

$$Z_{FO2}(s) = R_2 + L_2 s + \frac{1}{C_2 s^{\beta_2}} \tag{11}$$

As frequency increases, the real part of the impedance may increase its value in some patients. The real part of (11) depends on the resistance R_2 and the capacitance term C_2 (the latter being frequency-dependent). As frequency increases, the real part of the term in C_2 decreases, therefore unable to characterize correctly the impedance. However, if the model is evaluated in a frequency range in which the real part of the impedance is decreasing with frequency, the model has good performance.
The third model (FO3) proposed for evaluation is based on (11), but the FO term is in the inductance and not in the capacitance:

$$Z_{FO3}(s) = R_3 + L_3 s^{\alpha_3} + \frac{1}{C_3 s} \tag{12}$$

This model will provide good results for patients with increasing impedance values with frequency in the real part, since the term in L_3 is directly proportional to frequency.
The last model proposed for evaluation in this chapter is based on our previous work (Ionescu *et al*., 2009a; Ionescu & De Keyser, 2009b), i.e. the multi-fractal model:

$$Z_{FO4}(s) = L_4 s^{\alpha_4} + \frac{1}{C_4 s^{\beta_4}} \tag{13}$$

which does not contain the resistance term R_4. The decision to eliminate this term was taken as a result of previous investigation, showing that the identified values are negligible (Ionescu & De Keyser, 2009b). Physically, the term in R_4 is not necessary, since the theory of fractional order appearance in ladder networks shows that the effects of R_4 are indirectly captured in the values of the FO terms and FO coefficients (Oustaloup, 1995; Ionescu *et al.*, 2009d).

5. Results

The complex impedance values for the healthy and COPD patients have been obtained using (2) and they are depicted in figure 3 below. It can be observed that the healthy group has a resonant frequency (zero crossing in the imaginary part) around 8 Hz, whereas the COPD group around 16 Hz. This shows that the lung parenchyma in COPD patients is less elastic than in healthy subjects. The real part denotes mainly the mechanical resistance of the lung tissue, which is generally increased in the COPD group, resulting in higher work of breathing. Also, the resistance at low frequencies is much increased in the COPD group, suggesting increased damping of the lung parenchyma (viscoelasticity is mainly analyzed at low frequencies). In both cases, the real part of the impedance decreases with frequency until 10-15Hz, and the low frequency interval becomes more significant with pathology.

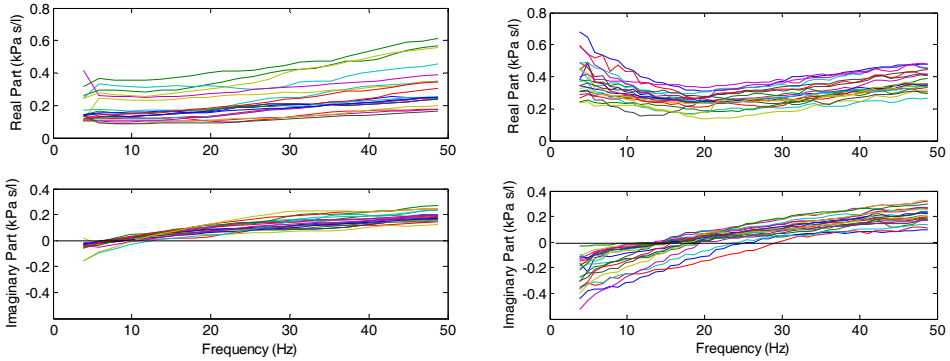


Fig. 3. Impedance plots for the healthy (left) and for the COPD (right) groups.

Next, the models from (10)-(13) are fitted to the complex impedance values. The results for model identification are obtained using the System Identification Toolbox within the MatLab platform, i.e. the *lsqnonlin* optimization function. The estimated parameter values along with the real, imaginary and total error values are given in Table 3 for the healthy subjects, respectively in Table 4 for the COPD patients.

Healthy	FO1	FO2	FO3	FO4
R	0.22±0.09	0.22±0.09	0.06±0.06	-
L	-	0.0007±0.0001	0.029±0.0302	0.0374±0.031
α	-	-	0.48±0.13	0.43±0.1
1/C	0	1.36±0.98	3.52±1.67	2.02±1.47
β	0.99±0.0006	0.99±0.01	-	0.79±0.16
$\mathbf{E_R}$	0.05±0.02	0.05±0.02	0.02±0.01	0.02±0.01
$\mathbf{E_X}$	0.12±0.02	0.01±0.006	0.015±0.0063	0.013±0.006
$\mathbf{E_T}$	0.13±0.03	0.05±0.02	0.02±0.01	0.02±0.01

Table 3. Estimated model parameters and modelling errors for the healthy group

COPD	FO1	FO2	FO3	FO4
R	0.18±0.08	0.26±0.08	0.27±0.05	-
L	-	0.0009±0.0001	0.0021±0.0014	0.015±0.008
α	-	-	0.87±0.1	0.59±0.09
1/C	1.73±3.32	5.20±2.49	8.9±3.79	2.94±1.54
β	0.18±0.36	0.83±0.16	-	0.52±0.11
$\mathbf{E_R}$	0.05±0.01	0.04±0.01	0.04±0.01	0.03±0.01
$\mathbf{E_X}$	0.14±0.02	0.02±0.006	0.03±0.011	0.02±0.006
$\mathbf{E_T}$	0.15±0.02	0.05±0.01	0.05±0.02	0.04±0.01

Table 4. Estimated model parameters and modelling errors for the COPD group

From the model parameters, one can calculate the tissue damping $G = \frac{1}{C}\cos\left(\frac{\pi}{2}\beta\right)$ and tissue elastance $H = \frac{1}{C}\sin\left(\frac{\pi}{2}\beta\right)$ (Hantos *et al*, 1992) and tissue histeresivity $\eta = G/H$ (Fredberg and Stamenovic, 1989). The relationship with (5) is found if the terms in *C* are re-written as:

$$\frac{1}{C\omega^{\beta}}\cos\left(\frac{\pi}{2}\beta\right) - j\frac{1}{C\omega^{\beta}}\sin\left(\frac{\pi}{2}\beta\right) = \frac{G - jH}{\omega^{\beta}} \tag{14}$$

From Tables 3 and 4 one may observe that the model FO4 gives the smallest total error. This is due to the fact that two FO terms are present in the model structure, allowing both a decrease and increase in values of the impedance with frequency. The FO2 model is the most commonly employed in clinical studies, with similar errors for the imaginary part, but higher error in the real part of the impedance than the FO4 model. The underlying reason is that the model can only capture a decrease in real part values of the impedance with frequency, whereas some patients may present an increase. As an example, figure 4 presents such a case, where one can visually compare the performance of the FO2 and FO4 models.

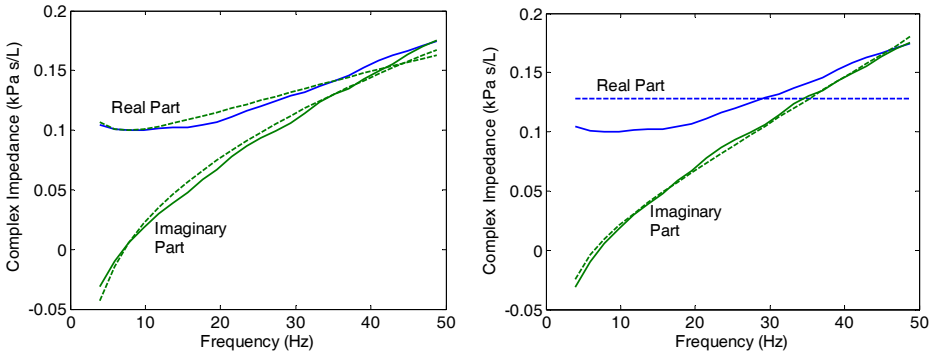


Fig. 4. A healthy subject data evaluated with FO4 (left) and with FO2 (right); continuous lines denote the measured impedance and dashed lines denote the identified impedance.

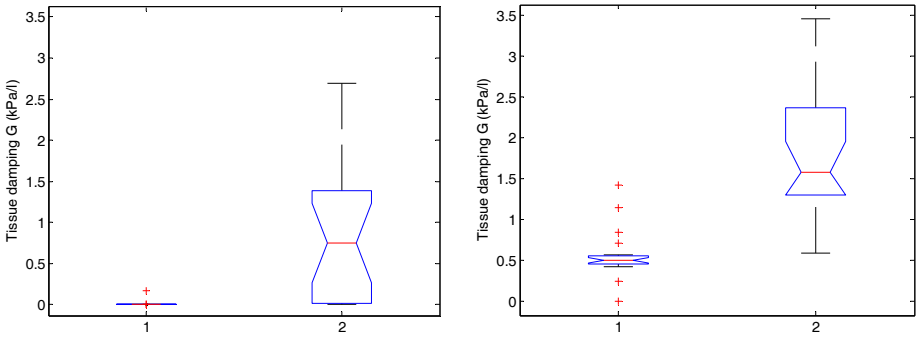


Fig. 5. Tissue damping G (kPa/l) with FO2, p<3e^{-5} (left) and with FO4, p<1e^{-8} (right); 1: Healthy subjects and 2: COPD patients.

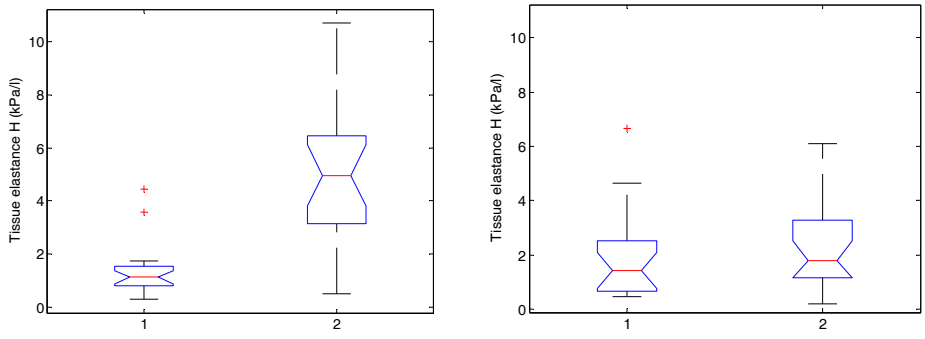


Fig. 6. Tissue elastance H (kPa/l) with FO2, p<0.0012 (left) and with FO4, p<0.0004 (right); 1: Healthy subjects and 2: COPD patients.

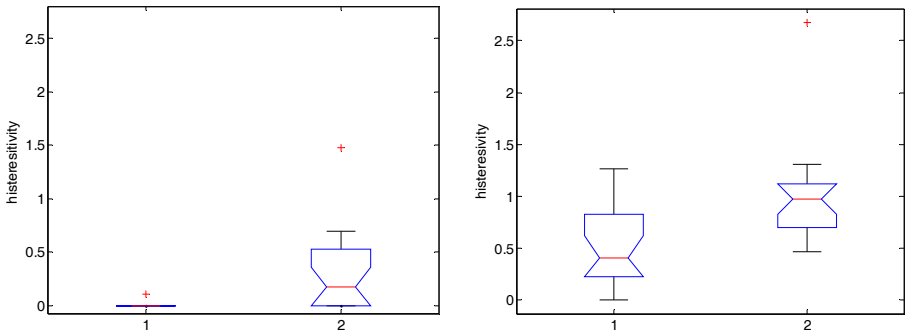


Fig. 7. Tissue hysteresivity η with FO2, p<0.0012 (left) and with FO4, p<0.0004 (right); 1: Healthy subjects and 2: COPD patients.

Figures 5, 6 and 7 depict the boxplots for the FO2 and FO4 for the tissue damping G, tissue elastance H and histeresivity η. Due to the fact that FO2 has higher errors in fitting the impedance values, the results are no further discussed. Although a similarity exists between the values given by the two models, the discussion will be focused on the results obtained using FO4.

Because FO are natural solutions in dielectric materials, it is interesting to look at the permittivity property of respiratory tissues. In electric engineering, it is common to relate permittivity to a material's ability to transmit (or *permit*) an electric field. By electrical analogy, changes in trans-respiratory pressure relate to voltage difference, and changes in air-flow relate to electrical current flows. When analyzing the permittivity index, one may refer to an increased permittivity when the same amount of air-displacement is achieved with smaller pressure difference. In other words, the hysteresivity coefficient incorporates this property for the capacitor, that is, the COPD group has an increased capacitance, justified by the pathology of the disease. Many alveolar walls are lost by emphysematous lung destruction, the lungs become so loose and floppy that a small change in pressure is enough to maintain a large volume, thus the lungs in COPD are highly compliant (elastic) (Barnes, 2000; Hogg, 2004; Derom *et al.*, 2007). The complex permittivity has a real part, related to the stored energy within the medium and an imaginary part related to the dissipation (or loss) of energy within the medium. The imaginary part of permittivity corresponds to:

$$\varepsilon = L \sin\left(\frac{\pi}{2}\alpha\right) \tag{15}$$

If the values are positive, (15) denotes the absorption loss. In COPD, due to the sparseness of the lung tissue, the air-flow in the alveoli is low, thus a low level of energy absorption is observed in figure 8. In healthy subjects, due to increased alveolar surface, higher levels of energy absorption are present, thus increased permittivity.

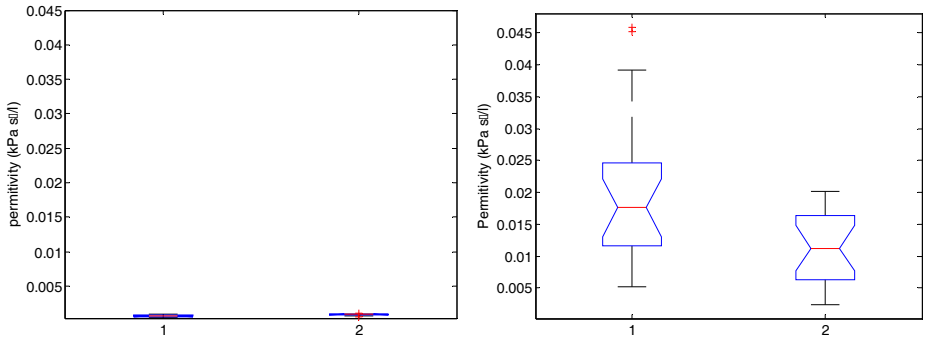


Fig. 8. Boxplots for the computed permittivity index ε in the FO2, p<0.0081 (left) and in FO4, p<0.0002 (right), in the two groups; 1: Healthy subjects and 2: COPD patients.

Another significant observation is that in general, FO4 identified more statistically significant model parameter values than FO2. In figures 5-7 FO4 parameters had identified similar variations between healthy and COPD groups. However, in figure 8, one can observe that FO4 identified a more realistic variation between healthy and COPD groups, i.e. a decreased permitivity index in COPD than in healthy.

6. Discussion

Tissue destruction (emphysema, COPD) and changes in air-space size and tissue elasticity are matched with changes in model parameters when compared to the healthy group. The physiological effects of chronic emphysema are extremely varied, depending on the severity of the disease and on the relative degree of bronchiolar obstruction versus lung parenchymal destruction (Barnes, 2000). Firstly, the bronchiolar obstruction greatly increases airway resistance and results in increased work of breathing. It is especially difficult for the person to move air through the bronchioles during expiration because the compressive force on the outside of the lung not only compresses the alveoli but also compresses the bronchioles, which further increase their resistance to expiration. This might explain the decreased values for inertance (air mass acceleration), captured by the values of L in the FO4. Secondly, the marked loss of lung parenchyma greatly decreases the elastin cross-links, resulting in loss of attachments (Hogg, 2004). The latter can be directly related to the fractional-order of compliance, which generally expresses the capability of a medium to propagate mechanical properties (Suki *et al.*, 1994).

The damping factor is a material parameter reflecting the capacity for energy absorption. In materials similar to polymers, as lung tissue properties are very much alike polymers, damping is mostly caused by viscoelasticity, i.e. the strain response lagging behind the applied stresses (Suki *et al.*, 1994;1997). In both FO models, the exponent β governs the degree of the frequency dependence of tissue resistance and tissue elastance. The increased lung elastance $1/C$ (stiffness) in COPD results in higher values of tissue damping and tissue elastance, as observed in Figures 5 and 6. The loss of lung parenchyma (empty spaced lung), consisting of collagen and elastin, both of which are responsible for characterizing lung elasticity, is the leading cause of increased elastance in COPD. The hysteresivity coefficient η

introduced in (Fredberg & Stamenovic, 1989) is G/H in this model representation. Given the results observed in Figure 7, it is possible to distinguish between tissue changes from healthy to COPD case. Since pathology of COPD involves significant variations between inspiratory and expiratory air-flow, an increase in the hysteresivity coefficient η reflects increased inhomogeneities and structural changes in the lungs.

It is difficult to provide a fair comparison between the values reported in this study and the ones reported previously for tissue damping and elastance. Firstly, such studies have been previously performed from excised lung measurements and invasive procedures (Suki *et al*. 1997; Brewer *et al.,* 2003; Ito *et al*., 2007), which related these coefficients with *transfer impedance* instead of *input impedance*. The measurement location is therefore important to determine mechanical properties of lungs. The data reported in our study, has been derived from non-invasive measurements at the mouth of the patients, therefore including upper airway properties. Secondly, the previously reported studies were made either on animal data (Hantos *et al*., 1992a;1992b; Brewer *et al*., 2003; Ito *et al*., 2007), either on other lung pathologies (Kaczka *et al.,* 1999).

Another interesting aspect to note is that in the normal lung, the airways and lung parenchyma are interdependent, with airway caliber monotonically increasing with lung volume. In emphysematous lung, the caliber of small airways changes less than in the normal lung (defining compliant properties) and peripheral airway resistance may increase with increasing lung volume. At this point, the notion of space competition has been introduced (Hogg, 2004), hypothesizing that enlarged emphysematous air spaces would compress the adjacent small airways, according to a nonlinear behavior. Therefore, the compression would be significantly higher at higher volumes rather than at low volumes, resulting in blunting or even reversing the airway caliber changes during lung inflation. This mechanism would therefore explain the significantly marked changes in model parameters in tissue hysteresivity depicted in figure 7. It would be interesting to notice that since small airway walls are collapsing, resulting in limited peripheral flow, it also leads to a reduction of airway depths. A correlation between such airway depths reduction in the diseased lung and model's non-integer orders might give insight on the progress of the disease in the lung.

The main limitation of the present study is that both model structures and their corresponding parameter values are valid strictly within the specified frequency interval 4-48Hz. Nonetheless, since only one resonant frequency is measured and is the closest to the nominal breathing frequencies of the respiratory system, we do not seek to develop model structures valid over larger frequency range. Moreover, it has been previously shown that one model cannot capture the respiratory impedance over frequency intervals which include more than one resonant frequency (Farré *et al*., 1989). A second limitation arises from the parameters of the constant-phase models. The fractional-order operators are difficult to handle numerically. The concept of modeling using non-integer order Laplace (e.g. $s^{\alpha}, \frac{1}{s^{\beta}}$) is rather new in practical applications and has not reached the maturity of integer-order system modeling. This concept has been borrowed from mathematics and chemistry applications to model biological signals and systems only very recently. Advances in technology and computation have enabled this topic in the latter decennia and it has captured the interest of researchers. Although the parameters are intuitively related to

pathophysiology of respiratory mechanics, the structural interpretation of the fractional-orders is in its early age.

Viscoelastic properties in lung parenchyma has been assessed in both animal and human tissue strips (Suki *et al*., 1994) and correlated to fractional-order terms. A relation between these fractional-orders and structural changes in airways and lung tissue has not been found (e.g. airway remodeling). In this line of thought, the mechanical properties of resistance, inertance and compliance have been derived from airway geometry and morphology (i.e. airway radius, thickness, cartilage percent, length, etc) (Ionescu *et al*., 2009b). These parameters have been employed in a recurrent structure of healthy lungs using analogue representation of ladder networks (Ionescu *et al*., 2009d). In the latter contribution, the appearance of a phase-lock (phase-constancy) is shown, supporting the argument that it represents an intrinsic property (Oustaloup, 1995). Its correlation to changes in airway morphology is an ongoing research matter. Experimental studies on various groups of patients (*e.g.* asthma *versus* COPD) to investigate a possible classification strategy for the parameters of this proposed model between various degrees of airway obstruction and lung abnormalities may also offer interesting information upon the sensitivity of model parameters.

7. Conclusions

This chapter presents a short overview on the properties of lung parenchyma in relation to fractional order models for respiratory input impedance. Based on available model structures from literature and our recent investigations, four fractional order models are compared on two sets of impedance data: healthy and COPD (Chronic Obstructive Pulmonary Disease). The results show that the two models broadly used in the clinical studies and reported in the specialized literature are suitable for frequencies lower than 15Hz. However, when a higher range of frequencies is envisaged, two fractional orders in the model structure are necessary, in order to capture the frequency dependence of the real part in the complex respiratory impedance. Since the real part may both decrease and increase within the evaluated frequency interval, there is need for both fractional order derivative and fractional order integral parameters.
The multi-fractal model proposed in this chapter provides statistically significant values between the healthy and COPD groups. Further investigations are planned in order to evaluate if the model is able to discriminate between various pathologies (e.g. asthma, cystic fibrosis and COPD).

Acknowledgements
C. Ionescu gratefully acknowledges the students who volunteered to perform lung function testing in our laboratory, and the technical assistance provided at University of Pharmacology and Medicine -"Leon Daniello" Cluj, Romania. This work was financially supported by the UGent-BOF grant nr. B/07380/02.

8. References

Adolfsson K., Enelund M., Olsson P., (2005), On the fractional order model of viscoelasticity, *Mechanics of Time-dependent materials,* Springer, 9, 15–34

Barnes P.J., (2000), Chronic Obstructive Pulmonary Disease, *NEJM Medical Progress*, 343(4), pp. 269-280

Birch M, MacLeod D, Levine M, (2001) An analogue instrument for the measurement of respiratory impedance using the forced oscillation technique, *Phys Meas*, 22, pp. 323-339

Brewer K., Sakai H., Alencar A., Majumdar A., Arold S., Lutchen K., Ingenito E., Suki B., (2003), Lung and alveolar wall elastic and hysteretic behaviour in rats: effects of in vivo elastase, *J. Applied Physiology*, 95(5), pp. 1926-1936

Craiem D., Armentano R., (2007) A fractional derivative model to describe arterial viscoelasticity, *Biorheology*, 44, pp. 251—263

Coleman, T.F. and Y. Li, (1996), An interior trust region approach for nonlinear minimization subject to bounds, *SIAM Journal on Optimization*, **6**, 418-445

Daröczy B, Hantos Z, (1982) An improved forced oscillatory estimation of respiratory impedance, *Int J Bio-Medical Computing*, **13**, pp. 221-235

Derom E., Strandgarden K., Schelfhout V, Borgstrom L, Pauwels R. (2007), Lung deposition and efficacy of inhaled formoterol in patients with moderate to severe COPD, *Respiratory Medicine*, 101, pp. 1931-1941

Desager K, Buhr W, Willemen M, (1991), Measurement of total respiratory impedance in infants by the forced oscillation technique*, J Applied Physiology*, 71, pp. 770-776

Desager D, Cauberghs M, Van De Woestijne K, (1997) Two point calibration procedure of the forced oscillation technique, *Med. Biol. Eng. Comput.,* 35, pp. 561-569

Diong B, Nazeran H., Nava P., Goldman M., (2007), Modelling human respiratory impedance, *IEEE Engineering in Medicine and Biology*, 26(1), pp. 48-55

Eke, A., Herman, P., Kocsis, L., Kozak, L., (2002) Fractal characterization of complexity in temporal physiological signals, *Physiol Meas,* 23, pp. R1-R38

Farré R, Peslin R, Oostveen E, Suki B, Duvivier C, Navajas D, (1989) Human respiratory impedance from 8 to 256 Hz corrected for upper airway shunt, *J Applied Physiology*, 67, pp. 1973-1981

Franken H., Clement J, Caubergs M, Van de Woestijne K, (1981) Oscillating flow of a viscous compressible fluid through a rigid tube, *IEEE Trans Biomed Eng*, 28, pp. 416-420

Fredberg J, Stamenovic D., (1989), On the imperfect elasticity of lung tissue, *J. Applied Physio*logy, 67:2408-2419

Gabrys, E., Rybaczuk, M., Kedzia, A., (2004) Fractal models of circulatory system. Symmetrical and asymmetrical approach comparison, *Chaos,Solitons and Fractals*, 24(3), pp. 707-715

Govaerts E, Cauberghs M, Demedts M, Van de Woestijne K, (1994) Head generator versus conventional technique in respiratory input impedance measuremenets, *Eur Resp Rev*, 4, pp. 143-149

Hantos Z., Daroczy B., Klebniczki J., Dombos K, Nagy S., (1982) Parameter estimation of transpulmonary mechanics by a nonlinear inertive model, *J Appl Physiol*, 52, pp 955--963

Hantos Z, Adamicza A, Govaerts E, Daroczy B., (1992) Mechanical Impedances of Lungs and Chest Wall in the Cat, *J. Applied Physiology*, 73(2), pp. 427-433

Hogg J. C., (2004), Pathophysiology of airflow limitation in chronic obstructive pulmonary disease, *Lancet*, **364**, pp. 709-21

Ionescu, C. & De Keyser, R. (2008). Parametric models for characterizing the respiratory input impedance. *Journal of Medical Engineering & Technology*, Taylor & Francis, 32(4), pp 315-324

Ionescu C., Desager K., De Keyser R., (2009a) Estimating respiratory mechanics with constant-phase models in healthy lungs from forced oscillations measurements, *Studia Universitatis Vasile Goldis Life Sciences Series, 19(1), pp. 123-132*

Ionescu C., Segers P., De Keyser R., (2009b) Mechanical properties of the respiratory system derived from morphologic insight, *IEEE Transactions on Biomedical Engineering*, April, 56(4), pp. 949-959

Ionescu C., De Keyser R., (2009c) Relations between Fractional Order Model Parameters and Lung Pathology in Chronic Obstructive Pulmonary Disease, *IEEE Transactions on Biomedical Engineering*, April, 56(4), pp. 978-987

Ionescu C., Oustaloup A., Levron F., De Keyser R., (2009d) "A model of the lungs based on fractal geometrical and structural properties", accepted contribution at the *15th IFAC Symposium on System Identification*, St. Malo, France, 6-9 July 2009

Ionescu C, Tenreiro-Machado J., (in press), Mechanical properties and impedance model for the branching network of the seiva system in the leaf of Hydrangea macrophylla, accepted for publication in *Nonlinear Dynamics*

Ito S., Lutchen K., Suki B., (2007), "Effects of heterogeneities on the partitioning of airway and tissue properties in mice", *J. Applied Physiology,* 102(3), pp. 859-869

Kaczka D., Ingenito E., Israel E., Lutchen K., (1999), "Airway and lung tissue mechanics in asthma: effects of albuterol", *Am J Respir Crit Care Med*, 159, pp. 169-178

Jesus I, Tenreiro-Machado J, Cuhna B., (2008), Fractional electrical impedances in botanical elements, Journal of Vibration and Control, 14, pp. 1389—1402

Losa G., Merlini D., Nonnenmacher T., Weibel E, (2005), *Fractals in Biology and Medicine*, vol.IV, Birkhauser Verlag, Basel.

Mandelbrot B. (1983) *The fractal geometry of nature*, NY: Freeman &Co

Machado, Tenreiro J., Jesus I., (2004), Suggestion from the Past?, *Fractional Calculus and Applied Analysis*, 7(4), pp. 403—407

Muntean I., Ionescu C., Nascu I., (2009) A simulator for the respiratory tree in healthy subjects derived from continued fraction expansions*, AIP Conference Proceedings vol. 1117: BICS 2008: Proceedings of the 1st International Conference on Bio-Inspired Computational Methods Used for Difficult Problems Solving: Development of Intelligent and Complex Systems, (*Eds): B. Iantovics, Enachescu C., F. Filip, ISBN: 978-0-7354-0654-4, *pp. 225-231*

Northrop R., (2002) *Non-invasive measurements and devices for diagnosis*, CRC Press

Oostveen, E., Macleod, D., Lorino, H., Farré, R., Hantos, Z., Desager, K., Marchal, F, (2003). The forced oscillation technique in clinical practice: methodology, recommendations and future developments, *Eur Respir J*, 22, pp 1026-1041

Oustaloup A. (1995) *La derivation non-entière* (in French), Hermes, Paris

Pasker H, Peeters M, Genet P, Nemery N, Van De Woestijne K., (1997) Short-term Ventilatory Effects in Workers Exposed to Fumes Containing Zinc Oxide: Comparison of Forced Oscillation Technique with Spirometry, *Eur. Respir. J.*, 10: pp. 523-1529

Podlubny, I. (1999). Fractional Differential Equations--*Mathematics in Sciences and Engineering*, vol. 198, Academic Press, ISBN 0125588402, New York.

Ramus-Serment M., Moreau X., Nouillant M, Oustaloup A., Levron F. (2002), Generalised approach on fractional response of fractal networks, *Chaos, Solitons and Fractals*, 14, pp. 479—488.

Suki, B., Barabasi, A.L., & Lutchen, K. (1994). Lung tissue viscoelasticity: a mathematical framework and its molecular basis. *J Applied Physiology,* 76, pp. 2749-2759

Suki B., Yuan H., Zhang Q., Lutchen K., (1997) Partitioning of lung tissue response and inhomogeneous airway constriction at the airway opening, *J Applied Physiology*, 82, pp. 1349--1359

Van De Woestijne K, Desager K, Duiverman E, Marshall F, (1994) Recommendations for measurement of respiratory input impedance by means of forced oscillation technique, *Eur Resp Rev*, 4, pp. 235-237

Weibel, E.R. (2005). Mandelbrot's fractals and the geometry of life: a tribute to Benoît Mandelbrot on his 80th birthday, in *Fractals in Biology and Medicine*, vol IV, Eds: Losa G., Merlini D., Nonnenmacher T., Weibel E.R., ISBN 9-783-76437-1722, Berlin: Birkhaüser, pp 3-16

Modelling of Oscillometric Blood Pressure Monitor – from white to black box models

Eduardo Pinheiro and Octavian Postolache
Instituto de Telecomunicações
Portugal

1. Introduction

Oscillometric blood pressure monitors (OBPMs) are a widespread medical device, increasingly used both in domicile and clinical measurements of blood pressure, replacing manual sphygmomanometers due to its simplicity of use and low price. A servo-based air pump, an electronic valve and the inflatable cuff are the main components of an OBPM, the nonlinear behaviour of the device emerges especially from this last element, in view of the fact that the cuff's expansion is constrained (Pinheiro, 2008).
The first sphygmomanometer developments and its final establishment, due to the works of Samuel von Basch, Scipione Riva-Rocci and Nicolai Korotkoff, are over a century old, but still are widely used by trained medical staff (Khan, 2006). In the Korotkoff sounds method, a stethoscope is used to auscultate the sounds produced by the brachial artery while the flow through it starts, after being occluded by the inflation of the cuff. The oscillometric technique is an alternative method which examines the shape of the pressure oscillations that the occluding cuff exhibits when the cuff's pressure diminishes from above systolic to below diastolic blood pressure (Geddes et al., 1982), and in recent times it has been increasingly applied (Pinheiro, 2008).
In the last decades, oscillometric blood pressure monitors have been employed as an indirect measurement of blood pressure, but have not been subject of deep investigation, and have been used as black-box systems, without explicit knowledge of their internal dynamics and features. Bibliography in this field is limited, (Drzewiecki et al., 1993) studied the cuff's mechanics while (Ursino & Cristalli, 1996) have concerned with biomechanical factors of the measurement, but both oblivious to the device's behaviour and performance.
The equations that govern both wrist-OBPM and arm-OPBM behaviour are the same, but wall compliances and other internal parameters assume diverse values, what may also happen between different devices of the same type. The knowledge of the relations ruling the internal dynamics of this instrument will help in the search for improvements in its measurement accuracy and in the device design, given that electronic controllers may be introduced to change the OBPM dynamics improving its sensibility. Moreover, since the OBPM makes discrete measurements of the blood pressure, the understanding of the device's characteristics and dynamics may allow taking a leap towards continuous blood pressure measurement using this inexpensive device.

Analyzing the OBPM, an insightful modelling effort is made to determine a white-box model, describing the dynamics involved in the OBPM during cuff compression and decompression and obtaining several non-ideal and nonlinear dynamics, using the results available on servomotors (Ogata, 2001) and compressible flows (Shapiro, 1953), obtained through electric, mechanic and thermodynamic principles. The approach taken was to divide the OBPM in two subsystems, the electromechanical, which receives electrical supply and outputs a torque in the crankshaft of the air pump, and the pneumatic subsystem, which establishes the evolution of the cuff pressure, separating the compression and decompression phases.
Subsequently blacker-box analysis is presented, in order to provide alternative models that require only the observation of the air pump's electric power dissipation, and pure identification methods to estimate a multiple local model structure. In this last approach the domain of operation was segmented in a number of operating regimes, identifying local models for each regime and fusing them using different interpolation functions thus providing better estimates and more flexibility in the system representation than a single global model (Murray-Smith & Johansen, 1997).

2. White-box model

The main dynamics that characterize the OBPM behaviour are the air pump's response to the command voltage, the air propagation in the device and the inflatable cuff mechanics.

2.1 Electromechanical section

An armature controlled dc servomotor coupled to a crankshaft that manages two cylinders that alternately compress the air are the components of the OBPM's air pump. The servomotor is controlled by V_a, the voltage applied to its armature circuit, while a constant magnetic flux is guaranteed. The armature-winding resistance is labelled R_a, the inductance L_a, and the current i_a, a depiction of the described command circuit is presented in Figure 1.

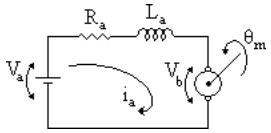


Fig. 1. Servomotor electrical control circuit

Due to the external magnetic field and the relative motion between the motor's armature, the back electromotive force, V_b, appears. At constant magnetic flux V_b is proportional to the motor's angular velocity, ω_m, being related through the back electromotive force constant of the motor, K_1, and with ω_m the derivative of θ_m, the angular displacement of the shaft in the motor, (1).

$$V_b(t) = K_1 \omega_m(t) \tag{1}$$

The current evolution in the circuit, (2), is obtained with Kirchhoff's laws.

$$L_a \frac{di_a(t)}{dt} + R_a i_a(t) + K_1 \omega_m(t) = V_a(t) \tag{2}$$

The transformation from electrical to mechanical energy is done relating the torque τ to the armature current, (3), where K_2 is the motor torque constant.

$$\tau(t) = K_2 i_a(t) \tag{3}$$

Regarding the mechanical coupling to the crankshaft, it will be considered that the servomotor and the crankshaft have moments of inertia J_m and J_c, rotate at angular velocities ω_m and ω_c, and have angular displacements of θ_m and θ_c respectively. The shaft coupling, the motor, and the crankshaft have non-homogeneous stiffness K_3 and viscous-friction b along the shaft (x-axis), Figure 2.

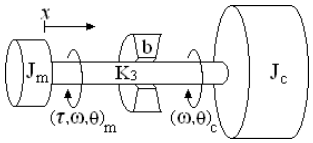


Fig. 2. Mechanical representation of the servomotor coupling to the crankshaft.

The torsion is intrinsically displacement-dependent and the rotational dissipation is velocity-dependent (Ljung & Glad, 1994), so, the equations of torque equilibrium will have to consider the velocity and stiffness in every point of the shaft to compute the torsion, regarding the friction along the shaft. This was dealt computing the product of the mean values of the friction and the angular velocity, which may be piecewise-defined functions. In (4) the angular velocity is defined as a function of time and location in the shaft, $\omega(t,x)$, with $\omega(t,m)$ matching $\omega_m(t)$ and $\omega(t,c)$ matching $\omega_c(t)$.

$$\begin{Bmatrix} J_m \dfrac{d\omega_m(t)}{dt} + \dfrac{\int_m^c b(x)dx \int_m^c \omega(t,x)dx}{(c-m)^2} + \int_m^c K_3(x)\omega(t,x)dx = \tau(t) \\ J_c \dfrac{d\omega_c(t)}{dt} + \dfrac{\int_c^m b(x)dx \int_c^m \omega(t,x)dx}{(c-m)^2} + \int_c^m K_3(x)\omega(t,x)dx = 0 \end{Bmatrix} \tag{4}$$

It should be noted that in the case of homogeneous rigidity the last term of the sum is simplified (5) just considering the angular displacements difference between θ_m and θ_c. Moreover, if the coupling between the inertias is perfectly inflexible, which is a good approximation if K_3 is very high, this term disappears.

$$\int_{m}^{c} K_3(x)\omega(t,x)dx = K_3\left[\theta_m(t) - \theta_c(t)\right] \tag{5}$$

Considering that the friction is applied in a single spatial point ($x = b$) and that the rigidity is homogeneous, the set of equations obtained, (4), is linearized to (6).

$$\left\{\begin{array}{l} J_m \dfrac{d\omega_m(t)}{dt} + \dfrac{K_3}{2}\left[\theta_m(t) - \theta_b(t)\right] = \tau(t) \\ b\dfrac{d\theta_b(t)}{dt} + \dfrac{K_3}{2}\left[\theta_b(t) - \theta_m(t)\right] + \dfrac{K_3}{2}\left[\theta_b(t) - \theta_c(t)\right] = 0 \\ J_c \dfrac{d\omega_c(t)}{dt} + \dfrac{K_3}{2}\left[\theta_c(t) - \theta_b(t)\right] = 0 \end{array}\right\} \tag{6}$$

2.2 Pneumatic section

The air pump output flows through a short piping system of circular cross section before entering in the cuff. The cylinders' output is generally composed of a number of orifices with very narrow diameter for example, three orifices with 0.5 mm which are linked through a minor connector to a plastic piping system, of about 5 mm internal diameter, conducing to the cuff. The modelling approach taken considers one-dimensional adiabatic flow, with friction in the ducts, regarding air as a perfect gas, and with the pneumatic connections represented by a converging-diverging nozzle, since the chamber-orifices passage is a contraction, succeed by a two-step expansion, first the passage to the pipes and next the arrival at the cuff.

The assumption of air as a perfect gas means that the specific heat is supposed constant and the relation $p = \rho RT / M$, is considered valid, with R the ideal gas constant, p and T its absolute pressure and temperature, M the gas molar mass and ρ its density. In view of the fact that at temperatures below 282 ºC the error of considering the specific heat constant is negligible, and that deviations from the perfect gas equation of state are also negligible at pressures below 50 atmospheres, the perfect gas approximation is found reasonable, (Shapiro, 1953).

The maximum velocity of the flow, v_{max}, may be determined considering the equation for adiabatic stagnation of a stream (7), where γ is the ratio of specific heats (isobaric over isochoric) and R the air constant, making the absolute temperature T null. It should be noticed that the deceleration's reversibility is not important since the stagnation temperature, T_0, will be the same.

$$v = \sqrt{\frac{2\gamma}{\gamma - 1} R\left(T_0 - T\right)} \tag{7}$$

Regarding the pressure, if the deceleration is irreversible the final pressure will be smaller than the isentropic stagnation pressure, p_0, which is function of the Mach number, M_a, the ratio of the flow velocity and the speed of sound, as seen in (8).

$$p_0 = p\left[1+\left(\frac{\gamma-1}{2}\right)M_a{}^2\right]^{\frac{\gamma}{\gamma-1}} \tag{8}$$

But these are very high limits, if one considers realistic γ, e.g. 1.4 of (Forster & Turney, 1986), even for very low temperature increases, the maximum velocity easily ascends at sonic values, which generates elevated stagnation pressures limits also.

Searching for tighter limits, it is possible to find the characteristics of the air pumps used in these applications. For instance, Koge KPM14A has an inflation time, from 0 to 300 mmHg, in a 100 cm^3 tank, of, about 7.5 seconds. Therefore, considering this inflation time representative, the mean volumetric flow is 13.333×10^{-6} m^3s^{-1} so, the mean air speed is 11.789 ms^{-1} in the three output orifices of the compression chamber, with 0.6 mm of diameter each. The most of the piping has 5 mm of internal diameter, reducing the mean speed to 0.170 ms^{-1}.

The Reynolds number of the flow, $\mathrm{Re} = \rho v D / \mu$, calculated in [20 ; 80] ^oC range to compensate heating of the fluid, considering air's dynamic viscosity μ and density ρ, at these temperatures, and the velocity v in both sections, with different diameter D, will cause the Reynolds number to be between 282 and 392 in the small orifices, and between 41 and 57 in the duct. Hence the Reynolds number is far from 2000, guaranteeing laminar flow in the orifices, even if the effective instantaneous speed achieves five times the mean speed calculated, and in the ducts even if the flow is 35 times faster.

Since the flow is laminar, the friction factor f may be calculated simply using $f = 16/\mathrm{Re}$. The use of the friction factor to represent the walls' shear stress, τ_w, according to $f = 2\tau_w / \rho v^2$, is correct if the flow is steady, but, in cases of velocity profile changes, f represents only an "apparent friction factor" since it also includes momentum-flux effects. In short pipes, which is clearly the case of the OBPM, the average apparent friction factor rises, (Shapiro, 1953) and (Goldwater & Fincham, 1981).

The air is fed into the 5 mm pipes from the three orifices of the compression chamber by an element of unimportant length, which will be assumed frictionless. Since the chamber leads to three 0.6 mm orifices converging to a 1 mm element, which introduces the flow in the 5 mm pipes, the piping profile is converging-diverging.

In view of the fact that the velocity would have to rise almost 29 times to produce sonic flow in the orifices, the flow is considered entirely subsonic, and this piece behaves as a conventional Venturi tube, introducing some losses in the flow (Benedict, 1980), with the flow rate being sensitive to the cuff pressure, what would not happen in the case of sonic or supersonic flow, where shock waves are present (Shapiro, 1953).

The effect of wall friction on fluid properties, considering one-dimensional (dx) adiabatic flow of a perfect gas in a duct with hydraulic diameter D and friction factor f, will rewrite the perfect gas, Mach number, energy, momentum, mass conservation, friction coefficient and isentropic stagnation pressure equations (Shapiro, 1953), creating the system of equations (9). The hydraulic diameter D changes along dx, and these changes must also be included in the model implementation.

$$\left\{\begin{array}{l}\dfrac{dp}{p}=-\dfrac{\gamma M_a{}^2\left[1+(\gamma-1)M_a{}^2\right]}{2\left(1-M_a{}^2\right)}4f\dfrac{dx}{D}\\ \dfrac{dM_a{}^2}{M_a{}^2}=\dfrac{\gamma M_a{}^2\left[2+(\gamma-1)M_a{}^2\right]}{2\left(1-M_a{}^2\right)}4f\dfrac{dx}{D}\\ \dfrac{dv}{v}=\dfrac{\gamma M_a{}^2}{2\left(1-M_a{}^2\right)}4f\dfrac{dx}{D}\\ \dfrac{dT}{T}=-\dfrac{\gamma(\gamma-1)M_a{}^4}{2\left(1-M_a{}^2\right)}4f\dfrac{dx}{D}\\ \dfrac{d\rho}{\rho}=-\dfrac{\gamma M_a{}^2}{2\left(1-M_a{}^2\right)}4f\dfrac{dx}{D}\\ \dfrac{dp_0}{p_0}=-\dfrac{\gamma M_a{}^2}{2}4f\dfrac{dx}{D}\end{array}\right\} \tag{9}$$

The inflatable cuff is an element whose mechanical performance is a determinant factor of the OBPM's response (Pinheiro, 2008). Due to the pressure-volume bond and since the constrictions to the cuff expansion introduce additional dynamics in the OBPM behaviour, the complete model must incorporate (10) the model of cuff's volume evolution with the pressure. It was followed (Ursino & Cristalli, 1996) line of thought, but disagreeing in some particular aspects, since it was considered cuff pressure perfectly equivalent to arm outer surface pressure, greatly reducing the number of biomechanical parameters involved (and their natural discrepancies when changing the subject's characteristics), and also, the ratio of specific heats γ was not considered constant, opposing to other, (Forster & Turney, 1986) and (Ursino & Cristalli, 1996), approaches.

$$\frac{1}{\gamma p_c{}^{1/\gamma}}q^{-1+1/\gamma}\frac{dq}{dt}-\frac{1}{\gamma p_c{}^{1+1/\gamma}}q^{1/\gamma}\frac{dp_c}{dt}=\frac{C_w}{p_c+p_w}\frac{dp_c}{dt} \tag{10}$$

In this equation, q represents the amount of air contained in the cuff, p_c is the cuff pressure (p after the total piping length) expressed in relative units, C_w is the wall compliance, $-p_w$ is the collapse pressure of the cuff internal wall (pressure at which the wall compliance goes infinite).

Finally, having characterized both fluid and structure equations, to complete this fluid-structure interaction model, coupling equations must be defined. One option is to consider fluid velocity inversely dependent on the crankshaft's inertia J_c, or alternatively, to consider that the velocity is dependent on the crankshaft's angular displacement θ_c.

This crankshaft-based coupling is justified taking into consideration the air pump operation cycle. The crankshaft is bicylindrical and each revolution makes the cylinders compress

once, since its construction is symmetrical, each revolution is the execution of the same movement cycle twice, and this cycle can be decomposed in forward (compression) and backward (recovery) movements, thus, the high frequency pulsatile air flow may have its velocity expressed depending only on θ_c or J_c, with an appropriate rational transformation. The modelling exercise is now complete, in the following section a greyer approach to subject is made, studying in more detail the relation of the crankshaft-related variables with the cuff pressure.

2.3 Greyer view – crankshaft load via power dissipation

A simplified way of modelling the mechanic-pneumatic connection will be to consider that all the dynamics of the flow and the inflatable cuff are manifested in the load of the servomotor. This way of thinking has the advantage of being assessed quite easily, by measuring the air pump's power dissipation, or the servomotor's vibrations using strain gages (Schicker & Wegener, 2002), with the latter requiring quite intrusive adjustments in the OBPM, while the first only requires secondary wire connections.

Given that the crankshaft operation cycle can be decomposed in two forward (compressions) and two backward (recoveries) movements, the inertia J_c may be expressed has a function dependent of θ_c, (11), to include the high-frequency dynamics previously described. However, since the dominant effect is unquestionably the filling of the cuff, J_c must be strongly bonded to the cuff pressure. Since it noticeable that the compression takes approximately $3\pi/4$ rad, and the decompression lasts for about $\pi/4$ rad, these are the key crankshaft's angular displacement values.

$$J_c(p,\theta_c)=\begin{cases} J_{comp}(p)\left|\sin\left(2\theta_c/3\right)\right|,\ \theta_c\in\left[0,\frac{3\pi}{4}\right]\cup\left[\pi,\frac{7\pi}{4}\right] \\ J_m(p)+J_{dec}(p)\left|\sin\left(2\theta_c-3\pi/2\right)\right|,\ \theta_c\in\left]\frac{3\pi}{4},\pi\right[\cup\left]\frac{7\pi}{4},2\pi\right[\end{cases} \tag{11}$$

The raise in J_c due to the cylinders' forward and backward movement is represented by the terms J_{comp} and J_{dec} correspondingly. The backward movement of the cylinders will add less inertia to J_c than the compression movement, and it is intuitive to suppose that both J_{comp} and J_{dec} will increase when the pressure in the cuff increases. Also, a minimum inertia J_m is added during decompression, since the inertia does not reduce to zero immediately after the compression ends. Subsequent Figure 3 shows the crankshaft's inertia estimative produced by (11) considering one cycle with a J_{comp} value of 0.15 kgm², J_{dec} valuing 0.025 kgm², and J_m 0.045 kgm².

Measurements made on a wrist-OBPM air pump, Koge KPM14A, registered an armature-winding resistance value of 3.9376 Ω and an inductance of 1.5893 mH, using an Agilent 4236B LCR meter (Pinheiro, 2008). Therefore, the implementation of a power measurement scheme based on a 0.111 Ω resistor in series with the supply circuit is innocuous to the OBPM's normal operation. The voltage in this resistor was acquired using a National Instruments DAQ Card 6024E data-acquisition board at a sampling rate of 100 kSamples/second. The power dissipation evolution obtained from these measurements is shown in Figure 4.

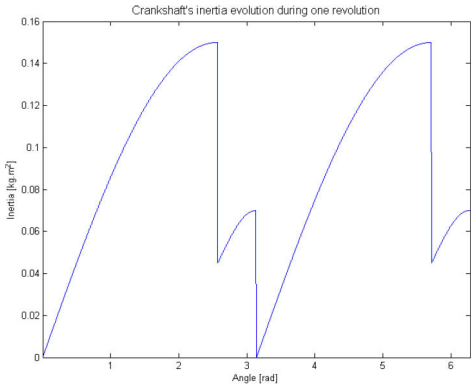


Fig. 3. Crankshaft inertia estimate characterization during one complete revolution, under J_{comp}, J_{dec}, J_m of 0.150, 0.025, and 0.045 kgm².

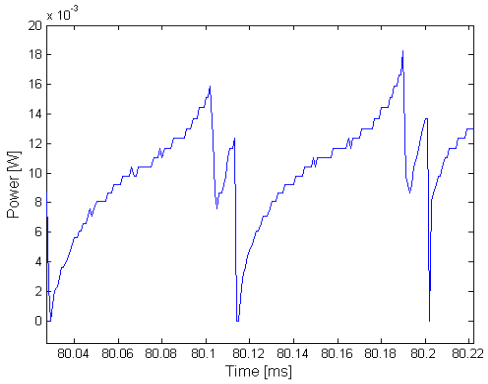


Fig. 4. Power dissipation in the air pump during one complete revolution of the crankshaft.

The abrupt dissipated power decreases after the local maximums are due to the conclusion of the forward and backward movements. The decompression conclusion practically leads to a zero power situation, while the compression conclusion is seen in previous Figure 4 to reduce the power to 50% of the maximum. The 50% proportion is approximately constant if the cuff pressure is below 20 centimetres of mercury column (cmHg), which is the nominal pressure range of OBPM's cuff. This means that in inertia terms, (11), J_m should be half of max{J_{comp}} calculated at the end of the compression.

The cuff pressure directly affects the terms J_{comp} and J_{dec}, since it is the variable ruling the effort of the air pump in each compression. Noticing that it is most important to measure the servomotor's power dissipation evolution and this high-frequency dynamic is not so significant, the curve in Figure 4 may be low-pass filtered in order to evaluate the power evolution once the air pumping changes the cuff pressure, instead of analysing every pump stroke.

To the acquisition hardware was added a Measurement Specialities 1451 pressure sensor, and it was implemented digitally a 3rd order Butterworth low-pass filter with 30 Hz cut-off

frequency. The results obtained are shown in Figure 5, where it is seen the power dissipation curves when compressing to the inflatable cuff, left, and to a constant volume reservoir with about the same capacity, right.

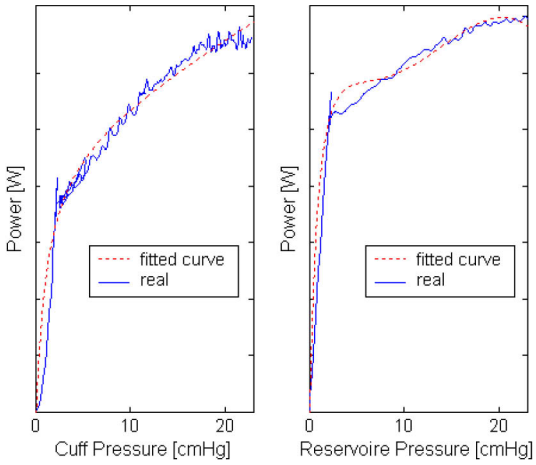


Fig. 5. Air pump's power dissipation dependence of cuff-pressure (blue) and approximating curve (red), when the air pump output is connected to the cuff (left) and to a constant-volume reservoir (right).

The air pump power dissipation, P, relation with the downstream pressure, p, was approximated by a rational function, (12), with coefficient of determination, R^2, of 0.984 when connected to the cuff and 0.967 when connected to the constant-volume reservoir, the a normalized root mean square deviation of the approximations was 2.17% and 2.26%, respectively. With these approximation functions, from the pressure measurements the power dissipation is calculated.

$$P(t) = \frac{a_4 p^4(t) + a_3 p^3(t) + a_2 p^2(t) + a_1 p(t) + a_0}{p(t) + b_0} \tag{12}$$

The inertia of the crankshaft J_c as been described as possible to be estimated from the power dissipation, this makes sense given that the major portion of the crankshaft's inertia is due to the cuff pressure, and the power-pressure relation has been established in (12). Thus, it will be assumed that low-pass filtering the crankshaft's inertia in a 30 Hz 3rd order Butterworth filter Ψ, makes it directly proportional to the power dissipation, (13), being K_4 the power-inertia conversion constant.

$$\Psi\left(J_c(t)\right) = K_4 P(t) \tag{13}$$

Aggregating the equations of the electromechanical section, with (12) and (13), and choosing state vector *X*, defined in (14), it is assembled a greyer and simpler space-state model of the OBPM.

$$X(t)^T = \left[\theta_c(t), \omega_c(t), \theta_m(t), \omega_m(t), i_a(t)\right]^T \tag{14}$$

It should be noticed that since power is the product of i_a the air pump's current (state variable) and V_a the voltage applied (input variable), both pressure and inertia can be estimated knowing only the power and applying (12) and (13) in that order.

2.4 Cuff decompression

The OBPM controls the air pump and the electronic valve in order to pressurize the cuff, until blood flow is cut off, and afterwards slowly reduces the cuff pressure, stopping the compression and letting the cuff's permanent leakage take effect, only opening the electronic valve to swiftly deplete the cuff when the blood pressure measurement is done. During the period while the air pump is stopped and the valve is closed, it is also necessary to evaluate the cuff pressure dynamic when the permanent leakage is the only influence.

In a constant-volume reservoir this dynamic is defined by an exponential decay, as seen in (Lyung & Glad, 1994), in this case, due to the expandability of the cuff, other parameters must infer in the exponential.

It were recorded twelve descents, by turning off the air pump at different pressures, p_{off}, from 9 to 20 cmHg, and then using the DAQ Card 6024E data-acquisition board at 1 kS/s to record the pressure fall curve. It was verified that the cuff pressure had an exponential decay, $p(t) = ae^{-bt}$, and that the exponential function parameters, a and b, were dependent on the pressure at which the inflation was stopped, p_{0ff}, as presented in (15), with corresponding coefficient of determination values of 0.980 and 0.877.

$$\left\{\begin{array}{l} a = 1.126 p_{off} - 3.35 \\ b = -0.3436 e^{-0.0732 p_{off}} \end{array}\right\} \tag{15}$$

3. Black-box model

The main nonlinearities involved in the OBPM operation refer to the dynamics of the air compression and flow, and the limitations to the cuff expansion. The black-box model approach will define a single-input single-output relation between the voltage supply to the OBPM's air pump, V_a, and the cuff pressure, p, applying system identification procedures (Ljung, 1999).

The OBPM, in its normal operating cycle, keeps the electronic valve always closed, by powering it, until cuff depletion is desired, and controls the air pump to compress the cuff until blood stops flowing. To do this, a National Instruments USB-6008 multifunction I/O board was used, with an acquisition rate and generation rate of 50 S/s, together with appropriate circuitry to allow supervision of the device's elements.

The identification procedure consisted of randomly deciding to power the air pump using white noise, but keeping the pressure in a defined range to maintain the device in the operating regime to be identified, thus in case of pressure range surpass the power was shut down and *vice versa*.
It was found by experience that the command voltage should be updated at a rate lower than the 50 S/s used to read the pressure sensor value, to permit the visualization of the effects of the voltage change, and so it was used a 10 S/s output update rate.
Besides connecting the air piping output to the wrist inflatable cuff, the OBPM identification tests were replied in the constant-volume reservoir, to observe the differences in the results due to the reservoir expansion.

3.1 At 5 regimes

The inflatable cuff's maximum nominal pressure is 19.5 cmHg, but, since an hypertensive person may have a systolic blood pressure higher than this limit, the maximum pressure considered was 22 cmHg, and the divisions were: [0 ; 6],]6 ; 10],]10 ; 14],]14 ; 18] and]18 ; 22] cmHg. These divisions arose from the analysis of the OBPM's behaviour when inflating the cuff, from which it was noticed that there are clearly different operating regimes, corresponding to the pressure ranges specified.
The identification tests had 30 minutes of duration, with the first 15 being used to estimate the models and the remaining to validate them. It were computed Output Error (OE), Autoregressive Exogenous Variable (ARX), Autoregressive Moving Average Exogenous Variable (ARMAX), and Box-Jenkins (BJ) models, using the formulation of (Lyung, 1999), of 3rd and 5th order (in all polynomials involved) without delay.
The fits of the various regimes were computed according to (16) (p_{av} is the average pressure and p_{est} the estimated pressure), and respecting the cuff and the constant-volume reservoir, are displayed in Table 1 and Table 2.

$$\text{fit}_{\%} = 100 \times \left| \frac{p_{est} - p}{p - p_{av}} \right| \tag{16}$$

Pressure [cmHg]	OE3	OE5	ARX3	ARX5	ARMAX3	ARMAX5	BJ3	BJ5
0-6	80.56	89.82	75.19	74.96	71.93	87.47	69.47	87.55
6-10	72.20	77.64	66.38	67.82	73.36	77.86	72.47	78.53
10-14	74.09	68.87	62.77	65.78	63.09	75.32	68.95	75.14
14-18	66.64	69.08	51.37	52.68	39.01	36.14	45.86	41.96
18-22	32.34	50.61	33.19	32.77	18.30	13.57	4.39	7.74

Table 1. Models' fit evolution with the air pump output connected to the inflatable cuff

From these results it is seen that the 5th order OE is the fittest model (highest average, μ=71.21, and lowest standard deviation, σ=14.33) with the 3rd order OE having the second highest μ of 65.16, showing the appropriateness of this model type, as it considers the error as white-noise, without estimating a noise model. The global μ is of 59.32 and σ of 25.63.

Pressure [cmHg]	OE3	OE5	ARX3	ARX5	ARMAX3	ARMAX5	BJ3	BJ5
0-6	68.35	68.72	70.78	70.03	68.46	73.79	69.26	69.74
6-10	66.81	70.51	60.50	60.17	60.51	69.01	60.48	71.27
10-14	51.21	51.71	51.57	51.69	58.89	48.06	59.37	60.01
14-18	64.11	62.81	56.22	56.44	57.85	50.85	57.93	51.05
18-22	48.89	9.76	40.00	39.23	3.64	33.78	2.29	36.90

Table 2. Models' fit evolution with the air pump output connected to the constant-volume reservoir

The results presented in Table 2 show the 3rd order OE as being the fittest model (highest μ=59.87, and lowest σ=9.13) while the 5th order BJ has the second highest μ of 57.79. The global μ decreases 9.44% to 53.71 and σ decreases 9.05% to 23.31, implying that although the fits were lower in average, their dispersion also diminished, given the general improvement in the two highest pressure regimes.
It is evident that for both cases the last regime]18 ; 22] cmHg is very difficult to represent using these models, since in the cuff tests the average fit for this regime was of 24.10 and in the reservoir 22.58. This regime is partially above the maximum nominal pressure, and the OBPM's dynamic is not homogeneous inside this pressure range, generating the poorest fit of all regimes.

3.2 At 22 regimes

The pressure range was divided in intervals with 1 cmHg of span, after the first which is [0 ; 2] cmHg, and the tests duration was reduced to 10 minutes. In subsequent Figure 6 and Figure 7 it is displayed the fits evolution, the first presents the results with air pump output connected to the cuff and the latter when connected to the constant-volume reservoir.

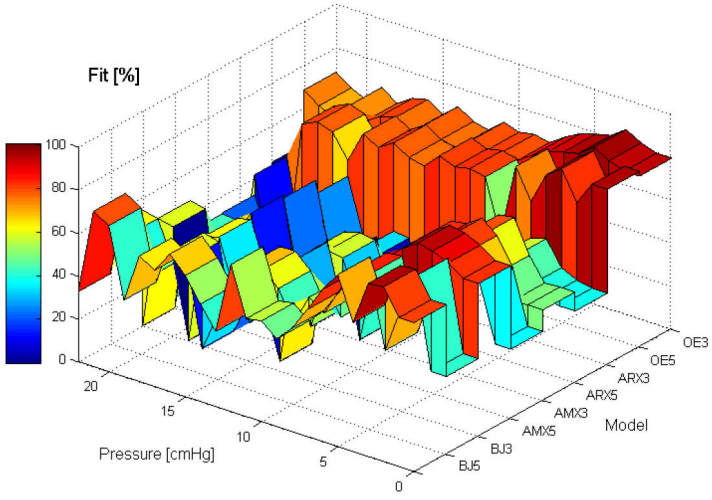


Fig. 6. Models' fit evolution when the air pump output is connected to the cuff

From these results it is seen that the 3rd order OE is the fittest model (highest average, μ=70.60, and lowest standard deviation, σ=6.88) with the 5th order OE having the second highest μ, 64.14. Such results show the suitability of this particular model type, as all other models have worse behaviour, namely the ARX models, with average fit below 30. Comparing with the 5 models approach, the global μ is of 48.61, a decrease of 18.06%, and σ of 26.62, a 3.84% increase.

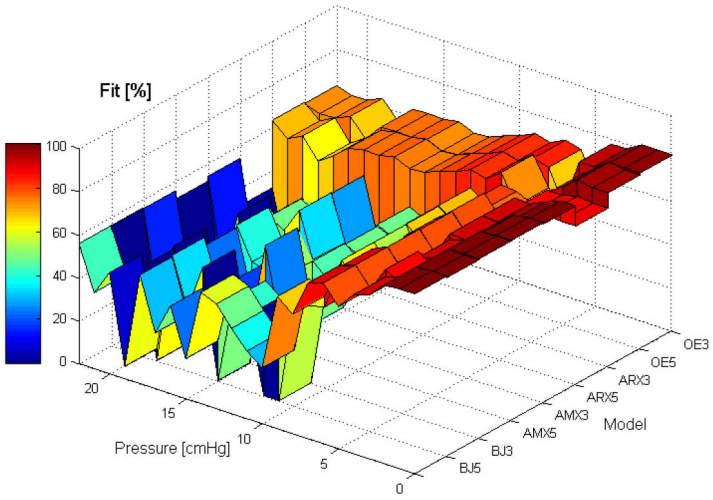


Fig. 7. Models' fit evolution when the air pump output is connected to the constant-volume reservoir

As happened with the cuff tests, the 3rd order OE is again the fittest model (μ_{OE3}=66.64, σ_{OE3}=7.52) while the 5th order OE is very near (μ_{OE5}=65.06). The global μ decreases 5.29% to 50.88 and σ increases 9.05% to 26.79 regarding the 5 models approach. Regarding the compression to the cuff with 22 models, global μ has increased 4.66% and global σ 0.66%.
It is discernible that for both cases the OE models have a regular fit, which does not decrease much in the higher pressure regimes. Moreover, comparing the average fit of the models that comprise the]18 ; 22] cmHg range, to the fit of the corresponding 5-regimes model, the division gains are evident.
Table 3 presents the fit increase for the three best models of the cuff and reservoir tests. The fit increase is the difference from the average fit of the 22-regimes models to the fit of 5-regimes model in the]18:22] cmHg pressure range.

Model	**OE3-cuff**	**OE5-cuff**	**BJ5-cuff**	**OE3-res**	**OE5-res**	**BJ5-res**
Fit increase [%]	31.80	8.01	50.69	11.39	47.55	-1.08

Table 3. Difference between the fit of the]18:22] cmHg regime and the average fits of the models that comprise this pressure range in the 22 divisions tests

3.3 Merging functions

The models correspondent to the different operation regimes should be connected in such way that the information available about a regime is somehow taken into account in a neighbour regime, instead of simply commuting between models (Narendra et al., 1995). To fuse the multiple models identified, a number of different solutions may be tested (Ljung, 2006), in this case, the fusion will be done using linear and Gaussian functions, with and without saturation in the interval centre, Figure 8, varying the dependence on the neighbour models, from the more neighbour-reliant linear without saturation to the most individualist Gaussian with saturation.

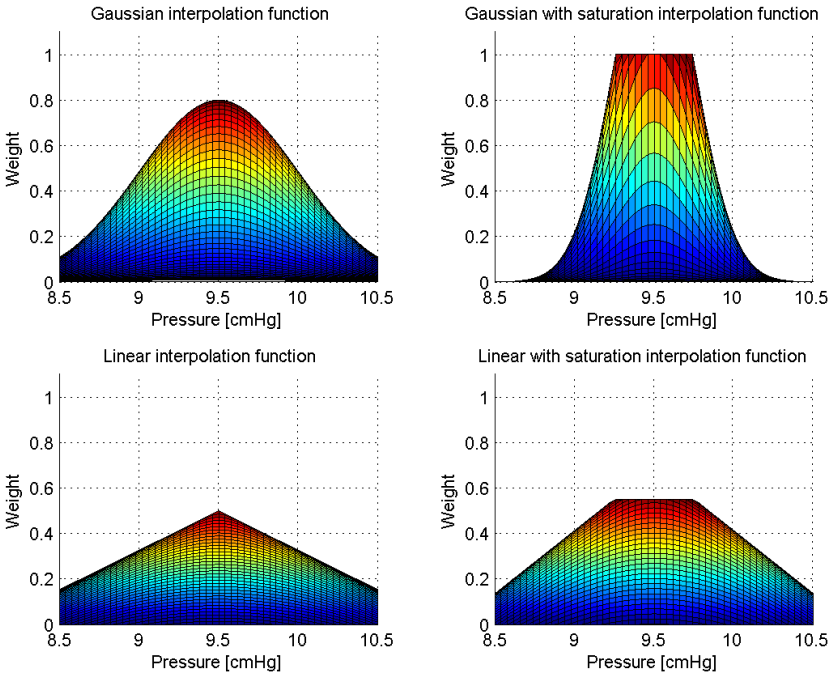


Fig. 8. Interpolation functions used to define the interpolation weights distribution on the global model of the]9 ; 10] cmHg local model, according to the measured pressure

The global model set up to assess the merging functions ability to reproduce the global OBPM behaviour used the 3rd order OE obtained when using 22 pressure divisions, since its fit was always above 60% and with the most homogeneous distribution, and the evaluation tests consisted of 22 trials, which were composed by the first minute of the identification input signal, thus focusing especially in one of the pressure divisions after an initial step input.

The transient response is especially dependent on the fusion quality, as the initial compression traverses many of the local models. The best global performances, regarding mean squared error, were found from 13 to 19 cmHg, Figure 9, although in some experiments the estimates presented overshoot in the transient response, this was rapidly

corrected, and in the remaining of the validation test, the models resemblance to the real behaviour was very truthful.

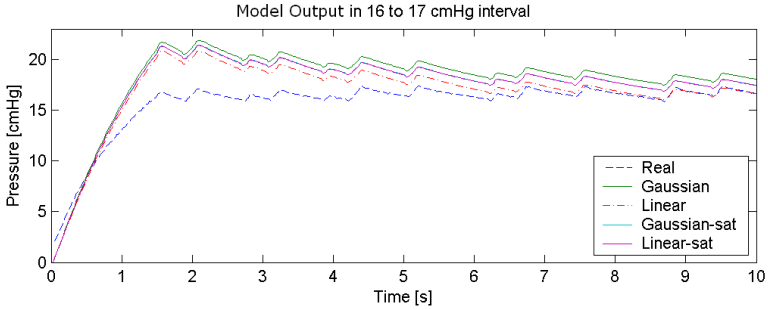


Fig. 9. Models' response in the]16 ; 17] cmHg range test, the blue dashed line is the actual pressure evolution, and the red dash-dot line the linear interpolation estimative, which presents the best transient

The Gaussian function was the most accurate in 10 of the 22 tests, particularly in the lower pressures, the linear and linear with saturation functions were the best solution in 5 tests each, and the Gaussian with saturation only in 2. In the 13 to 19 cmHg range, the linear interpolation function was the most accurate in four intervals, the majority of the cases.

4. Conclusions

A set of equations able to describe in detail the dynamics of an Oscillometric Blood Pressure Monitor, at different depth levels, were considered. Equations fully explaining the electromechanical and pneumatic behaviour of the device have been introduced, but also a more straightforward approach was followed, allowing the assembly of a greyer, yet simpler, model, assuming that the cuff's pressure increase limitations are reflected in the inertia of the air pump's crankshaft, and that this inertia may be estimated from the power dissipation on the air pump. Finally a multiple models identification procedure was described, offering a low computational complexity solution, while completely disregarding the model's physical interpretation, but allowing the compensation of local unsuitabilities while having a consistent global dynamic.

The whiter models developed considered several nonlinearities, such as non-homogeneous stiffness and viscous friction of the servomotor shaft, the flow restrictions in the various piping elements, and the relation between the cuff pressure and its volume. The black-box model approach was also flexible as it is possible to change the models merging functions, as well as the pressure ranges in which the models are used.

After these modelling steps, a number of different tools to obtain an OBPM model was introduced and tested, thus allowing a flexible application of the vast concepts involved in the device's behaviour, to build a model with customisable detail and accuracy. These models may help the search for improvements in the blood pressure measurement accuracy as design changes may improve the OBPM's characteristics, as well as the revision of the components used, to enhance OBPM's dynamics meliorating its performance.

5. Acknowledgements

Eduardo Pinheiro would like to thank the support of *Fundação para a Ciência e Tecnologia*, by means of its SFRH/BD/46772/2008 grant.

6. References

Benedict, R. (1980). *Fundamentals of Pipe Flow*, Wiley & Sons, ISBN 978-0-47-103375-2, New York.

Drzewiecki, G.; Bansal, V.; Karam, E.; Hood, R. & Apple, H. (1993). Mechanics of the occlusive arm cuff and its application as a volume sensor. *IEEE Transactions on Biomedical Engineering,* Vol. 40, No. 7, July 1993, 704-708, ISSN 0018-9294.

Geddes, L. A.; Voelz, M.; Combs, C.; Reiner, D. & Babbs, C. F. (1982). Characterization of the oscillometric method for measuring indirect blood pressure. *Annals of Biomedical Engineering,* Vol. 10, No. 6, November 1982, 271-280, ISSN 0090-6964.

Khan, M. (2006). *Encyclopedia of Heart Diseases*, Academic Press, ISBN 978-0-12-406061-6, USA.

Ljung, L. & Glad T. (1994). *Modeling of Dynamic Systems*. Prentice-Hall, ISBN 978-0-13-597097-3 Englewood Cliffs.

Ljung, L. (1999). *System Identification: Theory for the User (2nd Edition)*. Prentice-Hall, ISBN 978-0-13-656695-3 Englewood Cliffs.

Ljung, L. (2006). Identification of Nonlinear Systems, *Proceedings of the 9th International Conference on Control, Automation, Robotics and Vision*, Plenary paper, ISBN 1-4244-0342 1-06, Singapore, December 2006, IEEE.

Murray-Smith, R. & Johansen, T. (1997). *Multiple Models Approaches to Modelling and Control*. CRC Press, ISBN 978-0-74-840595-4 Boca Raton.

Narendra, K.; Balakrishnan, J.; & Ciliz, M. (1995). Adaptation and Learning Using Multiple Models, Switching and Tuning. *IEEE Control Systems Magazine*, Vol. 15, No. 3, June 1995, 37-51, ISSN 0272-1708.

Ogata K. (2001). *Modern Control Engineering (4th Edition)*, Prentice-Hall, ISBN 978-0-13-060907-6, Englewood Cliffs, New Jersey.

Pinheiro, E. C. (2008). Oscillometric Blood Pressure Monitor Modeling, *Proceedings of the 30th Annual Internacional Conference of the IEEE EMBS*, pp. 303-306, ISBN 978-1-4244-1814-5, Vancouver, August 2008, IEEE.

Shapiro, A. H. (1953). *The Dynamics and Thermodynamics of Compressible Fluid Flow*. Wiley & Sons, ISBN 978-0-47-106691-0, New York.

Schicker, R. & Wegener, G. (2002). *Measuring Torque Correctly*. Hottinger Baldwin Messtechnik, ISBN 978-3-00-008945-9, Darmstadt.

Ursino, M. & Cristalli, C. (1996). A mathematical study of some biomechanical factors affecting the oscillometric blood pressure measurement. *IEEE Transactions on Biomedical Engineering*, Vol. 43, No. 8, August 1996, 761-778, ISSN 0018-9294.

22

Arterial Blood Velocity Measurement by Portable Wireless System for Healthcare Evaluation: The related effects and significant reference data

Azran Azhim[1] and Yohsuke Kinouchi[2]
[1]*Tokyo Denki University*
[2]*The University of Tokushima*
[1]*Japan*
[2]*Japan*

1. Introduction

Arterial hemodynamic function is changed with aging, gender and regular exercise. Age-related decreases in cardiovascular function are evident. The hallmarks of cardiovascular aging are decreased for maximum heart rate, ejection fraction, maximal oxygen intake, maximum cardiac output and artery compliance (Lakatta, 2002; Tanaka et al., 2000). On the other hand, we have found that exercise could improve the age-related deterioration in common carotid blood velocity (Azhim et al., 2007).

Gender-related differences in arterial hemodynamic functions such as systolic blood pressure (SBP) are demonstrated in some previous studies (London et al., 1995; Mitchell et al., 2004). It is suggested that younger women have lower brachial and ankle systolic blood pressure (SBP) and a lower ankle-arm pressure index than age-matched men (London et al., 1995). It has been reported that the incidence of cardiovascular complications increases with SBP (Kannel and Stokes, 1985) and that an increases in the pulsatile components of blood pressure is associated with higher cardiovascular risk in postmenopausal women (Darne, et al., 1989). However, there are a few studies in blood flow and velocity. In this chapter, we present the impact of gender on blood velocity waveform in common carotid artery (CCA). It was found that there is significant gender difference in some velocity waveforms in CCA (Azhim et al., 2007).

The ability to measure and interpret variations of pressure and flow in humans depends on an understanding of physiologic principles and is based on a heritage well over 100 years old. Studies of pressure preceded those of flow, since reliable tools were available for pressure measurement almost 100 years ago but for flow only 50 years ago (Nichols and O'Rourke, 2005).

There are two kinds of noninvasive technique to measure blood flow for portable wireless applications, one is a Doppler ultrasound method and the other is an optical one. The Doppler ultrasound was widely used to measure hemodynamic in blood vessels as carotid

arteries that exist in the deep place from the human tissue (Prichard et al., 1979; Baskett et al., 1977; Gosling, 1977; He et al., 1992).
We have developed the telemetry measurement system in our laboratory which is capable to measure blood flow velocity in both aerial and aquatic environments. The device has enough performance for the measurement during physical exercise stress as well as at rest posture (Azhim et al., 2007; He et al., 1996; Jiang et al., 1994; Jiang et al., 1995). The measurement system with synchronized measurement of electrocardiogram and blood pressure will contribute to the extent of understandings in exercise physiology as well as further knowledge of arterial hemodynamic functions. We have shown that exercise has significant change to envelope waveform of CCA blood velocity in hemodynamic functions as evaluated from cross-sectional and intervention investigations (Azhim et al. 2007a).
Various telemetric techniques have been developed for ambulatory and noninvasive determination of bioelectrical and physiological signals in human subjects. Biomedical telemetry applications has brought numerous advantages such as comfort and portability, providing the critical information on improvements in quality of health care, efficiency in hospital administration capabilities and finally reduction at overall medical cost. Biomedical telemetry is a reliable tool for data gathering since the invention of integrated circuit (IC) technology which has enormous impacts on the contributions of microelectronics to biomedicine and health care applications (Azhim et al., 2009).
In the chapter, the usefulness of CCA velocities and the waveform indices after taking into account all relevant effects as reference value for clinical and healthcare applications are presented. In section 2, system is described the developed portable measurement system. Real-time monitor and data analysis are described in section 3. In section 4, data measurements and collections in the selected 202 healthy volunteers between the ages of 20 and 69 years are presented as a result by the following: Anthropometric data for the selected subjects, reference data for normal velocities in CCA and indices between the third to seventh decades after controlling for the effects of exercise training and gender, general age-related decrease in flow velocities and change in the velocity waveform, regular exercise training improved blood velocity waveforms which markedly different in regularly exercise-trained middle-aged and older age-groups compared to the sedentary age peers, and gender-related difference in velocity and its indices. In last section, it can be concluded that normal CCA blood velocity parameters which are determined in a total of 202 healthy volunteers between the third and seventh age decade after adjustment for gender and exercise effects may contribute to improved means of healthcare monitoring and clinical evaluation.

2. Measurement System

The last three decades has shown rapid increase in the use of Doppler ultrasound devices for monitoring cardiovascular functions. Developments in Doppler technology have led to a vast increase in the number of non-invasive blood velocity investigations carried out in many areas of medicine. As with many rapidly expanding technologies there have been a considerable number of types of instrument developed and used in their institutions of origin, whereas only a few are in widespread portable device for exercise use.
We have developed a portable wireless system for measurement of blood velocity spectra in CCA with synchronized measurements of ECG and BP as shown in Fig. 1. In our previous

studies, we have presented that the telemetry system has enough performance to get accurate data for estimation of blood circulation during physical exercise stress in both aerial and aquatic environments (Azhim et al., 2007a; Azhim et al., 2008; He et al., 1996; Jiang et al., 1994; Jiang et al., 1995). Measurements of blood velocity were noninvasively detected by using Doppler ultrasound method. The measurement system was consists of an ultrasound probe, a Doppler signal discriminator (DSD), an analog-digital (A/D) converter board and a laptop personal computer (PC), a wireless transmitter and a receiver as shown in Fig. 1. Data were transmitted using 315 MHz FM/FSK transmitter which has 28.8 kbps and ~0.5 mV/m (feeble wave) for transmission speed and output, respectively.

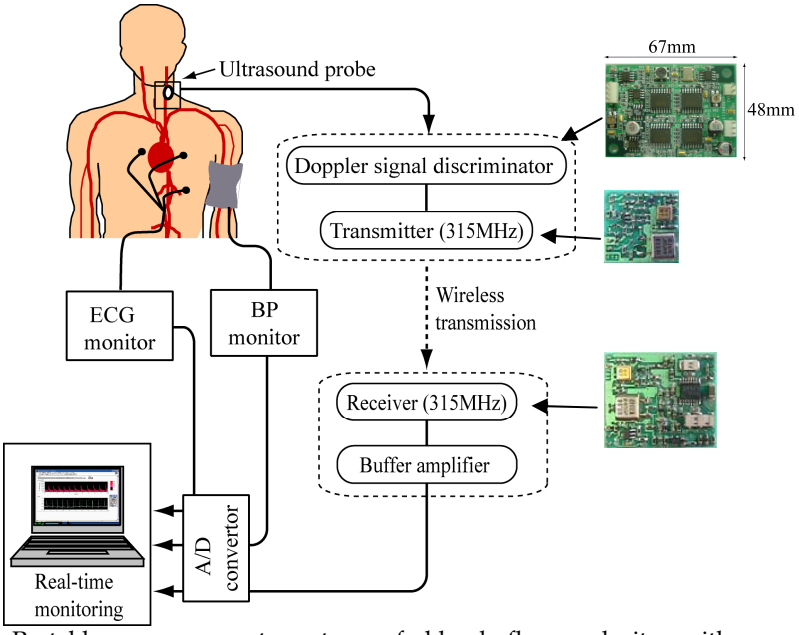


Fig. 1. Portable measurement system of blood flow velocity with synchronized measurements of electrocardiogram and blood pressure

Measurements of blood circulation during exercise stress or postural change was difficult. It provided high level artifacts caused by relative displacement of the ultrasound and artery and vibration of tissue, particularly measurement of blood flow in femoral artery (Dahnoun et al. 1990). Considering that we selected semicircular transducers to provide the wider and uniform transmitting ultrasonic beam (Zhang et al. 2002). Considering the depth of the arteries and the size of transducers, 2.0 MHz was chosen as the transmitting ultrasonic frequency. The probe was designed with a small size (W34×H20×D42 mm^3, approximately weighing 20 g) using two piezoelectric zirconate titanate transducers (PZT) with a diameter of 15 mm, where one was for transmitting ultrasound and the other was for receiving the Doppler echoes using continuous-wave ultrasound as shown in Fig. 2A. To make the emitting and receiving beams face to the target of blood vessels, a small angle of 184 degrees were set between the pair transducers for the carotid artery (Fig. 2B). The ultrasonic probe was attached to the left side of neck with 50 degrees of the Doppler angle of insonation as

shown in Fig. 2C and was fixed it with band wound around the neck. An exact attachment position in measuring blood flow velocity in CCA is between the sternocleidomastoid muscle and the throat at a level between the fourth and fifth cervical vertebrae.
A transmitter transducer chosen for clinical use had an intensity output of 8 mW/cm^2 spatial peak-temporal average (SPTA) as measured by a 0.4 mm diameter needle hydrophone (ONDA, model HNV-0400). The ultrasonic output intensity was safe for the human tissue.

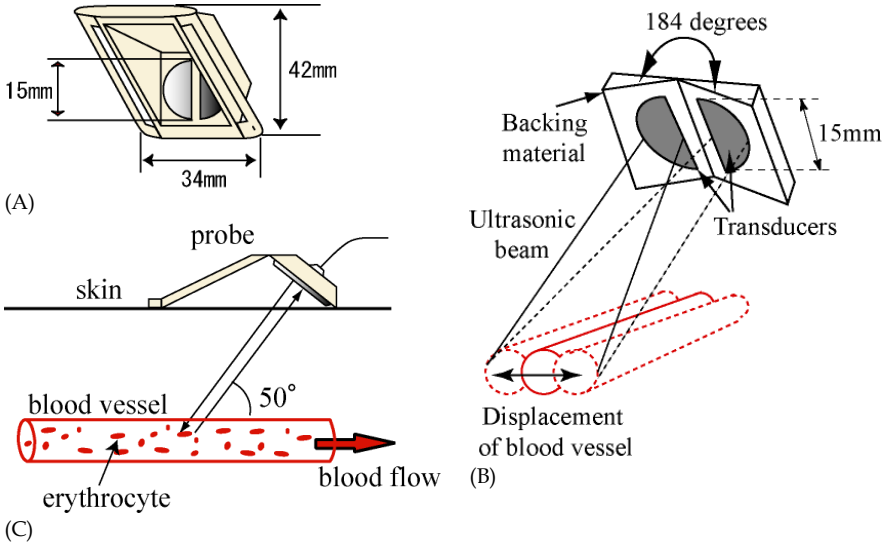


Fig. 2. (A) Dimension of ultrasound probe, (B) The design probe for wider transmitting ultrasonic beam, and (C) Doppler insonation angle

For the purpose use of healthcare application, the system was improved by miniaturizing the DSD, upgrading the probe with better attachment on the skin and developing the stand-alone real-time software package for measurement of blood flow velocity spectra with synchronized monitor of electrocardiogram (ECG) and blood pressure (BP). The DSD was miniaturized using mount-surface technique, and the downsized substrate size was 67×48 mm^2 as shown in Fig. 1. The substrate size of transmitter and receiver modules is 18.5×18.5 mm^2 and 28×24.5 mm^2, respectively. The power consumption of the DSD and transmitter modules was reduced to 2.1 W, therefore it was enabled battery installed in the portable system to be used approximately 10 hours.

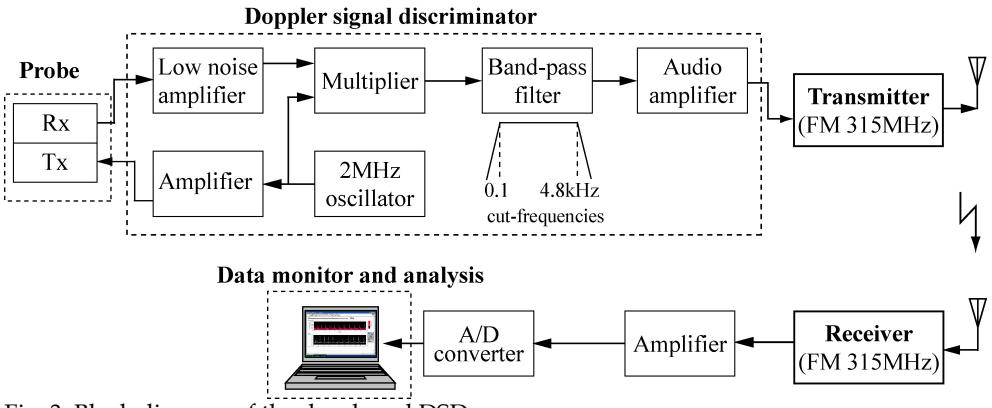


Fig. 3. Block diagram of the developed DSD

As illustrated in Fig. 3, square wave signal of 2 MHz was oscillated at crystal oscillator. The signal was converted to sinusoidal signal at LC circuit which consists of an inductor and a capacitor. Then, it was buffered at operational amplifier (OPA) before 2 MHz of ultrasound was emitted at the transmitting transducer. The emitted sinusoidal signal was set about 6.0 $V_{peak\text{-}to\text{-}peak}$. The detected signals at receiving transducer were including of Doppler shifted frequencies from the movement of blood flow (red blood cells), and vibration of tissues. After testing several kinds of circuits, we used the above DSD configuration to measure blood velocity for the purpose of telemetry (Azhim et al. 2007a).

A receiving Doppler signals were a few hundreds nanovolts which very small and weak signals. We used commercial available very-low-noise OPA (OP621, Burr Brown, US) to amplify the signals to tens of milivolts. After being amplified, the signals (ω_D) were synchronously detected with 2.0 MHz sinusoidal signals (ω_0) by commercial multiplier IC (AD835, Analog device, US). Typical outputs of multiplier generated the frequencies of sum ($\omega_0+\omega_D$) and difference ($\omega_0-\omega_D$), respectively. Consequently, Doppler signal could be derived from the output of difference frequencies, and the other unnecessary signals were filtered by band-pass filter (UAF4, Burr Brown, US). Considering of blood flow frequencies in particularly CCA, the cut-frequency (fc) of filter was set from 100 to 4800 Hz.

3. Real-time Monitor and Data Analysis

After processing the Doppler signal at DSD, the signals were sampled to digital signal by A/D converter and fed to the computer for monitor and data analysis. We developed the stand-alone software to monitor blood spectral velocity with real-time as described below.

3.1 Real-time Monitor

In the real-time processing monitor implementation, two main specifications had been taken in account. First, the execution time of signal processing must be low, to avoid overloading the available system resources. Second, the output latency time had kept as short as possible (less than 100 ms) to synchronize the sound with another output display (corresponding real-time spectrogram).

The signal processing had been implemented through a program written in Visual C++ ® for a stand-alone Windows ® application as shown in Fig. 4. The real-time spectrogram monitor was implemented by using loop timer method. Timer was set as 50 ms corresponding to the sampling data of 500 points. However, data were analyzed by using fast Fourier transforms (FFT) with multiples of 256 points of Hanning window. The spectrogram was processed using decimation in frequency (DIF) radix-2 FFT algorithm, which had smaller the discrete Fourier transform (DFT) computations in order to reduce the computation time. To increase the efficiency, decomposed of DFT was optimized by processing the real signal as called real-only FFT computation.

The average of CPU load, memory utilization, and computation time were measured in two laptop PCs of different performance as reported in previous study (Azhim et al. 2007a). Because of using set timer method, output latency was depended to timer rate, 50 ms. If the signal processing time was less than set timer rate, latency time was constantly maintained. The computation time was lower than 1 ms when tested in the recent laptop PC. However, the computation time was quite larger, about 16 ms in the quite obsolete laptop PC, so that latency time to rendering spectrogram was still maintained at rate of 50 ms.

Blood flow velocity spectra were measured for 1 minute in the relaxed sitting posture. Data collections were performed with synchronized monitoring of ECG and BP using a developed real-time processing monitor.

Systolic (SBP) and diastolic blood pressure (DBP) was collected at the left brachial artery by using the automatic blood pressure monitor (Tango, SunTech Medical, USA). Mean (MBP) and pulse blood pressure (PP) were calculated from DBP+ (SBP-DBP)/3 and SBP-DBP, respectively.

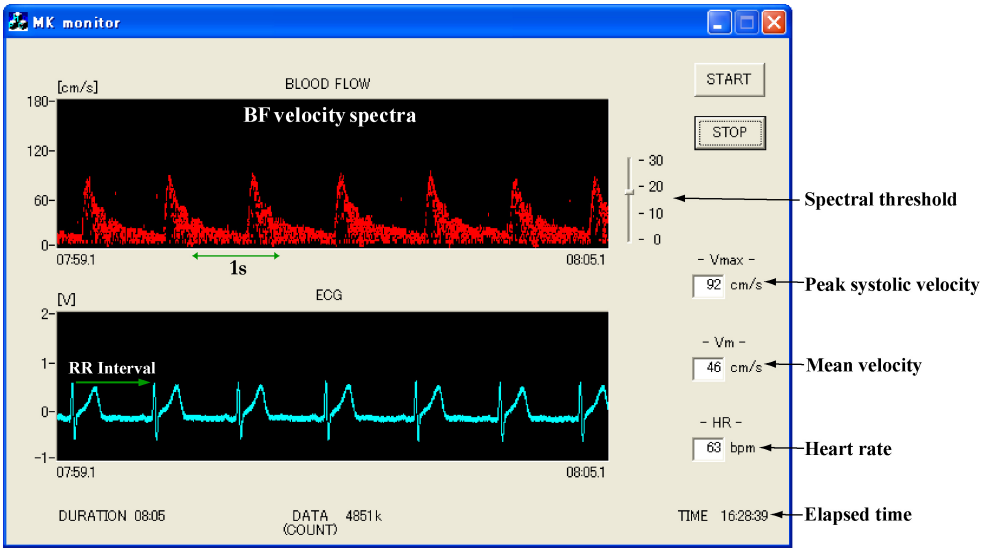


Fig. 4. The developed stand alone dialog software for real time monitoring of blood velocity spectra and ECG

3.2 Data Analysis

The measured signals which included a low-frequency noise and a harmonics noise were filtered by band-pass filter of 0.1 to 4.8 kHz in the DSD as shown in Fig. 3. From the frequency range, velocity spectra could be taken out by FFT analysis as monitored real-time in the developed dialog software (see Fig. 4). Estimation of flow velocity spectra (V_d) was performed from the detected Doppler shifted frequency (f_d) that was given by the classic equation: $V_d=cf_d/2f_0\cos\theta$, where, c=1540 m/s, sound speed in human tissue; f_0, an irradiated ultrasound frequency and θ=50 degrees, the angle insonation.

The signals were acquired at 10 kHz of fs through a 16 bit-A/D converter (CBI-360116TR, Interface JAPAN). It was repeatedly analyzed by using FFT with successive 25.6 ms, which were given by shifting 12.8 ms in turn. Therefore, an instantaneous spatial spectral frequency was calculated at 12.8 ms intervals with 39 Hz per point of frequency resolution.

Using a threshold method, flow velocity envelope (Vp) was extracted from it spectra as shown in Fig. 5. Blood flow velocities in CCA were characterize to 5 feature points of waveform; peak systolic (S1), second systolic (S2), insicura between systole and diastole (I), peak diastolic (D) and end-diastolic minimum (d) velocities (Azhim et al., 2007a). An ensemble-averaging method was used to characterize and calculate blood velocities and its indices (Azhim et al., 2007a). We selected 30 consecutive cardiac cycles of those to characterize the feature points as represented in Fig. 6. From these, velocity indices were calculated from 1-d/S1, S2/S1-1 and 1-I/D as resistive (RI), velocity reflection (VRI) and vascular elasticity indices (VEI), respectively (Azhim et al., 2007b).

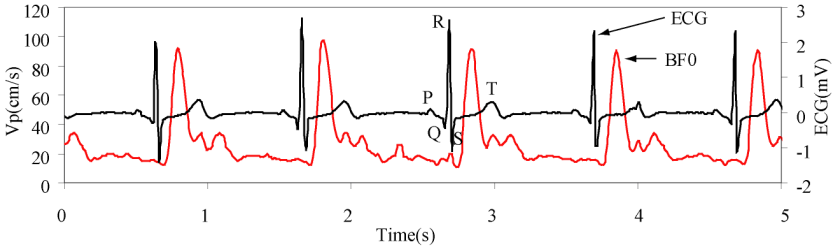


Fig. 5. Synchronization measurement of electrocardiogram (ECG) and blood velocity envelope (Vp).

As shown in Fig. 6, the velocity indices were calculated from the feature points of blood flow velocity. These indices were dimension-less and independent of the angle insonation. The RI was used as a typical peripheral vascular resistance index that can be calculated from waveforms as 1-d/S1, which the smaller RI the lower resistance and vice versa. The index was firstly used by Pourcelot on flow velocity waveform in CCA (Planiol et al., 1973; Pourcelot, 1976). Gosling et al. was firstly used velocity index of S1/S2 (as called A/B ratio) in CCA and superorbital arteries for detecting occlusive disease in the internal carotid artery (Gosling, 1977). The VRI was used in a manner similar to evaluate reflected wave velocity in second systolic velocity in CCA. In the previous study, the index of D/I was proposed and may be provided to evaluate the magnitude of vascular elastic recoil during cardiac diastole that is exerted by its smooth muscle cells (Azhim et al., 2007a). The VEI was used in the study to define the magnitude of vascular elasticity in a similar way. The velocity indices in

CCA were found that changed by aging, regular exercise effect and gender difference (Azhim et al., 2007a; Azhim et al. 2007b).

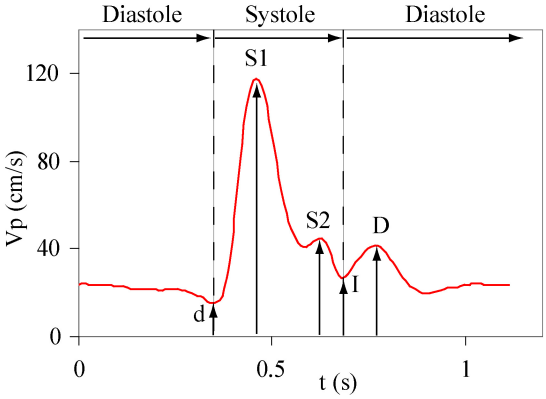


Fig. 6. Characteristic features on waveform: Ensemble-average velocities of 30 consecutive cardiac cycles in a young subject (21 years old, male). S1: the first peak systolic velocity wave (peak velocity), S2: the second systolic velocity wave, I: incisura between systole and diastole, D: the peak diastolic velocity wave, and d: the end-diastolic minimum velocity.

3.3 Statistical Analysis

Data were expressed as mean ± standard error (SE). For overall view, Pearson's correlation analysis was performed to confirm the relationship between all outcome variables and the factors. As previous studies, we have reported that blood velocity waveforms in CCA had influenced by aging and regular exercise (Azhim et al., 2007a) and had significantly difference in gender (Azhim et al., 2007b). Hence, multivariate ANOVA test was used to determine the outcome variables (velocity data) that influenced by the multiple effects. The designed age-groups of significant pairwise differences were determined using Tukey's post-hoc test. After controlled for the effects of exercise and gender, correlation coefficient, R of all variables with age was determined (Table 2). The effects of exercise training are compared by ANOVA, which provides P values. The P value represents the effects of exercise training on the entire group, combining the classified decade-group as a single group. Gender difference was determined after adjustment for aging and exercise effects by least significant difference method, and as a categorical variable gender was coded as 1 for men and 2 for women (Table 3). The significance level, P was set at 0.05. Statistical analysis was performed using statistical package for the social sciences (SPSS).

4. Measurement Data

Using the developed portable measurement system as described above, data were collected in the study. Study was performed in a total of 286 putatively healthy subjects. We studied the selected 202 subjects between the ages of 20 and 69 years, from a cohort which normotensive (systolic blood pressure≦140mmHg), normal body mass index (below 30 kg/cm^2), without any drug intake, and free of overt chronic diseases (including

hyperlipidemia, diabetes, arrythmia) as assessed by medical history. There were 5 classified groups from 20-year-old to 69-year-old in 10-year-old intervals as shown in Table 2. Subjects were classified as exercised who performed regular aerobic exercise training more than 3 times per week.

4.1 Anthropometric data for selected subjects

Table 1 shows the characteristics of selected 202 subjects between the ages of 20 and 69 years. Anthropometric data for height, weight, body mass index, systolic, diastolic, mean, and pulse blood pressure are represented in mean, standard error, minimum and maximum values.

	Mean	SE	Min	Max
Height (cm)	165	0.7	148	184
Weight (kg)	59	0.9	39	93
BMI (kg/cm^2)	22	0.2	16	29
SBP (mmHg)	119	0.9	83	140
DBP (mmHg)	74	0.7	50	96
MBP (mmHg)	89	0.9	62	110
PP (mmHg)	45	0.7	17	70
Heart rate (beats /min)	73	0.8	48	100

Table 1. Anthropometric data of the selected subjects. Data are mean, standard error (SE), minimum (Min) and maximum (Max). BMI, body mass index; SBP, systolic; DBP, diastolic; MBP, mean; PP, pulse blood pressure.

4.2 Reference data for normal velocities and the waveform indices

Table 2 represents the reference data for normal CCA velocities and the indices between the third to seventh decades. Correlation coefficients of all variables with age were determined after controlling for the effects of exercise and gender. The velocities in d and I waves were not changed with age (P=NS). There were significant negative correlations in S1 and D velocities with age (R=-0.615, P<0.001 and R=-0.248, P<0.001, respectively). However, S2 velocity significantly positively correlated with age (R=0.279, P<0.001). For velocity indices of RI and VEI, there were significant negative relation with age (R=-0.606 P<0.001 and R=-0.479, P<0.001, respectively), whereas there were significant positive relation with age for VRI (R=0.798, P<0.001). From the data, it clearly shown that velocity continuously changed with advancing age as compared to third decade-group. The S1 and D velocities in seventh decade-group decreased 65% and 83%, respectively compared to third decade-group. The S2 increased 16% in fifth decade-group and 13% in seventh decade-group compared to third decade one. The indices of RI and VEI decreased 89% and 67% in seventh decade-group. Due to markedly decreased in S1, the VRI increased 34% in seventh decade-group.

Decade (years old)		3rd (20~29)	4th (30~39)	5th (40~49)	6th (50~59)	7th (60~69)
Blood velocities (cm/s)	R, P					
d	0.084, NS	20.38±0.49	21.23±0.80	23.04±1.06	21.23±0.80	20.07±1.32
S1	-0.615, <0.001	108.6±1.87	102.0±3.05	84.18±4.02	76.0±3.048	70.54±5.00
S2	0.279, <0.001	52.42±1.27	59.21±2.06	61.0±2.72	59.96±2.06	59.44±3.38
I	0.097, NS	29.31±0.76	31.55±1.23	35.14±1.63	30.64±1.23	28.20±2.02
D	-0.248, <0.001	43.57±0.78	42.29±1.26	42.54±1.66	39.41±1.26	36.38±2.07
The indices (%)						
RI	-0.606, <0.001	80.8±0.5	78.6±0.8	72.4±1.1	72.0±0.8	72.2±1.3
VRI	0.798, <0.001	-51.0±0.1	-41.0±1.7	-27.4±2.3	-21.5±1.7	-17.1±2.9
VEI	-0.479, <0.001	33.1±1.0	26.2±1.6	0.18.0±2.1	21.8±1.6	22.2±2.7

Table 2. The reference data indicated velocity and the indices between the third to the seventh age decades. The values indicated mean and standard error. R indicates the values of partial correlation coefficient between variables and age after controlling for the effects of gender and exercise. P<0.001 and NS indicated significant level and not significant, respectively. RI, resistive index; VRI, velocity reflection index; VEI, vascular elasticity.

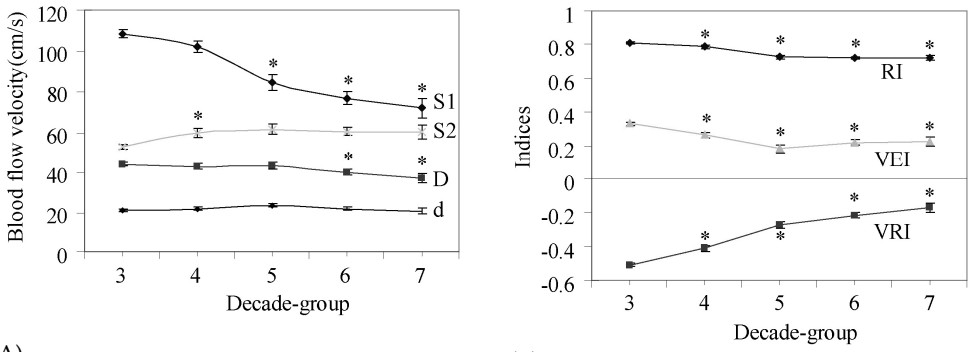
To our knowledge, there were only three analogous studies for age-related decline in blood flow velocity and for reference data of normal velocities in CCA between multiple age-groups. In a first paper, Gregova et al. reported the peak systolic and end-diastolic velocities of CCA in 199 subjects which in the age range of 20-92 years. They suggested peak systolic velocity in CCA decreased with aging as the following yearly rate: in the right 6.44 mm/s/year and left 7.39 mm/s/year. The yearly decrease of end-diastolic velocity on the both sides was in range between 1.72 and 2.28 mm/s (Gregova et al., 2004). In a second one, Scheel et al. presented the reference data of flow velocities in CCA and its index which performed in 78 healthy adults from 20 to 85 years old. They suggested that peak systolic velocity and end-diastolic minimum velocity decrease with age as following rate: 101±22 cm/s and 25±5 cm/s for age-group of 20-39 years, 89±17 cm/s and 26±5 cm/s for age-group of 40-59 years, and 81±21 cm/s and 20±7 cm/s for age-group of 60-85 years, respectively. In a third one, Fujishiro et al. measured blood velocity in the right common carotid artery in 140 normal healthy individuals in their teens to seventies using an ultrasonic quantitative flow measurement system (Fujishiro et al., 1982). As a result, they presented that S1 and d velocities markedly decreased with age, in which the values in the 70's were about 1/2 and 2/3 as small as that in the 20's, respectively (Fujishiro et al., 1982). These findings were consistent with age-related decrease in blood flow velocities in CCA (Tanaka et al., 2000; Gregova et al., 2004; Azhim et al., 2007a, Nagamoto et al., 1992; Fujishiro and Yoshimura, 1982).

Most of other studies were not taken into account for the multiple effects on flow velocities in CCA. In the study, we demonstrated that there are multiple effects that probably alter blood velocity waveforms in CCA (Azhim et al., 2007a; Azhim et al., 2007b). We also found that blood velocity data are more sensitive compared to BP data. In the study, we found that

not only S1 and d velocities, but also D velocity significantly decreased with age, whereas S2 velocity increased with age.

4.3 Age-associated changes in flow velocity waveforms

Fig. 8 represents the average values of five decade-groups as age-associated change blood velocity in CCA and the velocity indices. *Significant pairwise differences between third decade-group and the other designed decade-groups were determined using Tukey's post hoc test. The peak systolic and peak diastolic velocities were lower in the subjects over 40 years old (P<0.0001 and P<0.0001, respectively). The fourth decade-group had significant higher in S2 velocity compared to third decade-group. The d velocity had no significant different between the designed decade-groups. The fourth and over decade-groups had significant lower in the velocity indices of RI and VEI (P<0.0001 and P<0.0001, respectively). The VRI was higher in the decade-groups of third and over (P<0.0001).



(A) (B)

Fig. 8. Age-associated change in the feature points of velocity in CCA (A) and in its velocity indices (B). Data were mean and SE. *Significant pairwise differences versus third decade-group were determined by Tukey's post-hoc test.

As we found that flow velocities in CCA decreased with age, other groups had reported that a comparable decrease in flow velocities with age (Satomura, 1959; Nagamoto et al., 1992; Fujishiro et al., 1982). Most of their studies only focused velocity waveforms, i.e. on the S1 and d velocities, and demonstrated that it decreased with age (Satomura, 1959; Nagamoto et al., 1992; Fujishiro et al., 1982).

In our study, we found that S1 and D velocities decreased continuously with age (by 9.67 and 1.55 mm/s/year, respectively), and S2 velocity increased significantly wit h age (by 1.93 mm/s/year). In hemodynamic studies, characteristic blood flow velocities in S1, S2, and d were more focused on their relationship either between aging and carotid diseases (Nagamoto et al., 1992; Fujishiro and Yoshimura, 1982; Gregova, 2004; Schmidt-Truchsass et al., 1999; Johannes et al., 2001; Scheel et al., 2000). Our results suggest that not only S1 and D velocities (R=-0.612, P<0.001 and R=-0.248, P<0.001, respectively), but also its indices of RI, VRI and VEI decreased continuously with age (R=-0.606, P<0.001; R=-0.798, P<0.001; R=-0.478, P<0.001, respectively after adjustment for gender and exercise effects). Although the

latter velocities and the ratio could be measured easily using ensemble-average envelope velocities, only very few studies took this parameters into considerations with the intention of characterizing its associations with disease (Kaneko et al., 1978; Rutherford et al., 1977), aging and exercise training.

In the study, we also found that the decreases of D velocities and D/I parameter, which were depended on arterial elastic recoil, were the important determinant as predictor of age. The decrease may be homeostatically related to the reduction of arterial compliance and elasticity with age (Schmidt-Trucksass et al., 1999; Lakatta, 2002; Tanaka et al., 2000). Age-associated alterations in arterial properties comprised of structural, e.g. intima-media thickness (IMT), and functional, e.g. arterial elasticity, were related to changes of hemodynamic parameters particularly in peak blood flow velocity (Schmidt-Trucksass et al., 1999). Age was the strongest predictor in the decrease of S1, which may be a suitable parameter to evaluate the influence of aging or atherosclerotic risk factor on arterial structure and function (Schmidt-Trucksass et al., 1999).

Changes in the shape of velocity waveform may be quantified using RI as the most popular index. The index was originally used by Pourcelot on waveforms from CCA, as an indicator of peripheral vascular resistance beyond the measurement point (Planiol et al., 1973; Pourcelot, 1976), which the smaller RI the lower resistance and vice versa. It is dimension-less and independent of the angle insonation. It has consequently been widely used for the variety study of pathophysiological conditions including internal carotid stenosis (Pourcelot, 1976) and for the study of cerebral haemodynamics in neotanal and fetal (Permal, 1985; Donofrio et al., 2003). In the present study, we found that RI has significantly changed with age (see table 2).

The velocity ratio of S1/S2 is known that alters with aging and ICA disease (Baskett et al., 1977; Gosling, 1977; Prichard et al., 1979). Gosling demonstrated that S1/S2 ratio in sonograms from the CCA decreased with aging (Gosling, 1977). Baskett et al. used the ratio for screening and diagnosis of carotid junction disease (Baskett et al., 1977). They suggested that S1/S2 ratio less than 1.05 there is an 88 % probability of disease at carotid junction (Baskett et al., 1977). We found that the index of S1/S2 has same tendency that significantly decrease continuously with age ($R=-0.702$, $P<0.001$). In the study, velocity reflection index, as calculated from S2/S1-1, increased continuously with aging ($R=0.781$, $P<0.001$). This index seems to have been improved in all age-groups by the training (see details in section 4.4). It is expected that decreased of peak systolic velocity S1 is associated with increased of second systolic velocity S2 with advancing age. The increasing of S2 may be related to the increased of wave reflection properties cause an increase in augmentation index (AI) and wasted ventricular energy (see details in section 4.5).

We could not determine the hemodynamic variables for velocity data in the seventies and over because of there are no healthy volunteers that participated in this investigation. Thus, further studies are needed.

4.4 Effect of exercise on the entire groups

Fig. 9 shows the effect of exercise training on peak systolic velocity and the velocity indices (P by ANOVA). As anticipated results, HR is lower in exercise-trained than in sedentary ($P<0.0001$). The S1 velocities are significantly higher in exercise-trained than in sedentary ($P=0.009$). The others (S2, D, d) have no significant differences for the exercise effect (P=NS). Due to the significant increase in S1, it is reasonable that the RI and VRI in the exercise-

trained are significantly higher (P<0.008) and lower (P=0.0008), respectively. The VEI was not found significant difference between two groups (P=0.093).

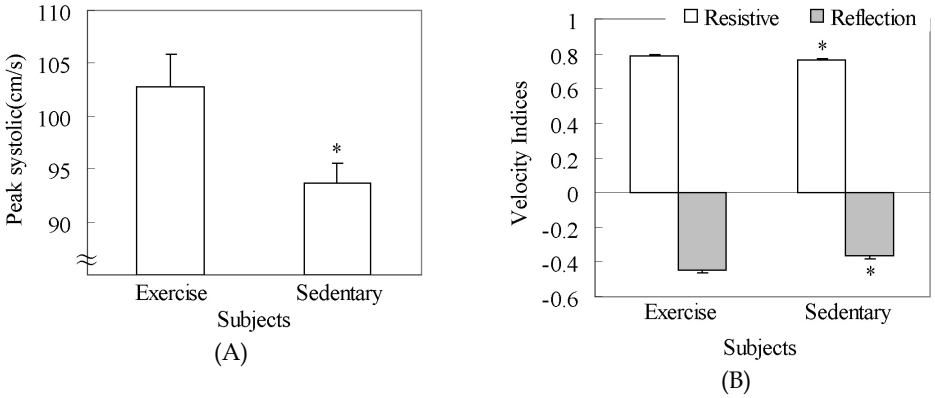


Fig. 9. Peak systolic velocity and the velocity indices in exercise-trained and sedentary subjects. *P<0.05 significance between exercise-trained and sedentary peers.

One of most pronounced cardiovascular adaptation to aerobic exercise training is a lowered HR in the resting (Chen et al., 1997; Goldsmith et al., 2000). For instance, it has been reported that trained runners often possess resting HR lower than 50 bpm (Costill, 1986). Similarly, lower resting HR have been shown to occur in sedentary individuals after they have been exposed to effective aerobic exercise training (Maciel et al., 1985). In our cross-sectional study, HR are lower in regular exercise trained than in the sedentary peers. As anticipated, the resting HR decreased in previously sedentary young subjects after exercise intervention.
As shown in Fig. 10, there are typical blood velocity waveforms in three age-groups. Peak systolic velocities are seen extremely sharper and higher in third decade-group, exercise-trained sixth decade- and seventh decade-persons. All blood flow velocity waveforms are not significantly changed in who regularly performed aerobic exercise training in three age-groups. The velocities are sharper and higher than those of sedentary peers in the especially, elderly age. The adaptations of blood flow to regular exercise are similarly changed with aging (P=NS by ANOVA). The patterns of blood flow waveforms are no significance difference between young and older exercise-trained. As we reported in previous study, there were no differential exercise effects in the designed age-groups (Azhim et al., 2007).
S1 velocities have a significant difference between exercise-trained and sedentary adults. Due to the increased of S1 velocities, the VEI and RI indicate a similar tendency change with the training. Thus, training exercise is a predictor of increasing S1 velocities as a suitable parameter to evaluate the effect of training exercise. The decreased of resting HR is associated with the increased of S1 velocities and RI and the decreased of VRI with training. We can only speculate on one of the adaptation of cardiovascular systems to regular aerobic exercise training improved arterial compliance and increased stroke volume in the present study. The increased of S1 velocities seem to reflect the changes of several structural and functional parameters including IMT and arterial compliance related to smooth muscle behavior in a similar way (Schmidt-Trucksass et al., 1999). Regular exercise training exerted a decreased HR with the association of increased blood velocity waveforms in CCA.

We also found that the ability of regular aerobic exercise to improve flow velocity waveform in particularly older population was not related on blood pressure. This was the first finding in age-associated deterioration in blood velocity waveforms of CCA could be modified favorably by regular exercise aerobic exercise (Azhim et al., 2007a).

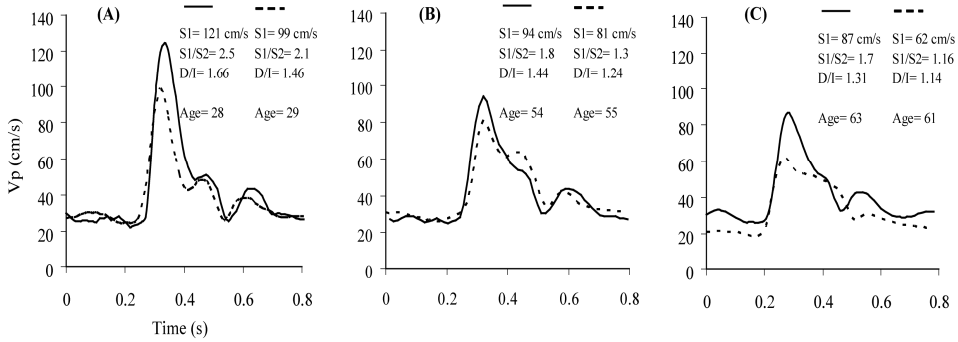


Fig. 10. Comparisons of typical blood flow velocity waveforms for regular aerobic exercise-trained (solid line) and sedentary adults (dashed line) in the third (A), sixth (B), and seventh (C) decade-groups. Regular exercise may retard age-associated diminishing in blood flow velocity waveforms in human.

4.5 Gender difference in the velocities and waveform indices

Table 3 represents the gender difference in the outcome hemodynamic variables and anthropometric data. Generally men had larger body stature and mass. As shown in the result, height, weight and BMI are significantly higher in men ($P<0.001$). Men had larger SBP and PP compared to women ($P<0.05$). However, we could not found the gender difference in DBP and MBP (P=NS). There was remarkable gender difference in flow velocities, expected for S1 velocity as shown in Table 3. It is shown that all velocity indices were significant different between men and women ($P<0.001$). There was no gender difference in heart rate (P=NS).

The findings present the effect of gender in the role body height and weight on arterial hemodynamics in carotid blood flow velocity waveforms. As we found in the study, the women had a lower brachial systolic pressure and carotid systolic velocity, whereas had higher end-diastolic and second systolic velocities. Therefore, the velocity indices in vascular resistive and elastic recoil were lower in women. However, wave reflection index was higher in women than in men.

The augmented secondary peak in the carotid systolic pressure waveform was attributed to wave reflection (Murgo et al., 1980). A major determinant of intensity of wave reflection was the distance to reflecting sites. Previous investigators suggested body height was a significance inverse determinant of augmentation due to earlier wave reflection (London et al., 1995). Some reports presented that the greater of wave reflection in women was associated with shorter body height, due to shorter distance to reflecting sites (Mitchell et al., 2004). They reported that augmentation wave pressure reflection was useful for assessments of cardiovascular risk and clinical potential (Marchais et al., 1993; Mitchell et al., 2004; Nichols et al., 2005). However, the importance of reflected flow waves was not emphasized

until relatively recently (Nichols et al., 2005). In the study, augmentation of reflected flow wave had highly correlated with that of reflected pressure wave. It was potential to determine wave reflection from its flow velocity components. Velocity wave reflections were highly correlated to its pressure components (R^2= 0.836) (Azhim et al., 2007c).

	Men	Women
Height (cm)	170±0.6	157±0.8**
Weight (kg)	64±0.9	51±1.1**
BMI (kg/cm^2)	22±0.3	20±0.4**
SBP (mmHg)	121±1.1	116±1.4*
DBP (mmHg)	75±0.9	73±1.2
MBP (mmHg)	90±0.9	88±1.1
PP (mmHg)	46±0.8	42±1.1*
Heart rate (beats /min)	73±0.8	73±1.3
Blood velocity (cm/s)		
d	18.8±0.5	23.0±0.6**
S1	98.8±1.9	99.4±2.4
S2	49.0±1.2	62.4±1.6**
I	27.0±0.7	34.0±0.9**
D	40.0±0.7	44.9±0.9**
Velocity indices (%)		
Resistive	80.1±0.5	76.4±0.6**
Reflection	-48.1±1.0	-36.0±1.3**
Vascular elastic recoil	31.8±0.9	24.4±1.2**

Table 3. Gender difference in the outcome variables and anthropometric data. The values were mean and SE after adjustment for the aging and exercise effects. BMI: body mass index; SBP: systolic, DBP: diastolic, MBP: mean, PP: pulse blood pressure. *P<0.05, **P<0.001 indicated the significant levels.

Epidemiological studies based on brachial artery pressure indicate that SBP were lower in young premenopausal women than in age-matched men (London et al., 1995; Kannel et al., 1985). The results were consistent finding that women had lower blood pressure in brachial artery than in men. Generally, SBP and PP increased as a pulse travels from aorta toward the peripheral, the increase being all the more pronounced as the distance of pulse propagation (London et al., 1995; Latham et al., 1985). Women had larger reflected waves than men, in part due to shorter body height and closer physical proximity between heart and reflecting sites. However, body height was not sufficient to fully explain higher reflected wave flow and pressure in women. In the study we indicated that the reflected wave had higher in women and was significantly correlated to body height and weight. Although, pressure wave reflection and propagation are known recently to correlate with body height (London et al., 1995; Latham et al., 1985; Mitchell et al., 2004), however, we also found that increased reflected wave was partially contributed by decreased body weight and increased heart rate level. The gender difference in arterial hemodynamics in carotid blood velocity waveforms is probably accounted for body height and weight.

It had been reported that women had lower carotid artery distensibility compared with men (Ylitalo et al., 2000). From the findings of present study, we agreed that women had lower arterial elasticity using the proposed velocity indices. The difference in the velocities and its indices were related to smaller body size in women that largely accounted for the gender differences. However, the difference in velocity indices was also influenced by concentrations of estrogen in hormone status of women (Krejza et al., 2001).
The gender difference in velocity waveforms in CCA found in this population was not depended on blood pressure. It was demonstrated that the gender difference in blood velocity waveforms of CCA are not directly linked to it pressure waveforms (Azhim et al., 2007b).
Although the finding in the effect of increased wave reflection in arterial system on body height was consistent, because the relation of body weight and body fat on the artery stiffness and flow velocities were largely unknown, further investigations are needed. The Doppler angle of insonation was important because it must be taken into account when calculating blood flow velocity from the Doppler shift frequency. However, the velocity indices of were independent of the insonating angle so that the assessments of hemodynamics were more accurate and reliable.

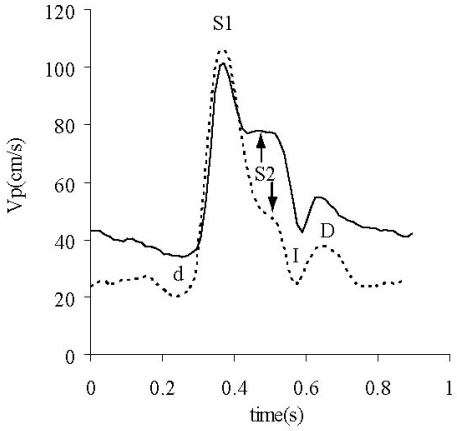


Fig. 11. Comparison of typical flow velocity waveforms in CCA for gender difference of man (dashed line) and woman (solid line). Subject's details were 171 cm, 65 kg, BMI: 22 kg/m^2, age: 23 years for man and 154 cm, 48 kg, BMI: 20 kg/m^2, age: 25 years for woman.

5. Conclusion

In the chapter, we have presented first, the portable measurement system has developed for ambulatory and nonivansive determination of blood circulation with synchronized of blood pressure and ECG signals, which has potential to provide the critical information in clinical and healthcare applications. Second, there are multiple factors which have effects on blood velocity waveforms in CCA. Regular exercise training is able to improve age-associated decrease blood velocity in CCA with similar effect between young and older exercise-trained. The velocity waveform patterns have no significantly change with age in entire

groups who regularly performed aerobic exercise. Gender-associated difference in the outcome of velocities and the indices is also found in the study. Reference data for normal velocities and the indices in CCA are determined after adjustment for the effects of age, gender, and exercise training. Reductions in blood flow velocities are believed to have contributed significantly to the pathophysiology of age-associated increase in not only cardiovascular but also cerebrovascular diseases. The findings have potentially important clinical and healthcare requirements for prevention of cardiovascular diseases.

6. References

Azhim, A.; Akioka, K.; Akutagawa, M.; Hirao, Y.; Yoshizaki, K.; Obara, S.; Nomura, M.; Tanaka, H.; Yamaguchi, H. & Kinouchi, Y. (2007c). Effects of aging and exercise training on the common carotid blood velocities in healthy men. *Conf. Proc. IEEE Eng. Med. Biol. Soc.*, vol. 1, pp. 989-99

Azhim, A.; Katai, M.; Akutagawa, M.; Hirao, Y.; Yoshizaki, K.; Obara, S.; Nomura, M.; Tanaka, H.; Yamaguchi, H. & Kinouchi, Y. (2008) Measurement of blood flow velocity waveforms in the carotid, brachial and femoral arteries during head-up tilt. *Journal of Biomedical & Pharmaceutical Engineering*, vol. 2-1, pp. 1-6

Azhim, A.; Akioka, K.; Akutagawa, M.; Hirao, Y.; Yoshizaki, K.; Obara, S.; Nomura, M.; Tanaka, H.; Yamaguchi, H. & Kinouchi, Y. (2007b). Effect of gender on blood flow velocities and blood pressure: Role of body weight and height. *Conf Proc IEEE Eng Med Biol Soc.*, pp. 967-970

Azhim, A.; Katai, M.; Akutagawa, M.; Hirao, Y.; Yoshizaki, K.; Obara, S.; Nomura, M.; Tanaka, H.; Yamaguchi, H. & Kinouchi, Y. (2007a). Exercise improved age-associated changes in the carotid blood velocity waveforms. *Journal of Biomedical & Pharmaceutical Engineering*, vol. 1-1, pp. 17-26

Azhim, A.; Kinouchi, Y. & Akutagawa, M. (2009). Biomedical Telemetry: Technology and Applications, In: *Telemetry: Research, Technology and Applications,* Diana Barculo and Julia Daniels, (Eds.), Nova Science Publishers, New York, ISBN: 978-1-60692-509-6 (2009)

Baskett, J. J.; XBeasley, J. J.; Murphy, G. J.; Hyams, D. E. & Gosling, R. G. (1977). Screening for carotid junction disease by spectral analysis of Doppler signals. *Cardiovasc Res.*, vol. 11(2), pp. 147-55

Chen, C. & Dicarlo, SE. (1997). Endurance exercise training-induces resting bradycardia. *Sport Med. Training Rehabil.*, vol. 8, pp. 37-77

Costill, D. (1986). *Inside running: basics of sports physiology*. Benchmark Press, Indianapolis, pp. 15

Dahnoun, N.; Thrush, A.J.; Fothergill, J.C. & Evans, D.H. (1990). Portable directional ultrasonic Doppler blood velocimeter for ambulatory use. *Med Biol Eng Comput*, vol. 28, pp. 474-482

Darne, B.; Girerd, X.; Safar, M.; Cambien, F. & Guize L. (1989) Pulsatile versus steady component of blood pressure: a cross-sectional analysis on cardiovascular mortality. *Hypertension*, vol. 13, pp. 392-400

Donofrio MT, Bremer YA, Schieken RM, Gennings C, Morton LD, Eidem BW, Cetta F, Falkensammer CB, Huhta JC and Kleinman CS. Autoregulation of cerebral blood

flow in fetuses with congenital heart disease: The brain sparing effect. *Pediatr Cardiol* 2003; 24: 436-443

Fujishiro, K. & Yoshimura, S. (1982). Haemodynamic change in carotid blood flow with age. *J. Jekeikai Med*, vol. 29, pp. 125-138

Goldsmith, R.L.; Bloomfeld, D.M. & Rosenwinkel, E.T. (2000). Exercise and autonomic function. *Coron. Artery Dis.*, vol. 11, pp. 129-135

Gosling, R.G. (1977). Extraction of physiological information from spectrum-analysed Doppler-shifted continuous wave ultrasound signals obtained non-invasively from the arterial system. In: *Institute of Electrical Engineers medical electronics monographs*, Hill D.W. & Watson B.W., (Eds), pp. 73-125, Peter Peregrinus, Stevenage

Gregova, D.; Termerova, J.; Korsa, J.; Benedikt, P.; Peisker, T.; Prochazka, B.; & Kalvach, P. (2004) Age dependence of flow velocities in the carotid arteries. *Ceska a Slovenska Neurologie a Neurochirurgie*, vol. 67 (6), pp. 409-414, 2004 (abstract in English)

He, J.; Kinouchi, Y.; Iritani, T.; Yamaguchi, H. & Miyamoto, H. (1992). Telemetering blood flow velocity and ECG during exercise. *Innov Tech Biol Med.*, vol. 13, pp. 567-577

He, J.; Pan, A. W.; Ozaki, T.; Kinouchi, Y. & Yamaguchi, H. (1996). Three channels telemetry system: ECG, blood velocities of the carotid and the brachial arteries. *Biomedical Engineering Applications Basis Communications*, vol. 8, pp. 364-369

Jiang, Z-L.; He, J.; Yamaguchi, H.; Tanaka, H. & Miyamoto, H. (1994). Blood flow velocity in common carotid artery in humans during breath-holding and face immersion. *Aviat Space Environ Med.*, vol. 65, pp. 936-943

Jiang, Z-L.; Yamaguchi, H.; Takahashi, A.; Tanabe, S.; Utsuyama, N.; Ikehara, T.; Hosokawa, K.; Tanaka, H.; Kinouchi, Y. & Miyamoto, H. (1995). Blood flow velocity in the common carotid artery in humans during graded exercise on a treadmill. *Eur J Appl Physiol*, vol. 70, no. 3, pp. 234-239

Johannes, S.; Michael, S.; Thomas, W.; Wolfgang, R.N.; Markus, V.; Markus, L. & Stefan F. (2001). Quantification of blood flow in the carotid arteries comparison of Doppler ultrasound and three different phase-contrast magnetic resonance imaging sequences. *Investigate Radiology*, vol. 36-11, pp. 642-647

Kaneko, Z.; Shiraishi, J.; Inaoka, H.; Furukawa, T. & Sekiyama, M. (1978). Intra- and extracerebral hemodynamics of migrainous headache. In: *Current concepts in migraine research*, Greene, R. (Ed.), pp. 17-24, Raven, New York

Kannel, W. B. & Stokes III, J. (1985). Hypertension as a cardiovascular risk factor. In: *Handbook of Hypertension. Clinical Aspects of Hypertension*, Robertson, J.I.S. (Ed.), pp. 15-34, Elsevier Science Publishing, New York

Krejza, J.; Mariak, Z.; Huba, M.; Wolczynski, S. & Lewko, J. (2001). Effect of endogenous estrogen on blood flow through carotid arteries. *Stroke*, vol. 32, pp. 30-36

Lakatta, E.G. (2002). Age-associated cardiovascular changes in health: Impact on cardiovascular disease in older persons. *Heart Fail Rev*, vol. 1, pp. 29-49

Latham, R. D.; Westerhof, N.; Sipkema, P.; Rubal, B. J.; Reuderink, P. & Murgo, J. P. (1985). Regional wave travel and reflections along the human aorta: A study with six simultaneous micromanometric pressures. *Circulation*, vol. 72, pp. 1257-1269

London, G.M.; Guerin, A.P.; Pannier, B.; Marchais, S.J. & Stimpel, M. (1995). Influence of sex on arterial hemodynamics and blood pressure: Role of body height. *Hypertension*, vol. 26, pp. 514-519

Maciel, B.C.; Gallo, L.; Marin-Neto, JA; Lima-Filho, E.C. & Mancoy, J.C. Parasympathetic contribution to bradycardia induced by endurance training in man. *Cardiovasc Res* 1985; 19: 642-648

Marchais, S.J.; Guerin, A.P.; Pannier, B.M.; Levy, B.I.; Safar, M.E. & London, G.M. (1993). Wave reflections and cardiac hypertrophy in chronic uremia: Influence of body size. *Hypertension*, vol. 22, pp. 876-883

Mitchell, G. F.; Parise, H.; Benjamin, E. J.; Larson, M. G.; Keyes, M. J.; Vita, J. A.; Vasan, R. S. & Levy, D. (2004). Changes in arterial stiffness and wave reflection with advancing age in healthy men and women: The Framingham Heart Study. *Hypertension*, vol. 43, pp.1239-1245

Murgo, J.; Westerhof, N.; Giolma, J. P. & Altobelli, S. (1980). Aortic impedance in normal man: relationship to pressure waveforms. *Circulation*, vol. 62, pp. 105-16

Nagatomo, I.; Nomaguchi M. & Matsumoto K. (1992). Blood flow velocity waveform in the common carotid artery and its analysis in elderly subjects. *Clin Auton Res.*, vol. 2(3), pp. 197-200

Nichols, W. W. & O'Rourke, M. F. (2005) *McDonald's Blood Flow in Arteries: Theoretic, Experimental and Clinical Principles*. Hodder Arnold, ISBN 0-340-80941-8, London

Permal JM. Neonatal cerebral blood flow velocity measurement. *Clin Perinatol* 1985; vol. 12, pp. 179-193

Planiol T and Pourcelot L. (1973). Doppler effects study of the carotid circulation, In: *Ultrasonics in medicine*, Vlieger, M.; White, D.N. & McCready, V.R. (Eds), pp. 141-147, Elsevier, New York

Pourcelot L. (1976). Diagnostic ultrasound for cerebral vascular diseases, In: *Present and future of diagnostic ultrasound*, Donald, I. & Levi, S., (Eds), pp. 141-147, Kooyker, Rotterdam

Prichard, D. R.; Martin, T. R. & Sherriff, S. B. (1979). Assessment of directional Doppler ultrasound techniques in the diagnosis of carotid artery diseases. *Journal of Neurology, Neurosurgery, and Psychiatry*, vol. 42, pp. 563-568

Rutherford, R.B; Hiatt, W.R. & Kreuter, E.W. (1977). The use of velocity wave form analysis in the diagnosis of carotid artery occlusive. *Surgery*, vol. 82-5, pp. 695-702

Satomura S. (1959). Study of the flow pattern in peripheral arteries by ultrasonics. *J. Acoust Soc Jpn*, vol. 15, pp. 151-158

Scheel, P.; Ruge, C. & Schoning, M. (2000). Flow velocity and flow volume measurements in the extracranial carotid and vertebral arteries in healthy adults: Reference data of age. *Ultrasound Med Biol.*, vol. 26, pp. 1261-1266

Schmidt-Trucksass, A.; Grathwohl, D.; Schmid, A.; Boragk, R.; Upmeier, C.; Keul, J. & Huonker M. (1999). Structural, functional, and hemodynamic changes of the common carotid artery with age in male subjects. *Arterioscler Thromb Vasc Biol.*, vol. 19, pp. 1091-1097

Tanaka, H.; Dinenno, F. A.; Monahan, K. D.; Christopher, M. C.; Christopher, A. D. & Seals, D.R. (2000). Aging, habitual exercise, and dynamic arterial compliance. *Circulation*, vol. 102, pp. 1270-1275

Ylitalo, A.; Airaksinen, K.E.; Hautanen, A. M.; Kupari, A.; Carson, M.; Virolainen, J.; Savolainen, M.; Kauma, H.; Kesaniemi, Y.A.; White, P.C. & Huikuri, H.V. (2000). Baroreflex sensitivity and variants of the renin angiotensin system genes. *J. Am. Coll. Cardiol.*, vol. 35, pp. 194-200

Yuhi, F. (1987). Diagnostic characteristics of intracranial lesions with ultrasonic Doppler sonography on the common carotid artery. *Med J Kagoshima Univ.*, vol. 39, pp. 183-225 (abstract in English)

Zhang, D.; Hirao, Y.; Kinouchi, Y.; Yamaguchi, H. & Yoshizaki, K. (2002). Effects of nonuniform acoustic fields in vessels and blood velocity profiles on Doppler power spectrum and mean blood velocity. *IEICE Transactions on Information and Systems*, vol. E85-D, pp. 1443-1451

Studying Ion Channel Dysfunction and Arrythmogenesis in the Human Atrium: A Computational Approach

Sanjay R. Kharche, Phillip R. Law, and Henggui Zhang
The University of Manchester, Manchester, UK

1. Introduction

Human atrial fibrillation (AF) is the most common sustained clinically observed cardiac arrhythmia causing mortality and morbidity in patients with increasing incidence in the elderly (Aronow 2009; Wetzel, Hindricks et al. 2009). It is prevalent in the developed world and a considerable burden on health care services in the UK and elsewhere (Stewart, Murphy et al. 2004; Aronow 2008a; Aronow 2008b). AF is a heterogeneously occurring disease often in complex with embolic stroke, thromboembolism, heart failure and other conditions (Novo, Mansueto et al. 2008; Bourke and Boyle 2009; Roy, Talajic et al. 2009). The treatment of paroxysmal AF includes pharmacological intervention primarily targeting cellular ion channel function (Ehrlich and Nattel 2009; Viswanathan and Page 2009). Persistent AF where episodes last for prolonged periods possibly requires electrical cardioversion (Wijffels and Crijns 2003; Conway, Musco et al. 2009) or repeated surgical interventions that isolate focal trigger sites that induce AF (Gaita, Riccardi et al. 2002; Saltman and Gillinov 2009; Stabile, Bertaglia et al. 2009). A better understanding of the underlying ion channel and structural mechanisms of AF will assist in design of improved clinical therapy at all stages of the disease.

The structure of the human atrium is shown in Fig. 1. Mechanisms underlying the genesis of AF are poorly understood yet. It is believed to be predominantly initiated by focal ectopic activity in the cristae terminalis of the right atrium, and pulmonary vein ostia in the left atrium (Haissaguerre, Jais et al. 1998). Spontaneous focal activities in the atrium could also be generated by intracellular calcium ($[Ca^{2+}]_i$) dysfunction (Chou and Chen 2009). The ectopic activity, under AF conditions, normally leads to a persistent single mother rotor of re-entrant excitation circuits. Upon interaction with anatomical obstacles along with intra-atrial electrical heterogeneity, the mother rotor wavefront breaks giving rise to smaller randomly propagating electrical wavefronts resulting in rapid erratic excitation of the atria (Moe, Rheinboldt et al. 1964) leading to uncoordinated contractions of the myocardium, which is reflected in the abnormal P-wave and R-R intervals of clinical ECG (Rosso and Kistler 2009). Recently a new mechanism, "AF begets AF" (Wijffels, Kirchhof et al. 1995) due to *AF induced electrical remodelling* (AFER), has been identified by which rapid excitation of atrial tissue gives rise to persistent AF AFER produces remarkable reduction in atrial *action potential (AP) duration* (APD) and effective refractive period (ERP), which are associated with

AF-induced changes in electrophysiology of ion channels. Several experimental studies have studied the effects of AFER on individual ion channels of human atrial myocytes (Bosch, Zeng et al. 1999; Workman, Kane et al. 2001; Bosch and Nattel 2002; Balana, Dobrev et al. 2003; Ravens and Cerbai 2008), and have identified several ion channels remodelled by chronic AF (Bosch, Zeng et al. 1999; Workman, Kane et al. 2001) .

Another mechanism underlying the genesis of AF is ion channel dysfunction arising from genetic mutations. There is growing interest in identifying genetic bases underlying familial AF following the first study by Chen et al. (Chen, Xu et al. 2003). In the rare but debilitating cases of familial AF, or lone AF, there is no apparent structural remodelling that precludes the onset of AF. However, several clinical studies have characterised the familial nature of several genetic defects that lead to AF (Chen, Xu et al. 2003; Xia, Jin et al. 2005; Makiyama, Akao et al. 2008; Restier, Cheng et al. 2008; Zhang, Yin et al. 2008; Li, Huang et al. 2009; Yang, Li et al. 2009). Hormonal imbalance during AF also causes electrical remodelling (Cai, Gong et al. 2007; Cai, Shan et al. 2009) that facilitates AF, but is not considered in this Chapter.

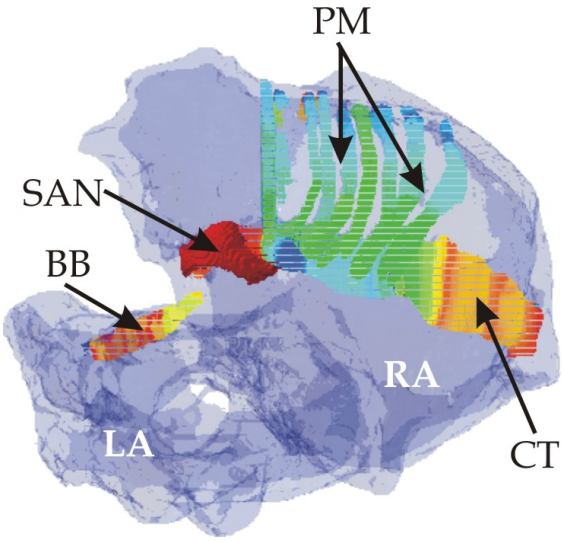


Fig. 1. 3D anatomical model of the human female atria showing internal structure and conduction pathways (figure adapted from our previous study (Zhang, Garratt et al. 2009)). Atrial tissue in the left (LA) and right (RA) atria is homogeneous (translucent blue). The sino-atrial node (SAN) is the pacemaker wherefrom cardiac electrical excitations originate. The main atrial conduction pathways, *i.e.* pectinate muscles (PM), cristae terminalis (CT) and the Bachman's bundles (BB), are the tissue types which possess electrical and structural heterogeneity and contribute to a small proportion of total atrial mass.

Experimental and clinical electrophysiological studies are vital to improve our understanding of AF and its underlying mechanisms. Such studies, however, require vast resources and involve ethical considerations. In addition, the effects of cellular level electrophysiological remodelling at multi-scale levels of cellular and spatially extended tissues is practically impossible in a clinical or physiology laboratory environment. Recently powerful biophysically detailed mathematical models of cardiac cells (Courtemanche,

Ramirez et al. 1998; Nygren, Fiset et al. 1998; Zhang, Holden et al. 2000; Pandit, Clark et al. 2001; ten Tusscher, Noble et al. 2004) and spatially extended tissues have been developed. Such biophysically detailed models of cardiac cells and tissues offer cost effective alternatives to experimental studies to investigate and dissect the effects changes in individual ion channels on cellular AP (Zhang, Garratt et al. 2005; Zhang, Zhao et al. 2007; Salle, Kharche et al. 2008) and tissue conduction properties (Kharche, Garratt et al. 2008; Kharche and Zhang 2008; Keldermann, ten Tusscher et al. 2009). With the ready availability of vast computational power, simulation offers an excellent complimentary method of studying AF *in silico* (Kharche, Seemann et al. 2008; Reumann, Fitch et al. 2008; Bordas, Carpentieri et al. 2009).

In this Chapter, we present a review of some of our recent works on studies of AFER and gene mutations in genesis and maintenance of AF. Comprehensive computational techniques for the quantification of the effects of AFER at cellular and tissue levels are described. Our simulation data at a multi-scale tissue level supported the "AF begets AF" hypothesis (Zhang, Garratt et al. 2005; Kharche, Seemann et al. 2007; Kharche, Seemann et al. 2008; Kharche and Zhang 2008), and demonstrated the dramatic pro-fibrillatory effects of Kir2.1 V93I gene mutation on the human atrium computational study (Kharche, Garratt et al. 2008). Techniques of high performance computing and visualisation of the computationally intensive 3D simulations are discussed.

2. Multi-scale simulation of the effects of AFER and Ione AF

In our studies of human atrial AF, we choose the widely used biophysically detailed cell model for human atrial AP developed by Courtemanche *et al.* (Courtemanche, Ramirez et al. 1998) (CRN). This 21 variable electrophysiological model consists of several sarcolemmal ion channel currents, pumps and exchanger currents, along with a sufficiently detailed intracellular ionic homeostasis mechanism. The model is able to reproduce human atrial AP accurately. Electrophysiological changes due to AFER and Kir2.1 V93I gene mutation can be immediately incorporated into this model allowing ready simulation of the resulting AP and $[Ca^{2+}]_i$ transients. Further, as described later in this section, the cellular models can be incorporated into multi-cellular tissue models using reaction diffusion formulations to simulate conduction propagation behaviour. To quantify the effects of AFER and Kir2.1 V93I gene mutation, a series of experimental protocols are computationally emulated quantifying their effects on atrial excitation at cellular and 3D anatomically detailed models.

2.1 Single cell modelling: electrophysiological changes due to AFER and monogenic AF

AFER and Kir2.1 V93I mutation both alter the biophysical properties of sarcolemmal ion channels underlying human atrial AP. Changes in ion channel current densities, time kinetics and steady state properties of ion channels have been quantified by experimental and clinical studies. The experimental data regarding AFER was obtained from two extensive studies wherein the effects of chronic human AF on atrial ion channels properties were studied. The study by Bosch et al. (Bosch, Zeng et al. 1999) considered patients with AF episodes lasting for more then 1 month (AF1), while the study by Workman et al. (Workman, Kane et al. 2001) considers patients with AF episodes lasting for more than 6 months (Workman, Kane et al. 2001) (AF2). In brief, remodelling in AF1 includes a 235% increase of the maximal conductance of the inward rectifier potassium current I_{K1}, 74%

reduction of the conductance of the L-type calcium current $I_{Ca,L}$, 85% reduction of conductance of the transient outward current (I_{to}), a shift of -16 mV of the I_{to} steady-state activation, and a -1.6 mV shift of sodium current (I_{Na}) steady state activation. Fast inactivation kinetics of $I_{Ca,L}$ is slowed down, and was implemented as a 62% increase of the voltage dependent inactivation time constant. Remodelling in AF2 includes a 90% increase of I_{K1}, 64% reduction of $I_{Ca,L}$, 65% reduction of I_{to}, 12% increase of the sustained outward potassium current (I_{Ksus}), and a 12% reduction of the sodium potassium pump ($I_{Na,K}$). Both AF1 and AF2 data have been incorporated into the CRN model in our previous study (Zhang, Garratt et al. 2005).

Simulation of Kir2.1 V93I gene mutation was based on the recent clinical data from Xia et al. (Xia, Jin et al. 2005) who examined several generations of a large family with hereditary AF associated with Kir2.1 V93I gene mutation. The Kir2.1 gene primarily regulates the I_{K1} channel current, which is modelled as

$$I_{K1} = g_{K1}\left(V - E_K\right) \tag{1}$$

$$g_{K1} = ag_{K1\max} + \frac{(1-a)g_{K1\max}}{1+e^{(b(V-c))}} \tag{2}$$

where V is the cell membrane potential; E_K the reversal potential of the channel; g_{K1max} the maximal channel conductance; "a" is the fraction of the channel conductance that is voltage-independent, (1-a) is the fraction of the channel conductance that is voltage-dependent, "b" the steepness of the g_{K1}-V relationship; "c" is the half point of the g_{K1}-V relationship. In simulations, we considered different conditions of the mutation from Control (Con), to heterozygous (Het) to homozygous (Hom) cases. Parametric values of equations 1 and 2 for different conditions of Kir2.1 V93I gene mutation are listed in Table 1, which were based on the experimental study of Xia et al. (Xia, Jin et al., 2005).

Experimental data sets of AFER and Kir2.1 V93I gene mutation as described above were then incorporated into the CRN human atrial AP model to simulate their effects on human atrial excitation at cellular and tissue models. A quantitative summary of all results is given in Table 2.

2.2 Quantifying the effects of AFER and Kir2.1 V93I gene mutation on atrial APs at cellular level

We first quantify the functional effects of AFER and Kir2.1 V93I mutation on atrial cellular APs. Excitable models, including human atrial cell models, are usually at resting state far away from the oscillating state and show rate dependent adaptation upon periodic pacing, similar to those seen experimentally (Workman, Kane et al. 2001; Cherry, Hastings et al. 2008). Therefore, the models have to be conditioned with several pulses before stable excitations can be elicited. In case of the CRN model, it was found that 10 pulses at a pacing cycle length (PCL) of 1 s was sufficient conditioning. Upon simulation, characteristics of AP profiles were quantified by measuring the resting potential and APD at 90% repolarisation (APD_{90}), the overshoot and the maximal upstroke velocity, dV/dt_{max}. APD_{90} reflects the overall changes in ion channel function during AP. dV/dt_{max} on the other hand, not only

Quantity	Con	Het	Hom
g_{K1max} (nS/pF)	0.09 (100%)	0.13 (141% ↑)	0.16 (173% ↑)
a	0.0	0.0355	0.0575
b (mV^{-1})	0.070	0.156	0.232
c (mV)	-80.0	-60.1	-54.7

Table 1. Parameters of I_{K1} equations (1-2) for various Kir2.1 V93I gene mutation conditions. Values were determined based on experimental data of Xia et al. (Xia, Jin et al. 2005) under Con, Het and Hom conditions.

influences cellular behaviour, but also the conduction properties at tissue level (Biktashev 2002). Due to the large increase in repolarisation potassium currents and reduction in depolarising currents, the AP profiles show large abbreviation in APD_{90} under AFER and Kir2.1 V93I gene mutation conditions. APD abbreviation under AFER conditions is due to a integral actions of remodelled ion channels. However, in the gene mutation condition, such an abbreviation is caused by gain-in-function of the I_{K1} channel. The effects of AFER and Kir2.1 V93I gene mutation on AP profiles are shown in Fig. 2.

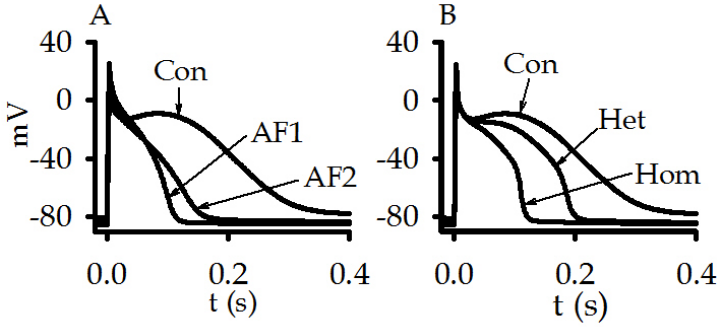


Fig. 2. AP profiles under AFER (A) and Kir2.1 V93I gene mutation (B) conditions. AFER and the mutation cause a dramatic abbreviation of APD.

APD restitution (APDr) measures the excitation behaviour of atrial cells subjected to premature pulses immediately after a previous excitation (Franz, Karasik et al. 1997; Qi, Tang et al. 1997; Kim, Kim et al. 2002; Burashnikov and Antzelevitch 2005; Cherry, Hastings et al. 2008). Recent experimental and modelling studies have shown the correlation between the maximal slope of APDr greater than unity and instability of re-entrant excitation waves in 2D and 3D tissues (Xie, Qu et al. 2002; Banville, Chattipakorn et al. 2004; ten Tusscher, Mourad et al. 2009). In our study, APDr is computed using a standard S1S2 protocol. A train of ten conditioning stimuli (S1) at a physiological PCL were applied before the premature pulse (S2) was applied. The time interval between the final conditioning excitation and onset of the premature excitation emulates atrial diastolic interval (DI), or the time the atrial organ has for recovery from the previous excitation. In the CRN model, S1 and S2 have stimulus amplitude of 2 nA and duration of 2 ms. A plot of the DI against APD_{90} gives APDr, as shown in Fig. 3 for Control, AFER and Kir2.1 V93I gene mutation conditions. At large DI,

APDr curves have negligible slopes and show AP profiles under physiological rates of pacing. At low DI, however, the slopes are noticeable. Under AFER conditions, the computed APDr slopes under various conditions are much greater than under Control conditions (Table 2).

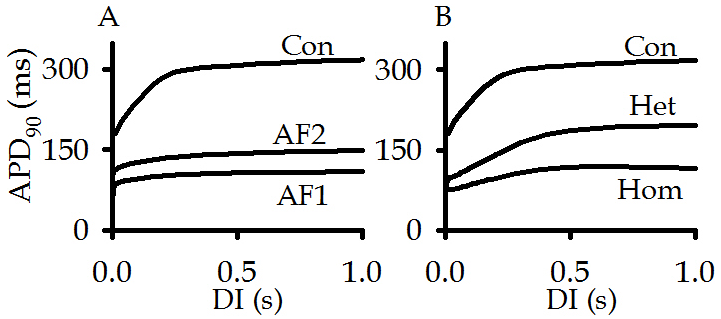


Fig. 3. APDr profiles under AFER (A) and Kir2.1 V93I gene mutation (B) conditions. At large DI, APDr curves reflect the changes in APD_{90} under Control (Con) and AF (AF1, AF2, Het and Hom) conditions. At low DI, the maximal slopes of APDr curves indicate the instabilities in 2D and 3D simulations. Quantitative details are given in Table 2.

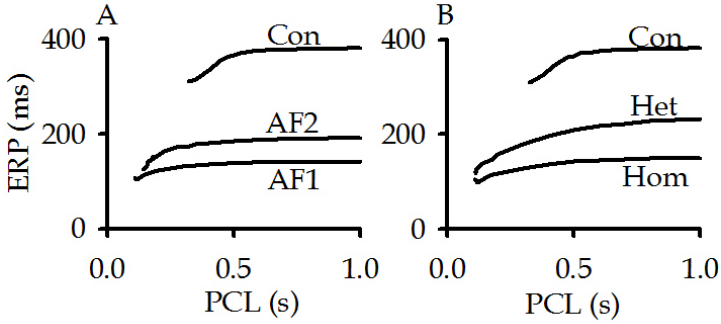


Fig. 4. ERP restitution curves under AFER (A) and Kir2.1 V93I gene mutation (B) conditions.

Shortening of atrial APD and effective refractory period (ERP) are well recognised features of atrial electrical activities during AF. ERP is generally measured by using cellular or tissue preparations (Workman, Kane et al. 2001; Laurent, Moe et al. 2008). In our studies, we adopted the cell based experimental protocol as described by Workman et al. (Workman, Kane et al. 2001) where the cell was stimulated 10 times at various PCLs. A premature stimulus S2 was then applied. The maximal time interval between S1 and S2 where the final excitation has AP amplitude of 80% as compared to the premature pulses is defined as the ERP. Due to the rate dependent adaptability of atrial AP, we usually compute ERP at several PCL values to obtain an ERP restitution curve. Results are shown in Fig. 4. It can be seen that AF reduces ERP (Table 2). Such a reduction is in qualitative agreement with experimental observations and clinical data (Workman, Kane et al. 2001; Li, Hertervig et al. 2002; Oliveira, da Silva et al. 2007).

2.3 1D and 2D tissue modelling

Human atrial tissue is spatially and electrically homogeneous tissue (Jalife 2003; Seemann, Hoper et al. 2006). The primary sources of heterogeneity in the human atrium are the conduction pathways as shown in Fig. 1, which contribute only a small fraction to total atrial mass. Therefore, it is reasonable to take human atrial tissue as homogeneous in simulations of the effects of AFER and Kir2.1 V93I gene mutation on atrial excitations (Kharche, Garratt et al. 2008; Kharche, Seemann et al. 2008).

To simulate atrial excitation at the tissue level, the CRN atrial cell AP model is incorporated into tissue models using a mono-domain reaction diffusion partial differential equation,

$$\frac{\partial V(r)}{\partial t} = -D\nabla^2 V(r) + I_{ion}(r) \tag{3}$$

where D is the homogeneous diffusion constant mimicking the intracellular gap junctional coupling, ∇^2 is the Laplacian operator and I_{ion} is the total reactive current at any given spatial location r in the tissue associated with the ion channels of the atrial cell at r. We take D to be 0.03125 mm²/ms to give physiological value of conduction velocity (CV) of 0.265 mm/ms, which falls in the range of physiological measurements. Such a formulation is sufficient for our purposes as we do not consider any extracellular potentials, fluids or indeed mechanical activity, for which more complex bi-domain formulations have to be adopted (Potse, Dube et al. 2006; Whiteley 2007; Vigmond, Weber dos Santos et al. 2008; Linge, Sundnes et al. 2009; Morgan, Plank et al. 2009).

To quantify the functional effects of AFER and Kir2.1 V93I gene mutation on atrial CV restitution (CVr) and temporal vulnerability (VW), models of 1D homogeneous atrial strand were used. CVr is computed by conditioning the 1D strand (S1) after which a premature pulse is applied. The CV of the second propagation as a function of the inter-pulse duration, or PCL, is termed as CVr. CV of propagations is computed from the central region of the strands as shown in Fig 5A. CVr for AFER and the gene mutation conditions are shown in Fig. 5, B and C, where the stimulation protocol is also illustrated. As can be seen, AF reduces solitary wave CV, i.e. CV at large PCL, or low pacing rates. Such CV reduction is not due to any changes in the inter-cellular coupling in the tissue, but solely due to the changes of atrial cell AP profiles. Our simulation data revealed that atrial tissue has better ability to sustain atrial conduction at fast pacing rates under AFER or gene mutation conditions than under Control conditions.

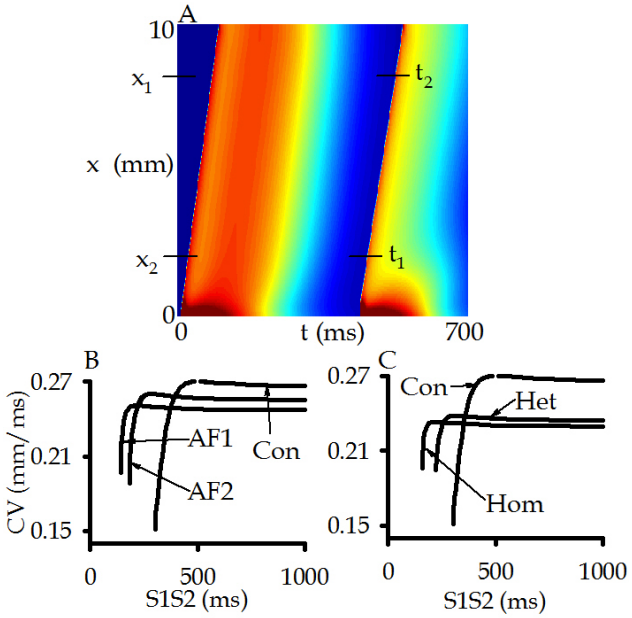


Fig. 5. (A) Electrical waves in a 1D strand where the first wave conditions the tissue, whilst the second wave is initiated after an interval S2. CV is computed according to when the second wave is at x1 (t1) and x2 (t2). (B) CVr under AFER conditions. (C) CVr under Kir2.1 V93I gene mutation conditions.

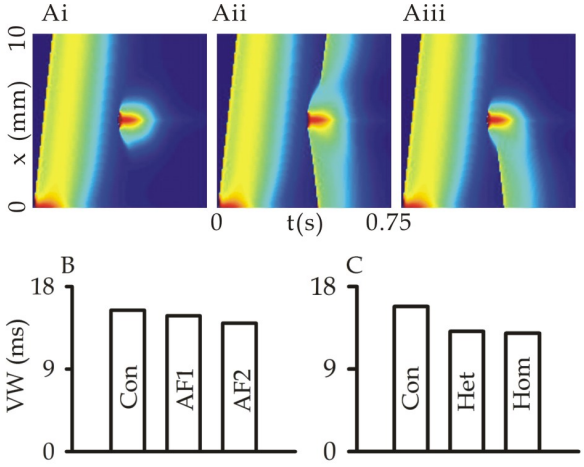


Fig. 6. Atrial excitation wave evoked by a S2 stimulus, applied at a time delay after the conditioning excitation wave, can be either bi-directional blocked (Ai) if the time delay is too soon, or bi-directional conduction (Aii) if the time delay is too late, or uni-directional conduction block (Aiii) if the time delay falls in the VW. Computed VW under AFER conditions (B) and Kir2.1 V93I gene mutation conditions (C).

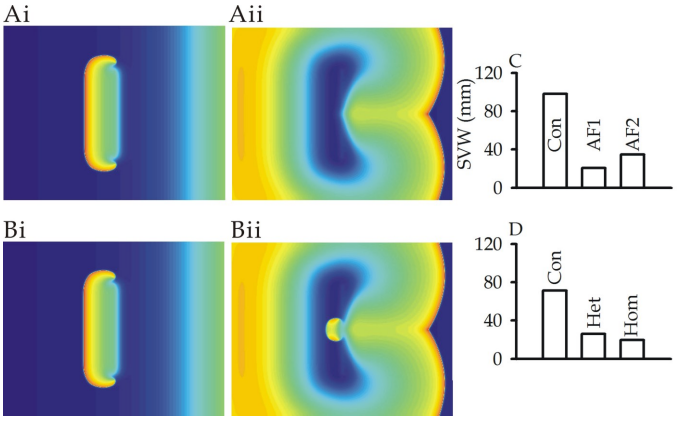


Fig. 7. Computed SVW from 2D tissue models by applying a premature stimulus in the repolarisation tail of a conditioning pulse so as to evoke a figure of 8 re-entry (Ai, Aii and Bi, Bii). The minimal length of the premature stimulus such that the evoked reentry sustains is termed as SVW. (C) SVW under AFER conditions. (D) SVW under Kir2.1 V93I gene mutation conditions. AFER and the gene mutation cause a dramatic reduction of SVW allowing the tissue to sustain re-entry with reduced substrate size.

Uni-directional conduction block in atria can lead to genesis of re-entrant excitation waves. Temporal vulnerability or vulnerability window (VW) measures the vulnerability of cardiac tissue to genesis of uni-directional conduction block. VW is computed by allowing a single solitary wave to propagate from one end of the 1D tissue to the other. After certain duration and in the repolarisation phase in the middle of the tissue, a premature pulse is applied. The time window during which the premature pulse elicits uni-directional propagation block is termed as the VW. Fig. 6 illustrates the protocol and also shows the measured VW under AFER and Kir2.1 V93I gene mutation conditions.

The effects of AFER and the Kir2.1 gene mutation on atrial tissue's spatial vulnerability are quantified by using 2D homogeneous models of human atrial tissue. Spatial vulnerability (SVW) is computed as the minimal atrial substrate size that can sustain re-entrant waves. To this end, a sufficiently long pulse as shown in Fig. 7 is applied in the repolarisation tail of the conditioning pulse, giving rise to a figure of "8" re-entrant waves. The minimum length that sustains such re-entry is termed as SVW. The results for AFER and gene mutation conditions are given in Fig. 7.

Effects of the AFER and Kir2.1 V93I gene mutation on the dynamical behaviours of re-entrant excitation waves are also studied. In 2D tissues, re-entrant wave simulations are performed in a tissue with a size of 37.5 cm x 37.5 cm. In simulations, re-entrant waves are initiated by using a cross-field stimulation protocol. After allowing a planar wave to sufficiently propagate through the 2D sheet, a cross-field stimulus is applied so as to initiate re-entry (Kharche, Seemann et al. 2007). Upon initiation of a re-entrant wave in the middle of the tissue, the re-entrant waves are allowed to evolve for several seconds. Results are shown in Fig. 8. Under Control conditions, the 2D re-entrant waves self-terminate. However, under AFER and Kir2.1 V93I gene mutation conditions, re-entrant waves become persistent. During the simulation, time series of APs from representative locations were also

recorded to allow analysis of dominant frequency of the re-entry. It is shown that the rates of atrial re-entrant excitation waves increased markedly from Control conditions to AF ER and gene mutation conditions. Traced trajectory of the core tips of re-entrant excitation illustrated the increased stability and persistence of the re-entrant waves under AFER and gene mutation conditions. These results are shown in Fig. 9.

2.4 Simulation of re-entrant waves in a 3D realistic geometry

The 3D anatomically detailed spatial model of human female atria as shown in Fig. 1 was developed in a previous study (Seemann, Hoper et al. 2006). It is based on the anatomical geometry of the human atria reconstructed from the visible human project (Ackerman, 1991; Ackerman and Banvard 2000). The anatomical model consists of electrically homogeneous atrial tissue, the SAN and conduction pathways. The SAN is the main pacemaker wherefrom cardiac electrical excitation originates. The conduction pathways are electrically and structurally heterogeneous and assist in normal conduction of electrical excitation in the human atrium. In our studies, we however study re-entrant waves and therefore do not consider SAN electrical activity, nor the heterogeneity associated with the conduction pathways. All cells in our 3D anatomical model simulations are considered to be electrically homogeneous.

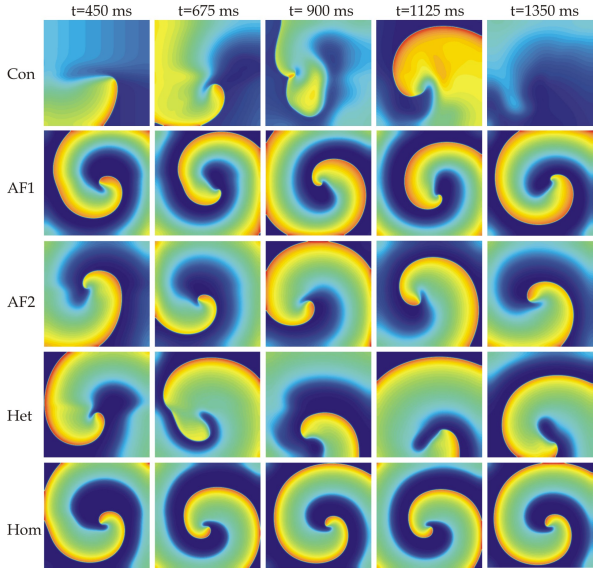


Fig. 8. Representative frames at regular intervals from 2D homogeneous re-entrant waves simulations under Control, AFER and Kri2.1 V93I gene mutation conditions. Re-entry self-terminates under Control conditions (top row), but becomes persistent under AFER and gene mutation conditions.

Re-entrant waves were initiated and allowed to propagate through the electrically and anatomically homogeneous model under Control, AFER and gene mutation conditions. The re-entrant waves were initiated using a protocol similar to the 2D case at a place in the right atrium to reduce boundary effects and interference from anatomical obstacles. The right

atrium was chosen to be ideal as it offers minimal anatomical defects interfering with the initial evolution of the re-entrant waves. Results from the 3D simulations under Control and AFER and gene mutation conditions are shown in Fig. 10.
Under Control conditions, re-entry self-terminated at around 4.2 s. AFER however rendered re-entry to be persistent. Again, if we study representative AP profiles during the simulation, we can see that AF increases the dominant frequency. The dominant frequency of oscillations in Control case is low at less than 3 Hz. In contrast, under AFER conditions, the re-entry is persistent with rapid excitation rate. AFER increases stability of the mother rotor under AF2 conditions. Due to the anatomical defects, the mother rotor degenerates into smaller persistent erratic propagating wavelets, with a dominant frequency more than 10 Hz. Similar results were obtained under the Kir2.1 V93I gene mutation conditions as shown in Fig. 11.

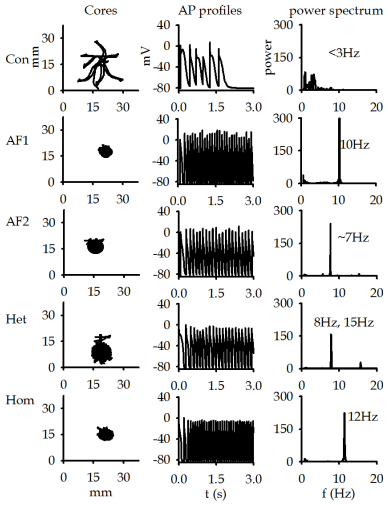


Fig. 9. Dynamical behaviours of 2D re-entrant waves as shown in Fig. 8 with core tip traces (left column), representative AP profiles (middle column) and dominant frequency of the AP profiles (right column) under various AFER and gene mutation conditions. Re-entrant waves are more stable and cause high rate of atrial tissue excitation under AFER and gene mutation conditions.

Our simulations have also shown another important mechanism by which re-entry becomes persistent without effects of AFER or gene mutation. Upon initiation of re-entry close to a blood vessel ostium, the electrical wave readily becomes anchored, as seen in Fig. 12. Such anchoring of an electrical propagation also gives rise to persistent and rapid excitation of atrial tissue.

2.5 Numerical considerations, algorithms and visualisation

Time integration of the CRN cellular models was carried out at a constant time step of 0.005 ms as given in the original CRN model. In the spatial 1D and 2D models, a space step of 0.1 mm was used in an explicit central Euler spatial integration scheme. The inter-node distance

of 0.1 mm represents human atrial size which is close to physiological values. In the 3D models, the space step was taken to be 0.33 mm, which allowed use of a time step of 0.5 ms. These choices gave stable solutions independent of integration parameters.

The 2D and 3D spatial models are large with 140625 and more than 26 x 10^6 nodes respectively. Parallelisation is therefore an important part of cardiac simulations. Solvers that used shared memory parallelism (OpenMP) and large distributed memory parallelism (MPI) were developed in our laboratory. Scaling of the solvers is shown in Fig. 13. In addition to parallelisation, novel cardiac specific algorithms that exploit peculiarities of the model were developed (Kharche, Seemann et al. 2008). The full geometrical model demands very large amounts of contiguous memory. 3D Atrial tissue geometry occupies about 8% geometry of the total data set, due to atrium being thin walled with large holes of atrial chambers and vena caves. We re-structured the computer code such that only atrial nodes, *i.e.* only 8% of the total 26 million nodes and related information are stored in the computer memory. This improved efficacy of memory usage. By re-numbering the real atrial nodes we are not storing any data points that are not atrium. The memory required is reduced to less than 10 GB in the 3D case, and the required computer floating point operations (flops) are also reduced.

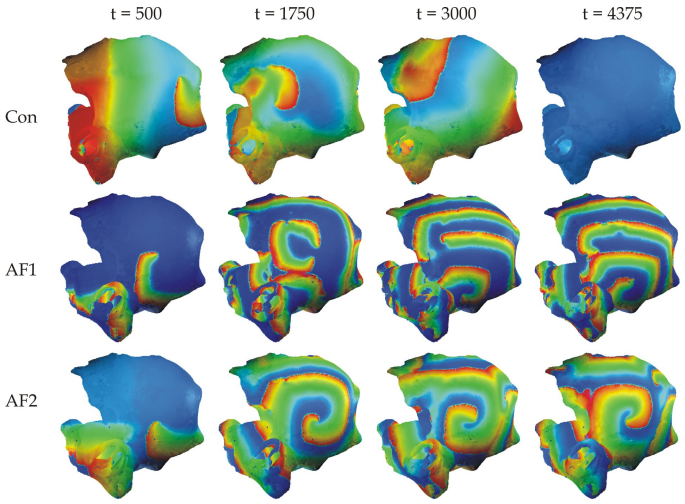


Fig. 10. 3D re-entry under Control (top panels), AF1 (middle panels) and AF2 (bottom panels). Re-entry self-terminates under Control conditions in 4.2 s. Under AF1 conditions, the narrow wavelength re-entrant wave breaks up due to interaction with anatomical obstacles and gives rise to rapid erratic electrical propagations which are persistent. AF2 caused the re-entrant rotor to be stable and gave rise to a mother rotor.

The 3D simulations produce large data sets of more than 30 GB. Traditionally this output is then post-processed to obtain measures quantifying the simulation, e.g. scroll wave filament meander, and to visualise the dynamics of the electrical propagations. Each output file consists of a binary data file of approximately 150 MB size. Efficient visualisation of the 3D data shown in Figs. 10 and 12 was carried using the RAVE package (Grimstead, Kharche et al. 2007) developed elsewhere. We have also developed visualisation techniques based on the visualisation package Advanced Visualisation System (AVS) developed by Manchester Visualisation Centre. This is versatile high level graphical software with a high level of

functionality. Images in Fig. 11 were produced using diamond shaped glyphs, each of which was colour coded with a scalar value, namely the value of voltage at that location.
For smaller visualisation jobs, *e.g.* 2D visualisation, we have used MATLAB due to its functionality and transparent scripting. Development of visualisation scripts using MATLAB is relatively straightforward with a high level of functionality. MATLAB is also available to our laboratory locally. Having successfully developed 2D visualisation pipelines using MATLAB, AVS as a high level visual programming environment is also versatile and the results obtained using MATLAB can be replicated by AVS.

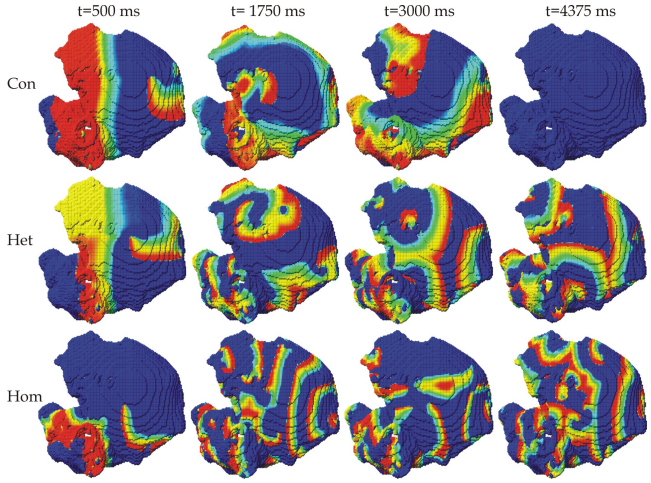


Fig. 11. 3D re-entry under Control (top panels), various Kir2.1 V93I gene mutation conditions (Het, middle panels; Hom, bottom panels). With Kri2.1 V93I gene mutation, condition, the re-entry became erratic leading to rapid excitation of atrial tissue.

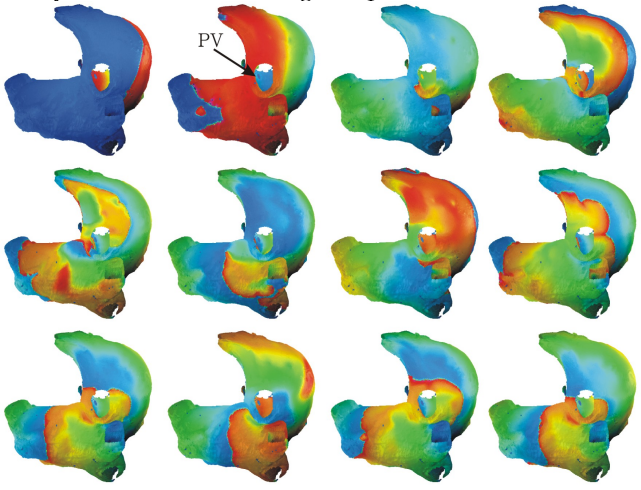


Fig. 12. Anchoring of re-entrant wave to pulmonary vein (PV). Location of PV is marked by the arrow in the first panel of the second column.

Model	Quantity	Con	AF1	AF2	Het	Hom
Cell	Resting potential (mV)	-80.5	-85.2	-83.8	-84.59	-85.07
	APD_{90} (ms)	313.0	108.5	147.6	196.2	137.2
	Overshoot (mV)	22.9	24.6	25.0	24.3	24.1
	dV/dt_{max} (mV/ms)	147.2	86.5	97.2	113.4	98.1
	APDr maximal slope	0.91	4.63	1.56	2.2	0.46
	ERP (ms) (stimulus interval ~ 1 s)	318.0	142.0	192.0	232.0	150.0
1D	CV (mm/ms)	0.27	0.25	0.26	0.26	0.25
	VW (ms)	15.4	14.8	14	13.1	12.9
	Wavelength (mm)	84.51	27.13	38.22	67.3	56.5
2D	LS (s)	1.8	> 10	> 10	>10	>10
	DF (Hz)	<3.0	10.0	7.0	8-15	12
	Tip meander area (mm^2)	615.0	48.0	72.0	101.2	76.1
	SVW (mm)	99.1	20.5	34.7	26.2	17.0
3D	LS (s)	4.2	> 6	> 6	>6	>6
	DF (Hz)	3.0	6.7	6.1	10.2	13.5

Table 2. Quantitative summary of the effects of AFER and Kir2.1 V93I gene mutation on atrial excitations.

3. Conclusions and future work

Our simulation results have shown that both the AFER and Kir2.1 V93I mutation shortened atrial APD and increased the maximal slopes of APDr. They reduced atrial ERP and the intra-atrial CV, all of which facilitated high rate atrial excitation and conduction as observe d experimentally and clinically in AF patients. Due to the large increase in repolarisation currents, the both the AFER and Kir2.1 V93I gene mutation reduced tissue's temporal VW. However, they also reduced the minimal substrate size required to sustain re-entry. Collectively of all these suggested the pro-arrhythmic effects of AFER and Kir2.1 gene mutation. Our results also showed AFER and the gene mutation increased the stability of re-entry, leading them to be persistent.

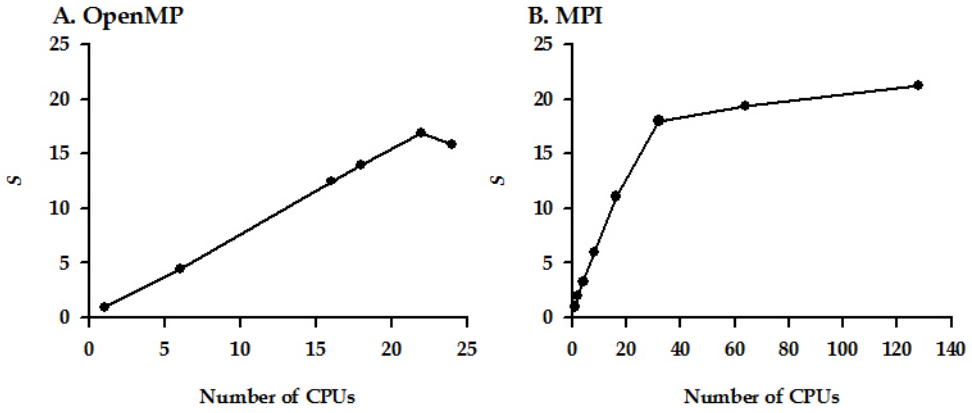


Fig. 13. (A) Scaling of the shared memory (OpenMP) solver. (B) Scaling of the distributed memory (MPI) solver.

These data have provided the first evidence in support of the hypothesis of "AF begetting AF". The methods described above characterise several aspects of the AFER and Kir2.1 gene mutation on generating and sustaining AF. Future studies may consider mechanism involving malfunctioning of intracellular $[Ca^{2+}]_i$ handling (Hove-Madsen, Prat-Vidal et al. 2006), spontaneous firing at atrial blood vessel ostia, interaction between SAN and atria. In addition to macro re-entrant waves, micro re-entry is also an important factor responsible for AF (Markowitz, Nemirovksy et al. 2007). Inclusion of the electrical and spatial heterogeneities in the various tissue sub-types in the atrium will further our understanding the genesis of AF, especially the micro-entry due to heterogeneity boundaries. Computational methods and algorithms can be further improved. This is especially relevant for patient specific simulations where real time results are vital. An immediate aspect of the current simulation-visualisation pipeline that can be addressed is that of incorporating the visualisation, at least partly, into the simulation process. This will enormously improve efficacy of the 3D simulations.

4. References

Aronow, W. S. (2008a). "Management of atrial fibrillation--Part 1." Compr Ther **34**(3-4): 126-33.

Aronow, W. S. (2008b). "Management of atrial fibrillation--Part 2." Compr Ther **34**(3-4): 134-42.

Aronow, W. S. (2009). "Management of atrial fibrillation in the elderly." Minerva Med **100**(1): 3-24.

Balana, B., D. Dobrev, et al. (2003). "Decreased ATP-sensitive K(+) current density during chronic human atrial fibrillation." J Mol Cell Cardiol **35**(12): 1399-405.

Banville, I., N. Chattipakorn, et al. (2004). "Restitution dynamics during pacing and arrhythmias in isolated pig hearts." J Cardiovasc Electrophysiol **15**(4): 455-63.

Biktashev, V. N. (2002). "Dissipation of the excitation wave fronts." Phys Rev Lett **89**(16): 168102.

Bordas, R., B. Carpentieri, et al. (2009). "Simulation of cardiac electrophysiology on next-generation high-performance computers." Philos Transact A Math Phys Eng Sci **367**(1895): 1951-69.

Bosch, R. F. and S. Nattel (2002). "Cellular electrophysiology of atrial fibrillation." Cardiovasc Res **54**(2): 259-69.

Bosch, R. F., X. Zeng, et al. (1999). "Ionic mechanisms of electrical remodeling in human atrial fibrillation." Cardiovasc Res **44**(1): 121-31.

Bourke, T. and N. G. Boyle (2009). "Atrial fibrillation and congestive heart failure." Minerva Med **100**(2): 137-43.

Burashnikov, A. and C. Antzelevitch (2005). "Role of repolarization restitution in the development of coarse and fine atrial fibrillation in the isolated canine right atria." J Cardiovasc Electrophysiol **16**(6): 639-45.

Cai, B., L. Shan, et al. (2009). "Homocysteine modulates sodium channel currents in human atrial myocytes." Toxicology **256**(3): 201-6.

Cai, B. Z., D. M. Gong, et al. (2007). "Homocysteine inhibits potassium channels in human atrial myocytes." Clin Exp Pharmacol Physiol **34**(9): 851-5.

Chen, Y. H., S. J. Xu, et al. (2003). "KCNQ1 gain-of-function mutation in familial atrial fibrillation." Science **299**(5604): 251-4.

Cherry, E. M., H. M. Hastings, et al. (2008). "Dynamics of human atrial cell models: restitution, memory, and intracellular calcium dynamics in single cells." Prog Biophys Mol Biol **98**(1): 24-37.

Chou, C. C. and P. S. Chen (2009). "New concepts in atrial fibrillation: neural mechanisms and calcium dynamics." Cardiol Clin **27**(1): 35-43, viii.

Conway, E. L., S. Musco, et al. (2009). "Drug therapy for atrial fibrillation." Cardiol Clin **27**(1): 109-23, ix.

Courtemanche, M., R. J. Ramirez, et al. (1998). "Ionic mechanisms underlying human atrial action potential properties: insights from a mathematical model." Am J Physiol **275**(1 Pt 2): H301-21.

Ehrlich, J. R. and S. Nattel (2009). "Novel approaches for pharmacological management of atrial fibrillation." Drugs **69**(7): 757-74.

Franz, M. R., P. L. Karasik, et al. (1997). "Electrical remodeling of the human atrium: similar effects in patients with chronic atrial fibrillation and atrial flutter." J Am Coll Cardiol **30**(7): 1785-92.

Gaita, F., R. Riccardi, et al. (2002). "Surgical approaches to atrial fibrillation." Card Electrophysiol Rev **6**(4): 401-5.

Grimstead, I. J., S. Kharche, et al. (2007). Viewing 0.3Tb Heart Simulation Data At Your Desk. EG UK Theory and Practice of Computer Graphics D. D. Ik Soo Lim.

Haissaguerre, M., P. Jais, et al. (1998). "Spontaneous initiation of atrial fibrillation by ectopic beats originating in the pulmonary veins." N Engl J Med **339**(10): 659-66.

Hove-Madsen, L., C. Prat-Vidal, et al. (2006). "Adenosine A2A receptors are expressed in human atrial myocytes and modulate spontaneous sarcoplasmic reticulum calcium release." Cardiovasc Res **72**(2): 292-302.

Jalife, J. (2003). "Experimental and clinical AF mechanisms: bridging the divide." J Interv Card Electrophysiol **9**(2): 85-92.

Keldermann, R. H., K. H. ten Tusscher, et al. (2009). "A computational study of mother rotor VF in the human ventricles." Am J Physiol Heart Circ Physiol **296**(2): H370-9.

Kharche, S., C. J. Garratt, et al. (2008). "Atrial proarrhythmia due to increased inward rectifier current (I(K1)) arising from KCNJ2 mutation--a simulation study." Prog Biophys Mol Biol **98**(2-3): 186-97.

Kharche, S., G. Seemann, et al. (2007). Scroll Waves in 3D Virtual Human Atria: A Computational Study. LNCS. F. B. S. a. G. Seemann. **4466:** 129–138.

Kharche, S., G. Seemann, et al. (2008). "Simulation of clinical electrophysiology in 3D human atria: a high-performance computing and high-performance visualization application." Concurrency and Computation: Practice and Experience **20**(11): 10.

Kharche, S. and H. Zhang (2008). "Simulating the effects of atrial fibrillation induced electrical remodeling: a comprehensive simulation study." Conf Proc IEEE Eng Med Biol Soc **2008**: 593-6.

Kim, B. S., Y. H. Kim, et al. (2002). "Action potential duration restitution kinetics in human atrial fibrillation." J Am Coll Cardiol **39**(8): 1329-36.

Laurent, G., G. Moe, et al. (2008). "Experimental studies of atrial fibrillation: a comparison of two pacing models." Am J Physiol Heart Circ Physiol **294**(3): H1206-15.

Li, Q., H. Huang, et al. (2009). "Gain-of-function mutation of Nav1.5 in atrial fibrillation enhances cellular excitability and lowers the threshold for action potential firing." Biochem Biophys Res Commun **380**(1): 132-7.

Li, Z., E. Hertervig, et al. (2002). "Dispersion of refractoriness in patients with paroxysmal atrial fibrillation. Evaluation with simultaneous endocardial recordings from both atria." J Electrocardiol **35**(3): 227-34.

Linge, S., J. Sundnes, et al. (2009). "Numerical solution of the bidomain equations." Philos Transact A Math Phys Eng Sci **367**(1895): 1931-50.

Makiyama, T., M. Akao, et al. (2008). "A novel SCN5A gain-of-function mutation M1875T associated with familial atrial fibrillation." J Am Coll Cardiol **52**(16): 1326-34.

Markowitz, S. M., D. Nemirovksy, et al. (2007). "Adenosine-insensitive focal atrial tachycardia: evidence for de novo micro-re-entry in the human atrium." J Am Coll Cardiol **49**(12): 1324-33.

Moe, G. K., W. C. Rheinboldt, et al. (1964). "A Computer Model of Atrial Fibrillation." Am Heart J **67**: 200-20.

Morgan, S. W., G. Plank, et al. (2009). "Low energy defibrillation in human cardiac tissue: a simulation study." Biophys J **96**(4): 1364-73.

Novo, G., P. Mansueto, et al. (2008). "Risk factors, atrial fibrillation and thromboembolic events." Int Angiol **27**(5): 433-8.

Nygren, A., C. Fiset, et al. (1998). "Mathematical model of an adult human atrial cell: the role of K+ currents in repolarization." Circ Res **82**(1): 63-81.

Oliveira, M. M., N. da Silva, et al. (2007). "Enhanced dispersion of atrial refractoriness as an electrophysiological substrate for vulnerability to atrial fibrillation in patients with paroxysmal atrial fibrillation." Rev Port Cardiol **26**(7-8): 691-702.

Pandit, S. V., R. B. Clark, et al. (2001). "A mathematical model of action potential heterogeneity in adult rat left ventricular myocytes." Biophys J **81**(6): 3029-51.

Potse, M., B. Dube, et al. (2006). "A comparison of monodomain and bidomain reaction-diffusion models for action potential propagation in the human heart." IEEE Trans Biomed Eng **53**(12 Pt 1): 2425-35.

Qi, A., C. Tang, et al. (1997). "Characteristics of restitution kinetics in repolarization of rabbit atrium." Can J Physiol Pharmacol **75**(4): 255-62.

Ravens, U. and E. Cerbai (2008). "Role of potassium currents in cardiac arrhythmias." Europace **10**(10): 1133-7.

Restier, L., L. Cheng, et al. (2008). "Mechanisms by which atrial fibrillation-associated mutations in the S1 domain of KCNQ1 slow deactivation of IKs channels." J Physiol **586**(Pt 17): 4179-91.

Reumann, M., B. G. Fitch, et al. (2008). "Large scale cardiac modeling on the Blue Gene supercomputer." Conf Proc IEEE Eng Med Biol Soc **2008**: 577-80.

Rosso, R. and P. Kistler (2009). "Focal atrial tachycardia." Heart.

Roy, D., M. Talajic, et al. (2009). "Atrial fibrillation and congestive heart failure." Curr Opin Cardiol **24**(1): 29-34.

Salle, L., S. Kharche, et al. (2008). "Mechanisms underlying adaptation of action potential duration by pacing rate in rat myocytes." Prog Biophys Mol Biol **96**(1-3): 305-20.

Saltman, A. E. and A. M. Gillinov (2009). "Surgical approaches for atrial fibrillation." Cardiol Clin **27**(1): 179-88, x.

Seemann, G., C. Hoper, et al. (2006). "Heterogeneous three-dimensional anatomical and electrophysiological model of human atria." Philos Transact A Math Phys Eng Sci **364**(1843): 1465-81.

Stabile, G., E. Bertaglia, et al. (2009). "Role of pulmonary veins isolation in persistent atrial fibrillation ablation: the pulmonary vein isolation in persistent atrial fibrillation (PIPA) study." Pacing Clin Electrophysiol **32 Suppl 1**: S116-9.

Stewart, S., N. F. Murphy, et al. (2004). "Cost of an emerging epidemic: an economic analysis of atrial fibrillation in the UK." Heart **90**(3): 286-92.

ten Tusscher, K. H., A. Mourad, et al. (2009). "Organization of ventricular fibrillation in the human heart: experiments and models." Exp Physiol **94**(5): 553-62.

ten Tusscher, K. H., D. Noble, et al. (2004). "A model for human ventricular tissue." Am J Physiol Heart Circ Physiol **286**(4): H1573-89.

Vigmond, E. J., R. Weber dos Santos, et al. (2008). "Solvers for the cardiac bidomain equations." Prog Biophys Mol Biol **96**(1-3): 3-18.

Viswanathan, M. N. and R. L. Page (2009). "Pharmacological therapy for atrial fibrillation: current options and new agents." Expert Opin Investig Drugs **18**(4): 417-31.

Wetzel, U., G. Hindricks, et al. (2009). "Atrial fibrillation in the elderly." Minerva Med **100**(2): 145-50.

Whiteley, J. P. (2007). "Physiology driven adaptivity for the numerical solution of the bidomain equations." Ann Biomed Eng **35**(9): 1510-20.

Wijffels, M. C. and H. J. Crijns (2003). "Recent advances in drug therapy for atrial fibrillation." J Cardiovasc Electrophysiol **14**(9 Suppl): S40-7.

Wijffels, M. C., C. J. Kirchhof, et al. (1995). "Atrial fibrillation begets atrial fibrillation. A study in awake chronically instrumented goats." Circulation **92**(7): 1954-68.

Workman, A. J., K. A. Kane, et al. (2001). "The contribution of ionic currents to changes in refractoriness of human atrial myocytes associated with chronic atrial fibrillation." Cardiovasc Res **52**(2): 226-35.

Xia, M., Q. Jin, et al. (2005). "A Kir2.1 gain-of-function mutation underlies familial atrial fibrillation." Biochem Biophys Res Commun **332**(4): 1012-9.

Xie, F., Z. Qu, et al. (2002). "Electrical refractory period restitution and spiral wave reentry in simulated cardiac tissue." Am J Physiol Heart Circ Physiol **283**(1): H448-60.

Yang, Y., J. Li, et al. (2009). "Novel KCNA5 loss-of-function mutations responsible for atrial fibrillation." J Hum Genet.

Zhang, H., C. J. Garratt, et al. (2009). "Remodelling of cellular excitation (reaction) and intercellular coupling (diffusion) by chronic atrial fibrillation represented by a reaction-diffusion system." Physica D: Nonlinear Phenomena **238**(11-12): 8.

Zhang, H., C. J. Garratt, et al. (2005). "Role of up-regulation of IK1 in action potential shortening associated with atrial fibrillation in humans." Cardiovasc Res **66**(3): 493-502.

Zhang, H., A. V. Holden, et al. (2000). "Mathematical models of action potentials in the periphery and center of the rabbit sinoatrial node." Am J Physiol Heart Circ Physiol **279**(1): H397-421.

Zhang, H., Y. Zhao, et al. (2007). "Computational evaluation of the roles of Na+ current, iNa, and cell death in cardiac pacemaking and driving." Am J Physiol Heart Circ Physiol **292**(1): H165-74.

Zhang, S., K. Yin, et al. (2008). "Identification of a novel KCNQ1 mutation associated with both Jervell and Lange-Nielsen and Romano-Ward forms of long QT syndrome in a Chinese family." BMC Med Genet **9**: 24.

Discovery of Biorhythmic Stories behind Daily Vital Signs and Its Application

Wenxi Chen
Biomedical Information Technology Laboratory, the University of Aizu
Japan

1. Introduction

The historical development of the study of biorhythms and the physiological background, as well as functionality of biorhythmic phenomena in human beings, is introduced. The latest achievements in modern chronomedicine, as well as their applications in daily health care and medical practice, are reviewed. Our challenges in monitoring vital signs during sleep in a daily life environment, and discovery of various inherent biorhythmic stories using data mining mathematics are described. Several representative results are presented. Finally, potential applications and future perspectives of biorhythm studies are extensively discussed.

1.1 Historical review

Biorhythmic phenomena are innate, cyclical biological processes or functions existing in all forms of life on earth, including human beings, which respond dynamically to various endogenous and exogenous conditions that surround us (Wikipedia, 2009b). The worldwide history of biorhythmic studies and their application in medical practice can be traced back more than 2000 years, to around a few centuries B.C. Since written records exist in China from more than 4000 years ago, numerous unearthed cultural relics and archaeological materials show that the philosophy of yin and yang and the concept of rhythmic alternation had dominated almost every aspect of Chinese society and people's behaviour (Sacred Lotus Arts, 2009).

Following the philosophy of yin and yang, the earliest existing medical book, "The Medical Classic of Emperor Huang", was formulated from a dialogue between Emperor Huang and a medical professional, Uncle Qi, based on the theory of yin and yang, and compiled from a series of medical achievements by many medical practitioners between 770–221 B.C. The first publication of the book was confirmed to have occurred no later than 26 B.C. and no earlier than 99 B.C. (Wang, 2005).

The book was a medical treatise consisting of a collection of 162 papers in two parts: "Miraculous Meridian and Acupuncture" and "Medical Issues and Fundamental Principles". Each part has nine volumes, and each volume has nine papers, because the number nine is the highest number in Chinese culture, and here, implies that the book covers all aspects of medical matters (Zhang et al., 1995).

This book provided a systematic medical theory and insights into the prevention, diagnosis, and treatment methodologies for diseases. At the same time, the interrelationship between meteorological factors, geographical conditions, and the health of human beings was established and rationalized as the theory of "The unity of heaven and humanity", which considered human beings an integral part of the universe.

This book laid the foundation for Traditional Chinese Medicine (TCM) in terms of fundamental concepts and a theoretical framework, including primary theories, principles, treatment techniques, and methodology. The advent of the book showed that TCM had matured enough to be an independent discipline, such as mathematics, astronomy, or geography, along with the many other scientific achievements in China.

Emperor Huang was considered to be the founder of Chinese civilization, and was the respected supreme authoritative as a Son from Heaven. Later work on the validation and further development of TCM remained to be carried out by many talented TCM successors.

One of the most eminent achievements was contributed by Zhang Zhongjing (ca. 150–219 A.D.) (Wikipedia, 2009g), who is known as the Chinese Aesculapius, and whose works "Treatise on Cold Pathogenic Diseases" and "Essential Prescriptions of the Golden Coffer" established medication principles and provided a summary of his medicinal experience based on his clinical practice and his interpretation of "The Medical Classic of Emperor Huang".

There are three important historical periods in the development and maturation of TCM following Zhang's pioneer work. The first period is from the 3rd to the 10th century, where the main works focused on inheritance, collation, and interpretation of the existing theories described in "The Medical Classic of Emperor Huang". Several milestones in the TCM system were reached in the second period, from the 10th to the 14th century, which is the most important period in the development of TCM. Many medical practitioners studied and annotated the ancient classic, and accumulated their own clinical experiences and proposed their own doctrines. The most eminent representatives were known as "the four great masters": Liu Wansu (1120–1200), Zhang Congzheng (1156–1228), Li Gao (1180–1251), and Zhu Zhenheng (1281–1358). Their contributions greatly enriched and accelerated the development of TCM. Further development and many practical medication approaches were matured in the third period, from the 14th to the 20th century.

Wu Youke (1582–1652) published "On Plague Diseases" in 1642, summarizing his successful fight against pestilence during periods of war, and proposed a theory on the cause of disease and pertinent treatments, which was a significant breakthrough in aetiology akin to modern microbiology.

Based on the "Herbal Classic of Shennong", which described medication using mainly herbal plants, as many as 365 components (252 plants, 67 animals, and 46 minerals), Li Shizhen (1518–1593) spent 29 years writing the "Compendium of Materia Medica", which identified herbal medication into 1892 classifications in 60 categories, and formulated more than 10,000 prescriptions.

The "Detailed Analysis of Epidemic Warm Diseases", written by Wu Jutong (1758–1836), was published in 1798. Many prescriptions described in this book are still considered to be effective, and are used in present clinical practice.

The more than 2000 years of TCM history were created and shaped by numerous medical practitioners through constant exploration and sustained innovation, starting with "The Medical Classic of Emperor Huang", which was built from a very simple philosophy, yin

and yang theory, just like modern computer science is built on a "one and zero" platform (Wikipedia, 2009f).

As shown in Figure 1, yin and yang represents two sides of everything, and governs all aspects of cosmic activities and phenomena in the universe. Constant alternation of the yin and yang status is the origin of universal dynamics. The two sides can coexist, be complementary, mutually inhibitable, mutual transformable, and inter-inclusive.

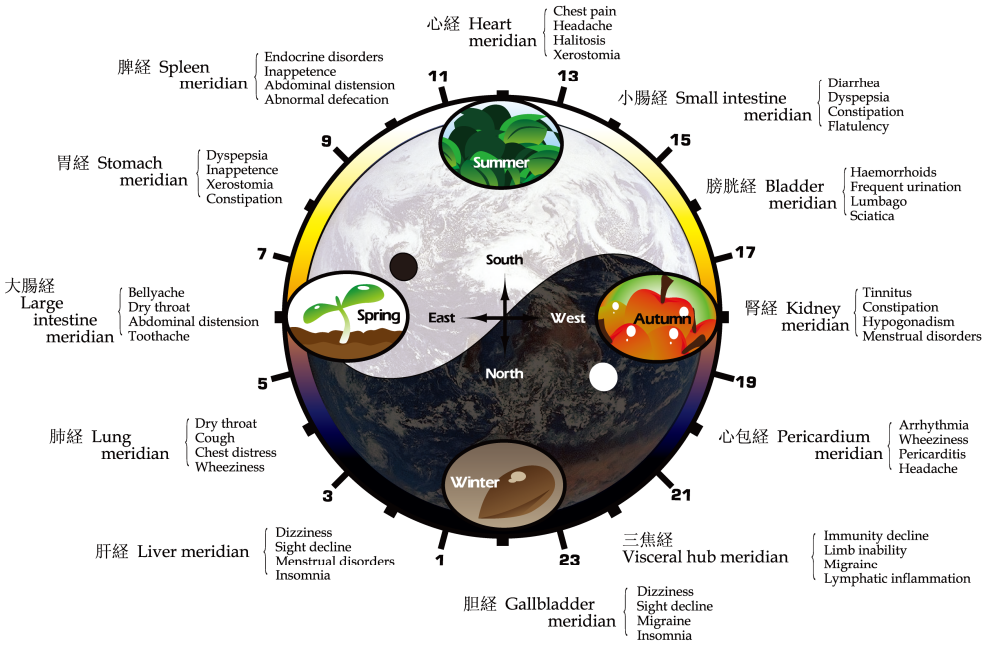


Fig. 1. A holistic overview of the TCM system. On-duty organic meridians in human beings, and disease vulnerabilities in different time domains, and their interaction with various exogenous aspects, such as meteorological, environmental, geographical, and temporal factors from daily to seasonal and yearly, are illustrated (visualization based on Wang, 2005 and Zhang et al., 1995).

TCM considers that a subtle energy ("Qi" in TCM) and blood kinetics in the human body can be expressed as yin and yang alternation corresponding to the waxing and waning periodicities of the sun and the moon. Human moods, health, and behaviour are modulated by the ebb and flow of yin and yang. Human behaviour must synchronize with the natural time sequence to maintain the "Qi" in a good dynamic balanced condition between the yin and yang status.

TCM insists that an unbalance between the yin and yang status is the essential cause of the incidence and exacerbation of disease. The goal of treatment is, in principle, to restore and maintain the balance between yin and yang among the visceral organs. A holistic balance between yin and yang indicates the health status. The yin and yang status can be affected by various endogenous and exogenous factors. The former includes emotional, psychological,

and behavioural aspects, and the latter includes meteorological, environmental, geographical, and temporal factors. Once the yin and yang falls into unbalance, i.e., excess or deficiency on either side, this induces disease. TCM persists from an integrative and holistic standpoint in terms of methodology and philosophy to explain health and disease issues as a result of interaction with many endogenous and exogenous aspects that surround us.

The theories of "syncretism of body and mind" and "the harmony of human with nature" in TCM consider that not only are mental and physical health interconnected, but also vital body functions are modulated by the environmental and seasonal variations due to the movement of the earth and sun and the waxing and waning of the moon over a year. For example, mental disorders, such as excess mood swings in joy, anger, worry, fright, shock, grief, and pensiveness, may affect the visceral organs directly. Depression disrupts the normal functions of the spleen and stomach. Marked changes in weather conditions, such as dryness, dampness, cold, heat, wind, and rain, can induce an unbalance in yin and yang and lead to disease.

TCM considers that an inseparable relationship exists between humans and nature, from birth, development, maturation, caducity, and death, just as seasonal alternations, waxing, and waning occur in the universe. Life activities must be synchronized with natural rhythms to reach harmonic status and maintain longevity. To obtain sufficient sunlight, to ward off chilly north winds, and to enjoy all amenities, the recommended habitation is for a house to sit the north and face the south, back onto mountains, and be close to water.

One of the most prominent features in TCM is the temporal concept in treating health and disease. Spring, summer, autumn, and winter imply burgeoning, growth, harvest, and reposition in nature, respectively. Following a seasonal alternation in work and life is the key to maintaining good health for human beings. Sleep is emphasized as being important as exercise, breathing, and meals in maintaining a normal life activity. A single night's sleeplessness may require 100 days to recover. The daily sleeping–waking cycle should follow the regular celestial mechanics. People should go to sleep late and get up early during the spring season, when all is recovering from the winter hibernation. Acupuncture treatments stipulate strict needle selection in terms of their geometric shape, position, and depth for different seasons using a series of precise instructions.

Because not only physiological and pathological functions, but also the severity of a disease and the effectiveness of its diagnosis/treatment are time-dependent from a TCM standpoint, a day is divided into four parts. From midnight to 6:00 a.m., yin begins to fade from its peak, and yang gradually increases. From 6:00 a.m. to noon, yin finally fades away and yang gradually reaches its peak. From noon to 6:00 p.m., yang begins to fade from its peak and yin gradually increases. From 6:00 p.m. to midnight, yang finally fades away and yin gradually reaches its peak. Most diseases become more severe after dusk when yin increases, and mitigate in daytime when yang dominates.

A day is further divided into 12 time slots. Individual organ-related meridians alternate in being on-duty in each time slot. As shown in Figure 1, many ailments and diseases have their own time-dependent features, which should be taken into consideration in diagnosis and treatment. Different diseases are related to different meridians, and the treatment should be targeted to the on-duty meridian. Patients with liver disease are usually better in the morning, exacerbate between 3:00–5:00 p.m., and become calmer at midnight. Patients with heart disease are calm in the morning, feel comfortable at noon, and become

exacerbated at midnight. Patients with spleen disease show severe symptoms at sunrise, are calm between 3:00–5:00 p.m., and feel better at sunset. Patients with lung disease show severe symptoms at noon, feel better between 3:00–5:00 p.m., and are calm at midnight. Patients with kidney disease feel better at midnight and are calm in the early evening, but become aggravated during four time slots (1:00–3:00 a.m., 7:00–9:00 a.m., 1:00–3:00 p.m., and 7:00–9:00 p.m.)(Wikipedia, 2009d; Ni, 1995; Veith, 2002).

Identifying the root cause of the disease is a very important part of TCM practice. TCM stresses that balance is the key to a healthy body. Any long-term imbalance, such as extreme climate change, undue physical exercise, heavy workload, excessive rest, too frequent or rare sexual activity, unbalanced diet, or sudden emotional changes can all cause disease (Xuan, 2006).

A holistic view of the human body is not the sole understanding of the TCM system. In approximately the same historical period on the other side of the earth, Hippocrates (ca. 460–ca. 370 B.C.), a Greek physician known as "the father of medicine", laid the foundations of Western medicine by freeing medical studies from the constraints of philosophical speculation and religious superstition.

"The Hippocratic Corpus" is a collection of about 60 treatises believed to have been written between 430 B.C. and 200 A.D. by different people under the name of Hippocrates (Naumova, 2006). The corpus describes many points of view on diseases related to temporal and environmental factors, such as:

- As a general rule, the constitutions and the habits of people follow the nature of the land where they live.
- Changes in the seasons are especially liable to beget diseases, as are great changes from heat to cold or cold to heat in any season. Other changes in the weather have similar severe effects.
- When the weather is seasonable and the crops ripen at regular times, diseases are regular in their appearance.
- Each disease occurs in any season of the year, but some of them occur more frequently and are of greater severity at certain times.
- Some diseases are produced by the manner of life that is followed, and others by the life-giving air we breathe.

As a pioneer in studying biorhythms, an Italian physician, Santorio Santorio (1561–1636), invented a large "weighing chair" to observe the weight fluctuations in his own body during various metabolic processes, such as digestion, sleep, and daily eating over a 30-year period (Wikipedia, 2009e). He reported the circadian variation both in body weight and in the amount of insensible perspiration in his book "On Statistical Medicine", published in 1614, which introduced a quantitative aspect into medical research, and founded the modern study of metabolism.

In 1729, a French astronomer named Jean Jacques Ortous de Mairan (1678–1771) devised a classical circadian experiment. He placed a heliotropic plant in the dark and observed that the daily rhythmic opening and closing of the heliotrope's leaves persisted in the absence of sunlight (Wikipedia, 2009c). We now understand that the circadian clock controls given processes, including leaf and petal movements, the opening and closing of stomatal pores, the discharge of floral fragrances, and many metabolic activities in plants.

Christopher William Hufeland (1762–1836), a German physician, published "The Art of Prolonging Life" in 1796. He stated that "The life of man, physically considered, is a peculiar

chemico-animal operation; a phenomenon effected by a concurrence of the united powers of Nature with matter in a continual state of change." He considered that the rhythmicity of twenty-four hours is formed by the regular revolution of our earth, and can be seen in all diseases, and all the other biorhythms are determined by it in reality (Hufeland, 1796).
In the early 19th century, identical conclusions from investigations into biorhythms from different approaches and from independent researchers in different fields, such as psychology and meteorology, were reached.
In his book "Die Perioden des Menschlichen Organismus (Periodicity in the Life of the Human Organism)", the Austrian psychologist Hermanna Swoboda stated that, "Life is subject to consistent changes. This understanding does not refer to changes in our destiny or to changes that take place in the course of life. Even if someone lived a life entirely free of outside forces, of anything that could alter his mental and physical state, still his life would not be identical from day to day. The best of physical health does not prevent us from feeling ill sometimes, or less happy than usual". By analysing dreams, ideas, and creative impulses of his patients, Swoboda noticed very regular rhythms with predictable variations in some artists' performances and the mental status of pregnant women (Biochart Com, 2009).
Even the influence of meteorological factors, such as sunspot activity, was associated with the acute chronic diseases of the heart, blood vessels, liver, kidney, and nervous system, ranging from mild to severe, such as excitability, insomnia, tiredness, aches, muscle twitches, polyuria, digestive troubles, jitteriness, shivering, spasms, neuralgia, neural crises, asthma, dyspnea, fever, pain, vertigo, syncope, high blood pressure, tachycardia, arrhythmia, and true angina pectoris (Vallot et al., 1922).
In 1924 and 1928, Alexander Chizhevsky (1897–1964) published "Epidemiological Catastrophes and the Periodic Activity of the Sun" and "Influence of the Cosmos on Human Psychoses", respectively, studying biorhythms in living organisms in their connections with solar and lunar cycles, stating that, "Life is a phenomenon. Its production is due to the influence of the dynamics of the cosmos on a passive subject. It lives due to dynamics, each oscillation of organic pulsation is coordinated with the cosmic heart in a grandiose whole of nebulas, stars, the sun and the planet", which is now formulated as the independent discipline of "heliobiology" (Wikipedia, 2009a).

1.2 Modern chrono-related studies

In the 1950s, Franz Halberg noticed that the eosinophil counts of both sighted and blinded groups of mice rose and fell cyclically with temperature variations. In the former group, this occurred at approximately the same time each day, and in the latter group, there was a slight shift and a shorter cycle. Neither group showed an exact 24-hour cycle, showing the existence of an endogenous mechanism (Halberg et al., 1959).
When the implications of these cycles were explored further, it was found that one group of mice developed seizures when exposed to an extremely loud noise at 10:00 p.m., the active period of their day, while another group that was exposed to the noise at noon, during their rest period, did not develop seizures. It was also found that when a potential poison or high doses of a drug were given to mice, whether they lived or died depended on the delivery time of the drug.
The study of the body's time structure was continued in the late 1960s by Halberg and his Indian co-researchers in medical practice by administering radiation therapy to patients

with large oral tumours. The tumour temperature was used as a marker to schedule treatments. Patients were divided into five groups and treated at a different time offset, –8, –4, 0, +4 and +8 hours, from their peak temperature. More than 60% of patients who received treatment when the tumour was at peak temperature were alive and disease-free two years later. This is perhaps because the highest metabolic activity at peak temperature enhanced the therapeutic effect (Halberg, 1969).

An increased swing in the amplitude of blood pressure, which develops before a rise in mean blood pressure, was found in rats (Halberg, 1983). In 1987, this phenomenon was confirmed to be a greater risk factor for ischemic stroke from a six-year study involving nearly 300 patients (Halberg & Cornélissen, 1993). This is now known as circadian hyper amplitude tension (CHAT). CHAT studies have shown that taking blood pressure medication at an undesirable time can cause CHAT, and can potentially lead to a stroke.

In addition to body temperature and blood pressure, biorhythmic variations in other vital signs, such as saliva, urine, blood, and heart rate, have been quantified to identify normal and risky patterns for disease, to optimize the timing of treatment, and to compare variations among subjects grouped by age and gender (Halberg et al., 2003; Halberg et al., 2006a; Halberg et al., 2006b).

In 1960, the nascent field of biorhythm studies celebrated its first symposium in New York, USA, and modern chrono-related studies are now expanding in both dimensional and functional scales, from the genome level to the whole-body level, and from fundamental chronobiology to medical applications, such as chronophysiology, chronopathology, chronopharmacology, chronotherapy, chronotoxicology, and chronomedicine. All of these topics are rooted in the study of biorhythmic events in living organisms and their adaptation to solar- and lunar-related rhythms, and are still in the exciting process of discovery.

Although rhythmic phenomena in many behavioural and life processes, such as eating, sleeping–waking, seasonal migration, heart-beat, and cell proliferation, had been observed in many aspects for a long time, little was known about their physiological background until recent advances in molecular biology and genetics. Scientists have now identified specific genes, proteins, and biochemical mechanisms that are responsible for spontaneous oscillations with rhythmic cycles extended from the molecular, cellular, tissue, and system levels on a spatial scale, from the millisecond intervals of neuronal activity to seasonal changes in the temporal scale (Martha & Sejnowski, 2005).

The suprachiasmatic nucleus (SCN), composed of 20,000 or so autonomous cells located in the hypothalamus, is now known to be responsible for controlling the timing of endogenous rhythms (Stetson & Watson-Whitmyre, 1976). The SCN receives an environmental input, such as light, a type of zeitgeber, from light receptors in the retina via the retinohypothalamic tract (RHT), and generates a rhythmic output to coordinate and synchronize body rhythms. The SCN is fundamental to each of the three major clock components in biological systems: entrainment pathways, pacemakers, and output pathways to effecter systems (Reppert & Weaver, 2001). Autonomous single-cell oscillators reside in peripheral tissues as well as in the SCN of the pineal gland. Peripheral oscillators may respond more directly to environmental factors, such as temperature, moisture, pressure, and sound. However, the SCN governs and coordinates the rhythms of the peripheral oscillators by both direct neural connections and diffusible biochemical processes (Balsalobre et al., 2000). As a result of such synchronization, the body as an entire system

maintains rhythms for not only the sleeping–waking cycle, but also for body temperature, heart rate, blood pressure, immune cell count, and hormone secretion levels, such as cortisol for stress and prolactin for immunity and reproduction. Rhythmic beating in the SCN is the timepiece not only for daily cycles, but also for the totality of lifelong personal patterns, potentially in a harmonic resonance with the environmental surroundings.

The clock genes are expressed not only in the SCN, but also in other brain regions and various peripheral tissues. The liver has been confirmed to be a biological clock capable of generating its own circadian rhythms (Turek & Allanda, 2002). A microarray analysis experiment has revealed that there are many genes expressing a circadian rhythm in the liver. The relative levels of gene expression in the liver of rats have been investigated from the viewpoint of the time of day. Sixty-seven genes were identified where their expression was significantly altered as a function of the time of day, and these were classified into several key cellular pathways, including drug metabolism, ion transport, signal transduction, DNA binding and regulation of transcription, and immune response according to their functions (Desai et al., 2004).

In the cases where exogenous cues (zeitgebers) for timing, such as light, temperature, or sound, are shielded, the SCN moves out of synchronization with the exogenous entrainment. However, the innate rhythm is not obliterated, because biorhythms are genetically built into cells, tissues, organs, and the whole-body system. The body still maintains its rhythms, but not in an organized tempo. The sleeping–waking cycle and body temperature variation will not follow an exact 24-hour cycle, which was entrained by the light–dark cycle or the sunset–sunrise cycle. Other biorhythms and daily activities could also be affected, although none has all its variables equal.

The broad spectrum of different biorhythms is classified into three categories, i.e., circadian rhythms, ultradian rhythms, and infradian rhythms.

The circadian rhythm is the most common biorhythm, alternates in an approximately daily cycle, and exists in most living organisms. The term "circadian" comes from circa, which means "about", and dies, which means "day".

Ultradian rhythms refer to those cyclic intervals that are shorter than the period of a circadian rhythm, exhibiting periodic physiological activity occurring more than once within a day, such as neuron firing, heart-beats, inhalation and expiration, and REM–NREM sleep cycles.

Infradian rhythms pertain to regular recurrences in cycles of longer than the period of a circadian rhythm, and occur on an extended scale from days to years. Some of these are listed below:

- Circasemiseptan rhythms have a cyclic length of 70 to 98 hours or 3.5 days, and exist in blood pressure and heart rate fluctuations. They can be found in patients with incidence of myocardial infarction and apoplexy.
- Circaseptan rhythms occur in periods of 140 to 196 hours or about one week, and are found in changes in body temperature and blood pressure.
- Circatrigintan rhythms behave in approximately monthly cycles. The most common is the female menstrual cycle, ranging from 25 to 35 days. Others include the emotional and physical stamina rhythms, which change over 28 days and 23 days, respectively. Intellectual rhythmicity was found to exhibit a regular 33-day cycle for mental agility and ability. The existence of a 21-day cycle related specifically to moods was uncovered.

Some vital signs, such as hormone secretion, blood pressure, and metabolic activity, have similar properties.

- Circannual rhythms occur over a period of between 305 to 425 days, or about a year. Most plants have a seasonal change from rootage, burgeon, blossom, and fructification. Migratory birds migrate in an annual pattern through regular seasonal journeys in response to changes in food availability, habitat, or weather.

Table 1 summarizes various known biorhythms ranging from periods of milliseconds to years that exist in living organisms.

Biorhythm		Cycle length	Related event
Ultradian		< 1 d	Neuron firing, heart beating, inhalation and expiration, REM–NREM sleep cycles
Circadian		1 d ± 4 h	Body temperature (BT), blood pressure (BP), heart rate (HR), hormone secretion
Infradian	Circadidian	2 ± 0.5 d	Body weight, urine volume
	Circasemiseptan	3.5 ± 1 d	Sudden death
	Circaseptan	7 ± 1.5 d	Rejection of heart transplant, activity, BP, BT
	Circadiseptan	14 ± 3 d	Body weight, grip strength
	Circavigintan	21 ± 3 d	Mood, 17-ketosteroid excretion
	Circatrigintan	28 ± 5 d	Emotional and physical stamina, mental agility and ability, menstruation
	Circannual	1 y ± 2 m	BP, aldosterone
	Circaseptennian	7 ± 1 y	Marine invertebrates
	Circaduodecennian	12 ± 2 y	BP
	Circadidecadal	20 y	BP

Table 1. Temporal definitions and the properties of diversified biorhythms ranging from periods of milliseconds to years (adapted from Halberg & Cornélissen, 1993; Koukkari & Sothern, 2006). Cycle length: h = hours; d = days; m = months; y = years.

Objective estimation of various biorhythmicities in different physiological vital signs and biochemical biomarkers, such as body temperature, heart rate, blood pressure, adrenocorticotropic hormone, and melatonin, is indispensable in medical practice. Many vital signs and biomarkers are usually modulated and interacted by multiple biorhythms. Similarly, multiple biorhythms are often interwoven within a vital sign or a biomarker as shown in Table 1. Because biorhythms are cyclic, recurring physiological events, their features in time structures are commonly expressed by parameters such as period, mesor,

amplitude and phase, zenith and nadir, onset of events, the minimum and maximum incidence of events, and the shape of the rhythmic pattern.

Mathematical approaches to quantifying biorhythms were classified into two categories in the early stages of their study: macroscopic and microscopic (Halberg, 1969). The former category employs many statistical techniques, such as histograms, mean, median, mode, and variance. The latter category uses chronograms, variance spectrum, auto/cross correlations, coherency, and the cosinor method.

The cosinor method uses least-squares criteria to fit raw data on a presumptive single sine wave model in the time domain. Its variants, such as population mean-cosinor, group mean-cosinor, multi-cosinor and non-linear cosinor methods, are similarly based on various compound models (Nelson et al., 1979). The multivariate method has also been used for the parameter estimation of biorhythms in human leukocyte counts in microfilariasis infection (Kumar et al., 1992).

In addition to living organisms, the biosphere and the solar system are good examples of self-tuning control systems. The laws governing the operation of control systems are incorporated in the development of mathematical methods for the identification of rhythms hidden in the dynamics of biological and heliogeophysical variables (Chirkova, 1995).

Fourier transformation and spectral analysis methods have also been developed to evaluate and analyse biorhythms regarding their general characteristics in terms of amplitude, phase, periodical frequency, and cyclic length (Chou & Besch, 1974).

The determination of biorhythms is helpful not only in clarifying their impact on the pathophysiology of diseases, but also in elucidating the pharmacokinetics and pharmacodynamics of medications.

Figure 2 shows the circadian properties of various physiological vital signs and biochemical markers, in alignment with time-dependent symptoms or events of diseases that are in either the severest timing or the most frequent incidence of the disease.

As shown in Figure 2, allergic rhinitis is typically worse in the early waking hours than later during the day. Asthma usually occurs more than 100 times more in the few hours prior to awakening than during the day. Angina commonly occurs during the first four to six hours after awakening. Hypertension typically occurs from late morning to middle afternoon. Rheumatoid arthritis is most intense upon awakening. Osteoarthritis worsens in the afternoon and evening. Ulcer disease typically occurs after stomach emptying, following daytime meals, and in the very early morning, often disrupting sleep. Epilepsy often occurs only at individual particular times of the day or night (Smolensky & Labrecque, 1997).

The daily variation pattern of the symptoms of diseases and biochemical-pathophysiological processes is now used to optimize treatment of various diseases, such as asthma, cancer, diabetes, fibromyalgia, heartburn, and sleep disorders. Chronopharmacokinetic studies have been reported for many drugs in an attempt to explain chronopharmacological phenomena, and these have demonstrated that the time of administration is a possible factor in the variation in the pharmacokinetics of a drug. Time-dependent changes in pharmacokinetics may proceed from the circadian rhythm of each process, e.g., absorption, distribution, metabolism, and elimination. Thus, circadian rhythms in gastric acid secretion and pH, motility, gastric emptying time, gastrointestinal blood flow, drug protein binding, liver enzyme activity and/or hepatic blood flow, glomerular filtration, renal blood flow, urinary pH, and tubular resorption play a role in such pharmacokinetic variations (Labrecque & Belanger, 1991). More than 100 drugs, such as cardiovascular agents, anti-

asthmatic agents, gastrointestinal agents, non-steroidal anti-inflammatory agents, and anti-cancer agents, have already been recognized as exhibiting circadian variations in pharmacokinetic and pharmacodynamic performance over a period of 24 hours (Lemmer, 1994). Chronotherapeutic principles are realized through innovative drug delivery technology in the safe and efficient administration of medications (Smolensky & Labrecque, 1997).

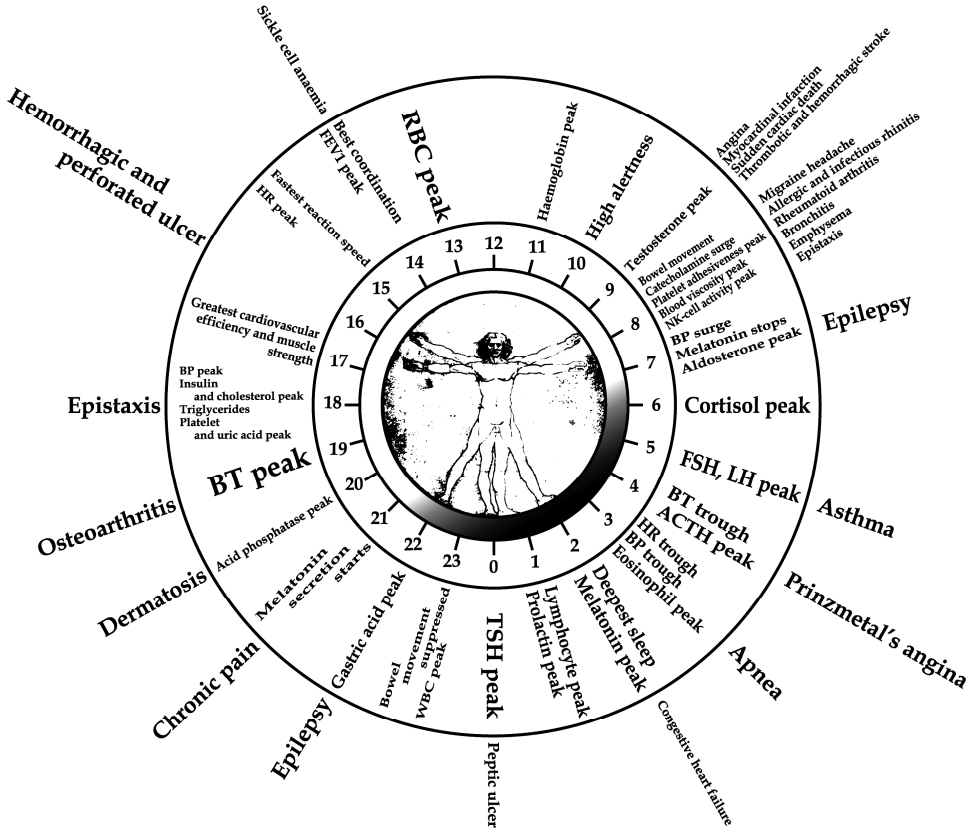


Fig. 2. Circadian rhythmic changes of physiological vital signs and biochemical markers, and symptoms or events of diseases in the case of worst timing or highest likelihood (adapted from Smolensky & Labrecque, 1997; Ohdo, 2007). The outer ring indicates the symptom and disease. The inner ring indicates vital signs and biomarkers.

Other applications utilizing biorhythms can be found in health care, human welfare, and behavioural administration domains. A conventional alarm clock is usually set in advance to sound a bell or buzzer at a desired hour. During the Stage 1 period of sleep, a person drifts in and out of sleep, and can be awakened easily. However, it is very difficult to be woken up during deep sleep periods, such as Stages 3 and 4. When a person is awakened during deep sleep stages, it is difficult for them to adapt immediately, and they often feel groggy and

disoriented for several minutes after waking. A biorhythm-based bell device, biological rhythm-based awakening timing controller (BRAC), was developed to estimate biorhythm changes in sleep cycles from fingertip pulse waves, and was used to optimize the alarm timing (Wakuda et al., 2007).

Jet lag is a malaise often associated with long-distance travel across several time zones. Some of the symptoms usually reported are fatigue, drowsiness, irritability, inability to concentrate during the day, difficulty in sleeping at night, and gastrointestinal discomfort (Katz et al., 2001). Shift workers, such as truck drivers and emergency medical personnel, who are obliged to work non-standard office hours, exhibit similar symptoms to those of jet lag.

Sufferers of jet lag and shift workers are affected by a transient misalignment of the circadian clock with the external clock. Both disorders have a common cause in aetiology, but a major difference exists between the two situations. A long-distance traveller can resynchronize their internal clock within a few days after their biorhythm is disturbed because their internal clock is out of phase with the external clock of sunrise and sunset. By contrast, as long as the daily work schedule of a shift worker cannot be synchronized with the natural biorhythms, they will be unable to truly adapt their biorhythms to the external clock. Although effective treatment has not been rigorously documented yet, the symptoms are usually treated using a light therapy method, for example, artificial light reversal of day and night, which can be attained by subjecting the patient to bright artificial light at night and avoiding photoic stimulation during sunlight hours by wearing sunglasses or closing window curtains (Smolensky & Lamberg, 2000).

It has also been shown that although the exact timing varies from individual to individual, performance in physical and intellectual activities exhibits a daily rhythmicity. The best performance is achieved around the peak body temperature time, which usually occurs in the late afternoon, although overall performance in real world situations can be affected by many other factors, such as innate and acquired skills, motivation, concentration, and spot exertion (Dunlap et al., 2004).

Biorhythmicities are recognized as affecting numerous physiological and behavioural processes. The daily pattern of human activity and stress amplifies the innate biological variability of biorhythms, and diseases can alter the expression and characteristics of circadian and other biorhythms. The outcomes of the chronotherapeutic treatment of several diseases that have predictable circadian variations, such as allergic rhinitis, angina pectoris, arthritis, asthma, diabetes, epilepsy, hypertension, dyslipidemia, cancer, and ulcers have been confirmed to be more effective than traditional homeostatic treatments (Elliott, 2001).

Such time-dependent biochemical processes and pathophysiological phenomena exist ubiquitously, from local cells to the whole body. In summary, the occurrence of biorhythms is physiologically indispensable in life processes, and provides several advantages (Moser et al., 2006):

- Stability maintenance in response to endogenous and exogenous variations by fine-tuning the characteristics at various levels, such as cellular, organic, and holistic systems, for controlling long-term physiological functionality;
- Synchronization and coordination of different visceral organs, enabling the system to function most efficiently;
- Temporal compartmentalization, mediating polar events, such as systole and diastole, inspiration and expiration, work and rest, waking and sleeping, which cannot happen simultaneously, to occur both in alternation and efficiently in the same physical space.

The discovery of biorhythmic patterns and their perturbation is essential not only for proper diagnosis and treatment of patients suffering from various diseases, but also for daily health management of healthy persons. The following section describes our studies and the results of the long-term monitoring of various biorhythms.

2. Our Studies

The natural world is teeming with cyclic patterns and sequential events, and biorhythms are known to be important in treating disease and managing health. However, monitoring vital signs continuously in a daily life environment over a long period is a tedious task indeed. People can put up with such unpleasant assignments without much complaint over a short time period if they are on a course of treatment. However, in cases where they have no obvious symptoms, and are asked to do so purely for health care purposes, such boring daily duties will soon cause people to run out of patience.
The purposes of our studies were twofold:

- To develop convenient ways to monitor vital signs that were suitable for utilization in daily life environments for any time period without much discomfort to the user.
- To assess biorhythms through various mathematical approaches from the large amount of physiological data collected daily over a long period.

Two modes of study model are presented below. The first part describes the detection of multiple biorhythms from a single vital sign, while the second part reports on the detection of a single biorhythm from multiple vital signs.

2.1 Discovery of multiple biorhythms from a single vital sign

Multiple biorhythms are usually interwoven within an identical vital sign. This section describes the detection of different biorhythms, i.e., sleep patterns, behavioural changes, and menstrual cycles using different mathematical approaches from heart rate data collected during sleep.

2.1.1 Data collection

Heart rate data were collected during sleep using the scheme shown in Figure 3. The subject slept wearing a wrist-type Bluetooth-enabled SpO2 sensor (Model 4100, Nonin Corp., USA). A bedside box situated nearby the bed was always on stand-by waiting for the SpO2 sensor to initiate. When the SpO2 sensor was switched on, the Bluetooth wireless connection between the bedside box and the SpO2 sensor device was established automatically. With the help of the bedside box, HR and SpO2 data were collected from the SpO2 sensor via the Bluetooth connection and were transmitted continuously to a database server by an HTTP connection through an ADSL LAN in the home during a given sleep episode. When the subject rose and removed the sensor in the morning, the Bluetooth connection was closed, the bedside box went into stand-by mode, and the data collection procedure was terminated. Although the sensor collected both HR and SpO2 data simultaneously, only the HR data were used in this study. A single night's sample of collected raw HR data is shown in the black trace in Figure 4. The frequent interruption of noise spikes was perhaps due to movement artefacts, or a misinterpretation of the transmitted data package. Such noise has to be suppressed before biorhythm detection is conducted.

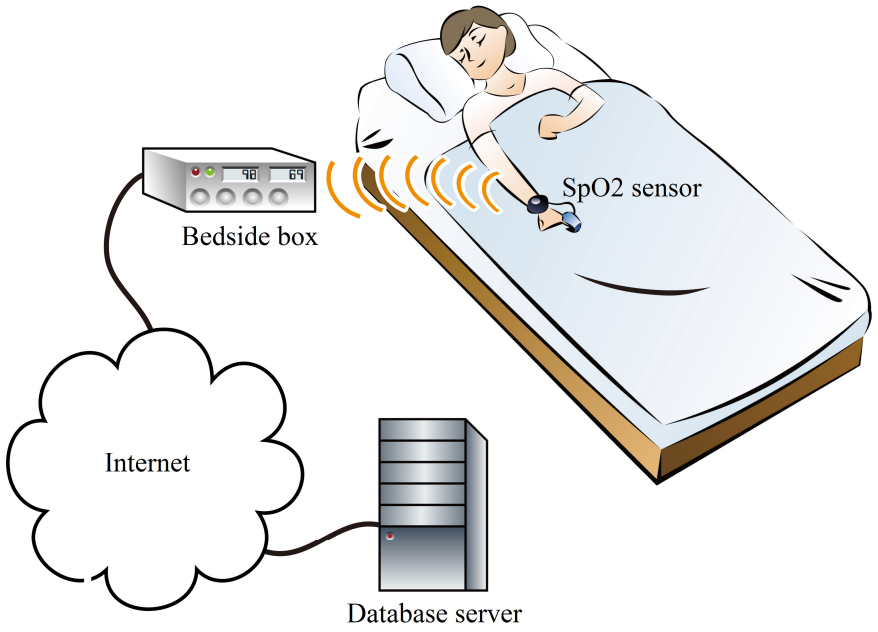


Fig. 3. Schematic drawing showing HR and SpO2 data collection during sleep. By attaching a Bluetooth-enabled SpO2 sensor to a fingertip, the nearby bedside box established a Bluetooth connection with the sensor automatically, and received HR and SpO2 data from the sensor simultaneously. These data were transmitted continuously to a database server via an HTTP connection.

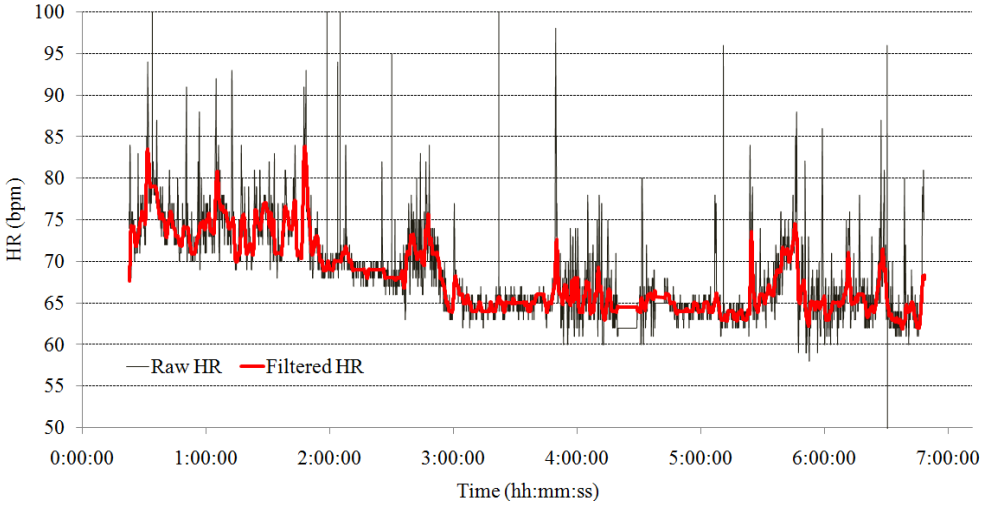


Fig. 4. Raw HR data (thin black trace) and filtered HR data (bold red trace) over a single night's sleep. Raw data were collected by a SpO2 sensor from a fingertip. Filtered data were obtained by applying a median filter and a Savitzky–Golay smoothing filter.

2.1.2 Noise suppression

Unless arrhythmia occurs, the premise of smoothing is that the HR varies slowly in nature, but its measurement is often contaminated by random noise or other artefacts. As shown in the black trace in Figure 4, the main source of noise in the raw measurement during sleep is a spike-like noise.

Noise suppression was implemented using two digital filters in two steps. A median filter was used in the first step to remove the spike-like noise, and a Savitzky–Golay filter was used in the second step to smooth the HR profile.

The median filter was a non-linear digital filtering technique, usually used in the image-processing field to remove speckle noise and salt/pepper noise from images. The idea was to represent the signal by replacing an extremely large or small value with a reasonable candidate value. This is realized using a window consisting of an odd number of data. The values within the window were sorted in numerical order, and the median value, the sample in the centre of the window, was selected as the output of the filter.

When the window was moved along the signal, the output of the median filter $y(i)$ at a moment i was calculated as the median value of the input values $x(i)$ corresponding to the moments adjacent to i ranging from $-L/2$ to $L/2$.

$$y(i) = median\left(x(i - L/2), x(i - L/2 + 1), ..., x(i), ..., x(i + L/2 - 1), x(i + L/2)\right), \quad (1)$$

where L is the length of the window.

The Savitzky–Golay filter was used to smooth the signal that was outputted from the median filter. The Savitzky–Golay filter segmented the signal as frames using a moving window, and approximated the signal frames one by one using a high-order polynomial, typically quadratic or quartic (Savitzky & Golay, 1964).

Each digital filter output $z(i)$ can be expressed by a linear combination of the nearby input points as

$$z(i) = \sum_{k=-n_L}^{n_R} c_k y(i + k), \quad (1)$$

where n_L is the number of points on the left-hand side of the data point i, and n_R is the number of points on the right-hand side of i.

The Savitzky–Golay filtering process is to find a proper polynomial to fit all n_L+n_R+1 points within each window frame on the least-squares meaning, and to produce a filter output $z(i)$ as the value of that polynomial at position i.

To derive filter coefficients, c_k, we considered fitting a polynomial of degree M in i, namely $a_0+a_1i+a_2i^2+\cdots+a_Mi^M$ to the values $y_{-nL},...,y_{nR}$. Then, $z(0)$ will be the value of that polynomial at $i = 0$, namely a_0. The design matrix for this problem is

$$A_{ij} = i^j, \qquad i = -n_L, ..., 0, ..., n_R, \qquad j = 0, ..., M, \quad (1)$$

and the normal equations for the polynomial coefficients vector, $\mathbf{a}=[a_0, a_1, a_2, \cdots, a_M]'$, in terms of the input data vector, $\mathbf{y}=[y_{-nL},...,y_{nR}]'$, can be written in matrix notation as below:

$$\mathbf{A} \cdot \mathbf{a} = \mathbf{y}, \quad (1)$$

The polynomial coefficients vector, $\mathbf{a}$, becomes

$$\mathbf{a} = \left(\mathbf{A}^T \cdot \mathbf{A}\right)^{-1} \cdot \left(\mathbf{A}^T \cdot \mathbf{y}\right), \quad (1)$$

We also have the specific forms

$$\left\{\mathbf{A}^T \cdot \mathbf{A}\right\}_{ij} = \sum_{k=-n_L}^{n_R} A_{ki} A_{kj} = \sum_{k=-n_L}^{n_R} k^{i+j} \ , \tag{1}$$

and

$$\left\{\mathbf{A}^T \cdot \mathbf{y}\right\}_{j} = \sum_{k=-n_L}^{n_R} A_{kj} y_k = \sum_{k=-n_L}^{n_R} k^{j} y_k \ , \tag{1}$$

Since the filter coefficient, c_k, is the component a_0 when $\mathbf{y}$ is replaced by the unit vector $\mathbf{e}_k$, we have

$$c_k = \left\{\left(\mathbf{A}^T \cdot \mathbf{A}\right)^{-1} \cdot \left(\mathbf{A}^T \cdot \mathbf{e}_k\right)\right\}_0 = \sum_{m=0}^{M} \left\{\left(\mathbf{A}^T \cdot \mathbf{A}\right)^{-1}\right\}_{0m} k^m \ , \ -n_L \le k < n_R \ , \tag{1}$$

When the filter coefficient vector $\mathbf{c}=[c_{-nL},\ldots,c_{nR}]$ was obtained using Equation (8), the signal shown in the black trace in Figure 4 could be smoothed using Equation (2), and the result is the red trace shown in Figure 4.

After these two filtering steps, the noise-suppressed HR data were used for the detection of three different biorhythms, as described in the following three sections.

2.1.3 Sleep cycle estimation

Sleep is clinically classified into two distinct states: the rapid eye movement (REM) state and the non-rapid eye movement (NREM) state. The NREM state is further divided into four stages, 1–4, indicating four depths of sleep from shallow to deep. When drifting into sleep, a normal sleep cycle moves in a sequential progress from Stage 1 through to Stage 4 and then Stage 3 and 2 of NREM, and finally to the REM state. Each sleep cycle lasts for 90 to 120 minutes.

During the REM sleep period, rapid eye movements occur, fluctuations in breathing movements and heart-beat become severe, blood pressure rises, and involuntary muscle jerks (loss of muscular tone) occur. However, the brain is highly active, and an EEG usually records high frequencies and low amplitudes, similar to those recorded during the awake state. Vividly recalled dreams mostly occur during REM sleep. There are three to five REM episodes per night. They occur at the end of each sleep cycle, and are not always constant in length, ranging from five minutes to over an hour.

NREM sleep is physiologically different from REM sleep, and is dreamless. As the NREM sleep advances from Stages 1 to 4, the EEG signal shows a slower frequency and a higher amplitude. Breathing and heart-beat become slower and more regular, the blood pressure and body temperature decrease, and the subject is relatively still.

About 75%–80% of sleep is NREM sleep, and almost half of the total sleep time is in Stage 2 NREM. REM sleep episodes account for 20%–25% of the total sleep period. The relative amount of REM sleep varies considerably with age. As age increases, the total sleep time becomes shorter, leading to shorter NREM sleep, but no significant change in REM sleep. By contrast, infants spend about half of their sleep time in REM sleep.

Rhythmic alternation of REM and NREM states during sleep is reflected in different physiological activities, such as eye movement, muscular tone, electroencephalogram,

respiration, heart rate, blood pressure, and body temperature. These features are clinically discernible in a polysomnogram measured by attaching more than 10 sensors to a subject.

This section describes a method for estimating the cyclic property of sleep based on HR only. Because variation in HR in the REM state is much larger than that in the NREM state, variance of the HR was used as a criterion to distinguish between REM and NREM sleep states.

The windowed local variance (WLV) method is used extensively in image processing for edge detection and pattern segmentation (Bocher & McCloy, 2006a, 2006b; Law & Chung, 2007). It is defined as the variance computed for pixel values within a window of size $w \times w$ from aggregated pixel data.

This study deals with one-dimensional HR data sequences, and defines the WLV_i at data point i as shown in Equation (9).

$$WLV_i = \frac{1}{w}\sqrt{\sum_{j=i}^{i+w}\left(x_j\right)^2 - \frac{1}{w}\left(\sum_{j=i}^{i+w} x_j\right)^2}\ , \tag{1}$$

where w is the window length and x_j is the input data within the window.

Figure 5 shows the noise-suppressed HR data in the red trace and the estimated result of a biphasic sleep cycle in the blue trace. The low-level phase indicates the period with lower HR perturbation, and the high-level phase corresponds to a period with increased HR fluctuation. Although it is not yet a conclusion that there is a relationship between the REM–NREM cycle and the estimated biphasic cycle, because no concomitant EEG was recorded, it is inferred that the low-level phase may imply the NREM state, while the high-level phase refers to the REM state.

As shown in Figure 5, the period length of the high-level phase gradually increased during the course of sleep, i.e., it was short at the beginning of the sleep period and longer towards the end of the sleep period, a behaviour similar to the REM state, although confirmation of this is required from an EEG.

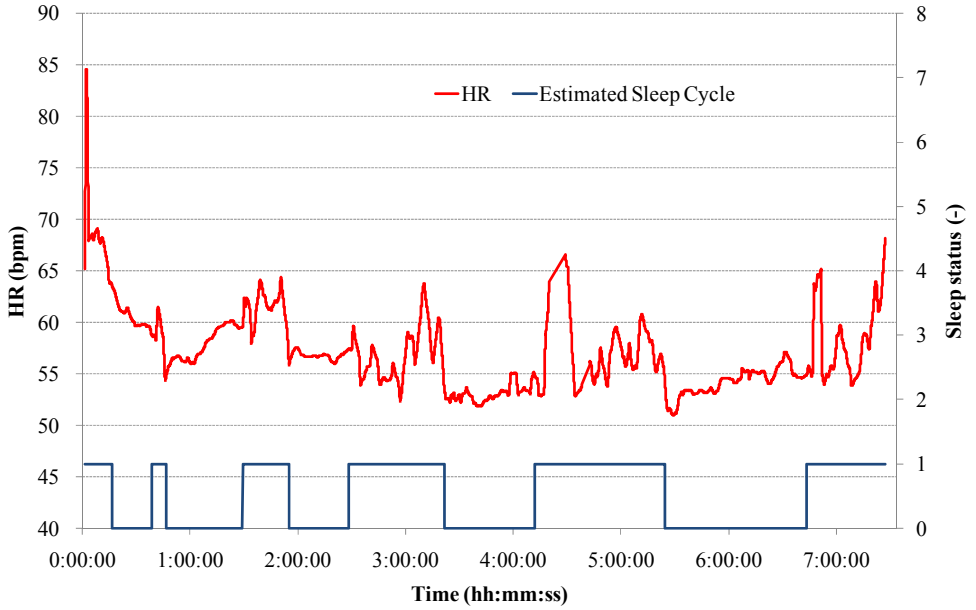


Fig. 5. HR profile of a single night's sleep and the estimated sleep cycle. Data were collected from a male student in his twenties. The red line is the profile of the noise-suppressed HR. The blue line is the estimated sleep cycle, in which the low-level phase indicates the period with low HR perturbation, and the high-level phase corresponds to the period with more HR fluctuations.

2.1.4 Detection of changes in daily behaviour

Because biorhythms are affected by endogenous and exogenous factors, any change in daily behavioural patterns can be reflected in biorhythmic changes. This study demonstrates the detection of behavioural changes during waking hours by applying the dynamic time warping (DTW) method to the HR data collected during sleep (Watanabe & Chen, 2009).

DTW is an algorithm used to measure the similarity between two data sequences that may differ in length. Well-known applications are in fields such as speech recognition and walking analysis, in which data sequences in either case generally vary in temporal span and rhythmic tempo.

The aim of DTW is to find the optimal alignment between two given data sequences under given criteria. Consider two given data sequences with variable length, the reference pattern R={r_1,...,r_M} with data length M, and the test pattern T={t_1,...,t_N} with data length N, as shown in Figure 6. The value of each black dot d_{ij} indicates the difference (distance) between the reference pattern data r_i and test pattern data t_j, as described by Equation (10).

$$d_{ij} = \sqrt{(i-j)^2 + (r_i - t_i)^2} \ , \ i=1, 2,\ldots, M; \ j=1, 2,\ldots, N \ , \tag{10}$$

Thus, a two-dimensional $N \times M$ distance matrix, $D_{N\times M}$, is constructed where the element d_{ij} is the distance between the ith data in the reference pattern and the jth data in the test pattern.

As a similarity measure, the shortest path from the start (the lower left-hand corner of the distance matrix) to the end (the upper right-hand corner of the distance matrix) of the data sequence must exist among multiple possible paths.
The shortest path is determined using the forward dynamic programming approach with a monotonicity constraint.

$$P_{ij} = \min_{k \geq j} \left\{ d_{jk} + P_{i+1,k} \right\}, \tag{1}$$

Here, P_{ij} denotes the distance from the ith and jth data node to the terminating node.
The overall minimum distance, $D(T, R)$, used as the similarity measure for two patterns (a smaller distance value indicates a higher similarity) is determined from

$$D(T,R) = P_{11}, \tag{1}$$

The Sleep Index (SI) is obtained by normalizing $D(T, R)$ between the values of 0 and 1, as below:

$$SI = \left(D(T,R) - D_{\min}\right) / \left(D_{\max} - D_{\min}\right), \tag{1}$$

where $D_{\min}$ and $D_{\max}$ indicate the minimum and maximum similarity values, respectively.
The smaller the SI value, the more regular the sleep is.

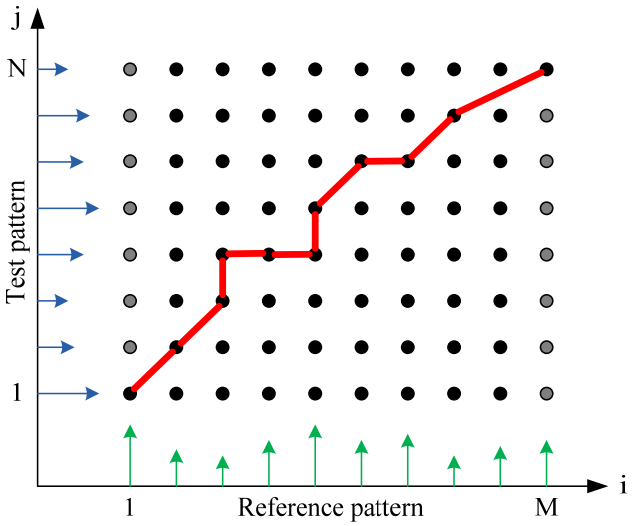


Fig. 6. The minimum distance trace (red line) from the beginning to the end of two data sequences. The black dot indicates the distance between the ith data in the reference pattern and the jth data in the test pattern. The value of the minimum distance, $D(T, R)$, is the sum of all the black dots along the red line, and indicates the similarity between the two patterns. The smaller the value of $D(T, R)$, then the greater the similarity is between the two patterns. All the data in the two data sequences were calculated to build a two-dimensional $N \times M$ distance matrix. Because the endpoint had constraints, the grey-coloured dots can be ignored.

The reference pattern was created by selecting one week's usual sleep data, and averaging these daily HR profiles after noise suppression and data length unification. The daily SI value was calculated using the daily HR profile and the reference pattern. The smaller the SI value, the greater the similarity was between the daily pattern and the reference pattern. Figure 7 shows the variation in the SI value over a period of seven weeks. SI values less than 0.6 indicate that daily sleep was relatively stable, but three days had SI values above 0.6. These three days were confirmed as coinciding with daily life behavioural changes, or heavy drinking in year-end and New Year parties. The HR data showed an increase as a whole and a marked variation pattern over these three days, suggesting that perhaps the use of alcohol stimulated the sympathetic nervous system and accelerated the heart-beat.

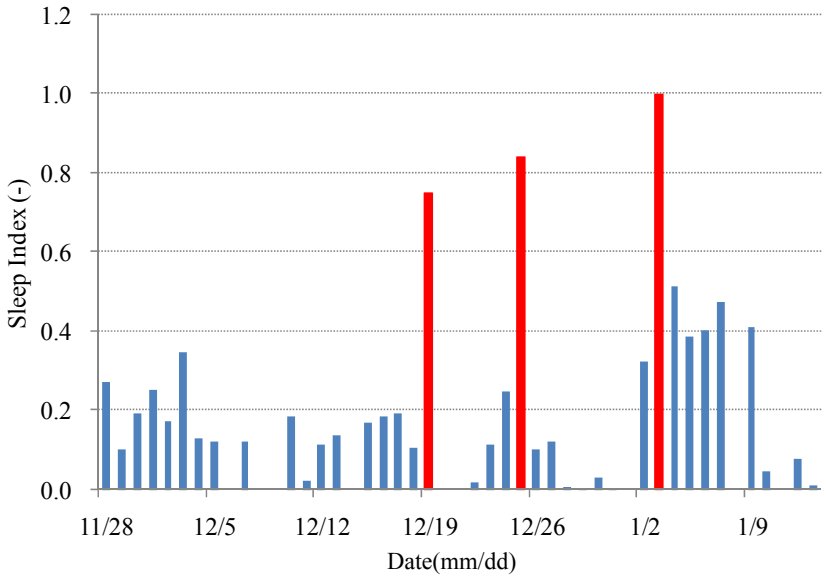


Fig. 7. Variation in SI over a period of seven weeks. The SI value was calculated using the DTW method from HR data collected from a male student in his twenties. The lower the value of the bar, the higher the similarity was, which in turn implies usual daily behaviour. The three higher red bars, whose values are greater than 0.6, indicate the days when the subject had a heavy intake of alcohol.

2.1.5 Estimation of the menstrual cycle

The menstrual cycle is usually estimated from the oral basal body temperature (BBT) in clinical practice. However, taking daily oral measurements is inconvenient for most women. By contrast, there are many convenient ways to measure HR. This study investigated whether the variation in HR measured during sleep could reveal menstrual rhythmicity, as oral BBT does.

The menstrual cycle was estimated by following three steps: calculation of the HR statistic profile, preprocessing of the profile, and analysis of the profile rhythmicity.

The first step was to calculate the daily HR mode value (most frequent value) from the noise-suppressed HR data over a single night, i.e., more than 20,000 HR data points during a 6–7-hour sleep episode. The second step had two tasks: (i) to smooth the daily HR mode profile using a Savitzky–Golay filter, and (ii) to remove any ultra-slow baseline deviations (which may imply seasonal biorhythmic changes and remain to be studied further in detail) using a multirate filtering approach. The final step was to estimate the rhythmicity from the detrended profile of the daily HR mode value using the cosinor analysis method.

The cosinor analysis method is often used to estimate biorhythms with regular cycle length from biological time series data (Nelson et al., 1979). Our aim was to look for the optimal parameter set (M, A, ω, φ) to represent the measured data using a cosine function, as shown in Equation (14)

$$f(t_i) = M + A\cos(\omega t_i + \varphi)\,, \tag{1}$$

where t_i represents the time of measurement of the ith data, M is the mean level (Midline Estimating Statistic Of Rhythm (MESOR)) of the cosine curve, A is the amplitude of the function, ω is the angular frequency (reciprocal of the cycle length) of the curve, and φ is the acrophase (horizontal shift) of the curve.

Considering the measurement of y_i to be the sum of $f(t_i)$ at time t_i and the residual error ε_i

$$y_i = M + A\cos(\omega t_i + \varphi) + \varepsilon_i\,, \tag{1}$$

The errors ε_i were assumed to be independent and normally distributed with a zero mean value and a common residual variance, σ^2.

The task was to find the optimal parameter set (M, A, ω, φ) that best fitted the measurement data y_i using Equation (15), and could be realized using the least-squares regression method. Equation (15) was rewritten as below.

$$y_i = M + A\cos\varphi\cos\omega t_i - A\sin\varphi\sin\omega t_i + \varepsilon_i\,, \tag{1}$$

Assigning surrogate parameters (β, γ), we obtain

$$\beta = A\cos\varphi \text{ and } \gamma = -A\sin\varphi\,, \tag{1}$$

$$x_i = \cos\omega t_i \text{ and } z_i = \sin\omega t_i\,, \tag{1}$$

Substituting Equations (17) and (18) into Equation (16), we get

$$y_i = M + \beta x_i + \gamma z_i + \varepsilon_i\,, \tag{1}$$

Supposing ω in Equation (18) has been suggested previously, and y_i in Equation (19) becomes a linear equation of M, β, and γ. Once M, β, and γ are calculated by applying the least-squares method to Equation (19), the optimal parameter set (M, β, γ) can be obtained.

The residual sum of squared (RSS) error is

$$RSS = \sum_{i=1}^{n}\left[y_i - (M + \beta x_i + \gamma z_i)\right]^2\,, \tag{1}$$

where n is the data length.

To minimize the value of RSS, Equation (20) is partially differentiated with respect to M, β, and γ. The following normal simultaneous equations can be established.

$$\begin{cases} nM + \left(\sum_{i=1}^{n} x_i\right)\beta + \left(\sum_{i=1}^{n} z_i\right)\gamma = \sum_{i=1}^{n} y_i \\ \left(\sum_{i=1}^{n} x_i\right)M + \left(\sum_{i=1}^{n} x_i^2\right)\beta + \left(\sum_{i=1}^{n} x_i z_i\right)\gamma = \sum_{i=1}^{n} x_i y_i \\ \left(\sum_{i=1}^{n} z_i\right)M + \left(\sum_{i=1}^{n} x_i z_i\right)\beta + \left(\sum_{i=1}^{n} z_i^2\right)\gamma = \sum_{i=1}^{n} z_i y_i \end{cases}, \quad (1)$$

After the parameter set (M, β, γ) is derived from the simultaneous equations (21), the value of *RSS* can be calculated using Equation (20). The values of A and φ can be calculated using Equation (17) as below:

$$A = \sqrt{\beta^2 + \gamma^2}\,, \quad (1)$$

$$\varphi = \arctan\left(\frac{\gamma}{\beta}\right), \quad (1)$$

Because the value of *RSS* depends on the proposed value of ω, the optimal value of ω, which has the minimum value of *RSS*, is chosen as the estimated cycle length.

Figure 8 shows the HR mode value and standard deviation profiles over a period of six months (upper subplot), and the menstrual cycle estimation procedure (lower subplot).

The HR data were collected from a female subject in her thirties during daily sleep from 8 October to 31 March. The data collection rate was 93.2% (i.e., 164 days collected out of a total 176-day period). The starting dates of the subject's menstruation were recorded by the subject as 15 October, 12 November, 9 December, 7 January, 5 February, 3 March, and 30 March. Each menstrual cycle over the six-month period could be deduced as being 28, 27, 29, 29, 26, and 27 days, respectively, and the average length ± the standard deviation of the self-recorded menstrual cycles was 27.7 ± 1.2 days.

The daily HR mode value and standard deviation were calculated from more than 20,000 HR data points during the 6–7-hour measurements of a single night's sleep episode. As shown in the upper subplot of Figure 8, the fluctuation of the raw HR mode value profile (MVP) shows no apparent regularity along the time axis.

The lower subplot shows the smoothed HR MVP data (bold blue line) obtained by applying the Savitzky–Golay smoothing filter to the raw HR MVP data. A slow wandering baseline in the smoothed HR MVP data was extracted using the multirate filter and subtracted from the smoothed HR MVP data to produce the detrended HR MVP data (dotted black line). The cosinor analysis method was used to calculate the best approximation (bold red line) of the detrended HR MVP data and to obtain the best-fitted menstrual cycle length of 24.9 days. This compares with the average self-recorded menstrual cycle length of 27.7 days, i.e., the mathematically estimated menstrual length induced an estimation error of 10.1%. It was observed that the timing of the self-recorded menstruation starting dates corresponded to the decrease phase in HR MVP data approximately, a similar characteristic which is shown in BBT biphasic data.

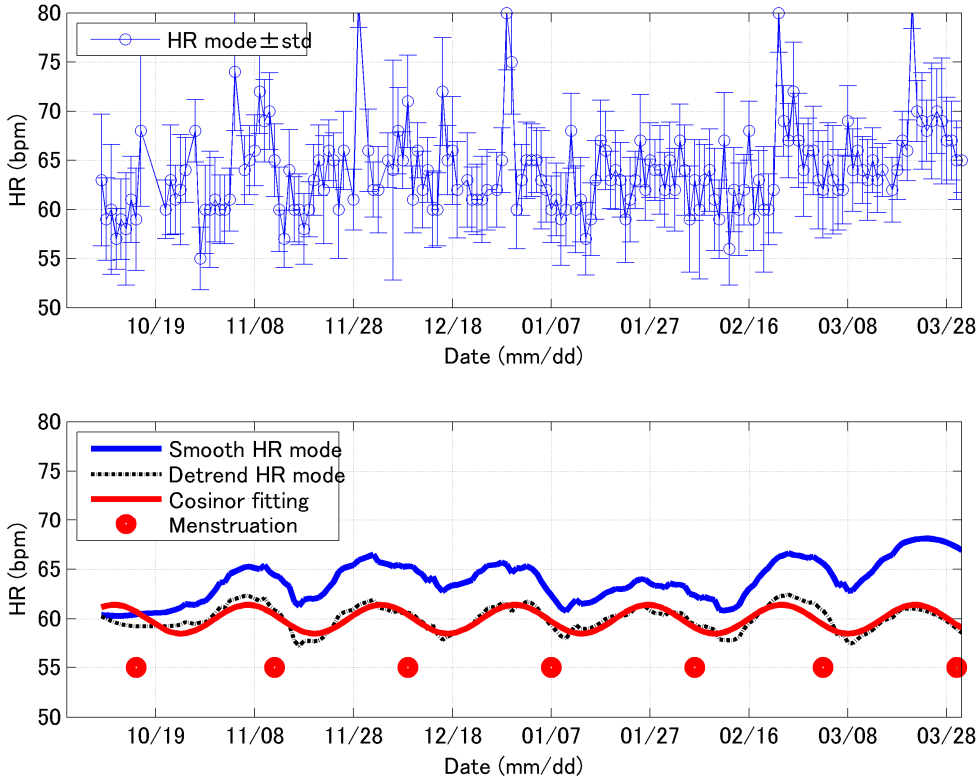


Fig. 8. HR mode value and standard deviation profiles (upper subplot), and menstrual cycle estimation procedure (lower subplot). Data are plotted based on the day-by-day data along the x-axis. The y-axis denotes HR in units of bpm. In the upper subplot, the data markers "o" and vertical bars "|", terminated at the upper and lower ends by short horizontal lines "-", show the mode values (most frequent values) and standard deviation of the HR data in daily sleep episodes. Some sporadic discontinuities can be seen, as no data were collected during those days. In the lower subplot, the bold blue line shows the smoothed profile of the daily HR mode values, and the black dotted line shows the detrended result of the smoothed HR mode values. The red line is the cosinor-fitted results of the black dotted line. Red circles denote the menstruation starting dates that were self-recorded by the subject.

The cosinor analysis method does not require that the data be sampled at equal intervals, and it also tolerates incidents of missing data. It provides an accessible means of estimating the periodic property of menstrual cycles. However, the cosinor analysis method postulates that the data should be reasonably represented in a deterministic cyclic form with a constant period. This prerequisite makes it unsuitable for those women with irregular menstrual cycles. To deal with irregular cycle cases, a hidden Markov model (HMM)-based method is presented in the next section.

2.2 Discovery of a single biorhythm from multiple vital signs

This section describes the estimation of a biphasic property, indicating ovulation and menstruation periods, in female menstrual cycles by applying the HMM method to three types of body temperature data: the oral basal body temperature (BBT), the skin body temperature (SBT), and the core body temperature (CBT).

Menstrual cycle dynamics, from ovum production to development, maturation, release, and fertilization, are one of the most important mechanisms in maintaining female mental and physical well-being, as well as reproductive function. This cyclic phenomenon is marked by changes in several physiological and hormonal signs. Throughout the menstrual cycle, changes occur in a variety of hormones, such as the luteinizing, follicle stimulating, progestational (luteal), and oestrogen (follicular) hormones, as shown in Figure 9. These changes are known to be reflected by changes in BBT measurements or in the chemical composition of the urinary metabolites of oestrogen and progesterone, cervical mucus, and saliva (Sund-Levander et al., 2002).

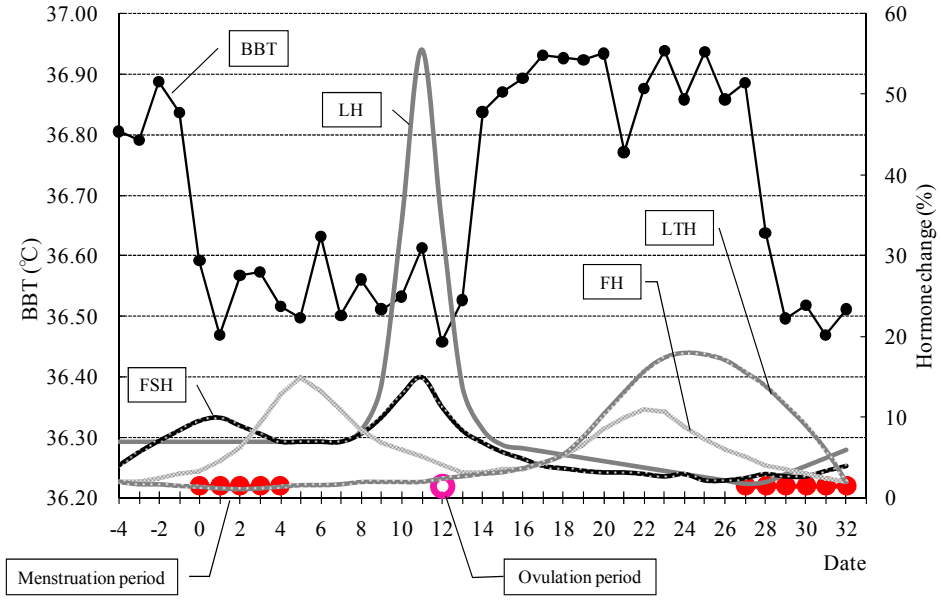


Fig. 9. Biphasic profile of the basal body temperature (BBT) and related hormone changes during a menstrual cycle. LH and LTH are the luteinizing hormone and luteal hormone, respectively, and FH and FSH are the follicular hormone and follicle-stimulating hormone, respectively. The menstruation period is indicated by the red dots, and ovulation day is marked by the pink circle. During the ovulation period, the surge in LH triggers the release of the ovum. If there is no chance of fertilization occurring within a period of one day, the ovum will shrink and form lutein cells. The concentration of LTH will increase and lead to a rise in BBT.

It is possible to correlate hormonal secretion with changes in the genital tissues during a normal menstrual cycle by employing modern bioassay techniques. Three different methods (biological, biochemical, and biophysical) have been developed to elucidate the cyclic

properties of ovulation day and the menstruation period during menstrual cycles (Collins, 1982). Urine and cervical mucus examination methods require chemical reagents and a complicated operation. The salivary method is vulnerable to influence from alcohol, smoke, and food. Cervical mucus and BBT are reported to be the most readily observable parameters among several physiological and hormonal signs (Owen, 1975; Royston, 1982).

Studies on the cause of body temperature changes in women, including BBT, CBT, and rectal temperature, can be traced back to the 1930s (Davis & Fugo, 1948; Lee, 1988; Zuck, 1938). Changes in body temperature, from lower to higher or vice versa, are indicative of the hormonal changes that lead to ovulation and menstruation. Because the BBT method only requires regular oral temperature measurements immediately following sleep, it is now widely accepted as a practical method for estimating the menstrual cycle. However, as BBT measurements are easily affected by any phlogistic illness, such as influenza or toothache, the biphasic property is often ambiguous, and it is difficult to decide the transition points from the temperature profile by visual observation. Therefore, the result largely depends on individual knowledge and a subjective judgement. Extreme caution is required in the interpretation of BBT data when evaluating menstrual cycle dynamics (Baker & Driver, 2007; Moghissi, 1980).

The aims of this study were twofold. The first was to examine whether an HMM-based method was applicable for estimating the biphasic property of menstrual cycles. The second was to investigate whether the same biorhythmic story can be told by different forms of body temperature data, which are measured at different times, at different sites, and using different techniques.

2.2.1 Data collection

Three forms of body temperature data were collected from each subject. As shown in Figure 10, both the SBT and the CBT were collected automatically by attaching two sensor devices (QOL Co. Ltd, 2009) on two sides of a drawers strap during sleep.

The SBT device (orange ellipse in Figure 10) was programmed to measure the skin body temperature at 10-minute intervals from midnight to 6:00 a.m. Measurement outliers above 40 °C or below 32 °C due to poor contact or movement artefacts were automatically disregarded. In the end, 37 data points at most can be collected during a six-hour period. The collected temperature data were encoded as a two-dimensional bar code, known as a "Quick Response" (QR) code (Denso Wave Inc., 2009) and displayed on an LCD window. A mobile phone built-in camera was used to capture the QR code image (Figure 10 (a)) on the device display (middle cycle). Once the QR code was captured on the mobile phone (Figure 10 (b)), the original temperature data (Figure 10 (c)) were recovered from the captured image and transmitted to a database server via HTTP protocol through a mobile network for data storage and analysis.

The CBT device (black cube in Figure 10) was developed using the zero-heat-flow principle (Kobayashi et al., 1975; Togawa, 1985; Nemoto & Togawa, 1988; Yamakage & Namiki, 2003). The device measured the deep tissue temperature at four-minute intervals following the first reading, which was obtained 90 minutes after the device was switched on. This was to ensure that the heat flow was balanced. The CBT data were collected using the electro-magnetic coupling method employing a docking station connected to a PC via an RS232 interface.

Oral BBT was measured by inserting a digital thermometer ("C520", Terumo Corp.) into the hypoglottis each morning immediately the subject wakes up.
Menstruation periods were recorded by the subject. Ovulation days were examined around the middle of the menstrual cycle using a diagnostic test kit ("Dotest LH", Rohto Pharmaceutical Co., Ltd), which identified the changes in concentration of LH, whose secretion increases suddenly before ovulation in a woman's urine. The day when a positive result was detected in the test was considered as the ovulation day.

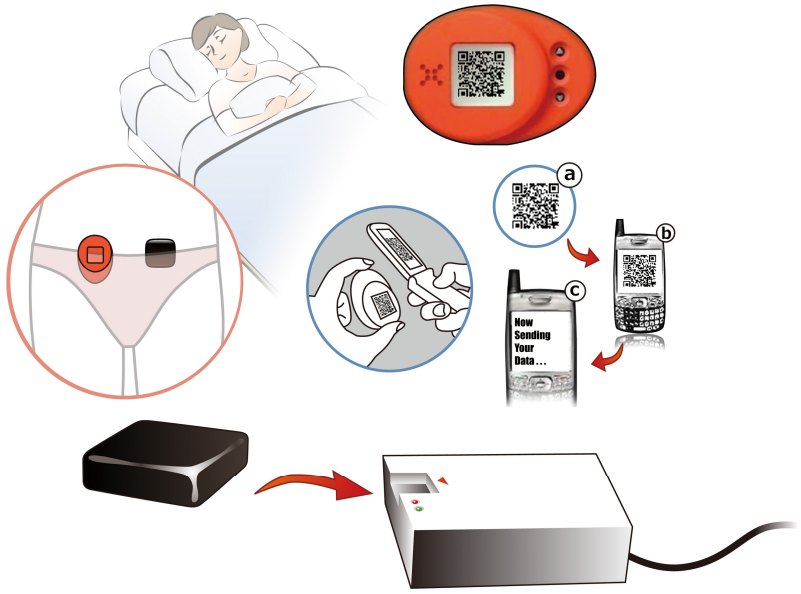


Fig. 10. A schematic drawing of the CBT and SBT data collection. Both the CBT and the SBT were measured automatically by clipping two wearable devices on a drawers strap on two sides of the subject's waist during sleep. The black cubic device was used for CBT data collection. The orange elliptic device was used to detect the SBT.

Figure 11 shows sample profiles of SBT and CBT data collected from a female subject in her thirties. The variation in the amplitude of the SBT even during sleep reached 1.5 °C, while the variation in the amplitude of the CBT was about half this value. This phenomenon is also shown in Figure 13. The different behaviour of the SBT and CBT is perhaps due to the SBT measurements being much more sensitive to the degree of contact with the skin, and this leads to many more artefacts.

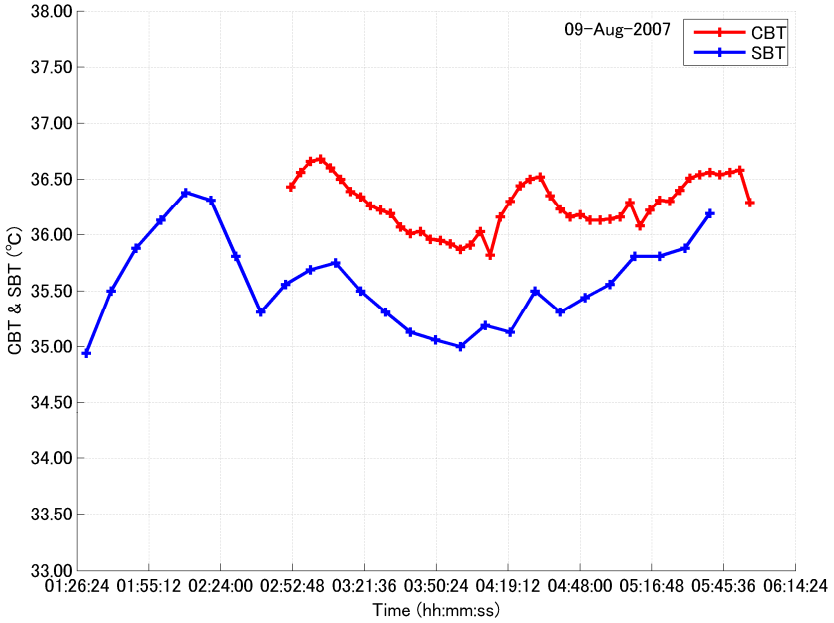


Fig. 11. A night's profile of the SBT (blue line) and CBT (red line) measured during sleep. The SBT data were collected every 10 minutes, and the CBT data were collected every four minutes after obtaining the first reading 90 minutes after the device was turned on.

2.2.2 Data analysis

The biphasic property of body temperature in a menstrual cycle was modelled using a discrete hidden Markov model (HMM) with two hidden phases: a lower temperature (LT) phase and a higher temperature (HT) phase, as shown in Figure 12.

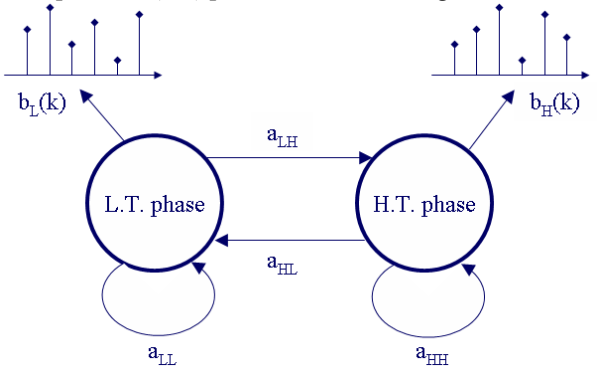


Fig. 12. A discrete HMM with two hidden phases for modelling the biphasic property of body temperature in a menstrual cycle. The term a_{LH} and similar terms represent the phase transition probability from the LT phase to the HT phase, and vice versa. The term $b_L(k)$ denotes the probability of a measurement data point *k* coming from the LT phase. The term $b_H(k)$ denotes the probability of a measurement data point *k* coming from the HT phase.

The biphasic property of the body temperature profile was estimated by finding the optimal HMM parameter set $\lambda(A,B,\pi)$ that determined the hidden phase from which each datum arose. The parameter set of an HMM is assigned randomly in the initial condition, and is optimized through the forward–backward iterative procedure until $P(O \mid \lambda)$, the probability of the measured temperature data originating from the HMM model with an assumed parameter set $\lambda(A,B,\pi)$, converges to a stable maximum value, or until the absolute logarithm of the previous and current difference in $P(O \mid \lambda)$ is not greater than the value of δ. The algorithms for calculating the forward variable, α, the backward variable, β, and the forward–backward variable, γ, are shown in Equations (24) to (26).

The forward variable, $\alpha_t(i)$, denotes the probability of phase q_i at time t based on a partial observed temperature data sequence, $O_1,O_2,\ldots,O_t$, until time t, and can be calculated using the following steps for a given set of $\lambda(A,B,\pi)$.

$$\alpha_t(i) = P_r(O_1,O_2,\ldots,O_t,i_t = q_i \mid \lambda)$$

$$\alpha_1(i) = \pi_i b_i(O_1), \quad 1 \le i \le N, t = 1$$

$$\alpha_{t+1}(j) = \left[\sum_{i=1}^{N} \alpha_t(i) a_{ij}\right] b_j(O_{t+1}), \quad 1 \le j \le N, t = 1,2,\ldots,T-1 \qquad (24)$$

The backward variable, $\beta_t(i)$, denotes the probability of phase q_i at time t based on a partial observed temperature data sequence, $O_{t+1},O_{t+2},\ldots,O_T$, from time t+1 to T, and can be calculated using the following steps for a given set of $\lambda(A,B,\pi)$.

$$\beta_t(i) = P_r(O_{t+1},O_{t+2},\ldots,O_T \mid i_t = q_i,\lambda)$$

$$\beta_T(i) = 1, \quad 1 \le i \le N, t = T$$

$$\beta_t(i) = \sum_{j=1}^{N} a_{ij} b_j(O_{t+1}) \beta_{t+1}(j), \quad 1 \le i \le N, t = T-1,T-2,\ldots,1 \qquad (25)$$

To find the optimal sequence of hidden phases for a measured temperature sequence, O, and a given model, $\lambda(A,B,\pi)$, there are multiple possible optimality criteria.

Choosing the phases q_t that are individually most likely at each time t, i.e., maximizing $P(q_t = i \mid O,\lambda)$, is equivalent to finding the single best phase sequence (path), i.e., maximizing $P(Q \mid O,\lambda)$ or $P(Q,O \mid \lambda)$. The forward–backward algorithm is then applied to find the optimal sequence of phases q_t at each time t, i.e., maximize $\gamma_t(i) = P(q_t = i \mid O,\lambda)$ for a measured temperature sequence, O, and a given parameter set, $\lambda(A,B,\pi)$.

$$\gamma_t(i) = P(q_t = i \mid O,\lambda),$$

$$= \frac{P(O, q_t = i \mid \lambda)}{P(O \mid \lambda)} = \frac{P(O, q_t = i \mid \lambda)}{\sum_{i=1}^{N} P(O, q_t = i \mid \lambda)}$$
$$= \frac{P(o_1 o_2 ... o_t, q_t = i \mid \lambda) P(o_{t+1} o_{t+2} ... o_T \mid q_t = i, \lambda)}{\sum_{i=1}^{N} P(o_1 o_2 ... o_t, q_t = i \mid \lambda) P(o_{t+1} o_{t+2} ... o_T \mid q_t = i, \lambda)}$$
$$= \frac{\alpha_t(i)\beta_t(i)}{\sum_{i=1}^{N} \alpha_t(i)\beta_t(i)} \quad , \tag{26}$$

The most likely phase q_t^* at time t can be found as

$$q_t^* = \arg\max_{1 \le i \le N} [\gamma_t(i)], \quad 1 \le t \le T \quad , \tag{27}$$

There are no existing analytical methods to optimize $\lambda(A,B,\pi)$, so that $P(O \mid \lambda)$ or $P(O,I \mid \lambda)$ is usually maximized (i.e., $\lambda^* = \arg\max_{\lambda} [P(O \mid \lambda)]$ or $\lambda^* = \arg\max_{\lambda} [P(O,Q \mid \lambda)]$,) using gradient techniques and an expectation-maximization method. In this study, the Baum–Welch method was used because of its numerical stability and linear convergence (Rabiner, 1989).

To update $\lambda(A,B,\pi)$ using the Baum–Welch re-estimation algorithm, we defined the variable $\xi_t(i,j)$ to express the probability of a datum being in phase i at time t and phase j at time t+1, given the model parameter set and the temperature data sequence, as

$$\xi_t(i,j) = P(q_t = i, q_{t+1} = j \mid O, \lambda) = \frac{P(q_t = i, q_{t+1} = j, O \mid \lambda)}{P(O \mid \lambda)} \quad , \tag{28}$$

From the definitions of the forward and backward variables, $\xi_t(i,j)$ and $\gamma_t(i)$ are related as

$$\xi_t(i,j) = \frac{\alpha_t(i) a_{ij} b_j(o_{t+1}) \beta_{t+1}(j)}{P(O \mid \lambda)} = \frac{\alpha_t(i) a_{ij} b_j(o_{t+1}) \beta_{t+1}(j)}{\sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_t(i) a_{ij} b_j(o_{t+1}) \beta_{t+1}(j)} \quad , \tag{29}$$

$$\gamma_t(i) = P(q_t = i \mid O, \lambda) = \sum_{j=1}^{N} P(q_t = i, q_{t+1} = j \mid O, \lambda) = \sum_{j=1}^{N} \xi_t(i,j) \quad , \tag{30}$$

where $\sum_{t=1}^{T-1} \gamma_t(i)$ indicates the expected number of transitions from phase i in O, and $\sum_{t=1}^{T-1} \xi_t(i,j)$ indicates the expected number of transitions from phase i to phase j in O.

Therefore, $\lambda(A,B,\pi)$ can be updated using Equations (31)–(33) as follows.

As π_i is the initial probability and indicates the expected frequency (number of times) in phase i at time t = 1, $\pi_i = \gamma_1(i)$, it can be used to calculate the forward and backward variables.

$$\pi_i = \frac{\alpha_1(i)\beta_1(i)}{\sum_{i=1}^{N}\alpha_1(i)\beta_1(i)} = \frac{\alpha_1(i)\beta_1(i)}{\sum_{i=1}^{N}\alpha_T(i)}, \tag{31}$$

where a_{ij} is the transition probability from phase i to phase j and can be calculated from the expected number of transitions from phase i to phase j divided by the expected number of transitions from phase i.

$$a_{ij} = \frac{\sum_{t=1}^{T-1}\xi_t(i,j)}{\sum_{t=1}^{T-1}\gamma_t(i)} = \frac{\sum_{t=1}^{T-1}\alpha_t(i)a_{ij}b_j(o_{t+1})\beta_{t+1}(j)}{\sum_{t=1}^{T-1}\alpha_t(i)\beta_t(i)}, \tag{32}$$

where $b_j(k)$ is the expected number of data arising from phase j, divided by the expected number of all measured data arising from phase j, and can be calculated by

$$b_j(k) = \frac{\sum_{t=1 \text{ s.t. } O_t = o_k}^{T}\gamma_t(j)}{\sum_{t=1}^{T}\gamma_t(j)}, \tag{33}$$

The initial input quantities are the known data N (number of hidden states), M (number of discrete temperature data), T (number of temperature data), O (symbolized temperature data), and the randomly initialized $\lambda(A,B,\pi)$. Once the values of α, β, and γ are calculated using Equations (24) to (26), then $\lambda(A,B,\pi)$ is updated using Equations (31) to (33) employing the newly obtained values of α, β, and γ. The search for the optimal parameter set, λ_{opt}, is terminated when $P(O|\lambda)$ converges to a stable maximum value, or when the absolute logarithm of the previous and current $P(O|\lambda)$ difference reaches δ. Thus, the most likely phase from which a datum is observed can be estimated using Equation (27).

2.2.3 Results

The HMM approach was applied to three types of body temperature data series, BBT, CBT, and SBT, which were measured using different techniques on different sites and at different times, respectively. Figure 13 shows the same story of female body rhythmicity (menstrual cycle) along with different body temperature measurements, and examines the algorithmic performance of the biphasic property estimation by comparing the menstruation records and ovulation test results.

Clinically, the transition point from the HT to the LT phase during a biphasic menstrual cycle corresponds to the menstruation period, and ovulation should occur with a coincidence of the transition point from the LT to the HT phase in time.

As shown in Figure 13, among the six menstruation periods and five days of ovulation over six months, the biphasic property estimated from the BBT data coincided with all the menstruation periods and four out of five ovulation days. The single mismatch error was one day later in the estimation result than the actual ovulation day. The biphasic property derived from CBT identified all six menstruation periods and all five ovulation days, but with errors in three out of five ovulations. Because a severe artefact occurred in the SBT

measurements (perhaps due to poor contact with skin), the daily variation in the SBT data was much higher than that in both the CBT and the BBT data. Six menstruation periods were identified, but three out of five ovulations were missed in the SBT measurements. Overall, the best estimation result was obtained from the BBT measurements. The CBT data were the second best in performance, and the SBT data showed the poorest result.

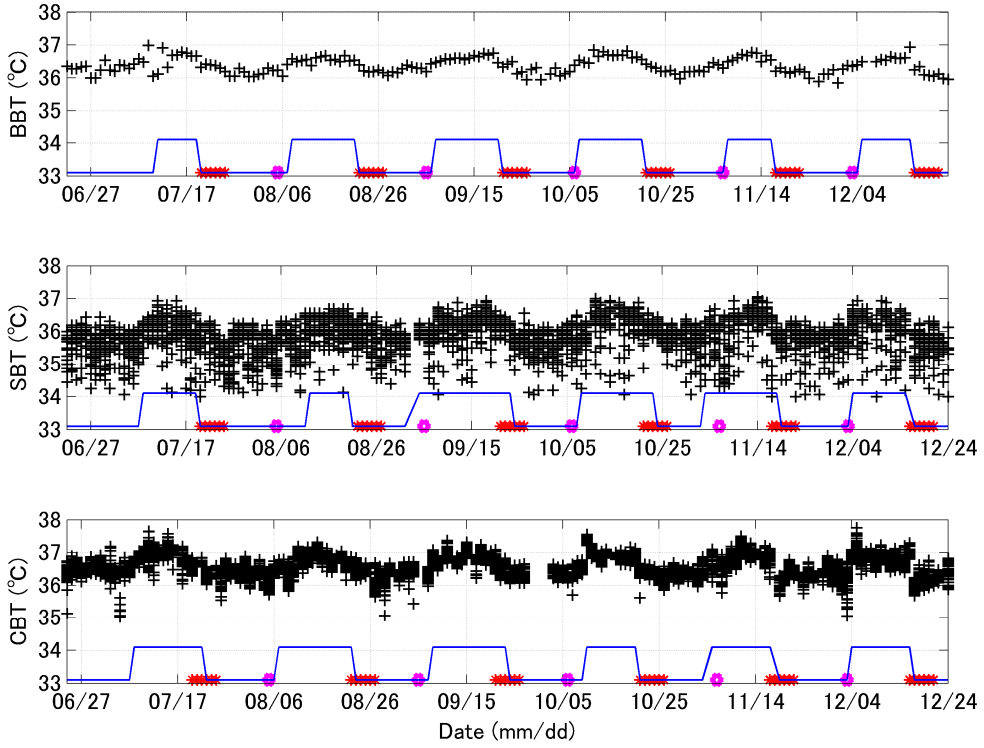


Fig. 13. Measurement of three types of body temperature data (SBT, BBT, and CBT), and the detected biphasic menstrual cycles over a period of six months. The BBT, SBT, and CBT data are plotted from top to bottom, respectively, and the data are indicated by the black markers "+". Each adjacent blue line indicates the detected biphasic menstrual cycles. The menstruation periods denoted by the red star "*" were recorded by subject. Ovulation days are denoted by the pink circle, and were determined by using a commercially available kit "Dotest LH" utilizing the LH secretion test. Physiologically, the transient point from the HT to the LT phase corresponds to a menstruation period, and the transient point from the LT to the HT phase corresponds to the ovulation day.

3. Discussion

Cosmic and hominine rhythms exist eternally and ubiquitously in the temporal and spatial domains. The rhythmic nature in the universe influences every aspect of animals and plants, commencing before conception and extending beyond death (Palmer, 2002; Foster &

Kreitzman, 2005). A large volume of academic understanding has been documented in the scientific literature (Refinetti, 2005).

Biorhythms are built-in genetically, and have evolved naturally through an interaction with the host's environmental factors. The range of biorhythm periods extends from milliseconds to more than a year. Individual clock cells exist genetically in many peripheral tissues, and oscillate semi-autonomously under the coordination of the SCN in the anterior hypothalamus (Shirakawa et al., 2001). Thus, a time-dependent alternation of homeostasis for different endocrine, physiological, and behavioural functions is controlled by the SCN (master clock) and the derived time instruction to the peripheral tissues (slave clocks) throughout the body (Dunlap, 1999). Biorhythmic functions within the human body demonstrate different peak–trough performances in different time slots in the course of a day, month, and year.

However, in most current clinical routines, when a sample of blood, urine, saliva, or tissue is taken, or when other vital signs, such as ECG, blood pressure, and body temperature are measured, they are evaluated as being normal or abnormal by simply comparing them with a range of values obtained statistically from a large population without any consideration of when the sample or measurement was taken, because it is supposed that human body functions are based on a homeostatic principle and are time-independent (Touitou & Haus, 1992). This doctrine has been criticized in that evaluating a patient's blood pressure using a single reading taken during an office visit is like trying to understand the plot of a film by looking at a single frame (Halberg, 1969).

Circadian rhythmic expression of most symptoms is known to worsen at night. For example, acute pulmonary oedema occurs more frequently at night (Manfredini et al., 2000). Allergic conditions such as asthma, allergic rhinitis, hay fever, and measles are exacerbated at night, during sleep, and in the early morning close to waking up compared with during the day. Similarly, symptoms of rheumatoid arthritis worsen at night and improve during the day (Bellamy et al., 2002; Cutolo et al., 2003). However, a reverse pattern can be found in migraine attacks, which occur more often in the morning immediately after waking and decrease at night (Solomon, 1992).

Circannual rhythms have also been identified in commonly used biomarkers in blood, urine, and saliva (Halberg et al., 1983). For example, sperm concentration is lowest in the summer and highest in the autumn and winter (Levine, 1999). Seasonal variations in immune defence systems, including all types of leukocytes, have been investigated (Touitou & Haus, 1994). Skin tends to be more hydrated in the summer and dryer in the winter, which is linked to an increase in the incidence of dermatitis (Mehling & Fluhr, 2006). The treatment recommended for "summer hypertension" and "winter hypertension", respectively may be opposite (Halberg et al., 2006b). In addition, human hair has been reported to reach a maximum growth rate in September and a minimum rate in January (Reinberg & Ghata, 1964).

Illnesses disrupt normal biorhythms and influence biorhythmic property parameters. Similarly, a perturbation in biorhythmic property parameters reflects a deviation from health status or the incidence of illnesses.

Since the kidney is a major organ of metabolism and detoxification, underlying circadian effects on daily biochemical and physiological processes play a key role in metabolism and detoxification. For example, one of the most important functions of antidiuretic hormone (ADH) is to regulate the body's retention of water. Its release causes the kidneys to conserve water, thus concentrating the urine and reducing urine volume. More ADH is normally

secreted at night than during the day in human beings, causing decreased urine production during the usual sleep episode. However, in older persons or patients with spinal cord injuries, there is a distorted diurnal rhythmic pattern in ADH secretion. Decreased nocturnal secretion of ADH causes increased urine production at night (nocturia) and interrupted sleep. Therefore, occurrence of an abnormal sleeping–waking pattern, i.e., frequent sleep disruption during the night may imply nocturia and kidney disorder (Szollar et al., 1997).

Recognizing biorhythms and their changes is important in interpreting and treating disease. To investigate a wide range of variations in biorhythms and their application in medicine and health care systematically, an innovative framework based upon sound definitions in the biomedical engineering domain is indispensable. It requires not only novel systematic theory and methodology for assessing complicated interactions of the time-dependent factors responsible for disease rhythmicity, but also inventive tools suitable for daily use by both medical professionals and a large population of untrained users.

It is essential that reliable information with fine temporal resolution is collected over a period long enough to allow objective characterization of an individual's periodic phenomena. Development of bioinstrumentation and sensory technologies to detect vital signs, and biochemical markers that do not present any inconvenience and are suitable for use in daily scenarios is of paramount importance.

Measurement of ECG or HR is now possible in various daily life situations. Whenever a person sits on a chair (Lim et al., 2006) or on a toilet (Togawa et al., 1989), sleeps on a bed (Kawarada et al., 2000; Chen et al., 2005), sits in a bathtub (Mizukami et al., 1989; Tamura et al., 1997), or even takes a shower (Fujii et al., 2002), then the ECG and HR data can be monitored without any inconvenience or discomfort.

The smart dress "wealthy outfit" weaves electronics and fabrics together to detect the wearer's vital signs, and transmits the data wirelessly to a computer. The built-in sensors gather information on the wearer's activities, ECG, and body temperature. Despite having nine electrodes and conductive leads woven into it, the suit looks completely normal and is worn without any discomfort (Marculescu et al., 2003; Rossi, 2008).

The wellness mobile phone, "SH706iw", has all the standard features of a mobile phone but also acts as a pedometer, a body fat meter, and a pulse rate and breath gas monitor. Moreover, daily data can be collected using a built-in game-like application (Sharp Corp., 2008; DoCoMo Corp., 2008).

Such innovations in sensory instrumentation technologies and physiological data collection schemes are indispensable for monitoring a wide range of biorhythms in everyday living environments that are oriented to mostly untrained users. Other key features should include zero administration, easy manipulation, automatic fault recovery, and the absence of unpleasantness or disturbance to everyday living to allow perpetual sustainable data collection. Related studies can be classified into three groups: invisible technology that requires no user intervention during operation, wearable technology that can be worn as if a part of the subject's underwear with little discomfort, and ubiquitous technology that is based on mobile devices for instant usage, at any time and in any place (Chen et al., 2008).

Moreover, automatic and continuous monitoring during sleep at night is worth paying special attention to, because not only is the sleeping–waking cycle important in keeping biorhythms in tune, but also much reliable physiological data can be obtained due to fewer movement artefacts. In addition, the attenuation in some biorhythms during sleep helps the

decoupling of overlapping multiple biorhythms and potential masking factors that usually only appear during diurnal measurements.

Further, not only the temporal alternations in the daily and seasonal domain, but also cycles of meteorological and geographical events, such as the solar wind, sun spots, and geomagnetic storms, have an important effect on human body functions (Halberg et al., 2006b). While keeping watch over diverse biorhythms, nature's clocks do not oscillate in isolation. Substantial improvements in daily data aggregation should include collective information, such as meteorological, environmental, and geographical aspects at the same time as the physiological data. This will facilitate the disentangling of diverse causal pathways of many endogenous and exogenous factors within biorhythms, as well as the interrelationships among different biorhythms and natural rhythms across the wide range of temporal and spatial factors.

The discovery of more biorhythms largely depends on learning how to take full advantage of the broad spectrum of accumulated data from a large population over a long period, and how to perform multiple modalities of data fusion through the integration of sound mathematical models and the implementation of robust computational algorithms.

As human beings evolve through an interaction with nature, human physiological functions become more organized and more complicated. Underlying biorhythmic processes in disease manifestation are complex and multifaceted. Although human beings have now become accustomed to the 24-hour light–dark cycle, circadian and other biorhythmic patterns are interwoven with each other and have been incorporated into a single vital sign or biochemical marker. These insights suggest that the separate computational models that have been developed for a single biorhythm will have to be integrated for solving multiple biorhythms. The emerging field of intelligent data mining and algorithm development for identification of hidden biorhythms and the separation of interlaced multiple biorhythms will complement established work in chrono-related medical science.

Chronopharmacology helps to explain the biorhythm dependencies of medications. Comprehensive investigation into biorhythms can assist us to synchronize the rhythmic variation in individual physiological functions while being related to pharmacokinetics, which will tell us how the body responds to a drug, and to pharmacodynamics, which will tell us how the drug affects the body (Redfern & Lemmer, 1997).

Treatment in the evening is associated with an elevation in the circadian amplitude of BP, which in turn may induce iatrogenic CHAT in some patients, thereby unknowingly increasing the risk of cardiovascular disease (Halberg et al., 2006b). It is not wise to lower the risk of hypertension, and instead introduce a higher risk of CHAT, which is like attending to one condition while worsening another.

Abnormalities in the variability of blood pressure and similar signs are difficult to find in a sporadic clinical examination, and the efficacy of treatment is difficult to optimize by relying on a spot check, which is driven by convenience rather than pertinence. Instead, through chronobiology, by interpreting their circadian or preferably longer rhythms, it is possible to comprehend the change of related illnesses in different temporal scales, over a day or over years, and to increase the impact on effectiveness of a treatment through scheduling the time of medication.

The optimal strategy for chronotherapy and administration of treatment for diseases requires clarification of each medication in terms of the best timing and dosage, such as when the drug is interacting with the body most efficiently, with the maximum positive

effect and the minimum adverse effect, and when the most appropriate amounts of the drug should be delivered to the desired target organ along a temporal course.
Timed-delay medication technology and automatic drug delivery devices play an important role in the optimal individualization of a treatment (Lemmer, 2007). Major approaches for drug delivery include oral, pulmonary, and transdermal routes. Microcapsules ingested by mouth can travel freely throughout the body, seek the target organ automatically, and deliver therapeutic agents at a desired time (Orive et al., 2004). Microneedles and electric field-driven polymers can diffuse therapeutic agents to target tissues from scheduled temporal profiles through the skin barrier as a means of penetrating plaques on vessel walls (Reed et al., 1998).
Prominent application of biorhythms in health care, disease prevention, and diagnosis, as well as the timing of treatments and drug regimens will gradually mature and be extensively recognized through diligent efforts and intense collaborations among multiple disciplines.

4. Conclusions

Historical endeavours in the study of biorhythms from both the oriental and the occidental worlds, especially the achievements over the last 60 years, and their application in medicine and health care, have been briefly reviewed in this chapter. A wide range of inherent biorhythm diversity exists and is subject to the influence of various endogenous and exogenous aspects. The destruction or asynchronism of biorhythms will harm human health. Likewise, any indisposition in health will be reflected in biorhythmic fluctuations. To identify various biorhythms and to facilitate their application in medical practice and daily life, convenient monitoring and comprehensive interpretation of long-term physiological data are indispensable. Our exploration has focused on the development of advanced sensory technology and data mining algorithms. These devices are suitable for the continuous monitoring of vital signs over long periods in a daily life environment. The algorithms developed for discovering a wide range of biorhythms were confirmed using long-term physiological data.
Through an investigation of the interplay among biorhythm behaviours, health status, and the intrinsic timing of disease development, a treatment strategy, such as dosage and dosing regimen to maximize the therapeutic effects, guarantee medication safety, and minimize adverse effects, can be optimized using automatic drug delivery technologies. In addition, health care performance and efficiency can be achieved by adapting human activity to the synchronization of organ physiological functions and environmental aspects.

5. Acknowledgements

The author thanks colleagues and students from universities and companies for co-work in the above studies, and thanks participants for their enduring efforts in long-term data collection. These studies were supported in part by several financial resources from: (a) The Innovation Technology Development Research Program under JST (Japan Science and Technology Agency) grant H17-0318; (b) MEXT Grants-In-Aid for Scientific Research No. 20500601; and (c) The University of Aizu Competitive Research Funding P-24.

6. References

Baker, F. C. & Driver, H. S. (2007). Circadian rhythms, sleep, and the menstrual cycle. *Sleep Medicine*, Vol. 8, No. 6, pp. 613-622.

Balsalobre, A.; Brown, S.A.; Marcacci, L.; Tronche, F.; Kellendonk, C.; Reichardt, H.M.; Schütz, G. & Schibler, U. (2000). Resetting of Circadian Time in Peripheral Tissues by Glucocorticoid Signaling. *Science*, Vol. 289. No. 5488, pp. 2344-2347

Bellamy, N.; Sothern, R.B.; Campbell, J. & Buchanan, W.W. (2002). Rhythmic variations in pain, stiffness, and manual dexterity in hand osteoarthritis. *Annals of the Rheumatic Diseases*, Vol. 61, pp. 1075-1080

Biochart Com. (2009). Biorhythms History. http://www.biochart.co.uk/biorhythms-history.shtml

Bocher, P. K. & McCloy, K. R. (2006a). The fundamentals of average local variance - part I: detecting regular patterns. *IEEE Transactions on Image Processing*, Vol. 15, No. 2, pp. 300-310.

Bocher, P.K. & McCloy, K. R. (2006b). The fundamentals of average local Variance-part II: sampling simple regular patterns with optical imagery. *IEEE Transactions on Image Processing*, Vol. 15, No. 2, pp. 311-318.

Chen, W.; Zhu, X.; Nemoto, T.; Kanemitsu, Y.; Kitamura, K. & Yamakoshi, K. (2005). Unconstrained detection of respiration rhythm and pulse rate with one under-pillow sensor during sleep. *Medical & Biological Engineering & Computing*, Vol. 43, No. 2, pp. 306–312.

Chen, W.; Zhu, X.; Nemoto, T.; Wei, D. & Togawa, T. (2008). A Scalable Healthcare Integrated Platform (SHIP) and Key Technologies for Daily Application, *Data Mining in Medical and Biological Research*, IN-TECH, 978-953-7619-30-5, Vienna, Austria, pp. 177-208.

Chirkova, É. N. (1995). Mathematical methods of detection of biological and heliogeophysical rhythms in the light of developments in modern heliobiology: A platform for discussion. *Cybernetics and Systems Analysis*, Vol. 31, No. 6, pp. 903-918.

Chou, B. J. & Besch, E. L. (1974). A computer method for biorhythm evaluation and analysis. *Biological Rhythm Research*, Vol. 5, No. 2, pp. 149-160.

Collins, W. P. (1982). Ovulation prediction and detection. *IPPF Med. Bull.*, Vol. 16, pp. 1–2.

Cutolo, M.; Seriolo, B.; Craviotto, C.; Pizzorni, C. & Sulli, A. (2003). Circadian rhythms in RA. *Annals of the Rheumatic Diseases*, Vol. 62, pp. 593-596.

Davis, M. E. & Fugo, N. W. (1948). The cause of physiologic basal temperature changes in women. *J. Clin. Endocrinol.*, Vol. 8, pp. 550–563.

Denso Wave Inc. (2009). QR code, http://www.denso-wave.com/qrcode/aboutqr-e.html.

Desai, V. G.; Moland, C. L.; Branham, W. S.; Delongchamp, R. R.; Fang, H.; Duffy, P. H.; Peterson, C. A.; Beggs, M. L. & Fuscoe, J. C. (2004). Changes in expression level of genes as a function of time of day in the liver of rats. *Mutation Res.*, Vol. 549, No. 1-2, pp. 115-129.

DoCoMo Corp., (2008). Wellness mobile phone. http://www.nttdocomo.co.jp/product/foma/706i/sh706iw/index.html

Dunlap, J.C. (1999). Molecular Bases for Circadian Clocks – Review. *Cell*, Vol. 96, pp. 271–290.

Dunlap, J.C.; Loros, J.J. & Decoursey, P.J. (2004). *Chronobiology: Biological Timekeeping*, Sinauer Associates, 087893149X, Massachusetts, USA.

Elliott, W. J. (2001). Timing treatment to the rhythm of disease: A short course in chronotherapeutics. *Postgraduate medicine*, Vol. 110, No. 2, pp. 119-122, 125-126, 129

Foster, R.G. & Kreitzman, L. (2005). *Rhythms of Life: The Biological Clocks that Control the Daily Lives of Every Living Thing*, 0-300-10969-5, Yale University, New Haven, USA.

Fujii, M.; Dema, H. & Ueda, T. (2002). Liquid Jet Electrode and Surface Potential Detector, *Japanese Patent* No. JP2002–65625A, 5 March 2002.

Halberg, F.; Halberg, E.; Barnum, C.P. & Bittner, J.J. (1959). Physiologic 24-hour periodicity in human beings and mice, the lighting regimen and daily routine. In: Withrow RB (ed). *Phototperiodism and Related Phenomena in Plants and Animals*. Ed. Publ. No. 55. Washington, D.C., pp. 803-878.

Halberg, F. (1969). Chronobiology. *Ann. Rev. Physiol*., Vol. 31, pp. 675-725.

Halberg, F. (1983). Quo vadis basic and clinical chronobiology: promise for health maintenance. *Am. J. Anat*., Vol. 168, pp. 543-594.

Halberg, F.; Lagoguey, M. & Reinberg, A. (1983). Human circannual rhythms over a broad spectrum of physiological processes. *International journal of chronobiology*, Vol. 8, No. 4, pp. 225-268

Halberg, F. & Cornélissen, G. (1993). Rhythms and blood pressure. *Ann. 1st. Super. Sanita*, Vol. 29, No. 4, pp.647–665.

Halberg, F.; Cornélissen, G.; Katinas, G.; Syutkina, E.V.; Sothern, R.B.; Zaslavskaya, R.; Halberg, F.; Watanabe, Y.; Schwartzkopff, O.; Otsuka, K.; Tarquini, R.; Frederico, P. & Siggelova, J. (2003). Transdisciplinary unifying implications of circadian findings in the 1950s. *Journal of Circadian Rhythms*, Vol. 1, No. 2, pp. 1-61.

Halberg, F.; Cornélissen, G.; Katinas, G.; Tvildiani, L.; Gigolashvili, M.; Janashia, K.; Toba, T.; Revilla, M.; Regal, P.; Sothern, R.B.; Wendt, H.W.; Wang, Z.; Zeman, M.; Jozsa, R.; Singh, R.B.; Mitsutake, G.; Chibisov, S.M.; Lee, J.; Holley, D.; Holte, J.E.; Sonkowsky, R.P.; Schwartzkopff, O.; Delmore, P.; Otsuka, K.; Bakken, E.E.; Czaplicki, J. & the International BIOCOS Group. (2006a). Chronobiology's progress. Part I, season's appreciations 2004–2005: time-, frequency-, phase-, variable-, individual-, age- and site-specific chronomics. *J. Appl. Biomed*., Vol. 4, pp. 1–38.

Halberg, F.; Cornélissen, G.; Katinas, G.; Tvildiani, L.; Gigolashvili, M.; Janashia, K.; Toba, T.; Revilla, M.; Regal, P.; Sothern, R.B.; Wendt, H.W.; Wang, Z.; Zeman, M.; Jozsa, R.; Singh, R.B.; Mitsutake, G.; Chibisov, S.M.; Lee, J.; Holley, D.; Holte, J.E.; Sonkowsky, R.P.; Schwartzkopff, O.; Delmore, P.; Otsuka, K.; Bakken, E.E.; Czaplicki, J. & the International BIOCOS Group. (2006b). Chronobiology's progress. Part II, chronomics for an immediately applicable biomedicine. *J. Appl. Biomed*., Vol. 4, pp. 73–86.

Hufeland, C.W. (1796). Art of Prolonging Life. http://www.archive.org/details/artofprolongingl00hufeuoft

Katz, G.; Durst, R.; Zislin, Y.; Barel, Y. & Knobler, H. Y. (2001). Psychiatric aspects of jet lag: review and hypothesis. *Medical Hypotheses*, Vol. 56, pp. 20-23

Kawarada, A.; Nambu, M.; Tamura, T.; Ishijima, M.; Yamakoshi, K. & Togawa, T. (2000). Fully automated monitoring system of health status in daily life, *Proceedings of the 22nd Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, Vol. 1, pp. 531–533.

Kobayashi, T.; Nemoto, T.; Kamiya, A. & Togawa, T. (1975). Improvement of Deep Body Thermometer for Man. *Annals of Biomedical Engineering*, Vol. 3, pp. 181-188.

Koukkari, W.L. & Sothern, R.B. (2006). *Introducing Biological Rhythms: A Primer on the Temporal Organization of Life, with Implications for Health, Society, Reproduction, and the Natural Environment*, Springer, 978-1-4020-3691-0, New York, USA

Kumar, K., Srivastava, M. & Mandal, S. K. (1992). A Multivariate Method for the Parameter Estimation in Biorhythms. *Biometrical Journal*, Vol. 34, No. 8, pp. 911–917.

Labrecque, G. & Belanger, P. M. (1991). Biological rhythms in the absorption, distribution, metabolism and excretion of drugs. *Pharmacol. Ther.*, Vol. 52, No. 1, pp. 95–107.

Law, M.W.K. & Chung, A.C.S. (2007). Weighted Local Variance-Based Edge Detection and Its Application to Vascular Segmentation in Magnetic Resonance Angiography. *IEEE Transactions on Medical Imaging*, Vol. 26, No. 9, pp. 1224–1241.

Lee, K. A. (1988). Circadian temperature rhythms in relation to menstrual cycle phase. *J. Biol. Rhythms.*, Vol. 3, pp. 255–263.

Lemmer, B. (1994). Chronopharmacology: time, a key in drug treatment. *Ann. Biol. Clin.*, Vol. 52, No. 1, pp. 1-7.

Lemmer, B. (2007). Chronobiology, drug-delivery, and chronotherapeutics. *Adv. Drug Deliv. Rev.*, Vol. 59, No. 9-10, pp. 825-827.

Levine, R.J. (1999). Seasonal variation of semen quality and fertility. *Scand J Work Environ Health*, Vol. 25 Suppl. 1, pp. 34-37

Lim, Y.G.; Kim, K.K. & Park, K.S. (2006). ECG measurement on a chair without conductive contact, *IEEE Trans. Biomed. Eng.*, Vol. 53, No. 5, pp. 956–959.

Manfredini, R.; Portaluppi, F.; Boari, B.; Salmi, R.; Fersini, C. & Gallerani, M. (2000). Circadian Variation in Onset of Acute Cardiogenic Pulmonary Edema is Independent of Patients' Features and Underlying Pathophysiological Causes. *Chronobiology International*, Vol. 17, No. 5, pp. 705-715.

Marculescu, D.; Marculescu, R.; Park, S. & Jayaraman, S. (2003). Ready to wear, *Spectrum, IEEE*, Vol. 40, No. 10, pp. 28–32.

Martha, U. G. & Sejnowski, T. J. (2005). Biological Clocks Coordinately Keep Life on Time. *Science*, Vol. 309, pp. 1196-1198.

Mehling, A. & Fluhr, J.W. (2006). Chronobiology: biological clocks and rhythms of the skin. *Skin Pharmacol. Physiol.*, Vol. 19, No. 4, pp. 182-189.

Mizukami, H.; Togawa, T.; Toyoshima, T. & Ishijima, M. (1989). Management of pacemaker patients by bathtub ECG, *Report of the Institute for Medical & Dental Engineering, Tokyo Medical and Dental University*, 23, pp. 113–119.

Moghissi, K. S. (1980). Prediction and detection of ovulation. *Fertil. Steril.*, Vol. 32, pp. 89–98.

Moser, M.; Fruhwirth, M.; Penter, R. & Winker, R. (2006). Why life oscillates - from a topographical towards a functional chronobiology. *Cancer Causes Control*, Vol. 17, No. 4, pp. 591-599.

Naumova, E.N. (2006). Mystery of Seasonality: Getting the Rhythm of Nature. *J. Public Health Policy*, Vol. 27, No. 1, pp. 2-12.

Nelson, W.; Tong, Y.L.; Lee, J.K. & Halberg, F. (1979). Methods for cosinor-rhythmometry. *Chronobiologia*, Vol. 6, No. 4, pp. 305-323.

Nemoto, T. & Togawa, T. (1988). Improved probe for a deep body thermometer. *Medical and Biological Engineering and Computing*, Vol. 26, No. 4, pp. 456-459.

Ni, M. (1995). *The Yellow Emperor's Classic of Medicine: A New Translation of the Neijing Suwen with Commentary*, Shambhala, 978-1570620805, Massachusetts, USA.

Ohdo, S. (2007). Chronopharmacology Focused on Biological Clock, *Drug Metabolism and Pharmacokinetics*, Vol. 22, No. 1, pp.3-14

Orive, G.; Hernández, R.M.; Gascón, A.R.; Calafiore, R.; Chang, T.M.S.; Vos, P.; Hortelano, G.; Hunkeler, D.; Lacík, I. & Pedraz, J.L. (2004). History, challenges and perspectives of cell microencapsulation. *Trends in Biotechnology*, Vol. 22, No. 2, pp. 87-92.

Owen, J. A. Jr. (1975). Physiology of the menstrual cycle. *Am. J. Clin. Nutr.*, Vol. 28, pp. 333–338.

Palmer, J.D. (2002). *The Living Clock: The Orchestrator of Biological Rhythms*, Oxford University Press, 978-0195143409, New York, USA.

QOL Co. Ltd., (2009). Ran's Night. http://rans-night.jp/

Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE*, Vol. 77, No. 2, pp. 257–286.

Redfern, P.H. & Lemmer, B. (1997). *Chronopharmacology: Physiology and Pharmacology of Biological Rhythms*, Springer, 978-3540615255, New York, USA.

Reed, M.L.; Wu, C.; Kneller, J.; Watkins, S.; Vorp, D.A.; Nadeem, A.; Weiss, L.E.; Rebello, K.; Mescher, M.; Smith, A.J.C.; Rosenblum, W. & Feldman, M.D. (1998). Micromechanical Devices for Intravascular Drug Delivery. *Journal of Pharmaceutical Sciences*, Vol. 87, No. 11, pp. 1387-1394.

Refinetti, R. (2005). *Circadian Physiology*, CRC, 2nd edition, 978-0849322334, FL. USA.

Reinberg, A. & Ghata, J. (1964). *Biological rhythms*, Walker, New York, USA

Reppert, S.M. & Weaver, D.R. (2001). Molecular analysis of mammalian circadian rhythms. *Annu. Rev. Physiol.*, Vol. 63, pp. 647-676.

Rossi, D. D. (2008). Ready to wear: clothes that look hip and track your vital signs, too. http://www.wired.com/techbiz/startups/magazine/16-02/ps_smartclothes, *Wired Magazine*, Vol. 16, No. 2, p. 55.

Royston, J. P. (1982). Basal body temperature, ovulation and the risk of conception, with special reference to the lifetimes of sperm and egg. *Biometrics*, Vol. 38, pp. 397–406.

Sacred Lotus Arts. (2009). The Origins of Traditional Chinese Medicine. http://www.sacredlotus.com/theory/yinyang.cfm

Savitzky, A. & Golay, M. J. E. (1964). Smoothing and Differentiation of Data by Simplified Least Squares Procedures. *Analytical Chemistry*, Vol. 36, No. 8, pp. 1627–1639.

Sharp Corp., (2008). Wellness mobile phone. http://plusd.itmedia.co.jp/mobile/articles/0805/27/news064.html

Shirakawa, T.; Honma, S. & Honma, K. (2001). Multiple oscillators in the suprachiasmatic nucleus. *Chronobiology International*, Vol. 18, No. 3, pp. 371-387.

Smolensky, M. H. & Labrecque, G. (1997). Chronotherapeutics. *Pharmaceutical News*, Vol. 4, pp. 10-16.

Smolensky, M. & Lamberg, L. (2000). *The Body Clock Guide to Better Health: How to Use Your Body's Natural Clock to Fight Illness and Achieve Maximum Health*, Henry Holt and Company, 978-0-8050-5662-4, New York, USA.

Solomon, G.D. (1992). Circadian rhythms and migraine. *Cleveland Clinic Journal of Medicine*, Vol. 59, No. 3, pp. 326-329.

Stetson, M.H. & Watson-Whitmyre, M. (1976). Nucleus suprachiasmaticus: the biological clock in the hamster? *Science*, Vol. 191, No. 4223, pp. 197-199.

Sund-Levander, M.; Forsberg, C. & Wahren, L. K. (2002). Normal oral, rectal, tympanic and axillary body temperature in adult men and women: a systematic literature review. *Scand J. Caring Sci.*, Vol. 16, No. 2, pp. 122-128.

Szollar, S.M.; Dunn, K.L.; Brandt, S. & Fincher, J. (1997). Nocturnal polyuria and antidiuretic hormone levels in spinal cord injury. *Archives of Physical Medicine and Rehabilitation*, Vol. 78, No. 5, pp. 455-458.

Tamura, T.; Yoshimura, T.; Nakajima, K.; Miike, H. & Togawa, T. (1997). Unconstrained heart-rate monitoring during bathing. *Biomedical Instrumentation & Technology*, Vol. 31, No. 4, pp. 391–396.

Togawa, T. (1985). Body temperature measurement. *Clin. Phys. Physiol. Meas.*, Vol. 6, pp. 83-108.

Togawa, T.; Tamura, T.; Zhou, J.; Mizukami, H. & Ishijima, M. (1989). Physiological monitoring systems attached to the bed and sanitary equipments. *Proceedings of the Annual International Conference of the IEEE Engineering in Engineering in Medicine and Biology Society*, Vol. 5, pp. 1461–1463.

Touitou, Y. & Haus, E. (1992). *Biologic Rhythms in Clinical and Laboratory Medicine*, Springer, 978-3540544616, New York, USA.

Touitou, Y. & Haus, E. (1994). Aging of the Human Endocrine and Neuroendocrine Time Structure. *Annals of the New York Academy of Sciences*, Vol. 719, pp. 378-397.

Turek, F. W. & Allanda, R. (2002). Liver has rhythm. *Hepatology*, Vol. 35, pp. 743–745.

Vallot, J.; Sardou, G. & Faure, M. (1922). De l'influence des taches solaires: sur les accidents aigus des maladies chroniques. *Gazette des Hôpitaux*, pp. 904–905.

Veith, I. (2002). *The Yellow Emperor's Classic of Internal Medicine*, University of California Press, 979-0520229364, California, USA.

Wakuda, Y.; Noda, A.; Sekiyama, K.; Hasegawa, Y. & Fukuda, T. (2007). Biorhythm-Based Awakening Timing Modulation. *IEEE International Conference on Robotics and Automation*, pp. 1232–1237.

Wang, H.T. (2005). *Lectures on "The Medical Classic of Emperor Huang"*, People's Medical Publishing House, 711704862X/R.4863, Beijing, China

Watanabe, H. & Chen, W. (2009). Detection of Biorhythm Change from Pulse Rate Measured during Sleep. *Proc. of 48th Japan Soc. ME & BE conference*, pp. 298.

Wikipedia. (2009a). Alexander Chizhevsky. http://en.wikipedia.org/wiki/Alexander_Chizhevsky

Wikipedia. (2009b). Circadian rhythm. http://en.wikipedia.org/wiki/Circadian_rhythm

Wikipedia. (2009c). Jean-Jacques d'Ortous de Mairan. http://en.wikipedia.org/wiki/Jean-Jacques_d'Ortous_de_Mairan

Wikipedia. (2009d). Medical Classic of Emperor Huang. http://en.wikipedia.org/wiki/Huangdi_Neijing

Wikipedia. (2009e). Sanctorius. http://en.wikipedia.org/wiki/Sanctorius

Wikipedia. (2009f). Yin and yang. http://en.wikipedia.org/wiki/Yin_yang

Wikipedia. (2009g). Zhang Zhongjing. http://en.wikipedia.org/wiki/Zhang_Zhongjing

Xuan, W.; Yuan, C.S.; Bieber, E.J. & Bauer, B. (2006). *Textbook of Complementary and Alternative Medicine*, 2nd Edition, Informa HealthCare, 978-1842142974, London, UK.

Yamakage, M. & Namiki, A. (2003). Deep temperature monitoring using a zero-heat-flow method. *Journal of Anesthesia*, Vol. 17, No. 2, pp. 108-115.

Zhang, C.J.; Xu, G.Q. & Zong, Q.H. (1995). *Commentaries of Medical Classic of Emperor Huang*, People's Medical Publishing House, 7-117-00789-3/R.790, Beijing, China

Zuck, T. T. (1938). The relation of basal body temperature to fertility and sterility in women. *Am. J. Obstet. Gynecol.*, Vol. 36, pp. 998–1005.

Linear and Nonlinear Synchronization Analysis and Visualization during Altered States of Consciousness

Vangelis Sakkalis[1] and Michalis Zervakis[2]
[1] *Institute of Computer Science, Foundation for Research and Technology, Greece*
[2] *Electronic and Computer Engineering Dept., Technical University of Crete, Greece*

1. Introduction

This chapter focuses on analyzing various established approaches in bio-medical signal analysis, with emphasis on the dynamic evolution of time series and their structural inter-relationships. The application area involves electroencephalographic (EEG) signals, but the analytic approaches can be extended to most physiological signals. Two major approaches will be discussed including i) linear synchronization methods and ii) nonlinear synchronization methods based on chaos theory. Comparisons and extensions will be thoroughly investigated as to highlight their advantages and limitations over traditional power-based techniques. In particular, this chapter will investigate traditionally formulated coherence methods (i.e., correlation, coherence, short time Fourier transform coherence) and modern linear analysis techniques (i.e., Wavelet coherence and the Partial Directed Coherence), as well as established nonlinear methods, such as phase and generalized synchronization algorithms. Applications will be referenced in the domains of cognitive awareness assessment and diagnosis of pathological cases, such as Schizophrenia, Alcoholism and Epilepsy. Furthermore, a link to graph theory and visualization of synchronization status will be briefly addressed in an attempt to better describe the functional characteristics of brain networks.

2. Linear synchronization techniques applied in EEG analysis

The revolutionary discoveries of Jean Baptiste Fourier (1768-1830), have had a major impact on the development of mathematics and are of great importance in an extremely wide range of scientific and engineering disciplines. Although, the original work of Fourier was focused on problems of mathematical Physics, which happen in the continuous time domain, there is also a rich variety of applications of the tools of Fourier analysis for discrete-time signals and systems. In particular, discrete-time concepts and methods are fundamental to the discipline of numerical analysis. Formulas for the processing of discrete sets of data points to produce numerical approximations for interpolation, integration, and differentiation were investigated in the early 1600's. This section starts with the mathematical foundation of the

traditionally formulated Magnitude Squared Coherence (MSCOH) that is based on the Fast Fourier Transform (FFT) theoretical foundation and extents to the most modern Wavelet and Partial Directed Coherence (PDC) ones.

2.1 Cross-correlation & Coherence

The study of functional relationships between two brain regions has been one of the main aims of the EEG. As early as 1951, the cross-correlation function was used to study the similarity between two EEG signals (Brazier and Casby, 1952). Its application in the area of brain analysis is based on the assumption that the higher the correlation, the stronger the functional relationships between the related brain regions (Shaw, 1981; Shaw, 1984). It has been extensively used in studying the interrelationships between different cortical regions in relation to sensory stimulation, voluntary movements, effect of drugs and a wide range of clinical and cognitive problems and tasks.

However, since the development of FFT, correlation has been replaced by an alternative mathematical method, the coherence spectrum. The latter, in addition to the correlation information, has the advantage of showing the co-variation between two signals as a function of frequency, thus allowing the study of spatial correlations between different bands (Gevins, 1987; Guevara et al., 1995). Coherence studies have been successfully conducted in many fields, including those dealing with cognitive functions and psychiatric disorders (French and Beaumont, 1984). To mention a few key articles that reviewed the applications of coherence to neural data the reader is referenced to (Dumermuth and Molinari, 1991; French and Beaumont, 1984; Shaw, 1984; Thatcher et al., 1986; Zaveri et al., 1999). Especially in epileptic seizures (Sakkalis et al., 2008a) when an abnormal synchronization of neurons takes place, coherence appears to be an ideal tool for measuring it (Sakkalis et al., 2009).

Both methods, correlation and coherence, may be considered as equivalent in that they evaluate the degree of similarity between two signals. However, there are important differences between them. Coherence is calculated by dividing the numerical square of the cross-spectrum by the product of the autospectra (eq. 3). Since, it is a complex measure; it is sensitive to both a change in power and a change in phase relationships. Consequently, if either power or phase changes in one of the signals, the coherence value is affected. Another important difference is that the value of coherence for a single epoch is always one, regardless of the true phase relationship and the differences in power between the two signals (Bendat and Piersol, 1993).

Therefore, the sample cross spectrum S_{xy} between two time series is defined by calculating the product of the Fourier transformed series as follows:

$$S_{xy}(f_k) = X(f_k) \cdot Y^*(f_k) \tag{1}$$

The latter measures the linear cross correlation between the two signals for any given frequencies. In general, the correlation function requires normalization to produce an accurate estimate and is given by:

$$C_{xy}(\tau) = \frac{1}{N-\tau} \sum_{k=1}^{N-\tau} x(k+\tau) y(k) \tag{2}$$

where N is the total number of samples and τ the time lag between the signals. Taking the square of the cross spectrum normalized by the auto spectra of each signal defines the magnitude squared coherence or simply the coherence Γ^2_{xy} as follows:

$$\Gamma^2_{xy}(f_k) = \frac{\left|\left\langle S_{xy}(f_k)\right\rangle\right|^2}{\left|\left\langle S_{xx}(f_k)\right\rangle\right|\left|\left\langle S_{yy}(f_k)\right\rangle\right|} \tag{3}$$

where $\langle\cdot\rangle$ indicates average over a number of equal in length signal segments or epochs, the so-called averaged modified periodogram typically using Welch's method.

Over successive epochs the coherence measure is dependent on power and phase of the two signals along the epochs. If there is no variation over time in the original relationship between the two signals, the coherence value remains unity. This means that coherence does not give direct information on the true relationship between the two signals, but only on the stability of this relationship with respect to power asymmetry and phase relationship. Correlation, on the other hand, may be calculated over a single epoch or over several epochs and it is sensitive to both, phase and polarity, independently of amplitudes. The calculation of coherence involves squaring the signal, thus producing values which go from 0 (linearly independent signals for a given frequency) to 1 (maximum linear correlation in given frequency), and a loss in polarity information. By contrast, correlation is sensitive to polarity and its values go from – 1 (complete linear inverse correlation) to +1 (complete linear direct correlation), with $C_{xy}(\tau) = 0$ suggesting lack of linear interdependence for a given time lag τ. However, under normal physiological conditions, no strong and abrupt power asymmetries would be expected to occur. Thus, the influence of power on coherence should be negligible and results similar to those produced by correlation also would be expected for the coherence measures.

Additionally one is able to define the phase function of the coherence as:

$$\Phi_{xy}(f_k) = \arctan\frac{-\Im S_{xy}(f_k)}{\Re S_{xy}(f_k)} \tag{4}$$

where $\Im$ and $\Re$ denote the imaginary and real part. Phase measure is meaningful only in the case of significant coherence, when one is interested in estimating the time delay between the two signals (Brazier, 1972) according to:

$$\tau_{xy}(f_k) = \frac{\Phi_{xy}}{2\pi f_k} \tag{5}$$

where Φ_{xy} is the phase difference in degrees at a frequency f_k.

Important concerns and limitations

It should be noted that EEG represents a set of continuous voltage/time values and may therefore be considered as a multivariate time series that belongs to the category of stochastic processes which may be described by probability distributions. The concept of temporal stochastic process may be extended, in the case of the EEG, to the spatial domain (Nunez, 1995). The stochastic nature (temporal and spatial) of the EEG allows the use of correlation and coherence analyses which presuppose independence between successive points in time. However, a trade-off has to be made regarding the length of the data segment for analysis, which on one hand must be short enough to satisfy the condition of

stationarity (required from Fourier Transform) and on the other hand must be long enough to provide good frequency resolution. EEG on the contrary is highly non-stationary.
Another concern is that Fourier transform loses the information concerning the time evolution of the signal, which in certain cases maybe useful.
Finally, another important point to take into account is that coherence is very sensitive to fluctuations of linearity in phase, relatively less so to nonlinear fluctuations of amplitude, and completely insensitive to linear fluctuations in amplitude. This becomes obvious if one rewrites equation 3 in terms of polar coordinates:

$$\Gamma^2_{xy}(f_k) = \frac{\left|\left\langle X(f_k)\cdot Y^*(f_k)\right\rangle\right|^2}{\left|\left\langle X^2(f_k)\right\rangle\right|\left|\left\langle Y^2(f_k)\right\rangle\right|} = \frac{\left|\left\langle |XY|\, e^{i(\arg X - \arg Y)}\right\rangle\right|^2}{\left|\left\langle |X|^2\right\rangle\right|\left|\left\langle |Y|^2\right\rangle\right|} \tag{6}$$

Both amplitude and phase information are evaluated with regard to their contribution to linear dependence. The relative importance of amplitude and phase covariance in this index is not altogether clear (Lachaux et al., 1999; Varela et al., 2001). Therefore, some authors have argued that amplitude should not contribute to the final measure at all. Hence, the relationship between the phases without any influence of the amplitudes is better calculated by alternative methods, i.e. Phase Locking Value (PLV) as described in section 3.1. Just to elucidate on this; PLV is essentially the square root of equation 6 with all the moduli set to 1.

2.2 Short time Fourier transform Coherence

The coherence measure, as described in section 2.1, assumes stationarity of the signals and is completely insensitive to the changes in coupling over time. Thus, a short time Fourier transform (STFT) approach is better suited and generally used to generate auto- and cross spectrograms, which are in turn utilized to produce the so called "coherogram". The coherogram is coherence calculated around a number of time instants. It results in a three dimensional matrix of time and frequency versus coherence.
However, while the assumption of wide sense stationarity is removed, stationarity is still required within each time interval for which coherence is calculated. In practice one should carefully decide on the optimal section length (over which each coherence estimate is measured), window length and overlapping (within each coherence estimate), which affects the resolution of the coupling measure. Hence, the STFT approach requires a priori information about the coupling range in time and frequency, in order to allocate the time-frequency resolution. The latter is constrained by the uncertainty principle: the wider the windows, the better the frequency resolution, at the expense of timing information, and vice versa. Lachaux et al. (2002) presents in detail this approach as applied to single-trial brain signals.

2.3 Wavelet Coherence

Coherence analysis, as noted before, ignores in a sense the temporal structure of the signal, thus cannot convey any information on dynamically varying or short-time dependence between the signals. To overcome this limitation and the implicit assumption that the signal is a process that remains stationary in time, wavelet coherence (WC) has been proposed. In contrast to the short time Fourier transform, which is just an extension of the regular Fourier transform, the wavelet transform is inherently a time frequency signal analysis. Continuous wavelet transform is preferred over the discrete counterpart in this case, as well. Since the

coherence estimator is most strongly influenced by linearity in phase, the wavelet should be complex.

Here again, the conceptually simple wavelet basis used, is the Morlet wavelet. This modulated Gaussian kernel has a simple and very smooth spectrum that allows for easy interpretation of the results achieved. Initially the cross wavelet transform (XWT) of two time series x_n and y_n is defined as:

$$W_n^{XY}(s) = W_n^X(s) W_n^{Y*}(s) \tag{7}$$

where W_n^X(s) and W_n^Y(s) are the WT of signals X and Y, respectively and * denotes complex conjugation. Furthermore, the cross wavelet power may be defined as $\left|W_n^{XY}(s)\right|^2$. The complex argument $\arg(W_n^{XY}(s))$ may be explained as the local relative phase between the two time series in time frequency space.

If one closely follows equation 3, then the wavelet coherence R_n^2 of two signals comes naturally and may be defined as:

$$R_n^2(s) = \frac{\left|S(s^{-1}W_n^{XY}(s))\right|^2}{S\left(s^{-1}\left|W_n^X(s)\right|^2\right) \cdot S\left(s^{-1}\left|W_n^Y(s)\right|^2\right)} \tag{8}$$

where S is a smoothing operator in time S_t (The time smoothing uses a filter given by the absolute value of the wavelet function at each scale, normalized to have a total weight of unity. For the Morlet wavelet this is just a Gaussian ($e^{-t^2/2s^2}$) and scale S_s such as $S(W) = S_s(S_t(W_n(s)))$ which for the Morlet wavelet is given by a Gaussian and a boxcar filter of width equal to 0.6, (the scale-decorrelation length) respectively (Grinsted et al., 2004; Torrence and Compo, 1998):

$$S_t(W,s) = \left(W_n(s) * c_1^{-t^2/2s^2}\right) \tag{9}$$

$$S_s(W,n) = \left(W_n(s) * c_2 \prod(0.6s)\right) \tag{10}$$

where c_1 and c_2 are normalization constants and Π is the rectangle function. The squared WC time-frequency transformed scalogram is depicted in figure 1.

Note that the Schwarz inequality (given using the notation of norm, as explained under norms on inner product spaces, as $\left|\langle x, y\rangle\right| \le \|x\|\|y\|$) forces the wavelet coherence to take on a value between 0 and 1. Moreover, the wavelet coherence is also a function of both time and (scale) frequency, measuring the coupling across both variables. The fact that the wavelet transform uses a shorter window for higher frequencies, and a longer one for lower frequencies makes this approach more suited to quantifying time varying coherence. This is accomplished via a frequency-adaptive tiling of the time frequency plane. In contrast, the constant size windows, and summation segments of the STFT coherence, force it to have the same resolution over all frequencies.

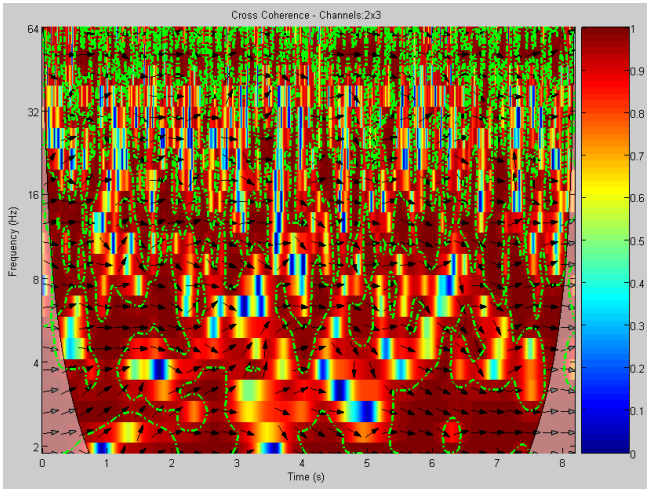


Fig. 1. The squared WC time-frequency transformed scalogram. The 5% significant regions over the time-scale transform are indicated by the contours (green dashed outline). The outer elliptical region at the edges of the second graph indicates the cone of influence in which errors (edge effects) may be apparent due to the transformation of a finite-length series EEG signal (Torrence and Compo, 1998). The relative phase relationship is shown as arrows (with in-phase pointing right, anti-phase pointing left).

In summary, the XWT identifies regions with high common power and can possibly convey information about the phase relationship. For example, if two series are physically related we would expect a consistent or slowly varying phase lag. On the other hand, WTC can be thought of as the local correlation between two CWTs. In this way, locally phase locked behavior is uncovered. The more desirable features of the WTC come at the price of being slightly less localized in time frequency space.

2.3.1 Significant Wavelet Coherence

The confidence levels for the cross wavelet spectrum can be derived from the square root of the product of two chi-square distributions (Jenkins and Watts, 1968; Torrence and Compo, 1998). If the two series have theoretical Fourier spectra B_s^X and B_s^Y, then the cross wavelet distribution is:

$$\frac{W_n^X(s)W_n^{Y*}(s)}{\sigma_X\sigma_Y} \Rightarrow \frac{Z_\nu(p)}{\nu}\sqrt{B_s^X B_s^Y} \tag{11}$$

where σ_X and σ_Y are the respective standard deviations and ν equals to 1 ($Z_1(95\%)=2.182$) or 2 ($Z_2(95\%)=3.999$) for real and complex wavelets, respectively.

In order to apply these ideas on real EEG signals one may define population specific background spectra for each channel pair (X, Y) and scale index s. Here again the background spectra or control-task spectra are defined as the mean time-averaged wavelet power spectrum for each channel and scale ($\overline{W}^X(s)$ and $\overline{W}^Y(s)$), which in the case of

population based analyses could be the average power spectrum over all subjects performing the control task (Sakkalis et al., 2006b). Hence, by calculating both background spectrums ($\overline{W}^X(s)$ and $\overline{W}^Y(s)$) for a pair of channels and setting $Z_v(p)=Z_2(p)=3.999$, since Morlet is a complex mother wavelet, one is able to define the background spectrum using equation 11.

However, in order to gain confidence in the interdependencies of the coherence findings one should estimate the statistical level of significance against a background spectrum using Monte Carlo methods (Grinsted et al., 2004). To adapt this idea to EEG signals one can generate a large (order of 1000) ensemble of mean time-averaged wavelet power spectrum for each channel and scale ($\overline{W}^X(s)$ and $\overline{W}^Y(s)$) using the bootstrap resampling procedure (You actually randomize the time signatures of each coherence measure). Then, for each pair of channels and each scale the wavelet coherence is calculated. Finally, the probability distribution of the calculated coherence values is used to define the 95% confidence level. After that one is able to indicate significant regions over the time-scale transform by contours (depicted in green dashed lines in Fig. 1).

Having derived this significant information, we are now able to form a single measure per scale which reflects the Significant Coherence. We are able to obtain the coherence values over those time- and band-localized regions where significant coherence is indicated by taking the coherence averages over certain bands and significant time points (contours in Fig. 1). An interesting study that successfully utilizes the latter approach in extracting the variability of neural interconnections in schizophrenia patients as compared to healthy ones is underlined in (Sakkalis et al., 2006a).

2.3.2 Partial Directed Coherence (PDC)

Extending interdependence methods to causal interaction relationship one should initially refer to the concept of Granger-causality (Granger, 1969), which is based on the commonsense idea that causes precede their effects in time and is formulated in terms of predictability. In a linear framework, Granger-causality is commonly evaluated by fitting Vector Autoregressive Models. Suppose that a set of N simultaneously observed time series $\mathbf{x}(t) = [x_1(t), \ldots, x_N(t)]^T$ is adequately represented by a Vector Autoregressive Model of order p (MVAR(p)):

$$\mathbf{x}(t) = \sum_{k=1}^{p} A_k \mathbf{x}(t-k) + \mathbf{w}(t) \tag{12}$$

where $\mathbf{A}_k = \begin{bmatrix} a_{11}(k) & \cdots & a_{1N}(k) \\ \vdots & \ddots & \vdots \\ a_{N1}(k) & \cdots & a_{NN}(k) \end{bmatrix}$ is the coefficient matrix at time lag k, and $\mathbf{w}(t) = [w_1(t), \ldots, w_N(t)]^T$ is the vector of model innovations having zero mean and covariance matrix Σ_w. The autoregressive coefficients $a_{ij}(k)$, $i,j=1,\ldots, N$ represent the linear interaction effect of $x_j(t-k)$ onto $x_i(t)$. In order to provide a frequency domain description of Granger-causality, Baccala and Sameshima (Baccala and Sameshima, 2001) introduced the concept of

Partial Directed Coherence (PDC), which has recently been generalized to the new PDC (Baccala et al., 2007), as follows:

Let
$$A(\lambda) = \sum_{k=1}^{p} A_k e^{-i2\pi\lambda k} \tag{13}$$

be the Fourier transform of the coefficient matrices, where λ is the normalized frequency in the interval [-0.5,0.5] and $i = \sqrt{-1}$. Then the new PDC is defined (Baccala et al., 2007) as:

$$\left|\pi_{i\leftarrow j}(\lambda)\right| = \frac{\frac{1}{\sigma_i}\left|\overline{A}_{ij}(\lambda)\right|}{\sqrt{\sum_{m=1}^{n}\frac{1}{\sigma_m^2}\overline{A}_{mj}(\lambda) * \overline{A}_{mj}^{H}(\lambda)}} \tag{14}$$

where $\overline{A}(\lambda) = I - A(\lambda)$ and σ_i^2 refers to the variance of the innovation processes $w_i(t)$. $\left|\pi_{i\leftarrow j}(\lambda)\right|$ ranges between 0 (indicating independence) and 1 (indicating maximum coherence). The latter is a linear method able to assess not only the independence of the brain regions, but also the direction of the statistically significant relationships. This method along with nonlinear ones (will be presented next) was successfully applied in studying the brain activity based on its independent components instead of the EEG signal itself. Both linear and nonlinear synchronization measures are applied to EEG components, which are free of volume conduction effects and background noise (Sakkalis et al., 2008b). More specifically, partial directed coherence was investigated in a working memory paradigm, during mental rehearsal of pictures.

3. Nonlinear synchronization techniques applied in EEG analysis

In the early 1980s, the notion of synchronization was extended to the case of interacting chaotic oscillators (Afraimovich et al., 1986; Fujisaka and Yamada, 1983; Pecora and Carroll, 1990; Pikovsky, 1984). A completely different approach in analyzing the nonlinear dynamics of an EEG signal started some decades after the discovery of deterministic chaos (Lorenz, 1963). According to Pikovsky et al. (2001) synchronization may be understood as an "adjustment of rhythms of oscillating objects due to their weak interaction". An interaction can be realized for instance through a unidirectional or bidirectional coupling. The latter case resembles mutual synchronization; both systems adjust their rhythms to each other, whereas the former case refers to synchronization initiated by an external force, called a driver, and the driven system is called a response. The rhythm of the response is adjusted to the rhythms of the driver. As soon as the interaction strength gets strong (very large values of coupling) one cannot speak of two interacting systems but rather of one combined system.

3.1 Phase Synchronization

Cognitive acts require the integration and constant interaction of widely distributed neuronal areas over the brain (Friston et al., 1997; Tononi and Edelman, 1998). Especially, gamma band is believed to reveal such large-scale oscillations that enter into precise phase-

locking over a limited period of time, often referred as phase-synchrony phenomena (Lachaux et al., 1999). Another example is thought to be the genesis of epileptic phenomenon, where synchrony has long been considered as an important factor (Niedermeyer and Lopes da Silva, 1999). Hence, some methods capable to obtain a statistical measure of the strength of phase synchronization in different areas of the brain have generated a lot of interest (Le Van Quyen et al., 2005; Mormann et al., 2000; Quian Quiroga et al., 2002).

One of the first experimental observations of synchronization was reported by Christiaan Huygens (1673) when two pendulum clocks hanging from the same beam got synchronized by attaining their maximal amplitudes at the same time but at opposite extremes thanks to the weak coupling provided by the vibration of the beam in response to their movement. However, these pendulum clocks were actually harmonic linear oscillators. But, it is well known at present that even the phases of two coupled nonlinear (noisy or chaotic) oscillators may also synchronize even if their amplitudes remain uncorrelated (Pikovsky et al., 2001; Rosenblum et al., 1996). The word "synchronization" arises from the Greek word "σύνγχρονος", which consists of two parts: συν- (syn-=common) and χρόνος (chronos=time). Hence, this word is capable of describing any phenomena that "happen at the same time".

The calculation of phase synchronization (PS) and specifically instantaneous phase calculation can be based on both the Hilbert (HT) (Le Van Quyen et al., 2005; Quian Quiroga et al., 2002) or wavelet transform (WT) (Le Van Quyen et al., 2005). Actually both approaches are essentially equivalent (Quian Quiroga et al., 2002). The main difference is that the HT needs band-pass pre-filtering (Angelini et al., 2004; Bhattacharya and Petsche, 2001; Koskinen et al., 2001), whereas WT do not, since filtering is intrinsic in the WT. The most commonly used approach among the two utilizes the HT. Hence, HT is involved in getting the analytical signal $A(t)$:

$$A(t) = x(t) + i\tilde{x}(t) \tag{15}$$

where $\tilde{x}(t)$ is the HT of the signal $x(t)$, defined as:

$$\tilde{x}(t) = \frac{1}{\pi} P\int_{-\infty}^{\infty} \frac{x(\tau)}{t-\tau} d\tau \tag{16}$$

with P denoting the Cauchy principal value. The instantaneous phase is the "unfolded" angle of the analytic signal, which is given by:

$$\phi(t) = \arctan \frac{\tilde{x}(t)}{x(t)} \tag{17}$$

Synchronization of noisy signals is marked by an appearance of horizontal plateaus in the phase difference across time. The phase synchronization is defined as the locking of the phases of two oscillating systems x and y at any time t:

$$\phi_{n,m}(t) = \left| n\phi_x(t) - m\phi_y(t) \right| \leq const \tag{18}$$

where $\phi_x(t)$ and $\phi_y(t)$ are the unwrapped phases of the signals associated to each system. In this case, as we aim to measure the synchronization between signals from within the same physiological system (i.e., the brain), we assume the phase locking ratio of n:m=1:1 as per Mormann et al. (2000).

Moreover, to quantify the strength of phase synchronization, we need to choose among a number of phase indices that have been used for intracranial and scalp EEG, such as Shannon entropy (Quian Quiroga et al., 2002) and phase locking value (PLV) (Lachaux et al., 1999; Le Van Quyen et al., 2005; Mormann et al., 2000; Quian Quiroga et al., 2002). The latter, which is also termed as mean phase coherence or synchrony factor (Tallon-Baudry et al., 2001), has been found to be more robust than the alternatives especially when applied to a low number of data samples (Quian Quiroga et al., 2002). Thus, it is an optimal choice when used for scalp EEG (Garcia Dominguez et al., 2005); thus was opted in our applications (Sakkalis et al., 2009).

$$PLV = \left| \left\langle e^{i\phi(t)} \right\rangle \right| = \left| \frac{1}{N} \sum_{j=0}^{N-1} e^{i[\phi_x(j\Delta t) - \phi_y(j\Delta t)]} \right| \tag{19}$$

where $\langle \cdot \rangle$ denotes the average over time and N is the time series. This can be expanded according to Euler's formula as follows:

$$PLV = \sqrt{\left(\frac{1}{N} \sum_{j=0}^{N-1} \sin[\phi_x(j\Delta t) - \phi_y(j\Delta t)] \right)^2 + \left(\frac{1}{N} \sum_{j=0}^{N-1} \cos[\phi_x(j\Delta t) - \phi_y(j\Delta t)] \right)^2} \tag{20}$$

Hence, PLV is a normalized value which in simpler words measures how the relative phase is distributed over the unit circle. In case of perfect synchronization the relative phase will occupy a small portion of the circle and PLV will be close to a value of 1, whereas when PLV has a value close to 0, relative phases spread out over the entire unit circle and the mean phase coherence is low.

Important notes and applications

Many studies have attempted to study synchronization based on the traditionally formulated coherence. However, PLV is more suited mainly because of the following reasons (coherence limitations):

- *Stationarity*: Coherence may be applied only to stationary signals. In EEG the assumption of stationarity is rarely met or validated. PLV on the other hand does not require any stationarity.
- *Strictly phase specific measurements*: Coherence also increases with amplitude covariance and the relative importance of amplitude and phase covariance in the coherence value is not clearly justified. However, since phase-locking on its own is adequate to indicate brain lobe interactions, PLV is superior because it is only based on the phase and does not consider the amplitude of the signals.

A number of key works in the field demonstrate the applicability of phase synchronization method especially in combination with the gamma frequency band. Synchronization in the gamma band of the EEG is thought to reflect the appearing of an integrative mechanism bringing together widely distributed sets of neurons to effectively carry out different cognitive tasks (Bendat and Piersol, 1993; Tallon-Baudry et al., 2001; Varela et al., 2001). Rodriguez et al. (1999) found increased PS with a latency of 260 ms after the stimulus in the frequency range between 35 and 45 Hz in a group of adult human subjects during visual perception of faces, as opposed to the no-perception situation. Phase synchrony has been found also during non-visual perception but in a different frequency band (Trujillo et al., 2005). Another interesting result reported by Rodriguez et al. (1999) is the existence of a

period of strong desynchronization with latency between 400 and 650 ms after the stimulus, which reflects the active uncoupling of the neural ensembles necessary to proceed from the visual perception cognitive state to the motor activation state (Rodriguez et al., 1999; Varela et al., 2001). Schnitzler and Gross (2005) have reviewed the latest findings on the concept of long-range neuronal synchronization and desynchronization in motor control and cognition, in normal as well as pathological conditions.
Phase synchrony and gamma desynchronization reflects the successful formation of declarative memory, as demonstrated by the analysis of the relationship between human EEGs from the hippocampus and the rhinal cortex (Fell et al., 2001). Fell et al. (2003) concluded that, whereas rhinal-hippocampal gamma EEG PS may be closely related to actual memory processes, by enabling fast coupling and decoupling of the two structures, theta coherence might be associated with slowly modulated coupling related to an encoding state.

3.2 Generalized Synchronization

Chaotic systems appear to have an apparently noisy behavior, which in fact is ruled by deterministic laws. In a purely deterministic system once its present state is fixed, the states of all future times are determined as well. Deterministic chaos is characterized by sensitivity to initial conditions, which means that trajectories starting from very close points may give results that diverge exponentially after some time. Non linear neural time series analysis was motivated by the fact that many crucial neural processes enclose nonlinear characteristics, i.e. the regulation of voltage-gated ion channels corresponds to a steep nonlinear step-function relating membrane potential to current flow. Human EEG data was reported to exhibit chaotic dynamics in waking states (Accardo et al., 1997; Mayer-Kress and Layne, 1987; Pritchard and Duke, 1992) and in the alpha rhythm (Gallez and Babloyantz, 1991; Soong and Stuart, 1989). In addition, chaotic descriptors like the fractal dimension of the EEG were found to change through the different sleep stages (Fell et al., 1993) and during the performance of various cognitive tasks (Gregson et al., 1990; Lutzenberger et al., 1992). However, even if neurons are highly nonlinear devices, further studies did not find any strong evidence of chaos in EEG (Theiler and Rapp, 1996). Hence, at present there is a wide consensus that EEG signal are not (low-dimensional) chaotic (Lehnertz et al., 2000). In spite of this, nonlinear measures in EEG are still used since there is evidence that even if no chaotic behavior is mathematically evident, the use of phase space representations (see next paragraph) of the signals may reveal nonlinear structures hidden to standard linear approaches (Stam, 2005).
In a deterministic system it is important to establish a vector space, known as state space or phase space for the system such that specifying a point in this space specifies the state of the system, and vice versa. Even for nondeterministic systems the state concept is also powerful in the sense that a system maybe described by a set of states (possibly infinite) and a set of transition rules, which specify how a system proceeds from one state to the next.
After stressing the importance of phase space for the study of the dynamic nature of a deterministic system, there is a need to link this concept with experiments where one acquires time series and not phase space portraits. We therefore need to convert the time series samples observations into state vectors. This problem of state space reconstruction is solved by the method of delays (Takens, 1980).

If we consider the time series to be xn then a delay reconstruction vector in m dimensions is formed by the vectors:

$$\vec{x}_n = (x_n, x_{n-\tau}, \ldots, x_{n-(m-2)\tau}, x_{n-(m-1)\tau}) \tag{21}$$

where n=1,…,N (N equals the signal length), m is the embedding dimension and τ denotes the delay time or lag. Note that for $\tau > 1$, only the time window covered by each vector is increased, while the number of vectors constructed from the scalar time series remains roughly the same. This is because we create a vector for every scalar observation, with $n>(m-1)\tau$. Finding a good embedding is always a matter of great concern and is related to the proper selection of the embedding dimension m and the time lag τ. The right parameter selection depends on the underlying dynamics in the data and on the kind of the analysis intended. Generally, these parameters are optimized by either using statistics for their determination (Cao, 1997), or by starting the intended analysis right away and further increasing the values of m and τ until the results are optimized. A precise knowledge of m is beneficial in order to exploit determinism with minimal computational cost and no redundancy. On the other hand a good estimate of the time lag τ is more difficult to obtain. If τ is too small compared to the internal time scales of the system under consideration, successive elements of the delay vectors are strongly correlated. On the contrary, if τ is very large, successive elements are already almost independent. However, the first zero of the autocorrelation function of the signal often yields a good trade-off between the former extreme cases.

The simplest form of synchronization occurs if the states of systems exactly coincide in time. This type of synchronization is usually referred to as identical synchronization. It can be observed if the coupling strength between identical systems is high enough (Pikovsky, 1984). Two systems X and Y with state vectors $\vec{x}(t)$ and $\vec{y}(t)$ are identically synchronized ($X \Leftrightarrow Y$) if:

$$\lim_{t\to\infty}[\vec{x}(t) - \vec{y}(t)] = 0 \tag{22}$$

In order to achieve identical synchronization the parameters of the coupled systems must be identical. Otherwise, even if there is a slight mismatch the states of the systems may come close but still remain different. Obviously, identical synchronization is defined in a conceptual basis. In practice, even if the systems are not identical, weaker synchronization is possible when one system (response) is driven by the other (driver). This phenomenon is characterized as generalized synchronization (Pecora and Carroll, 1990; Rulkov et al., 1995) and is met if the state of the response $\vec{y}$ is completely defined by the state of the driver $\vec{x}$ ($X \to Y$). Hence, there exists a transformation function G such that:

$$\vec{y} = G(\vec{x}) \tag{23}$$

Strong or weak synchronization is realized in the case of smooth or non-smooth transformation function (Hunt et al., 1997; Pyragas, 1996). Rulkov et al. (1995) defined this function to be continuous one-to-one, meaning that not only the response orbit may be predicted by the driver's orbit, but also the neighborhoods in the driver space are mapped into the neighborhoods of the response space. The latter property is extensively utilized for experimental purposes.

Important concerns

Phase synchronization and generalized synchronization are not clearly related. Initially, the phase synchronization was found to appear before the generalized synchronization pops in and as a result it was thought that generalized synchronization implies phase synchronization. However, later studies on several examples had shown the reverse order (Zheng and Hu, 2000). Another thought was to apply generalized synchronization methodologies to the phases of the systems rather than the systems themselves (Lee et al., 2003). The latter is often described as generalized phase synchronization.

A challenging application for measures of synchronization is the study of neuronal dynamics, since synchronization phenomena have been increasingly recognized as a key feature for establishing the communication between different regions of the brain (Fell et al., 2001; Varela et al., 2001). Pathological cases such as Parkinson's disease or epilepsy have also taken advantage of synchronization methods in revealing synchronization and desynchronization events. Especially, when epilepsy patients undergo pre-surgical diagnostics (Lopes da Silva, 1999), intracranial recordings are acquired in order to provide sufficient localization information for the epileptic focus in the brain. This motivated us to carry out a comprehensive comparison of different measures of synchronization, as presented in Sakkalis et al. (2009).

In the experimental setting the instantaneous states of a neurophysiological process and system are not accessible. However, considering that scalp EEG is able to provide us a set of m channels equation 23 may be extended as:

$$y_j(t) = G(\vec{x}_s(t)) + \varepsilon_j(t), \quad j = 1, 2, \ldots, m \tag{24}$$

where $\varepsilon_j(t)$ denotes extraneous noise. In the simplest case, the transformation function G may be a linear superposition of a subset (denoted as s) of the $\vec{x}_s$ (such as the oscillating membrane potentials), or may involve a more complex transformation. However, the primary aim of interdependence analysis is not to recover the underlying dynamics, but only to determine whether there is nonlinear interdependence between the observables y_j due to nonlinear interdependence between the local subsystems $\vec{x}_s$.

In brief, synchronization measures quantify how well the phase space trajectory can be predicted, knowing the trajectory of the other. Clearly, such an attempt makes sometimes easier or more difficult to predict the own dynamics of each system alone. Schiff et al. (1996) initially developed the first algorithm that exploited such considerations and demonstrated nonlinear interdependence in a spinal cord preparation. The idea was to first locate the k-nearest neighbors of each point along an orbit in one phase space and then project these points into the other state space. Next, the center of the resulting cloud of projected points is used to predict the state of that system. Finally, the Euclidean distance between the predicted and observed vectors yields an error, which is then normalized by reference to a randomly chosen vector (the predicted error is, on average, no closer than random to the observed vector). The minimum error value of 0 or the maximum of 1 are taken when the systems experience maximum interdependence or when the systems are completely independent, respectively. Variants of this idea have been proposed in order to improve predictions (Feldmann and Bhattacharya, 2004; Terry and Breakspear, 2003).

Alternatively, another set of in principle more robust measures have been proposed, where instead of looking for predictions one quantifies how neighborhoods in one attractor maps into the other. In the following section this kind of algorithms will be discussed thoroughly.

Arnhold et al. (1999) similarly use k-nearest neighbors, but calculate the average radius of the cloudlike region formed by these neighbors. The ratio of the cloudlike radius to that of the cloud in the other system's reconstructed phase space defines a measure of system interdependence. This idea turned out to be a robust and reliable way of assessing the extent of GS (Sakkalis et al., 2009; Schmitz, 2000).
First, we reconstruct delay vectors (Takens, 1980) out of our time series;

$$x_n=(x_n,\ldots,x_{n-(m-1)\tau}) \text{ and } y_n=(y_n,\ldots,y_{n-(m-1)\tau}) \tag{25}$$

where n=1…N, and m, τ are the embedding dimension and time lag, respectively. Let $r_{n,j}$ and $s_{n,j}$, j=1,…,k, denote the time indices of the k nearest neighbors of x_n and y_n, respectively. For each x_n the mean squared Euclidean distance to its k neighbours (blue arrows, Fig. 2) is defined as:

$$R_n^{(k)}(X)=\frac{1}{k}\sum_{j=1}^{k}(\mathbf{x}_n-\mathbf{x}_{r_{n,j}})^2 \tag{26}$$

and the Y-conditioned squared mean Euclidean distance $R_n^{(k)}(X|Y)$ is defined by replacing the nearest neighbors by the equal time partners of the closest neighbors of y_n (red arrows, Fig. 2):

$$R_n^{(k)}(X|Y)=\frac{1}{k}\sum_{j=1}^{k}(\mathbf{x}_n-\mathbf{x}_{s_{n,j}})^2 \tag{27}$$

If the set of reconstructed vectors (point cloud x_n) has an average squared radius $R(X)=(1/N)\sum_{n=1}^{N}R_n^{(N-1)}(X)$, then $R_n^{(k)}(X|Y)\approx R_n^{(k)}(X)<<R(X)$ if the systems are strongly correlated, while $R_n^{(k)}(X|Y)\approx R(X)>>R_n^{(k)}(X)$ if they are independent. Hence, an interdependence measure $S^{(k)}(X|Y)$ is defined as (Arnhold et al., 1999):

$$S^{(k)}(X|Y)=\frac{1}{N}\sum_{n=1}^{N}\frac{R_n^{(k)}(X)}{R_n^{(k)}(X|Y)} \tag{28}$$

Since $R_n^{(k)}(X|Y)\geq R_n^{(k)}(X)$ by construction, it is clear that S ranges between 0 (indicating independence) and 1 (indicating maximum synchronization).
Following equation 28 another nonlinear interdependence measure $H^{(k)}(X|Y)$ is possible to be defined as (Arnhold et al., 1999; Quian Quiroga et al., 2000):

$$H^{(k)}(X|Y)=\frac{1}{N}\sum_{n=1}^{N}\log\frac{R_n(X)}{R_n^{(k)}(X|Y)} \tag{29}$$

This is zero if X and Y are completely independent, while it is positive if nearness in Y implies also nearness in X for equal time partners. It would be negative if close pairs in Y would correspond mainly to distant pairs in X. This measure if applied to coupled chaotic systems was found to be robust against noise (Quian Quiroga et al., 2000), but with the drawback that it is not normalized.

Hence, another normalized and more robust version of S maybe defined as in equation 30 (Quian Quiroga et al., 2002) and is the one actually used in our work (Sakkalis et al., 2009):

$$N^{(k)}(X|Y) = \frac{1}{N}\sum_{n=1}^{N}\frac{R_n(X) - R_n^{(k)}(X|Y)}{R_n(X)} \tag{30}$$

In fact, all three measures described are just different ways of normalizing the point cloud ratio of distances.

The opposite interdependences $S^{(k)}(Y|X)$, $H^{(k)}(Y|X)$ and $N^{(k)}(Y|X)$ are defined in complete analogy but in general are not equal to $S^{(k)}(X|Y)$, $H^{(k)}(X|Y)$ and $N^{(k)}(X|Y)$, respectively. However, this asymmetry as opposed to other symmetrical nonlinear measures such as the phase synchronization may be advantageous, since it can give information about driver response relationships (Arnhold et al., 1999; Quian Quiroga et al., 2000; Schmitz, 2000). A schema representing the idea of this algorithm is represented in figure 2.

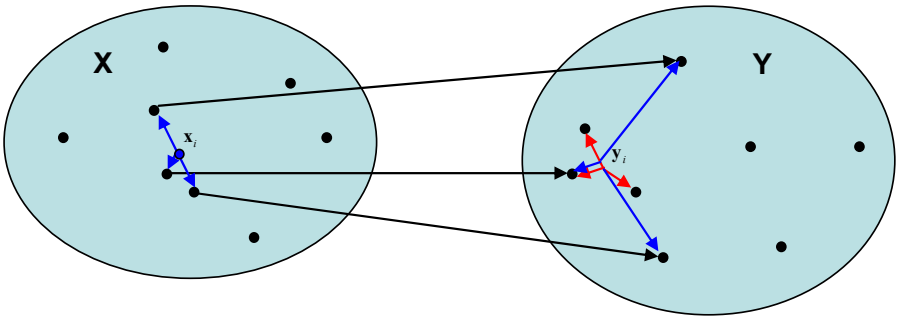


Fig. 2. Scheme representation of the basic idea of the synchronization method described by Arnhold et al. (1999). The blue and red arrows reflect an exemplar case of three nearest neighbors of x_i and y_i, respectively. The black arrows denote the respective positions of the neighbors of x_i in the Y domain.

Stam & van Dijk (2002) proposed a method which, by normalizing to a reference neighborhood, yields a synchronization likelihood (SL) measure that is not biased by the dimension of the systems' phase spaces (which are not necessarily equal). It is closely related to the previous idea and represents a normalized version of mutual information.
Supposing that x_i, x_j and y_i, y_j be the time delay vectors, SL actually expresses the chance that if the distance between x_i and x_j is very small (bounded by the white circular area in Fig. 3), the distance between the corresponding vectors y_i and y_j in the state space will also be very small. For this, we need a small critical distance ε_x, such that when the distance between x_i and x_j is smaller than ε_x, x will be considered to be in the same state at times i and j. ε_x is chosen such that the likelihood of two randomly chosen vectors from x (or y) will be closer than ε_x (or ε_y) equals a small fixed number p_{ref}. p_{ref} is the same for x and y, but ε_x need not be equal to ε_y. Now SL between x and y at time i is defined as follows:

$$SL_i = \frac{1}{N'} \sum_{\substack{j=1 \\ w_1 < |i-j| < w_2}}^{N'} \theta(\varepsilon_{y,i} - |\mathbf{y}_i - \mathbf{y}_j|)\theta(\varepsilon_{x,i} - |\mathbf{x}_i - \mathbf{x}_j|) \quad (31)$$

Here, $N'=2(w_2-w_1-1)P_{ref}$, $|\cdot|$ is the Euclidean distance and θ is the Heaviside step function, $\theta(x)=0$ if $x\leq 0$ and $\theta(x)=1$ otherwise. w_1 is the Theiler correction for autocorrelation effects and w_2 is a window that sharpens the time resolution of the synchronization measure and is chosen such that $w_1 << w_2 << N$ (Theiler, 1986). When no synchronization exists between x and y, SL_i will be equal to the likelihood that random vectors y_i and y_j are closer than ε_y; thus $SL_i=p_{ref}$. In the case of complete synchronization $SL_i=1$. Intermediate coupling is reflected by $p_{ref} < SL_i < 1$. Finally, SL is defined as the time average of the SL_i values.

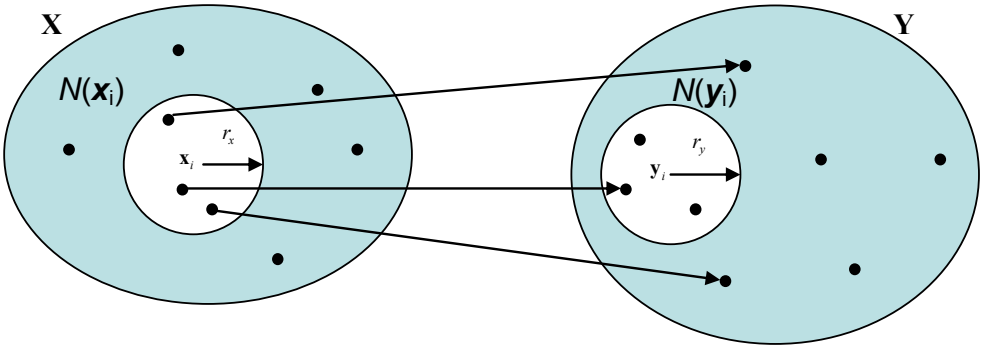


Fig. 3. Scheme representation of the basic idea of the synchronization method described by Stam et al. (2002). SL expresses the chance that if the distance between x_i and x_j (neighboring delay vectors) is less than r_x, the distance (r_y) between the corresponding vectors y_i and y_j in the state space will also be very small.

4. Surrogate time series analysis

So far we have discussed about linear and nonlinear methods for detecting synchronization in bivariate EEG signals. But how can one decide on whether a linear or nonlinear model better describes the data under study? A possible answer lies in the surrogate data testing method. In other words, to demonstrate that the synchronization methods addressed are sensitive in detecting nonlinear structures and thus reliable, surrogate data testing is used. The surrogate data method was introduced about a decade ago and the basic idea is to compute a nonlinear statistic Q for the original data under study, as well as for an ensemble of realizations of a linear stochastic process, which mimics "linear properties" of the studied data the surrogate data (Theiler, Eubank et al. 1992). If the computed nonlinear statistic for the original dataset is significantly different from the values obtained from the surrogate set, one can infer that the data is not generated by a linear process; otherwise the null hypothesis, that a linear model fully explains the data is accepted.

The surrogating procedure preserves both the autocorrelation of the signals and their linear cross-correlation, but the nonlinear individual structure of the individual signals, as well as their nonlinear interdependence, if any, is destroyed. This simply means that an ensemble of

"surrogate data" has the same linear characteristics (power spectrum and coherence) as the experimental data, but is otherwise random.

In practice, a set of p time series (surrogates) is constructed, which share the same characteristics, but lack the property we want to test, the nonlinearity in our case. Using the newly created surrogates the same index $Q_{surrogates}$ is repeatedly calculated leading to p+1 estimations of this. This procedure allows testing of the null hypothesis H_0 that the original value of the statistic belongs to the distribution of the surrogates, hence H_0 is true. In other words, one has to determine whether H_0 can be rejected at the desired level of confidence. By estimating the mean and the standard deviation of the distribution of the statistic from the surrogates and then comparing them with its value from the original signals Z-score is calculated:

$$Z = \frac{\left| Q - \overline{Q}_{surrogates} \right|}{\sigma_{surrogates}} \tag{32}$$

Z-score reveals the number of standard deviations Q is away from the mean Qs of the surrogates. Assuming that Q is approximately normally distributed in the surrogates ensemble, H_0 is rejected at the p<0.05 significance level when Z>1.96 (one-sided test). If, in addition, no other possible causes of such a result can be accounted for, then it is reasonable to conclude that the tested measure accounts for any nonlinear phenomena.

However, it should be noted that, although the above surrogating procedure preserves both the autocorrelation of the signals and their linear cross-correlation, the nonlinear individual structure of the individual signals, if any, is also destroyed. In other words, any nonlinearity not only between but also within the signals is not present in the surrogates. Therefore, these surrogates only test the hypothesis that the data are bivariate stochastic time series with an arbitrary degree of linear auto and cross-correlation (Andrzejak, Kraskov et al. 2003). Nevertheless, if the two signals studied do have any nonlinear structure, it is not possible to ascribe a rejection of the hypothesis that the interdependence is nonlinear due to the nonlinearity of the interdependence, because the nonlinearity of the individual signals may also play a role. Hence, the generation of surrogate data preserving all the individual structure but destroying only the nonlinear part of the interdependence is currently one of the most challenging tasks in the field, and it is a subject of ongoing research (Andrzejak, Kraskov et al. 2003; Dolan 2004).

Pure nonlinear interdependence can contribute to linear correlations, but cannot be detected by linear methods alone. It signifies the formation of macroscopic, dynamic neural cell assemblies and transient low-dimensional interactions between them. Nonlinear interdependence informs that the underlying dynamics are governed by nonlinear processes, or that they are linear but evolving in the vicinity of a non-linear instability and driving noise. Nonlinearities generate correlations that cannot be generated by stochastic processes, such as coupling between oscillations with different frequencies (Friston 1997; Breakspear and Terry 2002).

The most widely used method to obtain surrogate data is to randomize the phases of the signal in the Fourier domain (Theiler, Eubank et al. 1992). Recent advances such as employing iterative loops (Schreiber and Schmitz 1996), simulated annealing (Schreiber 1998) and others (Schreiber and Schmitz 2000) are all aimed to improve the goodness of the fit between the linear properties of the experimental data and surrogate ensemble.

Unforunately, as noted beforehand, no surrogate technique is perfect (Schreiber and Schmitz 2000).

To conclude the whole nonlinearity section it should be stressed that even nonlinear techniques look promising one should be cautious in practice. Many findings may have been premature in that apparent nonlinear effects were in fact caused by limitations of the data such as the sample length (Ruelle 1990). During the previous years there was a general notion that EEG is chaotic, but nowadays there is a wide consensus and it is certainly no longer generally accepted that the healthy EEG is a chaotic signal.

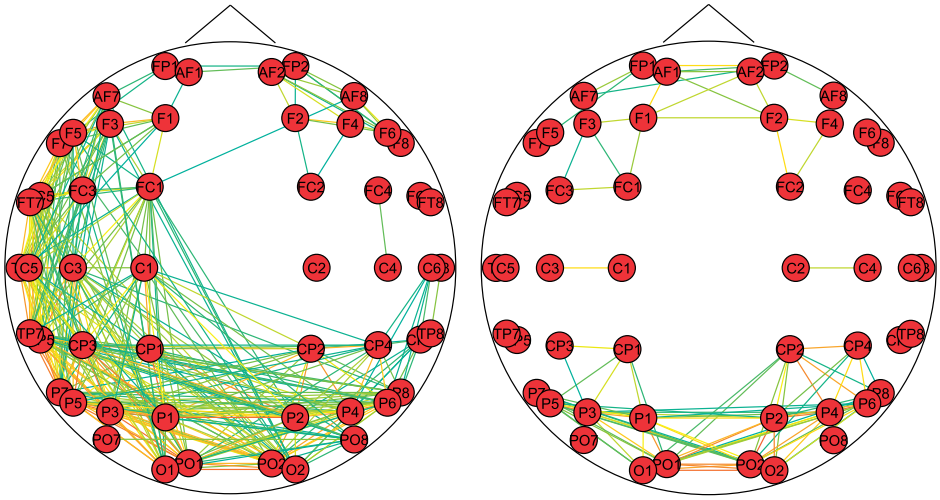


Fig. 4. A "healthy" network (left graph) appears to exhibit strong lateralization compared to the "alcoholic" one (right graph) which exhibits interhemispheric symmetry, when the broadband signals are analyzed.

5. Graph Theory in EEG analysis

An alternative approach to the characterization of complex networks is the use of graph theory (Strogatz 2001; Sporns, Chialvo et al. 2004; Sporns and Zwi 2004). A graph is a basic representation of a network, which is essentially reduced to nodes (vertices) and connections (edges) as illustrated in Fig. 4. Both local and long distance functional connectivity in complex networks may alternatively be evaluated using measures and visualizations derived from graph theory. Special interest in using graph theory to study neural networks has been in focus recently, since it offers a unique perspective of studying local and distributed brain interactions (Varela, Lachaux et al. 2001; Fingelkurts, Fingelkurts et al. 2005).

Using the interdependence methods and measures analyzed in the previous sections one is able to measure (in terms of 0 to 1) the coupling between different channels. If such interdependence measures are constructed for every possible channel pair a coherence matrix (CM) (i.e. 30x30, if 30 channels are used) with elements ranging from 0 to 1. Next, in order to obtain a graph from a CM we need to convert it into an *N*x*N* binary adjacency

matrix, A. To achieve that we define a variable called threshold T, such that $T \in [0,1]$. The value $A(i,j)$ is either 1 or 0, indicating the presence or absence of an edge between nodes i and j, respectively. Namely, $A(i,j)=1$ if $C(i,j) \geq T$, otherwise $A(i,j)=0$. Thus we define a graph for each value of T, i.e., for the purposes of our work, we defined 1000 such graphs, one for every thousandth of T (Sakkalis et al., 2006a). After constructing A, one is able to compute various properties of the resulting graph. These include the average degree K, the clustering coefficient C and the average shortest path length L of our graph, which will be presented in the next section. Figure 4 illustrates an example graph that resembles a "healthy" network (left graph) compared to the "alcoholic" one, in both broadband and lower beta frequency bands (Sakkalis et al., 2007).

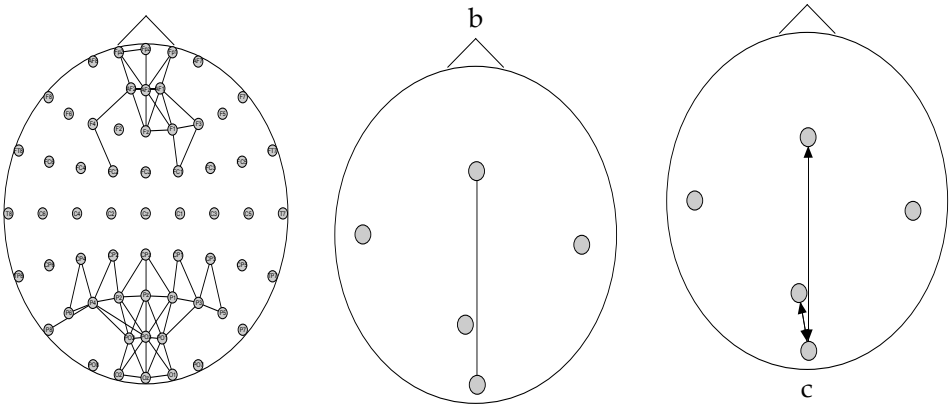


Fig. 5. a) Aerial view of the scalp with the position of electrodes. The depicted average network reflects a local prefrontal and occipitoparietal synchrony, as identified in gamma band using the nonlinear synchronization method on the actual electrode signals in a working memory paradigm. The next parts of this figure (b, c) are considering cross-regional synchrony. b) The nonlinear synchronization method is applied in gamma band ICs reflecting the underlying activity in the different brain regions (prefrontal (upper node), temporal (left and right lateral nodes), parietal (lower middle node) and occipital (lowest central node)). This figure focuses on the inter-region connectivity between the prefrontal and occipital brain areas. c) Similarly to the middle graph but using PDC; again ICs in gamma band exhibit significant linear coupling between the prefrontal and occipital areas, as well as between the occipital and parietal areas. Directionality is also identified. The apparent bidirectional coupling indicates no single influence between the "cause" and "effect" relationship. The illustrated graphs are averaged over all subjects.

Another study (Sakkalis et al., 2008b) was able to identify and visualize the established brain networks in gamma band by means of both linear and nonlinear synchrony measures, in working memory paradigm. The nonlinear GS method was initially applied on all the actual electrode recordings. The scalp map obtained (Fig. 5a) identified a network tendency to localize synchronization activity mostly at frontal and occipitoparietal regions. However, no linking between the two regions is evident. When we focus on the independent components (instead of the actual electrodes themselves), the prominent inter-region connectivity in gamma band between the prefrontal and occipital brain areas becomes evident (Fig. 5b).

Finally, a similar network topology is also derived by the linear PDC method (Fig. 5c). The latter method is able to derive additional information on the "driver and response" significant relationship between observations, denoted by arrows in Fig. 5c. However, the bidirectional arrows denote no single one-way interconnection, but a significant pathway connecting the prefrontal and occipital areas, as well as the occipital and parietal areas, is identified (Fig. 5c).
Graph theory is for sure an emerging field in EEG analysis and coupling visualization. Recent articles illustrate that graph properties maybe of particular value in certain pathologies, i.e., alcoholism (Sakkalis et al., 2007) and Alzheimer disease (Stam, Jones et al. 2006).

6. Conclusion

Throughout this chapter both linear and nonlinear interdependence measures are discussed. Even if the complex nature of EEG signals justify the use of nonlinear methods there is no evidence to support and prejudge that such methods are superior to linear ones. On the contrary, the information provided by nonlinear analysis does not necessarily coincide with that of the linear methods. In fact, both approaches should be regarded as complementary in the sense that they are able to assess different properties of interdependence between the signals. In addition the linear ones most of the times appear to be robust against noise, whereas nonlinear measures are found to be rather unstable. Stationarity is again a main concern, since it is a prerequisite which is not satisfied in practice. The selection of an adequate method will depend on the type of signal to be studied and on the questions expected to be answered. One should also bear in mind that all nonlinear methods presented require stationary signals. If this is not the case, one is better off using a linear alternative like wavelet coherence, due to its inherent adaptive windowing scaling. Another alternative is phase synchronization calculation, PLV method in specific, which requires neither stationarity nor increases with amplitude covariance like coherence. In addition, since phase-locking on its own is adequate to indicate brain lobe interactions, PLV is superior because it is only based on the phase and does not consider the amplitude of the signals. However, an interesting extension in identifying the most significant regions, in terms of increased coherence, as compared to background signals is possible using the significant wavelet coherence.
Visual ways to illustrate the results and possibly fuse them together are the topographic maps and graphs. Topographic colour maps may be used in visualizing the power spectral-based estimations, where different colourings reflect altering brain activity. In addition, interdependencies may be illustrated using graph visualizations, where channel pairwise coupling is visualized using edges of increasing thickness with respect to increasing coupling strength.
As noted throughout this chapter most of the methods presented, traditional linear or nonlinear, must assume some kind of stationarity. Therefore, changes in the dynamics during the measurement period usually constitute an undesired complication of the analysis, which in EEG may represent the most interesting structure in identifying dynamical changes in the state of the brain. Hence, a fundamental topic for further research should be the formation of a powerful test for stationarity able to indicate and reject, with increased certainty, the sections of the EEG raw signal that experience stationary behavior.

Another active research direction focuses on extending current interdependence analysis from bivariate to multivariate signals. This is important since pairwise analysis is likely to find plasmatic correlations in special cases where one driver drives two responses. In this case both responses may found to have a common driver component, even if the responses might be fully independent.

7. References

Accardo A, Affinito M, Carrozzi M, Bouquet F. Use of the fractal dimension for the analysis of EEG time series. Biol. Cybern. 1997; 77: 339-350.

Afraimovich VS, Verichev NN, Rabinovich MI. Stochastic synchronization of oscillations in dissipative systems. Radiophys. Quantum Electron. 1986; 29: 795.

Andrzejak RG, Kraskov A, Stogbauer H, Mormann F, Kreuz T. Bivariate surrogate techniques: necessity, strengths, and caveats. Phys. Rev. E 2003; 68: 066202.

Angelini L, de Tommaso M, Guido M, Hu K, Ivanov P, Marinazzo D, et al. Steady-state visual evoked potentials and phase synchronization in migraine patients. Phys Rev Lett 2004; 93: 038103.

Arnhold J, Lehnertz K, Grassberger P, Elger CE. A robust method for detecting interdependences: Application to intracranially recorded EEG. Physica D 1999; 134: 419.

Baccala L, Sameshima K, Takahashi DY. Generalized partial directed coherence. 15th Intern. Conf. Digital Signal Processing 2007, 163-166.

Baccala LA, Sameshima K. Partial directed coherence: a new concept in neural structure determination. Biological Cybernetics 2001, 84(6): 463-474.

Bhattacharya J, Petsche H. Musicians and the gamma band: a secret affair? Neuroreport 2001; 12: 371-4.

Bendat JS, Piersol AG. Engineering applications of correlation and spectral analysis. New York: J. Wiley, 1993.

Brazier MA. Spread of seizure discharges in epilepsy: anatomical and electrophysiological considerations. Exp Neurol 1972; 36: 263-72.

Brazier MA, Casby JU. Cross-correlation and autocorrelation studies of electroencephalographic potentials. Electroencephalogr Clin Neurophysiol Suppl 1952; 4: 201-11.

Cao L. Practical method for determining the minimum embedding dimension of a scalar time series. Physica D 1997; 110: 43-50.

Dolan K. Surrogate analysis of multichannel data with frequency dependant time lag. Fluct. Noise Lett. 2004; 4: L75-L81.

Dumermuth G, Molinari I. Relationships among signals: cross-spectral analysis of the EEG. In: Weitkunat R, editor. Digital Biosignal Processing. Vol 5. Amsterdam: Elsevier Science Publishers, 1991: 361-398.

Feldmann U, Bhattacharya J. Predictability improvement as an asymmetrical measure of interdependence in bivariate time series. Int. J. of Bifurcation and Chaos 2004; 14: 505-514.

Fell J, Klaver P, Elfadil H, Schaller C, Elger CE, Fernandez G. Rhinal-hippocampal theta coherence during declarative memory formation: interaction with gamma synchronization? Eur J Neurosci 2003; 17: 1082-8.

Fell J, Klaver P, Lehnertz K, Grunwald T, Schaller C, Elger CE, et al. Human memory formation is accompanied by rhinal-hippocampal coupling and decoupling. Nat Neurosci 2001; 4: 1259-64.

Fell J, Roschke J, Beckmann P. Deterministic chaos and the first positive Lyapunov exponent: a nonlinear analysis of the human electroencephalogram during sleep. Biol Cybern 1993; 69: 139-46.

Fingelkurts AA, Fingelkurts AA, Kahkonen S. Functional connectivity in the brain--is it an elusive concept? Neurosci Biobehav Rev 2005; 28: 827-36.

French CC, Beaumont JG. A critical review of EEG coherence studies of hemisphere function. Int J Psychophysiol 1984; 1: 241-54.

Friston KJ, Stephan KM, Frackowiak RSJ. Transient phase-locking and dynamic correlations: Are they the same thing? Human Brain Mapping 1997; 5: 48-57.

Fujisaka H, Yamada T. Stability theory of synchronized motion in coupled dynamical systems. Prog. Theor. Phys. 1983; 69: 32-47.

Gallez D, Babloyantz A. Predictability of human EEG: a dynamical approach. Biol. Cybern. 1991; 64: 381-391.

Garcia Dominguez L, Wennberg RA, Gaetz W, Cheyne D, Snead OCa, Perez Velazquez JL. Enhanced synchrony in epileptiform activity? Local versus distant phase synchronization in generalized seizures. J Neurosci 2005; 25: 8077-8084.

Gevins AS. Overview of computer analysis. In: Gevins AS and Rémond A, editors. Handbook of electroencephalography and clinical neurophysiology ; rev. ser., v. 1. Vol I. NY, USA: Elsevier, 1987: 31-83.

Granger J. Investigating causal relations by econometric models and cross-spectral methods. Econometrica 1969, 37(3): 424-438.

Gregson RA, Britton LA, Campbell EA, Gates GR. Comparisons of the nonlinear dynamics of electroencephalograms under various task loading conditions: a preliminary report. Biol Psychol 1990; 31: 173-91.

Grinsted A, Moore JC, Jevrejeva S. Application of the cross wavelet transform and wavelet coherence to geophysical time series. Nonlinear Processes in Geophysics 2004; 11: 561-566.

Guevara MA, Lorenzo I, Arce C, Ramos J, Corsi-Cabrera M. Inter- and intrahemispheric EEG correlation during sleep and wakefulness. Sleep 1995; 18: 257-65.

Hunt BR, Ott E, Yorke JA. Differentiable generalized synchronization of chaos. Phys. Rev. E 1997; 55: 4029-4034.

Huygens C. Horoloquium Oscilatorium. Paris, 1673.

Jenkins GM, Watts DG. Spectral Analysis and Its Applications. San Francisco, CA: Holden-Day, Inc., 1968.

Koskinen M, Seppanen T, Tuukkanen J, Yli-Hankala A, Jantti V. Propofol anesthesia induces phase synchronization changes in EEG. Clin Neurophysiol 2001; 112: 386-92.

Lachaux JP, Lutz A, Rudrauf D, Cosmelli D, Le Van Quyen M, Martinerie J, et al. Estimating the time-course of coherence between single-trial brain signals: an introduction to wavelet coherence. Neurophysiol Clin 2002; 32: 157-74.

Lachaux JP, Rodriguez E, Martinerie J, Varela FJ. Measuring phase synchrony in brain signals. Hum Brain Mapp 1999; 8: 194-208.

Lehnertz K, Arnhold J, Grassberger P, Elger C. Chaos in Brain? World Scientific. Singapore, 2000.

Le Van Quyen M, Soss J, Navarro V, Robertson R, Chavez M, Baulac M, et al. Preictal state identification by synchronization changes in long-term intracranial EEG recordings. Clin Neurophysiol 2005; 116: 559-68.

Lee D-S, Kye W-H, Rim S, Kwon T-Y, Kim C-M. Generalized phase synchronization in unidirectionally coupled chaotic oscillators. Physical Review E 2003; 67: 045201.

Lopes da Silva FH. EEG Analysis: theory and practice. In: Niedermeyer E and Lopes da Silva FH, editors. Electroencephalography : basic principles, clinical applications, and related fields. Baltimore: Williams & Wilkins, 1999: 1097-1123.

Lorenz EN. Deterministic non-periodic flow. J. Atmos. Sci. 1963; 20: 130.

Lutzenberger W, Birbaumer N, Flor H, Rockstroh B, Elbert T. Dimensional analysis of the human EEG and intelligence. Neurosci Lett 1992; 143: 10-4.

Mayer-Kress G, Layne S. Dimensionality of the human EEG. Annals New York Acad. Sci. 1987; 504: 62-87.

Mormann F, Lehnertz K, David P, Elger CE. Mean phase coherence as a measure for phase synchronization and its application to the EEG of epilepsy patients. Phys. D 2000; 144: 358--369.

Niedermeyer E, Lopes da Silva FH. Electroencephalography : basic principles, clinical applications, and related fields. Baltimore: Williams & Wilkins, 1999.

Nunez PL. Quantitative states of neocortex. In: Nunez PL, editor. Neocortical Dynamics and Human EEG Rhythms. Oxford ; New York: Oxford University Press, 1995: 33-39.

Pecora LM, Carroll TL. Synchronization in chaotic systems. Phys. Rev. Lett. 1990; 64: 821.

Pereda E, Quiroga RQ, Bhattacharya J. Nonlinear multivariate analysis of neurophysiological signals. Prog Neurobiol 2005; 77: 1-37.

Pikovsky A, Rosenblum M, Kurths J. Synchronization : a universal concept in nonlinear sciences. Cambridge: Cambridge University Press, 2001.

Pikovsky AS. On the interaction of strange attractors. Z. Phys. B: Condens Matter 1984; 55(2): 149.

Pritchard W, Duke D. Dimensional analysis of no-task human EEG using the Grassberger-Procaccia method. Psychophysiol. 1992; 29: 182-192.

Pyragas K. Weak and strong synchronization of chaos. Phys. Rev. E 1996; 54: 4508-4511.

Quian Quiroga R, Arnhold J, Grassberger P. Learning driver-response relationships from synchronization patterns. Physical Review E 2000; 61: 5142.

Quian Quiroga R, Kraskov A, Kreuz T, Grassberger P. Performance of different synchronization measures in real data: a case study on electroencephalographic signals. Phys Rev E Stat Nonlin Soft Matter Phys 2002; 65: 041903.

Rosenblum MG, Pikovsky AS, Kurths J. Phase synchronization of chaotic oscillators. Physical Review Letters 1996; 76: 1804-1807.

Ruelle D. Deterministic chaos: The science and the fiction. Proc. of the Royal Society of London 1990; 427A: 241-248.

Rulkov NF, Sushchik MM, Tsimring LS, Abarbanel HDI. Generalized synchronization of chaos in directionally coupled chaotic systems. Phys. Rev. E 1995; 51(2): 980-994.

Sakkalis V, Giurcăneanu CD, Xanthopoulos P, Zervakis M, Tsiaras V, Yang Y, Micheloyannis S. Assessment of linear and nonlinear synchronization measures for analyzing EEG in a mild epileptic paradigm. IEEE Trans. Inf. Tech. 2009; 13(4):433-441 (DOI: 10.1109/TITB.2008.923141).

Sakkalis V, Oikonomou T, Pachou E, Tollis I, Micheloyannis S, Zervakis M. Time-significant Wavelet Coherence for the Evaluation of Schizophrenic Brain Activity using a Graph theory approach. Engineering in Medicine and Biology Society (EMBC 2006). New York, USA, 2006a.

Sakkalis V, Zervakis M, Micheloyannis S. Significant EEG Features Involved in Mathematical Reasoning: Evidence from Wavelet Analysis. Brain Topography 2006b; 19: 53-60.

Sakkalis V, Cassar T, Zervakis M, Camilleri KP, Fabri SG, Bigan C, Karakonstantaki E, Micheloyannis S. Time-Frequency Analysis and Modelling of EEGs for the evaluation of EEG activity in Young Children with controlled epilepsy. Comput Intell Neurosci. CIN 2008a: 462593 (DOI: 10.1155/2008/462593).

Sakkalis V, Tsiaras V, Michalopoulos K, Zervakis M. Assessment of neural dynamic coupling and causal interactions between independent EEG components from cognitive tasks using linear and nonlinear methods. 30th IEEE-EMBS, Engineering in Medicine and Biology Society (EMBC 2008), Vancouver, Canada, August 20-24. 2008b.

Sakkalis V, Tsiaras V, Zervakis M, Tollis I. Optimal brain network synchrony visualization: Application in an alcoholism paradigm. 29th IEEE-EMBS, Engineering in Medicine and Biology Society (EMBC 2007), Lyon, France, August 23-26, 2007.

Schiff SJ, So P, Chang T, Burke RE, Sauer T. Detecting dynamical interdependence and generalized synchrony through mutual prediction in a neural ensemble. Physical Review E 1996; 54: 6708.

Schmitz A. Measuring statistical dependence and coupling of subsystems. Physical Review E 2000; 62: 7508.

Schnitzler A, Gross J. Normal and pathological oscillatory communication in the brain. Nat Rev Neurosci 2005; 6: 285-96.

Schreiber T. Constrained randomization of time series data. Phys. Rev. Lett. 1998; 80: 2105-2108.

Schreiber T, Schmitz A. Improved surrogate data for nonlinearity tests. Phys. Rev. Lett. 1996; 77: 635-638.

Schreiber T, Schmitz A. Surrogate time series. Physica, D 2000; 142: 346-382.

Shaw JC. An introduction to the coherence function and its use in EEG signal analysis. J Med Eng Technol 1981; 5: 279-88.

Shaw JC. Correlation and coherence analysis of the EEG: a selective tutorial review. Int J Psychophysiol 1984; 1: 255-66.

Soong A, Stuart C. Evidence of chaotic dynamics underlying the human alpharhythm electroencephalogram. Biol. Cybern. 1989; 42: 55-62.

Sporns O, Chialvo DR, Kaiser M, Hilgetag CC. Organization, development and function of complex brain networks. Trends Cogn Sci 2004; 8: 418-25.

Sporns O, Zwi JD. The small world of the cerebral cortex. Neuroinformatics 2004; 2: 145-62.

Stam CJ. Nonlinear dynamical analysis of EEG and MEG: review of an emerging field. Clin Neurophysiol 2005; 116: 2266-301.

Stam CJ, Jones BF, Nolte G, Breakspear M, Scheltens P. Small-World Networks and Functional Connectivity in Alzheimer's Disease. Cereb Cortex 2006.

Stam CJ, van Dijk BW. Synchronization likelihood: an unbiased measure of generalized synchronization in multivariate data sets. Physica D: Nonlinear Phenomena 2002; 163: 236-251.

Strogatz SH. Exploring complex networks. Nature 2001; 410: 268-76.

Takens F. Detecting strange attractors in turbulence. In: Rand D and Young L, editors. Dynamical Systems and Turbulence. Vol 898. Warwick: Springer-Verlag, 1980: 366-381.

Tallon-Baudry C, Bertrand O, Fischer C. Oscillatory synchrony between human extrastriate areas during visual short-term memory maintenance. J Neurosci 2001; 21: RC177.

Terry J, Breakspear M. An improved algorithm for the detection of dynamical interdependence in bivariate time-series. Biol Cybern. 2003; 88: 129-136.

Thatcher RW, Krause PJ, Hrybyk M. Cortico-cortical associations and EEG coherence: a two-compartmental model. Electroencephalogr. Clin. Neurophysiol. 1986; 64: 123-143.

Theiler J. Spurious dimension from correlation algorithms applied to limited time-series data. Phys. Rev. A 1986; 34: 2427.

Theiler J, Eubank S, Longtin A, Galdrikian B, Farmer J. Testing for nonlinearity in time series: the method of surrogate data. Physica D 1992; 58: 77-94.

Theiler J, Rapp P. Re-examination of the evidence for low-dimensional, nonlinear structure in the human EEG. Electroenceph. Clin. Neurophysiol. 1996; 98: 213-222.

Tononi G, Edelman GM. Consciousness and complexity. Science 1998; 282: 1846-51.

Torrence C, Compo G. A practical Guide to Wavelet Analysis. Bull. Am. Meteorol. Soc. 1998; 79: 61-78.

Trujillo LT, Peterson MA, Kaszniak AW, Allen JJ. EEG phase synchrony differences across visual perception conditions may depend on recording and analysis methods. Clin Neurophysiol 2005; 116: 172-89.

Varela F, Lachaux JP, Rodriguez E, Martinerie J. The brainweb: phase synchronization and large-scale integration. Nat Rev Neurosci 2001; 2: 229-39.

Zaveri HP, Williams WJ, Sackellares JC, Beydoun A, Duckrow RB, Spencer SS. Measuring the coherence of intracranial electroencephalograms. Clin. Neurophysiol. 1999; 110: 1717-1725.

Zheng Z, Hu G. Generalized synchronization versus phase synchronization. Phys. Rev. E 2000; 62: 7882-7885.

RFId technologies for the hospital. How to choose the right one and plan the right solution?

Ernesto Iadanza
Department of Electronics and Telecommunications – Università degli Studi di Firenze
Italy

1. Introduction

RFId is an acronym for Radio Frequency Identification. Many different technologies are gathered under this abbreviation, each optimized for some particular tasks. Factories can take advantage of RFId for managing and optimizing their supply-chains, inspecting the content of a pack without actually opening it. Stores use RFId as a substitute to barcode labels because it works even without any lines of sight. Many offices and car parks use some RFId based solutions to allow the access for authorized people only. Recently, RFId technology has been used to implement fast and secure payment services, using disposable wristbands that stop functioning once removed from the wrist and cannot be put back together.

Besides military systems, the first spread use of RFId technology dates back to the late 1960s, when the first Electronic Article Surveillance (EAS) systems where implemented against shopliftings. They were based on simple transponders transmitting a single bit just to signal their presence.

We must wait for the 1990s to see some modern RFId equipments, thanks to the great miniaturization of the electronics and to the resulting reduced power requirements.

Nowadays, also the healthcare world is rapidly approaching to RFId, both for increasing the automation level and for reducing the overall clinical risk for patients. Following, a few examples.

Passive RFId tags are used on surgical tools to read the composition of a sterile surgical kit prior to start the operation.

RFId wristbands can be worn by patients for reducing identification errors and for tracking their therapies or treatments. If the wristbands are equipped with active RFId tags, the patient position inside the hospital can also be easily monitored and tracked: this is particularly useful to caregivers for managing children or patients with reduced cognitive functions.

Blood transfusion errors can be heavily reduced by using RFId in the blood supply chain: patients and bags of blood can be tagged to make sure every patient receives the right blood product.

Similarly, the pharmaceutical supply chain could take advantage of RFId technology both for replacing barcodes and for implementing single-dose delivery automated systems.

2. RFId technology

An RFId system is typically composed by at least two components: tag and reader. In the simplest functioning mode, when the reader "wakes up" the tag (forward link), this responds by transmitting its own unique ID code (reverse link). If the tag is passive, i.e. is not provided with a battery power, the reader itself must energize the tag. The communication between the reader and the tag can hence be only initiated by the reader.

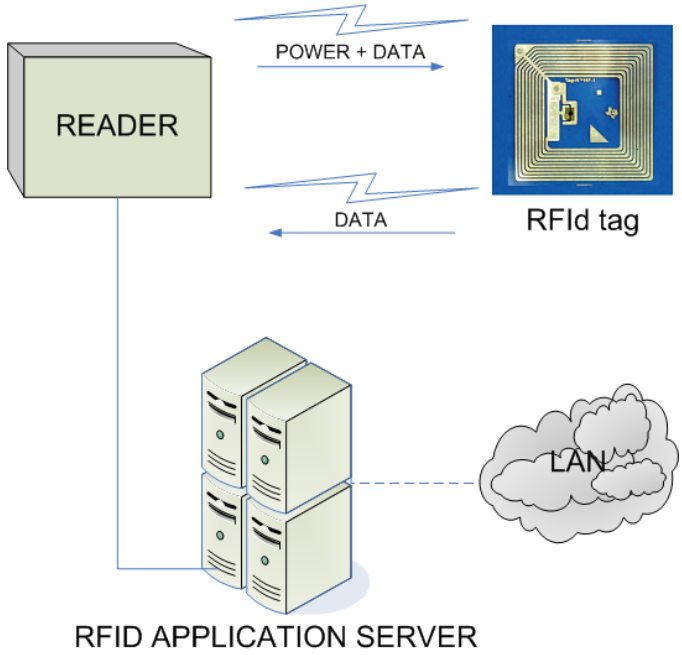


Fig. 1. A simple RFId system

A simple reader can be made by the following parts:

- *rx/tx antenna*
- *modulator*, used to query or to transfer data to the tag
- *demodulator,* to decode the received data
- *control unit*, a microcontroller used to manage the link with the tag and to transfer the read data to some external devices like a PC
- *power adaptor* or *battery*

The tag, or transponder, incorporates at least the following four components:

- *antenna*, used both to receive the power by the reader (if the tag is not provided with a battery) and to exchange data with the reader

- *microchip*, that is used to manage the data link implementing the desired protocol, frequency and modulation.
- *memory* (sometimes internal to the microchip)
- *package*, that keeps together and protects all the components; this part can be very variant depending on the intended use of the tag (labels, wristbands, glass cylinders, etc.)

The tag types are usually classified basing upon their powering modes: passive, semi-passive and active tags.

2.1 Passive tags

Passive tags are fed directly by the reader. This can be achieved by using an inductive coupling (LF or HF) or even a backscatter coupling (UHF).

In the first case, both the reader and the tag are provided with coil antennas. The inductive coupling between the two antennas, assimilable to the primary and the secondary coils in an electric transformer, transfers energy to the tag for the operation of the microchip. This can happen if the two devices are close enough: the tag must be within 0.16λ meters from the reader's antenna in order to be in the near field region. Typical frequencies used are 13.56 MHz and 135KHz, hence the wavelengths are much greater than the distance between the reader's coil antenna and the tag (22.109 m for 13.56 MHz; 2220.7 m for 135 KHz systems). Therefore the electro-magnetic field may be treated as a simple magnetic alternating field.

An alternate voltage is generated by induction in the tag's coil antenna leadings, and is then rectified by means of a simple diode and used to power up to the microchip. The antenna coil inductance is used, together with a capacitor connected in parallel, to obtain an LC parallel resonant circuit. The resonant frequency is chosen same as the reader's transmission frequency.

The reverse link communication is obtained modulating the voltage of the tag's antenna by switching on and off a load resistance with a very high frequency f_S (*load modulation*). These controlled variations create two spectral lines at a distance of $\pm f_S$ around the transmission frequency of the reader and are reflected as an amplitude modulation of the *subcarrier* f_S to the "primary coil" on the reader. This method can be used to send back data from the transponder to the reader. [www.rfid-handbook.com]

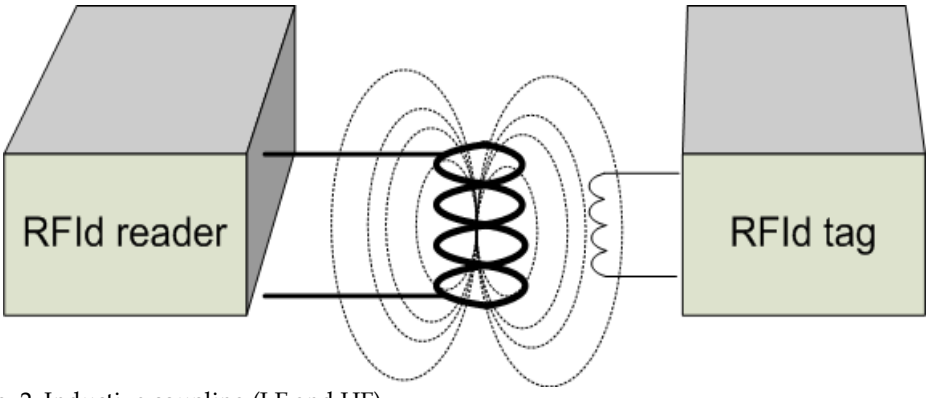


Fig. 2. Inductive coupling (LF and HF)

UHF passive RFId systems use dipole antennas both on the tag and on the reader. The typical work frequencies are 868MHz (EU), 915MHz (US) and above (microwave). Since the higher is the frequency the smaller is the wavelength, these system make it simple to design smaller antennas. These are called long-range systems since the distance between the reader and the tag can be greater than 1m. The tag is fed by the reader using electromagnetic coupling.

A backscattering phenomenon is used to allow the tag to perform the reverse link. Here is how it works: a fraction of the power that comes from the reader is reflected by the transponder dipole antenna back to the reader depending on the tag's antenna reflection *cross-section*. This characteristic parameter can be altered by switching on and off a load resistor connected in parallel to the transponder antenna. You can take advantage of this phenomenon to transmit data from the tag to the reader by modulating the power fraction reflected back.

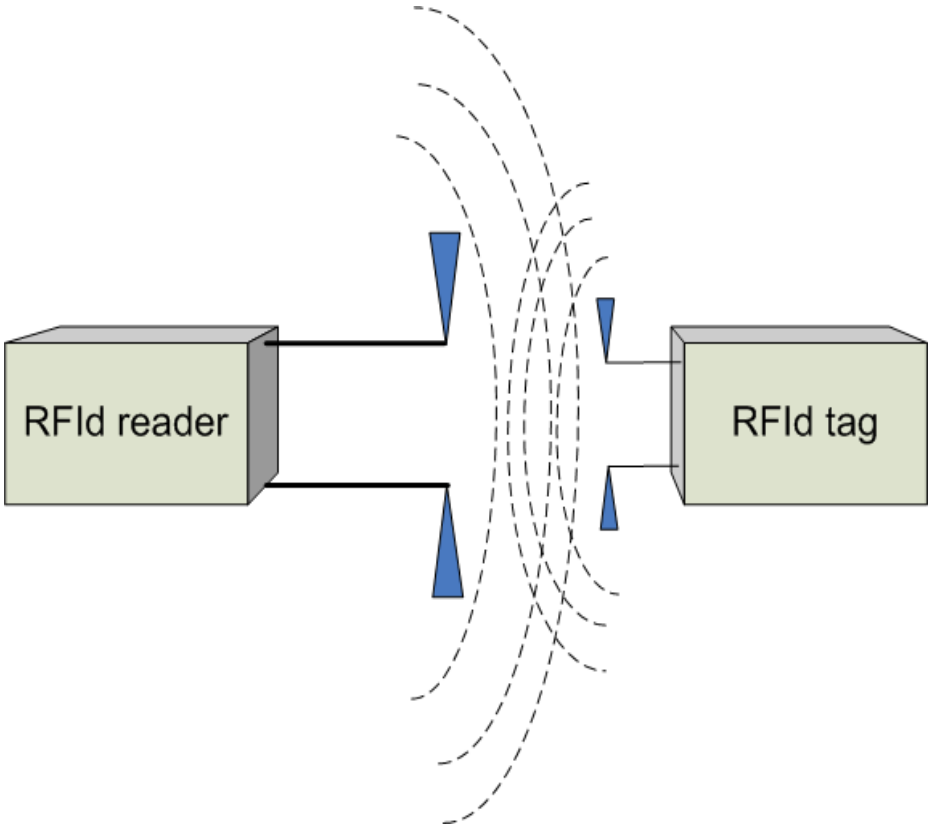


Fig. 3. Backscattering coupling (UHF)

2.2 Semi-passive tags

These tags are in all similar to passive ones, but are provided with an onboard battery used for feeding some sensors (accelerometers, temperature sensors, pressure sensors, etc.). Anyway, the battery is not used to feed the tag microchip or antenna; this means that these transponders cannot start a communication autonomously. The reliability is of course affected by the presence of the battery itself.

2.3 Active tags

An RFID tag is called "active" when it is equipped with a battery, to be used to feed the tag's microchip and antenna and also as a source of power for onboard sensors. These tags are proper transceivers, therefore they are able to start a transmission even if not queried by any readers.

Some typical work frequencies are 433MHz, 868MHz, 915MHz, 2.45GHz and 5.8GHz. The higher bandwidth gives you the chance to implement a real complete communication system.

The maximum communication distance can reach tens or even hundreds of meters, according to the work frequency used and to the output power (according to national regulations).

Active RFId technology gives you the opportunity to implement a real tracking system, provided that the tag's spatial position can be calculated using some RTLS (Real Time Location System) algorithm or some other source of spatial information.

The main drawbacks are the transponder end user price, tens of times higher if compared to passive tags, the increased size and weight, and the necessity for maintenance.

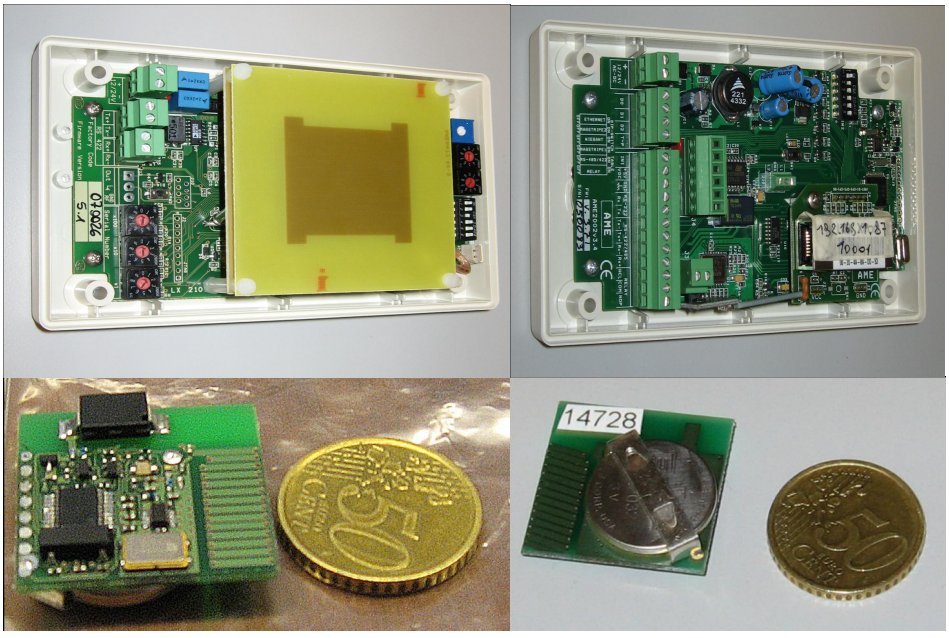


Fig. 4. Active RFId system (courtesy of AME, www.ameol.it)

2.4 UWB (Ultra Wide Band)

UWB (Ultra Wide Band) is a technique that makes use of a broad frequency range (3,1 GHz - 10,6 GHz). This is often obtained by using radiofrequency impulses with a very low time duration, few tens of picoseconds, that translates in a very wide spectrum. Also, since the time-pulse is so short, the UWB is slightly sensible to interferences caused by wave reflections. The energy needed to generate such narrow time-pulses is very low: this is a great plus of this technology because it can at once save the tag's battery life and generate few electromagnetic interferences.

All this makes UWB very good for use in "noisy" environments like factories or hospitals.

This technology has been widely used in military field, in the last 20 years for telecommunications and geolocalization. After 2004 US government has allowed the use of UWB for civil scopes.

UWB can show its potential in healthcare applications, because of the following issues:

- the short duration of time pulses reduces the possible interferences due to reflected signals, since such a short signal is correctly received and processed before any mirrored out-of-phase signals can be;
- tag battery life is preserved since tag's total power consumption (Tx+Rx) is reduced down to 1 mW;
- reduced or no interferences at all with other narrow band communications in the same range (3,1 GHz - 10,6 GHz);
- if combined to recent powerful Real Time Location System algorithms (RTLS), UWB allows for very good performances in locating assets, patients or personnel, in terms of precision and accuracy;
- high data rates
- high insensitivity to obstacles, fluids and metals if compared to other narrow band active RFId systems
- simplified tag circuitry, compared to narrow band RFIds: pure digital signals can be generated and transmitted by UWB transponders without having any DAC/ADC onboard or any analog modulators/demodulators.

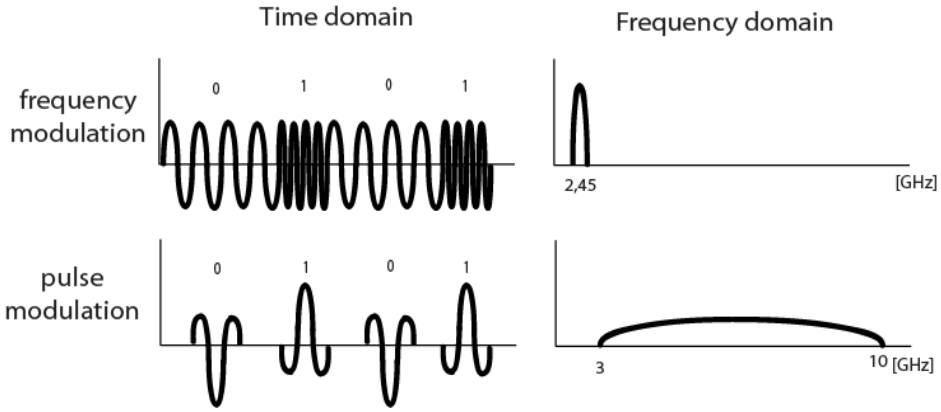


Fig. 5. Narrowband vs UWB functioning principles

3. Applications of RFId in healthcare

This paragraph summarizes some applications of RFId to healthcare. The listed experiences are an abstract of the investigation performed by the author together with Dr. Roberto Bonaiuti, former member of his research team.

3.1 Drugs management

RFId technologies, alone or combined with others like barcodes, are used for the automation of the drugs management process. Many steps can be managed: drugs production and packaging inside the factories, deliveries to the hospital pharmacy, automation of the pharmacy storage and retrieval, patient's bedside therapy preparation and tracking.
The Ospedale "G.B.Morgagni-L.Pierantoni" di Forlì is an Italian public owned hospital (Azienda Unità Sanitaria Locale di Forlì, Servizio Sanitario Regionale Emilia-Romagna) counting about 550 beds. It has been equipped with a Pillpick system by Swisslog (www.swisslog.com).
The solution consists of an automated management of the pharmacy, combined with an interface to the prescription software (CPOE - Computerized physician order entry) and to the Hospital Information System and an unit dose process in the wards. The drugs are placed in holders tagged with RFId and managed using an automated robot. The data recorded in the tag are about operator, drug type and posology, drug's expiring date and more. These data are read by nurses at the bedside using handheld passive RFId readers; then this data are coupled with patient's ID using barcode wristbands. (Bianchi, 2008)

3.2 Tracking of biopsic specimens

The Mayo Clinic (www.mayoclinic.com Rochester, MN, USA) uses passive RFId tags to track biological gastrointestinal tissue specimens, from their collection in one building to the pathology laboratory in another.
The system has been developed by 3M (http://solutions.3m.com/en_US/). It uses ISO 18000-3 compliant passive RFId tags operating at 13,56MHz attached to the sample holders. Each tag's unique ID is linked to patient's data from the EPR in the HIS central database. These data also include the sample coded description coming from a surgical database. (Bacheldor, 2007 a)

3.3 Tracking of blood bags for transfusion

Blood bags for transfusion are an important field of application for RFId in healthcare. If fact blood, plasma and blood products are stored at low temperatures for cryopreservation. This causes ice on the bag's surface. Therefore optical based identification technologies like barcodes are useless for this scope.
The hospital of Saarbrüken (Germany) uses RFId to track blood bags, record transfusions and perform a matching of patients and blood bags. Patients are provided with a passive RFId wristband. Blood bags are tagged with self-adhesive passive RFId labels operating at 13,56MHz. The labels are equipped with a 2KB memory to store an unique ID and some informations about the blood composition.
Both these tags are read using an handheld PDA equipped with a passive RFId reader. The data matching is then verified by a central software. Hence, the operator is able to verify the

correct coupling between patient and blood bag, thus reducing significantly the occurrence of errors. (Wessel, 2006)

3.4 Asset tracking

The Harmon Medical Center (Las Vegas, NV, USA) uses an asset localization solution developed by Exavera Technologies (www.exavera.com).
The system makes use of active RFId tags and readers operating at 915MHz frequency. A custom software lets you locate the assets using a cartographical map.
Every room is equipped with active readers that locate the assets and send their ID to the central software via LAN. These data are then linked to the information coming from the clinical engineering department like datasheets and maintenance operations performed. (Bacheldor, 2007 b)

The Spartanburg Regional Medical Center (Spartanburg, SC, USA) uses an 802.11g solution to locate more than 550 intravenous infusion pumps.
The system is developed by McKesson (www.mckesson.com) using hardware and RTLS (Real Time Location System) by Ekahau (www.ekahau.com). The whole hospital is covered using more than 300 Wi-Fi access points. The active tags "beep" once in an hour to communicate their unique ID. Each time a tag detects a change of position, thanks to movement sensors mounted onboard, it communicates its ID waiting just six seconds. This behaviour lets the batteries go on for even two years.
A web based software shows the pump positions over a plan of the hospital. The system is as well capable of sending alarms in case some pumps enter particular areas.
(Bacheldor, 2007 c)

The Washington Hospital Center (District of Columbia, USA) uses an UWB RFId system from Parco (www.parcomergedmedia.com) to track and localize medical devices, mainly devices used to move patients like litters, wheelchairs, wheel beds and portable radiographs. The UWB transponders are shaped in cubes 2.5cm wide screwed or glued to the device to be tracked. Tags can be located by readers within a 180m radius with a pretty good accuracy of less than half meter. The tag's battery can last 4 years with a pulse frequency of 1 Hz. Every transponder is provided with a 32 bytes of data memory and is capable of transmitting its ID number together with some more info about battery life and manumissions.
A GIS software is included to show the position of every tag on a map of the hospital.
(Bacheldor, 2007 d)

4. RFId and electromagnetic interferences (EMI): case study

Radio Frequency Identification (RFId) technology is quickly entering hospitals, as shown in the above chapter 3, often close to the patient himself.
Some of the outlined tasks can be done having recourse to simple passive RFId tags: mother-baby matching with wristbands to avoid mix-ups; patient-drug tracking using RFId tagged packaging; blood bags tracking; sterile surgical tools tracking, etc.
On the other hand, active RFId systems allow some tasks not achievable with passive ones or using older technologies like barcodes, video-cameras or else. Some studies show that the

active technology is particularly suitable for tasks such as the location of patients or assets (Iadanza, 2008; Fry, 2005; Davis, 2004; Wicks, 2006, Sangwan, 2005)
RFId use in healthcare is also receiving much attention to assess the implications in terms of patient safety. (Ashar, 2007; Van der Togt, 2008)
The possible EMI on medical equipment is a concern, primarily when the life of the patient is related to the medical device correct function. Some recent studies showed contrasting results, pointing out the need for further investigations to be done case by case (Van der Togt, 2008; Christe, 2008). The focus of this paragraph is examining the EMI between an active RFId system and the critical care equipment in a children's ICU.
As mentioned above, an active RFId system consists of three main devices: illuminator, receiver and tag. The system is then connected to a data network and is managed by a master software. The tag is battery-powered and is normally in stand-by mode; when entering an illuminator field cone, it wakes up and it starts to transmit its ID code together with the illuminator's ID code to a receiver. The various systems on the market use many different transmitting frequencies and modes of operation, also depending on the different national regulations.
The electrical medical equipment must comply to UL/EN/IEC 60601 standard plus some national deviations. In particular the collateral standard TE 60601-1-2 applies to electromagnetic compatibility of medical electrical equipment and medical electrical systems. Nevertheless many medical devices, still widely used in hospitals, only meet older versions of the standard that required lower immunity test levels over the frequency range 26 MHz to 1 GHz.
Here is why it is important to test the RFId system for possible EMIs on the hospital medical electrical equipment.
The tested RFId hardware is an active dual frequency system, LNX®, by Advanced Microwave Engineering S.r.l. (www.ameol.it., Florence, Italy). The LNX system includes three devices: the illuminator, the tag and the reader. (Iadanza, 2008)
The major EMI source in the system is the illuminator. The system is tested for its possible use in a children's hospital intensive care department.
In this application the footprint of its antenna is designed to cover a single ICU room. It consists of a 2.45 MHz PLL oscillator cascaded with a OOK modulator and a medium power MMIC amplifier. The radiation pattern of the antenna has 120 degrees –8 dB angular aperture. Circular polarization is employed because the orientation of the tag, that uses a linear polarized antenna, is unpredictable in many applications. The signal transmitted by the illuminator provides a programmable ID code and few more setting commands that are used for programming the operation mode of the tag entering its field pattern. The RF output power of the illuminator can be set from 0 dBm to 20 dBm (Biffi Gentili, 2008). For each test , the maximum power of 20dBm has been used.
The RFId tag is a battery-powered dual frequency device that can be activated and programmed by the illuminator. It comes with a 4 Kbytes memory board and it is in a low power consumption stand-by mode until it is activated. Then it transmits its own ID code and the illuminator code to a receiver unit, using a 433 MHz centred band and a maximum output power of 0 dBm.

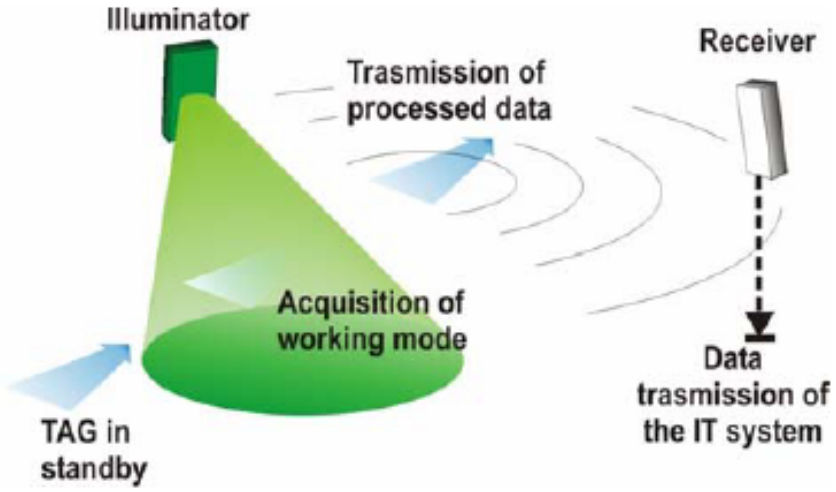


Fig. 6. LNX System working scheme

The tested critical care devices are a typical equipment for a children's resuscitation. An ICU room, away from the patients area, was set up with a moveable RFId illuminator and some active RFId tags.
The medical equipment was operated by healthcare personnel, trained to manage it in everyday use. Table 1 shows a list of the 16 devices, tested in two different times.

DEVICE	N. OF DEVICES
Ventilator	4
Syringe pump	4
Volumetric Infusion pump	3
Defibrillator / Monitor	3
Multi-parametric monitor	2

Table 1. Tested critical care equipment

EMC assessment on all the medical equipment has been performed starting from their documentation. For each medical device it has been developed a particular checklists containing the tests to perform.
Only the two ventilators were compliant to the latest IEC 60601-1-2:2003 standard, that specifies a general immunity test level to radiated RF noise of 10 V/m. The remaining 14 devices, according to their manuals, were compliant to previous versions of the same standard, that required a level of immunity of just 3 V/m.
All the tests were performed switching on a single appliance at a time in a fully operating critical care room without any patients.

The test method was based on the American National Standards Institute recommendation ANSI C63.18 to assess the electromagnetic immunity of the medical devices by the RFId illuminator and an active tag (IEEE; 1997). The standard has been integrated with checklists, as stated above, designed for each medical device after the analysis of its operational and maintenance documentation.
Each electrical medical device was first checked using its own internal test procedure and by healthcare staff. If necessary the devices were connected to the provided simulators.
Then the illuminator was turned on. The distance between the illuminator and the device was reduced, according to ANSI C63.18 standard, in three following steps from 2m to 0,6m to 0.01m (indicating illuminator on top of the device, below the minimal distance for the RF immunity tests imposed by IEC 60601-1-2). For each step the device was turned off and then on, the device internal test procedures were performed and the performances were evaluated by the healthcare personnel.
Each test was repeated having a battery powered transmitting tag attached to the device body. At the minimal distance, the illuminator was moved in three different positions on the axes x (frontal), y (lateral) and z (above the device).
No malfunctions spotted on the ventilators in Paw, flow, respiratory frequency or other parameters for any of the tested modes:

1. IPPV (Intermittent Positive Pressure Ventilation);
2. SIMV (Synchronized Intermittent Mandatory Ventilation);
3. MMV (Mandatory Minute Volume Ventilation);
4. CPAP (Continous Positive Airway Pressure);
5. ASB (Assisted Spontaneous Breathing);
6. BIPAP (Biphasic Positive Airway Pressure);
7. APRV (Airway Pressure Release Ventilation);
8. PPS (Proportional Pressure Support).

No malfunctions in the alarms, tested simulating alert situations, neither for the older devices of the set, that conformed just to the first version of the IEC 60601-1-2.
None of the tested pumps, set to give 5 mL/h, revealed malfunctions during the tests. Alarms correct functioning was assessed by simulating an occlusion, then waiting for the alarm beeps and for the error message, both disappeared as soon as the shrinkage was eliminated.
No anomaly as well for the defibrillators. Tests were performed using device's 'User Test' mode with no actual defibrillator shots. Also the ECG trace, obtained by connecting the electrodes to a test subject, showed no errors: the ECG curve has not revealed distortions and heart rate remained constant. The 'signal absence alarm' functioning was verified, after removing an electrode. The alarm stopped as soon as the electrode was repositioned.
Also the Siemens multi-parametric monitors, tested detecting the ECG and the pulse oximetry signal, worked properly during all the performed tests.

Eventually, our study found no evidence that the use of an active low power microwave RFId system does affect the performances of the neighboring medical devices.
The set of devices tested does certainly not cover the broad spectrum of the devices on the market. Nevertheless it is heterogeneous, composed by critical devices and containing many outdated models.

Therefore it is feasible enough to extend these results to a generic hospital ward equipment.

5. A multi-layer method to design a technical RFId solution for in-patients tracking

In this paragraph we will discuss a design method to correctly identify the technical solution that best responds to aims and imposed constraints. (Iadanza 2008)
The method consists in four consecutive steps. The first thing to do is to focus all the project aims. This step has to be performed together with the client, that in this case is the top management of the hospital. The higher is the position, on the hospital organization chart, of your interlocutor, the best are the chances to spot the final wishes of the client.
The second step is addressed to the translation of the upper level aims in functional requirements that the final system must satisfy. Also in this phase it is important to maintain a close connection to the people that will actually take advantage of the system, like head of department, caregivers or technical personnel.
Afterwards, the functional requirements must be transformed in technical constraints; this step can be completely managed by the designer himself.
Eventually, many technical solutions are compared in order to assess which one best fits the upper levels constraints.
Therefore, as in a four layer planning architecture, the top layer (layer I: "project aims") must be satisfied as much as possible be the lower layer ("functional requirement" - layer II). Similarly, the "technical constraints" layer III is created to satisfy as much as possible its upper level (layer II). Using this operating mode, the technical solution comes out very coherently to the main project aims.

LAYER	DESCRIPTION
LAYER I	PROJECT AIMS
LAYER II	FUNCTIONAL REQUIREMENTS
LAYER III	TECHNICAL CONSTRAINTS
LAYER IV	TECHNICAL SOLUTION

Table 2. The four layers in the proposed multilayer design method

5.1 Layer I: Project aims

Project aims can be divided in three main categories: functionality, economical efficiency, compliance to standards and laws.

- *Functional* aims: the system must have both a good spatial and time resolution. It must be thought to be used by medical and paramedical staff. It must ease the duty of staff when they want to inform patient relatives about the progress of cares to their kinsman (especially in wards like Emergency Department). It must be provided with alarm procedures for danger situations in which the in-patient, possibly confused, could be. It must be open to an interface towards Hospital information System (HIS).
- *Economical* aims: the system must provide a total cost reduction for patients logistics management. It must lower costs due to clinical errors by reducing error probability.

- *Safety and laws*: the system does not have to be an obstacle to clinical practice and must guarantee a good cohabitation with EMI sensitive devices.

5.2 Layer II: Functional requirements

According to Layer I aims, 28 different requirements were spotted. Some of these are shown here:

- real-time tracking
- indoor and outdoor tracking capability
- coverage range
- system-to-HIS integration
- alarms
- procedure to associate tag and patient
- design and ergonomics of tag support
- tag support resistance, cost efficiency and duration
- interaction with Medical Devices
- patient privacy

5.3 Layer III: Technical constraints

Technical specifications come directly from the upper layer II functional requirements and are the base to assess the technological solution. Many specifications categories are provided to cover the whole range of constraints just showed:

- Technology
  - i. Active RFId
  - ii. Functional range illuminator-tag > 10m
  - iii. Functional range tag-reader > 20 m;
  - iv. Each illuminator must define an area of interest
  - v. Site survey and plan of fixed devices location
  - vi. RFId provided with a robust anti-collision algorithm
  - vii. Not disposable reprogrammable tags
  - viii. 32 Kbit or more on the tag
  - ix. Frequency range: 433 MHz – 2,45 GHz, in compliance with Italian laws
  - x. Tag battery duration > 2 years
- Interface
  - i. Reader must become a node of the HIS (Hospital Information System) via LAN or wireless-LAN
  - ii. Custom software to process tag information and show it on the hospital floor plan;
  - iii. Critical areas recognition and link to some alarm system
- Standard and laws
  - i. Privacy protection
  - ii. accordance of the RFID system with electro-magnetic compatibility and safety guidelines

- Economics
  i. Reusable tag (re-programming tag separable from the support);

- Device package
  i. Both indoor and outdoor use
  ii. Easily wearable tag (custom solutions also for children and injured people)
  iii. Use of an armlet as support for the tag (not for use in Neonatal Intensive Care Units)
  iv. Total weight 'support + tag' < 35 g
  v. Cut-to-remove armlet
  vi. Possibility to separate the tag from the support for sterilization purposes

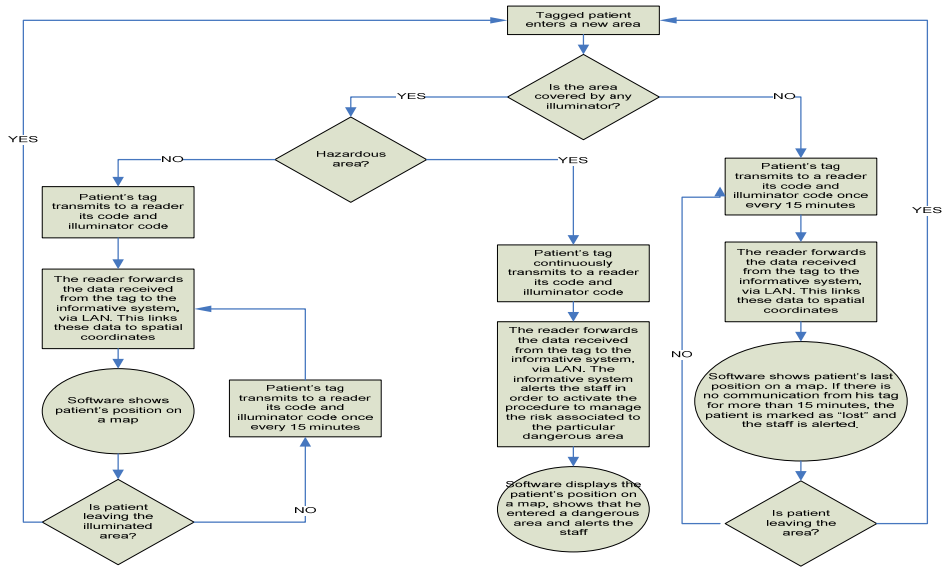


Fig. 7. Patient Tracking Process flow chart

5.4 Layer IV: Technical solution and system configuration

The technical solution that best fits all the above constrains is the same as described in the above section 4 and, therefore, is not described once more in this paragraph.

For patient tracking scope it comes useful a particular operational mode provided by the system: the "beeper fast" operating mode. It allows you to implement a system not just based on a "gate" working model but on an "area recognition" model. In beeper fast mode the tag, once activated, transmits the same signal (tag factory code and illuminator area code) at programmable intervals of time until it enters a different illuminator's antenna lobe. This functioning mode is useful to extend tag's battery life: when the tag enters a closed area (e.g. a room) it could be programmed to transmit once every 15 minutes or more to save energy, whereas the owner patient can't be but inside the room. Out of the room the tag will be enlightened by a different illuminator: this will cause an immediate new transmission.

6. Case study: RFId for children and newborns in intensive care

This paragraph shows how to apply the above multilayer design method (see par. 5) in designing a system to identify and know the actual position of patients in a children's Intensive Care Unit (ICU).

As a first step you must spot the system purposes. In this case the main objective is to provide the ICU with a system to lower down the clinical risks related to misidentification of patients or unintentional rooms swaps.

Still, this does not exhaust the correct definition of all the project scopes. In the first layer, "project aims", you must also take into account many aspects that are linked to functional aims as well as to financial aspects and to quality and standards. The whole system must be designed following standards and laws about privacy and data protection in healthcare.

Also, the technical solutions may be very different according to the patients type, age and cognitive conditions and according to the hospital building types and shapes. A hospital with separate pavilions requires solutions that may be very different from multilevel monoblock buildings. Similarly, in designing a tracking solution for non cooperative patients, you will face requirements very different from surgical patients or newborns.

Furthermore, if delicate healthcare tasks are involved in the process that the system is called to manage, like drugs administrations and Electronic Medical Record (EMR) updates, it will have to deal with many other requirements such as the drugs inventory system, the identification of caregivers, the interface with the Hospital Information System (HIS). (Iadanza 2008)

Children in a resuscitation ward have, in most cases, no cognitive abilities and a wide range of variability in age, weight and dimensions (from newborns to overweight children). This makes it hard their identification by healthcare personnel. They could have no alive parents at all; newborns are often similar one to another; senseless patients simply cannot tell you their name, etc.

The diagnostic and therapeutic process for these children involves frequent movements to other hospital departments for diagnostic tests, surgeries or ward transfers. This raises the risk of rooms misplacement when they come back to the ICU.

The described active RFId solution is intended to identify ward rooms, cradles/beds and patients with unique ID numbers. It also lets the caregivers trace the patients movements on a wide screen, giving warnings and alerts to the nurses in case of dangerous situations.

The proposed system addresses all the constraints induced by the particular environment. Critical care children are of course stationary in their bed, but they can often be moved in a new bed for many reasons (cleaning up, going out, coming from thermal cradle, etc.). They are sometimes not well recognizable one from another, therefore if we use some RFId identifying tags we must lower down as much as possible the need for tags removing and replacement.

Wristbands are not a suitable solution, since these patients can be very weak, small and delicate, hence we must be aware of it in designing the tag case.

The system is composed of five different hardware devices and a tracking software, purposely designed and realized in collaboration with Advanced Microwave Engineering (AME, www.ameol.it). (Biffi Gentili, 2008)

6.1 Reader

This is the only standard device. It is an AME LX 2002 433 MHz receiver provided with an omnidirectional monopole antenna. The signal received from the CRADLE_TAG (see below) is forwarded to the central software system, once added the internal date and time of the receiver, using an Ethernet or IEEE 802.11g wireless interface. (Biffi Gentili, 2008)

6.2 Illuminator

It provides a programmable ID code that can be use to spot the spatial position, once linked to a cartographic map. It is able as well of programming the operation mode of the CRADLE_TAG entering in its field pattern. The illuminator is a custom device obtained modifying a commercial unit AME LX2101 (Biffi Gentili, 2008), changing the antenna and modifying the firmware. It consists of a 2.45 MHz PLL oscillator cascaded with an OOK modulator and a medium power MMIC amplifier. It is attached to the ceiling inside the area you want to outline (room, corridor, nurse space, exits etc.) and it is used to create a confined area underneath itself, in which a tag, provided with a 2.45 GHz receiving section, can receive the Illuminator's ID and some more information that you can set via software.
To narrow the area covered by the Illuminator, obtaining a good separation between two adjoining rooms, we introduced an uniform linear array of eight planar patch antennas. The radiation pattern of this antenna has about 30 degrees –8 dB angular aperture. Circular polarization is employed because the orientation of the CRADLE_TAG, that uses a linear polarized antenna, could vary depending on how it is fixed to the bed.
Placing two Illuminators on a line we obtain a well defined "dark zone" where there is no signal: we could use this zone to discriminate two different areas in an open-plan ICU environment or, simply, two adjoining rooms.

6.3 CRADLE_TAG

This device is used as a bridge between the BABY_TAG and the environment since the system is thought to let the children be moved in a new bed or cradle whenever it is needed. It must be fixed on the bed/cradle side. It gives a link between the BABY_TAG and the system (Readers and Illuminators). The unit is battery powered to be easily mounted on hospital beds or cradles in accordance to safety and convenience of movements. Indeed, this unit is used also when the bed or the cradle are moved elsewhere in the hospital, to follow and track its way.
It is made of four sections:

- a 433 MHz receiver to get Baby-ID from the BABY_TAG;
- a 2.45 GHz receiver to get the Illuminator ID
- a 433 MHz transmitter to send to the Reader unit a string containing Baby-ID, Illuminator-ID and Cradle-ID;
- a passive RFId reader capable of reading passive RFId tags compatible to ISO15693 standard using a frequency range of 13.56 MHz ± 7KHz.

6.4 BABY_TAG

Small, active (battery powered) patient tag fixed underneath the child foot. It can transmit a Patient-ID code to the CRADLE_TAG using a 433 MHz centred band. The device has on board, in addition to a microcontroller, a 3V CR2032 battery, an external 433 MHz

transmitting loop antenna and a miniaturized 2.45 GHz receiving antenna. The receiving section is used to program the unit using an Activator (see below) during the admission of the patient to the ICU. The BABY_TAG could also incorporate a passive RFId tag with the same Patient-ID. This lets the caregivers accurately identify the patient when administering drugs or treatments, even in wards not covered by the active infrastructure.

6.5 ACTIVATOR

It consists of a standard portable Illuminator (see above). The only differences are the antenna (a small single patch model) and the ID number: every activator ID starts with '99'. This device is used to initialize the BABY_TAG once fixed to the patient. It also lets the tracking software to link the Patient-ID to the actual patient identity and EPR.

6.6 SYSTEM MANAGEMENT SOFTWARE

A custom software has been developed to implement every step of the identification and tracking process, from the admission of the baby to the ICU, to his discharge from hospital. The software gets UDP packets coming from the Reader unit via Ethernet or wireless LAN and monitors the ward searching for anomalies like room swaps, empty cradles or beds, empty rooms, etc. It is provided with a graphical interface showing the ward map and can log every warning or alarm coming from possible danger situations (see fig. 5). The software is supplied with a "track mode" that is able to find and follow a single patient moving from his room to a surgical block or a diagnostics department covered by the active RFId infrastructure.

7. Conclusion

This chapter showed that there is a broad field of healthcare applications in which RFId technology can seriously improve efficiency, speed, accuracy, safety, risk management.
In planning an RFId solution it is vital to contemplate an accurate design phase, in order to establish the final characteristics of the system working together with the actual final users.
The adoption of an off-the-shelf solution is not a good idea for healthcare systems, since you must cut out the best system and choose the right technology after assessing all the particular requirements and constrains for the specific environment you are working with.
There is a need for rigorous evaluations of safety related issues, like electromagnetic compatibility and privacy, much more central in healthcare than elsewhere. In fact medical electrical equipment is involved as well as patient's personal data.
Bioengineers must be the chief architects of these systems; they must coordinate the team of experts that should include clinicians, electromagnetism experts, IT solution providers, hospital designers.
This is a relatively new field of bioengineering, hence bioengineers should be encouraged – also during their course of studies – in acquiring all the necessary competences to be leading actors in designing RFId systems for healthcare.

8. References

Ashar B. S., Ferriter A. (2007). Radiofrequency Identification Technology in Health Care: Benefits and Potential Risks, *JAMA*. 2007; 298(19):2305-7

Bacheldor B. (2007 a). At Mayo Clinic RFID Tracks Biopsies, RFId Journal www.rfidjournal.com/article/articleview/2955

Bacheldor B. (2007 b). Harmon Hospital Implements RFID to Track Assets, RFId Journal www.rfidjournal.com/article/articleview/2933

Bacheldor B. (2007 c). Spartanburg Medical Center Deploys Wi-Fi RFID System, RFId Journal, www.rfidjournal.com/article/view/2949/1

Bacheldor B. (2007 d). Washington Hospital Center to Quadruple Its RFID Expansion, RFIdJournal, www.rfidjournal.com/article/articleview/3009

Bianchi, S et al. (2008). Medication management in an italian hospital: Forli hospital experience, *Proceedings of HEPS 2008*, 25-27 Sep 2008, Strasbourg

Biffi Gentili G., Salvador C. (2008). A New Versatile Full Active RFID System, in *Proc. RFIDays 2008 - Workshop on Emerging Technologies for Radio-frequency Identification*, Roma, 2008, pp 30 – 33

Christe B. Et al. (2008). Testing potential interference with RFID usage in the patient care environment, *Biomed Instrum Technol*. 2008;42(6):479-84

Davis S. (2004). Tagging along. RFID helps hospitals track assets and people. Health Facil Manage 2004; 17(12):20–4

Fry E., Lenert L (2005). MASCAL - RFID tracking of patients, staff and equipment to enhance hospital response to mass casualty events, Amia Symposium, Bethesda (USA), American Medical Informatics Association, 2005. Available: https://www.wiisard.org/papers/command_center/events.pdf

Iadanza E., et al. (2008). Patients tracking and identifying inside hospital: A multilayer method to plan an RFId solution, *Engineering in Medicine and Biology Society, 2008. EMBS 2008. 30th Annual International Conference of the IEEE*, pp.1462-1465, 20-25 Aug. 2008

IEEE - Institute of Electrical and Electronics Engineers (1997). American National Standard Recommended Practice for On-site Ad Hoc Test Method for Estimating Radiated Electromagnetic Immunity of Medical Devices to Specific Radio-frequency Transmitters (Standard C63.18). Piscataway, NJ: IEEE; 1997

Sangwan R.S. et al. (2005). Using RFID tags for tracking patients, charts and medical equipment within an integrated health delivery network, in Networking, *Sensing and Control, 2005. Proceedings. 2005 IEEE*, pp. 1070-1074

Van der Togt R et al. (2008). Electromagnetic Interference From Radio Frequency Identification Inducing Potentially Hazardous Incidents in Critical Care Medical Equipment, *JAMA*. 2008;299(24):2884-2890

Wessel R. (2006). German Clinic uses RFID to track blood, RFId Journal www.rfidjournal.com/article/articleview/2169/1/1

Wicks A. M. et al. (2006). Radio Frequency Identification Applications in Hospital Environments. *Hospital topics* 2006;84(3):3-9

Improvement of Touch Sensitivity by Pressing

Hie-yong Jeong[1], Mitsuru Higashimori[2], and Makoto Kaneko[2]
[1]*SAMSUNG Heavy Industries,* [2]*Osaka University*
[1]*Republic of Korea,* [2]*Japan*

1. Introduction

Human can perceive a touch sensation, when it is enough for the source of all touch information which is the spatial-temporal distribution of mechanical loads on the skin at the contact interface to get to be larger than the just noticeable difference (Srinivasan, 1996). If mechanical stimuli of external surroundings are smaller than the difference threshold, human cannot recognize any information of touch sensation during tactile exploration. The difference threshold is an important factor for evaluating the touch sensation, whereas it is well known that the just noticeable difference of human is easily affected according to physical environment as well as mental condition.

There have been a number of studies on finding factors to have an influence on a touch sensation (Johansson; Mountcastle; Bolanowski, 1983; 1972; 1982). Among various studies, especially we are interested in making sure whether the change of blood flow changes the touch sensitivity of human or not. Skin sympathetic nerve activity controls blood flow and sweating. Stimulation of skin sympathetic nerve activity causes vasoconstriction, which in turn decreases in skin blood flow. We have started to answer a question that the blood flow in the skin might alter physical property of skin and modify sensitivity of mechanosensitive receptors. However, the characteristic of blood flow is simply changed by mental as well as physical conditions. Accordingly, it is difficult to measure the touch sensitivity while getting the blood flow rate to be kept up with the settled amount in order to have a statistically meaningful data. The lack of information regarding the effect of skin blood flow on neurosensory mechanisms has been due to no appropriate measurement method.

To cope with this issue, our approach is to observe what happens when we compulsorily block the blood flow to the finger tip by pressing the proximal phalange of finger as shown in Fig. 1. Under such a condition the blood flow rate of fingertip can get to be kept up with a small and constant value. Fig. 1 shows a conceptual image of the experimental result how the touch sensitivity varies with respect to time. The goal of this work is to examine how the touch sensitivity alters under the pressed condition as shown in Fig. 1 through the weight discrimination test of non-invasive method. The sensory tissue will eventually get a serious damage after all for continuously pressing the finger. This situation will make us lose any touch sensitivity due to the necrosis. We would like to confirm whether the touch sensitivity temporarily increases and then decreases, or just start to decrease under the pressed condition.

As a result of performed experiments, unexpectedly, we discovered that the touch sensitivity improves temporarily by approximately 3.6 times with the statistical significance test of below 0.1 %, when the proximal phalange of finger is pressed. That is, the human difference threshold is more sensitive than that before pressing. We believe that this is a discovery on how the human touch sensitivity changes under the pressed condition.
This work is organized as follows. In Section 2, we review related works. In Section 3, we show what experimental equipments are used for this main purpose, and then we describe the experimental results in detail in Section 4. In Section 5, we discuss the experimental results with measurement of skin physical property and vibrotactile perception thresholds, so that we can estimate which receptor mainly contributes to improving the touch sensitivity. Finally in Section 6, we conclude this work.

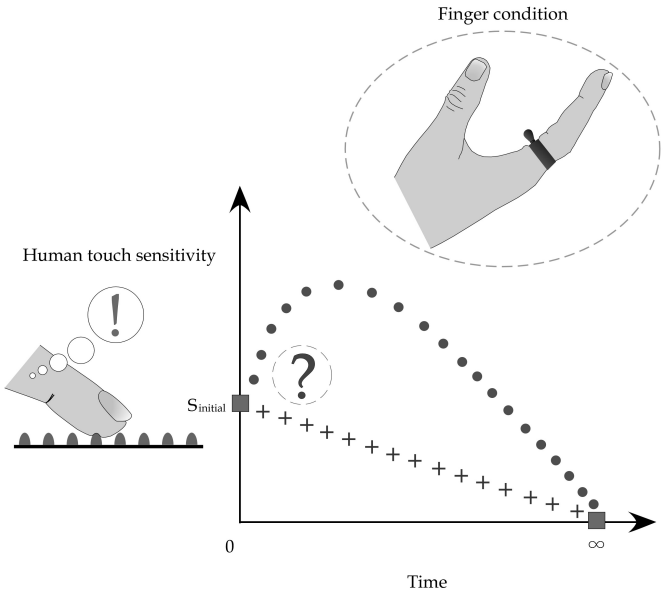


Fig. 1. Human touch sensitivity with respect to time after pressing.

2. Related Works

Johansson *et al*. have discovered the response characteristics and receptive field of skin mechanoreceptors when the human skin surface is stimulated by recording single cutaneous afferent fibres with the electrode (Johansson & Vallbo, 1983). Srinivasan *et al*. have examined the relationship between the skin structure and mechanoreceptors, and then have shown that mechanoreceptors make the nerve impulse when the deformation of the skin surface exceeds the stimulus threshold (Srinivasan, 1996). On the other hand, Gaydos *et al*. have found that the significant decrement of finger dexterity performance is observed when the fingertip skin temperature goes down from the experimental result of the relationship between the skin temperature and the dexterity performance of the fingertip (Gaydos & Dusek, 1958). Brajkovic *et al*. have shown experimentally that the finger dexterity is

unaffected when the fingertip skin temperature is maintained between 28 ° C through 35 ° C despite the decrement of the blood flow according to the temperature decrement in the experimental room (Brajkovic & Ducharme, 2003). It has been indicated that not only the fingertip temperature but also the blood flow has an influence on the fingertip dexterity performance by these studies. However, although the fingertip touch sensitivity supports the dexterity performance, as far as we examined, there has been no study on the relationship between the human touch sensitivity and the blood flow.

3. Experimental System

3.1 Definition of Human Touch Sensitivity

In this work, the human touch sensitivity is defined as the percentage of correct answer that is measured by whether subjects wearing an eye mask can notice the difference or not, through the weight discrimination test based on Weber's Law.

3.2 Experimental System

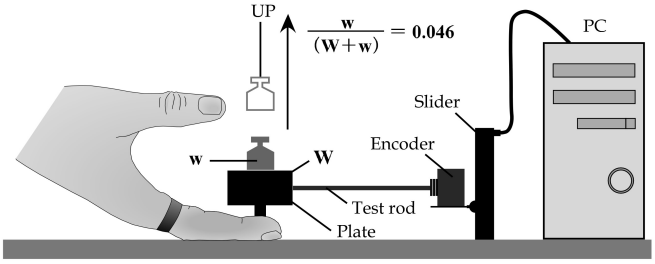
Three kinds of experimental equipments in this paper are used as shown in Fig. 2; an instrument of weight discrimination test for measuring the human touch sensitivity, a non-contact point-typed stiffness sensor for measuring the fingertip stiffness according to the change of blood flow, and a laser tissue blood flow meter for monitoring the blood flow and mass.

The weight difference threshold based on Weber's Law is used for evaluating the touch sensitivity, Ernst Weber, the nineteenth century experimental psychologist, observed that the size of the difference threshold appears to be lawfully related to initial stimulus magnitude (Weber, 1978). This relationship based on Weber's Law can be expressed by:

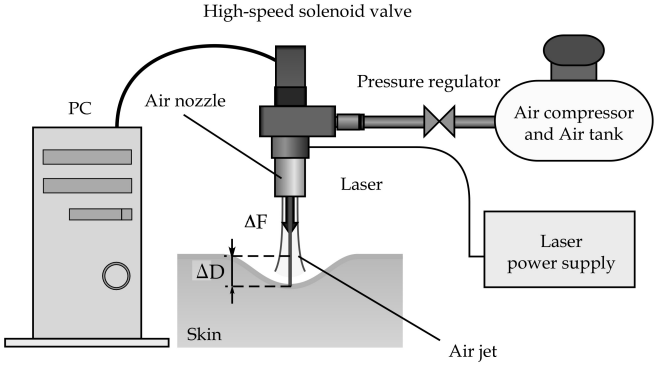
$$I = \frac{\Delta S}{S} \tag{1}$$

where ΔS, S, and I express the difference threshold, the initial stimulus intensity, and the constant ratio despite variations in the S term, respectively. For example, if the ratio is 0.05, we can notice that the increment threshold for detecting the difference from S = 100 g is ΔS = 5 g. When S = 200 g, ΔS = 10 g according to Equation. (1). Fig. 2 (a) shows an overview of the apparatus based on Weber's Law, where it is assembled with the slider to control the height, the test rod connected with the friction-free encoder from the slider, and the plate connected with the tip of the test rod in order to put on the weight, respectively. The different height according to the finger size of the subject is adjusted by monitoring the output of the encoder. The encoder is for making both the plate and the test rod horizontally. The probe contacting with the fingertip has the circular shape whose diameter is 5 mm.

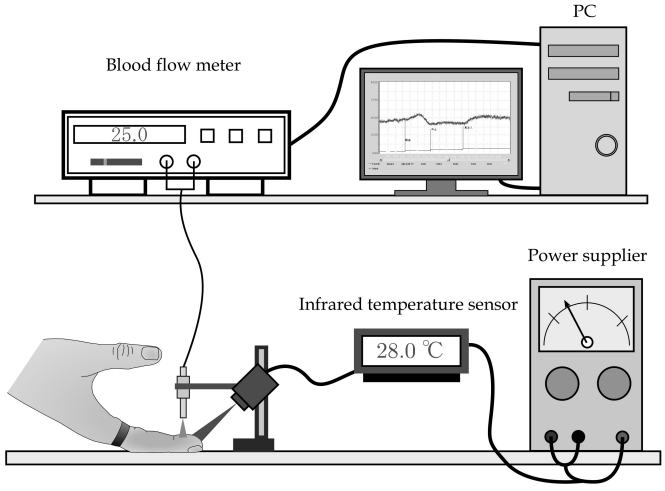
Fig. 2 (b) shows an overview of the non-contact point-typed stiffness sensor. This sensor consists of both the laser displacement sensor of OPTEX FA CO., LTD. and the air-pressure unit to control the airflow, respectively. The air-pressure unit is composed of an air nozzle and a high-speed electromagnetic valve to be able to make a switching motion with the maximum frequency of 500 Hz. The laser beam for sensing the displacement passes through the inside of the air nozzle. It has a common axis between the air nozzle and the laser beam.



(a) Apparatus for measuring weight difference thresholds.



(b) Non-contact stiffness sensor.



(c) Laser blood flow meter and infrared temperature sensor.

Fig. 2. Experimental system.

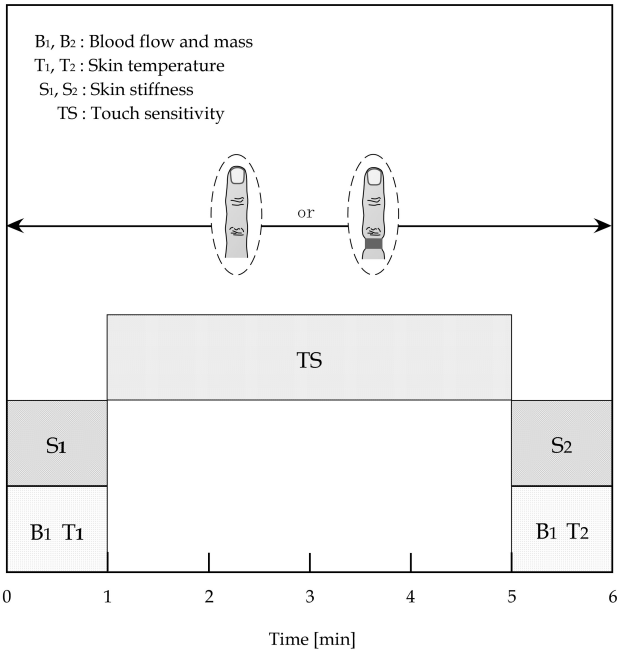


Fig. 3. Time schedule for experiments.

This mechanical configuration is for making the sensing point coincide with the force applying point at all time, so that we can achieve a consistent stiffness sensing for a deformable object like a human skin (Tanaka, 2007). We set the displacement between the laser sensor and human skin with 10 mm and the applied air force with 0.179 N during 200 ms. Then we can compute the average stiffness K by obtaining the value of the laser displacement sensor by the following equation:

$$K = \frac{\Delta F}{\Delta D} \tag{2}$$

where ΔF and ΔD represent the change of force and skin surface displacement, respectively. As for the applied force, it is hard to measure the force during the stiffness measurement. Instead, we calibrate the force characteristic by changing the distance between the nozzle exit and the object in advance. We suppose that the pushing force by air is the function of the distance between the tip of nozzle and the surface of object.

Fig. 2 (c) shows a laser blood flow meter and an infrared temperature sensor. FLO-N1 of OMEGAWAVE, INC. as the blood flow meter is used to measure the accumulated mass of blood as well as the blood flow rate at the fingertip with the sampling time of 1 ms. The length of laser wave is 780 nm. The ultra-compact digital radiation temperature sensor of KEYENCE Corporation is used for examining the temperature. This sensor utilizes the set emissivity to convert the amount of received infrared radiation into temperature.

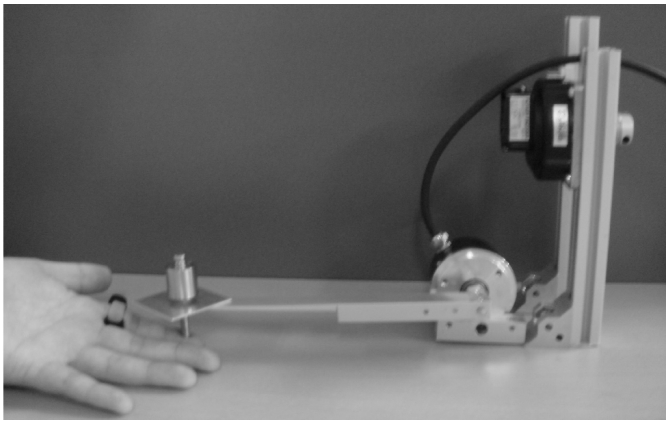


Fig. 4. An overview of experimental environment.

4. Experiment

4.1 Experimental Method

We first perform experiments under the non-pressed condition to obtain the touch sensitivity, the stiffness, the skin temperature, and the blood flow and mass at the time of zero. Then, we obtain three parameters under the pressed condition according to the time schedule as shown in Fig. 3. Fig. 4 shows an overview of weight discrimination test under the pressed condition. For acting a force on the fingertip, we put on the 103 g weight. The weight can be adjusted to load the 5 g weight on the plate and then the 5 g weight is lifted up. This test is done randomly 5 times/min, and totally 20 times during 4 min, while each subject wears an eye mask. The ratio I of Weber's Law is 0.046. During the experiment under the pressed condition, we monitor how much blood flow is, while we start the experiment with roughly 30 % of the blood flow under the non-pressed condition. All of subjects are 24 persons, the male with the age distribution of 21-25, where the blood flow rate under the non-pressed condition is 18 ± 7 ml/min/100g and that under the pressed condition is 6 ± 3 ml/min/100g, respectively. We executed all experiments under the room temperature of 26 $^{\circ}$C. The test point is the middle intersection of the distal phalange of the index finger of left hand.

4.2 Experimental Results

Fig. 5 shows the main result of this work to confirm how the touch sensitivity changes under the pressed condition, where each point is computed with the time average during every 1 min. The horizontal and the vertical axes denote the time and the touch sensitivity, respectively. As you can see from Fig. 5, when the proximal phalange of finger is pressed, unexpectedly, we found that the touch sensitivity is improved from 12 % to 43 % by approximately 3.6 times, while the continuous pressing results in the numbed finger because of the necrosis of sensory receptors. In other words, this result means that human under the pressed condition are able to perceive the smaller stimulus than that under the non-pressed condition because of the more sensitive discrimination threshold. These results are guaranteed with the statistical significance test (p-value) of below 0.1 %.

Fig. 6 shows the measured results of skin physical property with respect to time. Fig. 6 (a) shows the change of stiffness during the experiment. We would note that the stiffness is measured every 15 s during 1 min after pressing and before stopping the experiment as shown in Fig. 3. From Fig. 6 (a), we can see that the stiffness noticeably increases within 15 s under the pressed condition. Compared with that under the non-pressed condition, the fingertip stiffness under the pressed condition gets approximately 1.6 times harder than that under the non-pressed condition. We would note that the tendency is very reliable because the differences between two conditions are approved with the statistical significance test (*p*-value) of below 0.1 %. This tendency comes from the accumulation of the blood mass at the fingertip. This fact of harder stiffness under the pressed condition can be proved by the experiment as shown in the appendix where the skin surface deformation is captured with an assistance of high-speed camera system. Fig. 6 (b) shows the change of fingertip skin temperature during experiment under the non-pressed and the pressed conditions. We can see that the skin temperature under the pressed condition just linearly decreases compared with that under the non-pressed condition. Fig. 6 (c) shows the change of the accumulated blood mass. As we expected, the accumulated blood mass increases at the fingertip under the pressed condition.

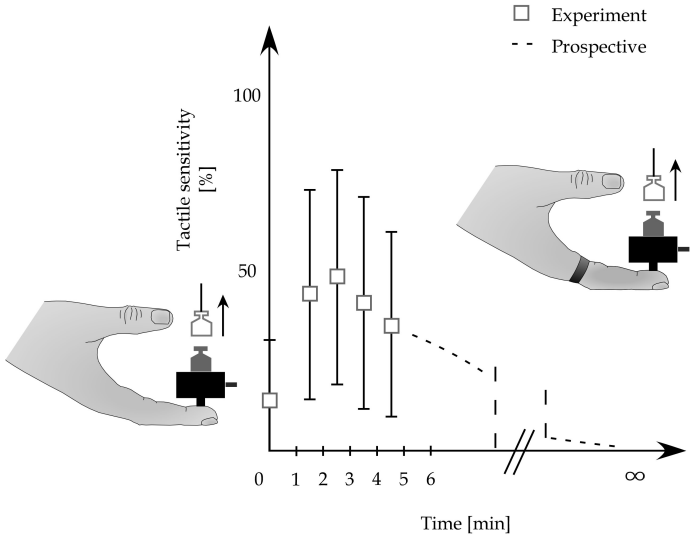
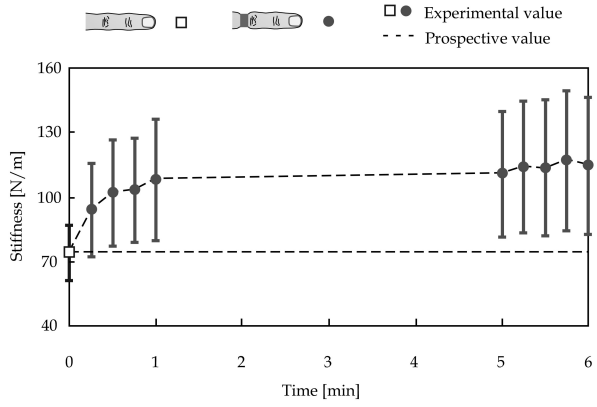


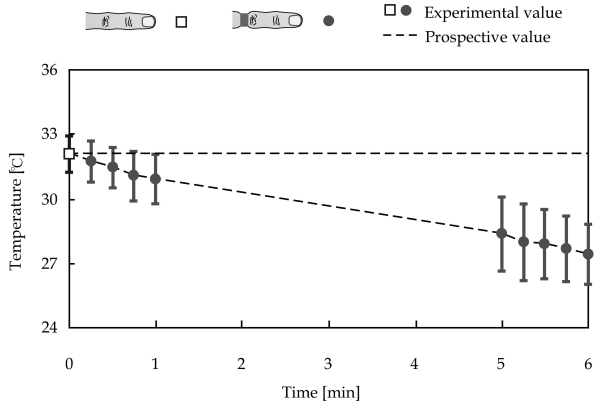
Fig. 5. Experimental results of touch sensitivity with respect to time.

5. Discussion

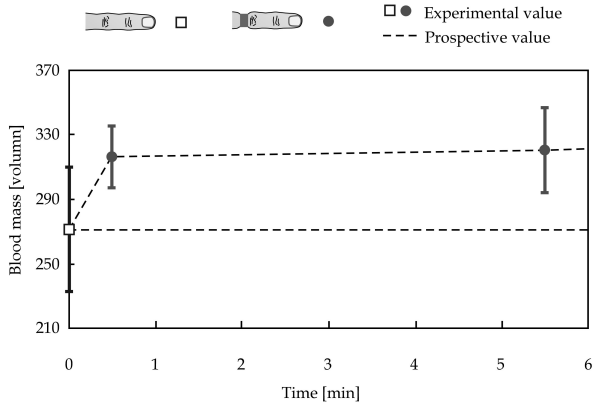
Fig. 7 shows two possible reasons why the touch sensitivity improves under the pressed condition. At the first route, in order to confirm the change of neural activity, we examine the touch sensitivity of other parts in finger. At first, we measure the point of A and B to



(a) Stiffness.



(b) Skin temperature.



(c) Blood mass.

Fig. 6. Experimental results of skin physical property with respect to time.

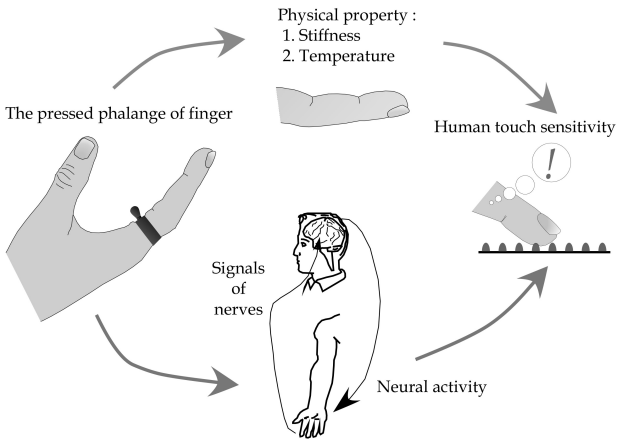


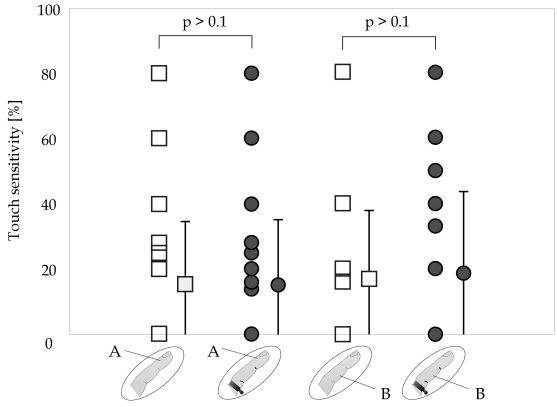
Fig. 7. Two possible routes on improving touch sensitivity under the pressed condition.

confirm whether the touch sensitivity improves or not when we execute the same experiment for the other part where mechanoreceptors have less distributed (Johansson & Vallbo, 1983). As shown in Fig. 8 (a), there is no difference of touch sensitivity at two points, although the finger stiffness is changed with the significance test (*p*-value) of below 1 % as shown in Fig. 8 (b). Now, let us consider two cases which measure the middle finger by the same experiment under the non-pressed and the pressed index finger condition. If the neural route is more active under the pressed condition, we can expect that the touch sensitivity of other parts be also improved. According to the experiment in Fig. 8 (c), a small improvement can be observed for the touch sensitivity of the middle finger while the difference is not enough. This effect may be partially due to the effect of neural activity (Wallin, 1990).

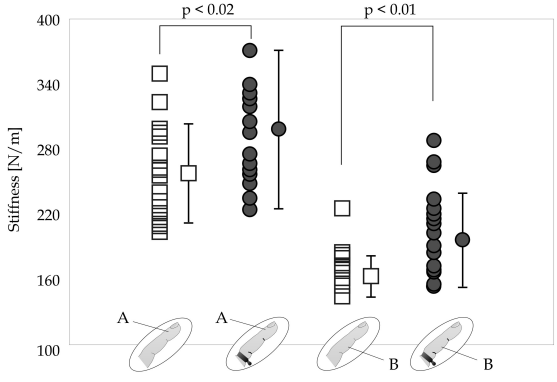
Now, let us consider the change of physical property. It is well known that skin temperature is a factor to change the touch sensitivity (Hayward; Harazin; Nelson, 1986; 2007; 2004). Fig. 6 (b) shows the fingertip skin temperature with respect to time. From this result, we can see that it goes down linearly with respect to time under the pressed condition. However, as you can see from Fig. 5 and Fig. 6 (b), it is difficult for the skin temperature to affirm only to make an influence on the touch sensitivity. This is because the touch sensitivity is changed with the time constant of 20 s under the pressed condition, although the skin temperature shows almost linearly decreasing. This could suggest that another factor except the skin temperature works the change of the touch sensitivity. Let us now consider the relationship between the fingertip stiffness and the touch sensitivity with respect to time. From Fig. 5 and Fig. 6 (a), we can see that the change patterns of both results are similar each other with respect to time, especially in the initial phase under the pressed condition. There have been a couple of reports where the fingertip stiffness affects the touch sensitivity (Dellon, 1995; 1981; 1978). We believe that the increase of fingertip stiffness under the pressed condition makes the increase of touch sensitivity.

In order to explain results appropriately, it is important to examine the frequency response of sensory cell in the fingertip skin. It is well known that there are four kinds of mechanoreceptors, Ruffini ending (below 10 Hz), Merkel cell (below 10 Hz), Meissner

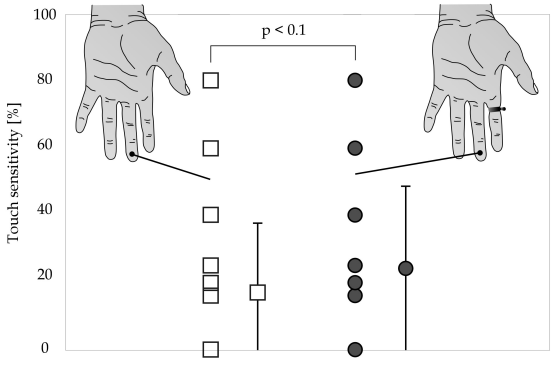
corpuscle (20-40 Hz), and Pacinian corpuscle (over 40 Hz) in fingertip (Johansson; Mountcastle; Bolanowski; Bolanowski, 1983; 1972; 1982; 1984). Each receptor has



(a) Touch sensitivity at the point of A and B.



(b) Stiffness at the point of A and B.



(c) Touch sensitivity at the middle finger under two different conditions.

Fig. 8. Touch sensitivity of the other part finger.

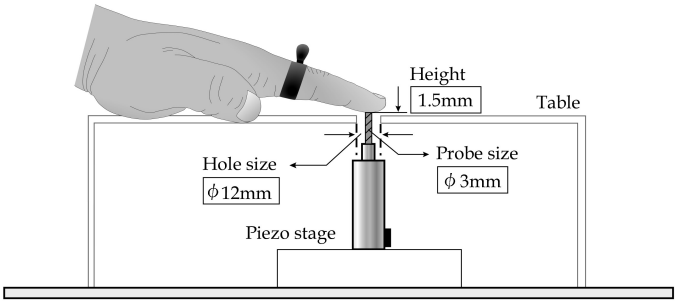


Fig. 9. Apparatus for measuring vibrotactile perception thresholds.

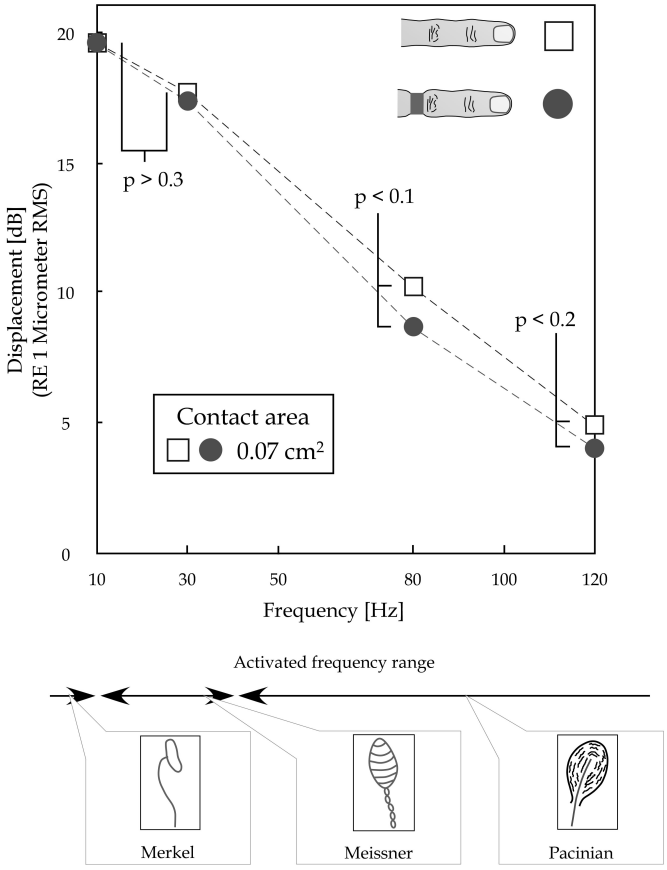


Fig. 10. Experimental results of vibrotactile perception thresholds.

individual different frequency characteristic. In order to evaluate the frequency characteristic of touch sensitivity, we examined vibrotactile perception thresholds by using

the probe whose frequency and amplitude are controllable as shown in Fig. 9 where an active rod made by MESS-TEK Corporation is used. Through experiments, we found an interesting result where the vibrotactile perception threshold is done for the frequency of 80-120 Hz where the Pacinian corpuscle is working dominantly. Fig. 10 explains these results where the dotted line and the one-point dotted line are referred by (Verrillo, 1962; 1963). No statistically significant effect is found on the response of Merkel's and Meissner's mechanoreceptors to vibration stimuli at the frequency of 10 Hz and 30 Hz, and also Ruffini receptors can be excluded for this study because this sensory is revealed to activate for the stretch force of skin. According to these experimental results, it must be the proper evaluation that Pacinian receptors get more sensitive under the pressed condition compared with the response characteristic of the non-pressed condition. This hypothesis makes sense, since the contact force can transmit to the Pacinian corpuscle more directly through a harder tissue caused by the blocked blood. In order to really make sure, we need to adopt the invasive method for assuring our results, which is the way looking into the response with piercing the stimulus probe to mechanoreceptors directly (Toma & Nakajima, 1995).

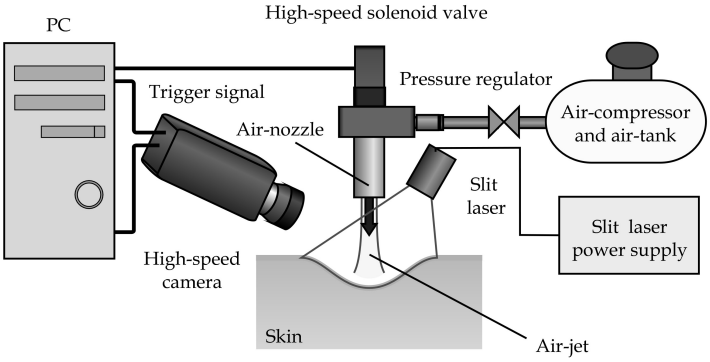


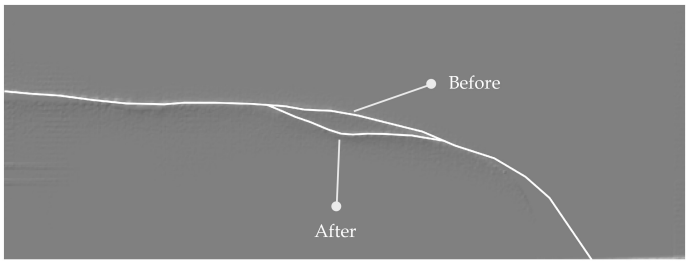
Fig. 11. High-speed camera system.

6. Conclusion

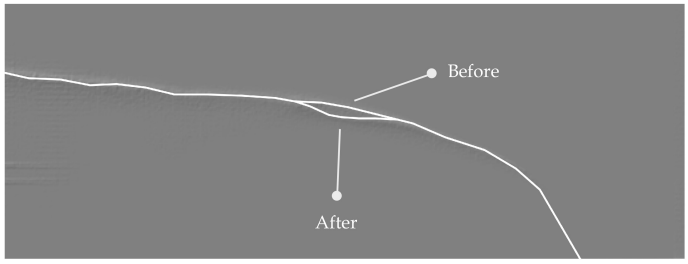
We performed how the touch sensitivity was changed under the pressed condition through the weight discrimination test based on Weber's Law for 24 subjects. Based on these experiments, we concluded this paper as follows.

1) We discovered that the touch sensitivity improved temporarily when the proximal phalange of finger was bound and pressed.
2) We also confirmed that the accumulated blood caused the stiffness of fingertip to get to be harder, and found that the tendency of touch sensitivity was similar to that of fingertip stiffness while skin temperature was decreased linearly.
3) Through the vibrotactile perception threshold, we suggested that the Pacinian corpuscle be a candidate to bring about the improvement of touch sensitivity under the pressed condition.

We would like to investigate the responsiveness of mechanoreceptors in the glabrous skin of the fingertip to vibratory stimuli by using a microneurographic technique under the pressed condition when the frequency and applied pressure to the skin are varied in the future.



(a) Under the non-pressed condition.



(b) Under the pressed condition.

Fig. 12. Captured deformation by a high-speed camera.

7. References

Srinivasan, M.A. & Dandekar, K. (1996). *Journal of Biomechanical Engineering*, Vol.118, 48-55.

Johansson, R.S. & Vallbo, A.B. (1983). *Trends in Neuroscience*, Vol.6, 27-32.

Mountcastle, V.B.; LaMotte, R.H. & Carli, G. (1972). *Journal of Neurophysiology*, Vol.35, 122-136.

Bolanowski, S.J. & Verrillo, R.T. (1982). *Journal of Neurophysiology*, Vol.48, 836-855.

Bolanowski, S.J. & Zwislocki, J.J. (1984). *Journal of Neurophysiology*, Vol.51, 793-811.

Gaydos, H.F. & Dusek, E.R. (1958). *Journal of Applied Physiology*, Vol.12, 377-380.

Brajkovic, D. & Ducharme, M.B. (2003). *Journal of Applied Physiology*, Vol.95, 758-770.

Weber, E.H. (1978). *Academic Press*, ISBN-13: 978-0127405506, New York.

Tanaka, N. & Kaneko, M. (2007). Direction Dependent Response of Human Skin, *Proceedings of International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 1687-1690, Lyon, France, August 2007.

Johansson, R.S. & Vallbo, A.B. (1983). *Trends in Neuroscience*, Vol.6, 27-32.

Wallin, B.G. (1990). *Journal of the Autonomic Nervous System*, Vol.30, 185-190.

Hayward, R.A. & Griffin, J. (1986). *Scandinavian Journal of Work Environment and Health*, Vol.12, 423-427.

Harazin, B. & Harazin-lechowska, A. (2007). *International Journal of Occupational Medicine and Environmental Health*, Vol.20, 223-227.

Nelson, R.; Agro, J.; Lugo, E. ; Gasiewska, H.; Kaur, E.; Muniz, A.; Nelson, A. & Rothman, J. (2004). *Electromyogr. Clin. Neurophysiol*, Vol.44, 209-216.

Dellon, E.S.; Keller, K.; Moratz, V. & Dellon, A.L. (1995). *The Journal of Hand Surgery*, Vol.20B, No.1, 44-48.
Dellon, A.L. (1981). Baltimore, Williams and Wilkins.
Dellon, A.L. (1978). *Journal of Hand Surgery*, Vol.3, No.5, 474-481.
Verrillo, R.T. (1962). *The Journal of the Acoustical Society of America*, Vol.34, No.11, 1768-1773.
Verrillo, R.T. (1963). *The Journal of the Acoustical Society of America*, Vol.35, No.12, 1962-1966.
Toma, S. & Nakajima, Y. (1995). *Neuroscience Letters*, Vol.195, 61-63.

Appendix

In order to observe the skin surface deformation under the pressed condition compared with that under the non-pressed condition, we set up the high-speed camera system as shown in Fig. 11. Fig. 12 shows the result of the skin surface deformation where Fig. 12 (a) and (b) are under the non-pressed and the pressed condition, respectively. The ``Before" and ``After" in Fig. 12 denote the surface profiles before and after the force impartment, respectively. The finger deformation during the force impartment is obtained by chasing the slit laser with an assistance of the high-speed camera. The result provides us with the sufficient information on the deformation under both conditions, compared with the point-typed stiffness sensor. An interesting observation is that the deformed shape keeps the similarity between two conditions. This means that the deformation in the lateral direction due to the force impartment is proportional to the deformation in the depth direction. Through Fig. 12, we can confirm that the diameter of deformed area under the pressed condition is almost 2 times less than that under the non-pressed condition for this particular experiment.

Acknowledgment

This work was supported by 2007 Global COE (Centers of Excellence) Program「A center of excellence for an *In Silico* Medicine」in Osaka University.

Modeling Thermoregulation and Core Temperature in Anatomically-Based Human Models and Its Application to RF Dosimetry

Akimasa Hirata
Nagoya Institute of Technology
Japan

1. Introduction

There has been increasing public concern about the adverse health effects of human exposure to electromagnetic (EM) waves. Elevated temperature (1-2°C) resulting from radio frequency (RF) absorption is known to be a dominant cause of adverse health effects, such as heat exhaustion and heat stroke (ACGIH 1996). According to the RF research agenda of the World Health Organization (WHO) (2006), further research on thermal dosimetry of children, along with an appropriate thermoregulatory response, is listed as a high-priority research area. The thermoregulatory response in children, however, remains unclear (Tsuzuki *et al.* 1995, McLaren *et al*. 2005). Tsuzuki suggested maturation-related differences in the thermoregulation during heat exposure between young children and mothers. However, for ethical reasons, systemic work on the difference in thermoregulation between young children and adults has not yet been performed, resulting in the lack of a reliable thermal computational model.

In the International Commission on Non-Ionizing Radiation Protection (ICNIRP) guidelines (1998), whole-body-averaged specific absorption rate (SAR) is used as a metric of human protection from RF whole-body exposure. In these guidelines, the basic restriction of whole-body-averaged SAR is 0.4 W/kg for occupational exposure and 0.08 W/kg for public exposure. The rationale of this limit is that exposure for less than 30 min causes a body-core temperature elevation of less than 1°C if whole-body-averaged SAR is less than 4 W/kg (e.g., Chatterjee and Gandhi 1983, Hoque and Gandhi 1988). As such, safety factors of 10 and 50 have been applied to the above values for occupational and public exposures, respectively, to provide adequate human protection.

Thermal dosimetry for RF whole-body exposure in humans has been conducted computationally (Bernardi *et al*. 2003, Foster and Adair 2004, Hirata *et al*. 2007b) and experimentally (Adair *et al*. 1998, Adair *et al*. 1999). In a previous study (Hirata *et al*. 2007b), for an RF exposure of 60 min, the whole-body-averaged SAR required for body-core temperature elevation of 1°C was found to be 4.5 W/kg, even in a man with a low rate of perspiration. Note that the perspiration rate was shown to be a dominant factor influencing the body-core temperature due to RF exposure. The SAR value of 4.5 W/kg corresponds to a

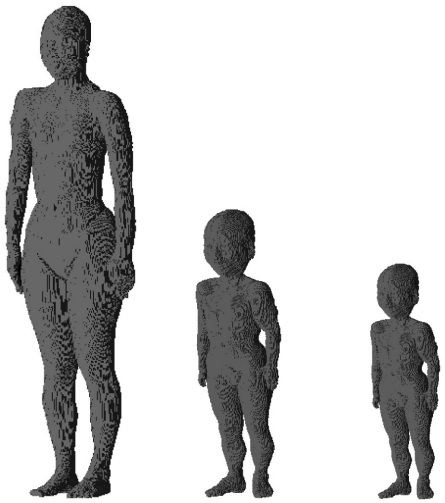
safety factor of 11, as compared with the basic restriction in the ICNIRP guidelines, which is close to a safety margin of 10. However, the relationship between the whole-body-averaged SAR and body-core temperature elevation has not yet been investigated in children.
In this chapter, a thermal computational model of human adult and child has been explained. This thermal computational model has been validated by comparing measured temperatures when exposed to heat in a hot room (Tsuzuki *et al.* 1995, Tsuzuki 1998). Using the thermal computation model, we calculated the SAR and the temperature elevation in adult and child phantoms for RF plane-wave exposures.

2. Model and Methods

2.1 Human Body Phantom

Figure 1 illustrates the numeric Japanese female, 3-year-old and 8-month-old child phantoms. The whole-body voxel phantom for the adult female was developed by Nagaoka *et al.* (2004). The resolution of the phantom was 2 mm, and the phantom was segmented into 51 anatomic regions. The 3-year-old child phantom (Nagaoka *et al.* 2008) was developed by applying a free-form deformation algorithm to an adult male phantom (Nagaoka *et al.* 2004). In the deformation, a total of 66 body dimensions was taken into account, and manual editing was performed to maintain anatomical validity. The resolution of these phantoms was kept at 2 mm. For European and American adult phantoms, e.g., see the literatures by Dimbylow (2002, 2005) and Mason et al. (2000). These phantoms have the resolution of a few millimeter.

In Section 3.1, we compare the computed temperatures of the present study with those measured by Tsuzuki (1998). Eight-month-old children were used in her measurements.



(a) (b) (c)

Fig. 1. Anatomically based human body phantoms of (a) a female adult, (b) a 3-year-old child, and (c) an 8-month-old child.

	H[m]	W[kg]	S[m²]	S/W[m²/kg]
Female	1.61	53	1.5	0.029
3 year old	0.90	13	0.56	0.043
8 month old	0.75	9	0.43	0.047

Table 1. Height, weight, and surface area of Japanese phantoms.

Thus, we developed an 8-month-old child phantom from a 3-year-old child by linearly scaling using a factor of 0.85 (phantom resolution of 1.7 mm). The height, weight, and surface area of these phantoms are listed in Table 1. The surface area of the phantom was estimated using a formula proposed by Fujimoto and Watanabe (1968).

2.2 Electromagnetic Dosimetry

The FDTD method (Taflove & Hagness, 2003) is used for calculating SAR in the anatomically based human phantom. The total-field/scattered-field formulation was applied in order to generate a proper plane wave. To incorporate the anatomically based phantom into the FDTD method, the electrical constants of the tissues are required. These values were taken from the measurements of Gabriel (1996). The computational region has been truncated by applying a perfectly matched layer-absorbing boundary. For harmonically varying fields, the SAR is defined as

$$\mathrm{SAR} = \frac{\sigma}{2\rho} |\mathbf{E}|^2 = \frac{\sigma}{2\rho}(|\hat{E}_x|^2 + |\hat{E}_y|^2 + |\hat{E}_z|^2) \tag{1}$$

where $\hat{E}_x$, $\hat{E}_y$, and $\hat{E}_z$ are the peak values of the electric field components, and σ and ρ are the conductivity and mass density, respectively, of the tissue.

2.3 Thermal Dosimetry

The temperature elevation in numeric human phantoms was calculated using the bioheat equation (Pennes 1948). A generalized bioheat equation is given as:

$$\begin{aligned} C(\mathbf{r})\rho(\mathbf{r})\frac{\partial T(\mathbf{r},t)}{\partial t} &= \nabla\cdot(K(\mathbf{r})\nabla T(\mathbf{r},t)) + \rho(\mathbf{r})SAR(\mathbf{r}) + A(\mathbf{r},t) \\ &\quad - B(\mathbf{r},t))(T(\mathbf{r},t) - T_B(\mathbf{r},t)) \end{aligned} \tag{2}$$

where $T(\mathbf{r},t)$ and $T_B(\mathbf{r},t)$ denote the temperatures of tissue and blood, respectively, C is the specific heat of tissue, K is the thermal conductivity of tissue, A is the basal metabolism per unit volume, and B is a term associated with blood perfusion. The boundary condition between air and tissue for Eq. (2) is expressed as:

$$-K(\mathbf{r})\frac{\partial T(\mathbf{r},t)}{\partial n} = H(\mathbf{r})\cdot(T_s(\mathbf{r},t) - T_e(t)) + S \quad \left(\text{at } \mathbf{r}, T_s(\mathbf{r},t)\right) \tag{3}$$

$$SW\left(\mathbf{r},T_s(\mathbf{r},t)\right)=P_{ins}+SW_{act}\left(\mathbf{r},T_s(\mathbf{r},t)\right) \tag{4}$$

where H, T_s, and T_e denote, respectively, the heat transfer coefficient, the body surface temperature, and the air temperature. The heat transfer coefficient includes the convective and radiative heat losses. SW is comprised of the heat losses due to perspiration SW_{act} and insensible water loss P_{ins}. T_e is chosen as 28°C, at which thermal equilibrium is obtained in a naked man (Hardy & Du Bois 1938).

In order to take into account the body-core temperature variation in the bioheat equation, it is reasonable to consider the blood temperature as a variable of time $T_B(\mathbf{r},t)=T_B(t)$. Namely, the blood temperature is assumed to be uniform over the whole body, since the blood circulates throughout the human body in 1 min or less (Follow and Neil 1971). The blood temperature variation is changed according to the following equation (Bernardi *et al*. 2003, Hirata & Fujiwara, 2009):

$$T_B(t)=T_{B0}+\int_t \frac{Q_{BT}(t)-Q_{BT}(0)}{C_B \rho_B V_B}\mathrm{d}t \tag{5}$$

where C_B (= 4,000 J/kg·°C) is the specific heat, ρ_B (= 1,050 kg/m^3) is the mass density, and V_B is the total volume of blood. V_B is chosen as 700 ml, 1,000 ml, and 5,000 ml for the 8-month-old and 3-year-old child phantoms and the adult phantom (ICRP 1975), respectively. Q_{BT} is the rate of heat acquisition of blood from body tissues given by the following equation;

$$Q_{BT}(t)=\int_V B(t)(T_B(t)-T(\mathbf{r},t))dV. \tag{6}$$

Thorough discussion on blood temperature variation in the bioheat equation can be found in Hirata & Fujiwara (2009).

2.4 Thermal Constants of Human Tissues

The thermal constants of tissues in the adult were approximately the same as those reported in our previous study (Hirata *et al*. 2006a), as listed in Table 2. These are mainly taken from Cooper and Trezek (1971). The basal metabolism was estimated by assuming it to be proportional to the blood perfusion rate (Gordon *et al*. 1976), as Bernardi *et al*. did (2003). In the thermally steady state without heat stress, the basal metabolism is 88 W. This value coincides well with that of the average adult female. The basal metabolic rate in the 8-month-old and 3-year-old child phantoms were determined by multiplying the basal metabolic rate of the adult by factors of 1.7 and 1.8, respectively, so that the basal metabolism in these child phantoms coincides with those of average Japanese (Nakayama and Iriki 1987): 47 W and 32 W for 3-year-old and 8-month-old children. Similarly, based on a study by Gordon *et al*. (1976), the same coefficients were multiplied by the blood perfusion rate. The specific heat and thermal conductivity of tissues were assumed to be identical to those of an adult, because the difference in total body water in the child and adult is at most a few percent (ICRP 1975).

The heat transfer coefficient is defined as the summation of heat convection and radiation. The heat transfer coefficient between skin and air and that between organs and internal air are denoted as H_1 and H_2, respectively. Without heat stress, the following equation is maintained:

tissue	K[W m^{-1} ^{o}C]	C[J kg^{-1} ^{o}C]	ρ[kg m^{-3}]	B[W m^{-3} ^{o}C]	A[W m^{-3}]
air	0	0	0	0	0
Internal air	0	0	0	0	0
skin	0.27	3600	1125	1700	1620
muscle	0.40	3800	1047	2000	480
fat	0.22	3000	500	1500	300
bone (cortical)	0.37	3100	1990	3400	610
bone (cancellous)	0.41	3200	1920	3300	590
nerve (spine)	0.46	3400	1038	40000	7100
gray matter	0.57	3800	1038	40000	7100
CSF	0.62	4000	1007	0	0
eye (aqueous humor)	0.58	4000	1009	0	0
eye (lens)	0.40	3600	1053	0	0
eye (sclera/wall)	0.58	3800	1026	75000	22000
heart	0.54	3900	1030	54000	9600
liver	0.51	3700	1030	68000	12000
lung (outer)	0.14	3800	1050	9500	1700
kidneys	0.54	4000	1050	270000	48000
intestine (small)	0.57	4000	1043	71000	13000
intestine (large)	0.56	3700	1043	53000	9500
gall bladder	0.47	3900	1030	9000	1600
spleen	0.54	3900	1054	82000	15000
stomach	0.53	4000	1050	29000	5200
pancreas	0.52	4000	1045	41000	7300
blood	0.56	3900	1058	0	0
body fluid	0.56	3900	1010	0	0
bile	0.55	4100	1010	0	0
glands	0.53	3500	1050	360000	64000
bladder	0.43	3200	1030	9000	160
testicles	0.56	3900	1044	360000	64000
lunch	0.56	3900	1058	0	0
adrenals	0.42	3300	1050	270000	48000
Tendon	0.41	3300	1040	9000	1600

Table 2. Thermal constants of biological tissues.

$$\int_v A(\mathbf{r})dv = \int_S P_{ins}(\mathbf{r})dS + \int_S H(\mathbf{r},t)(T(\mathbf{r},t) - T_a)dS \tag{6}$$

where T_a is the air temperature. The air temperature was divided into the average room temperature T_{a1} (28°C) and the average body-core temperature T_{a2}, corresponding to H_1 and H_2, respectively.

Insensible water loss is known to be roughly proportional to the basal metabolic rate: 20 ml/kg/day for an adult, 40 ml/kg/day for a 3-year-old child, and 50 ml/kg/day for an 8-month-old child (Margaret *et al*. 1942). For the weight listed in Table 1, the insensible water

	P_{ins1}[W]	P_{ins2}[W]	H_1[W m^{-2} ^{o}C]	H_2[W m^{-2} ^{o}C]
Female	20.3	8.7	4.1	26.0
3 year old	10.7	4.6	4.0	13.1
8 month old	8.9	3.8	3.9	13.3

Table 3. Insensible water loss and heat transfer rate in the adult female and 3-year-old and 8-month-old children.

losses in the phantoms of a female adult, a 3-year-old child, and an 8-month-old child are 29 W, 15.3 W, and 12.7 W, respectively. Note that the insensible water loss consists of the loss from skin (70%) and the loss from the lungs through breathing (30%) (Karshlake 1972). The heat loss from the skin to the air P_{ins1} and that from the body-core and internal air P_{ins2} are calculated as listed in Table 3.

For the human body, 80% of the total heat loss is from the skin and 20% is from the internal organs (Nakayama & Iriki 1987). Thus, the heat loss from the skin is 68 W in the adult female, 37.6 W in the 3-year-old child, and 25.6 W in the 8-month-old child. Similarly, the heat loss from the internal organs is 17 W in the adult female, 9.4 W in the 3-year-old child, and 6.4 W in the 8-month-old child. Based on the differences among these values and the insensible water loss presented above, we can obtain the heat transfer coefficients, as listed in Table 3.

In order to validate the thermal parameters listed in Table 3, let us compare the heat transfer coefficients between skin and air obtained here to those reported by Fiala *et al.* (1999). In the study by Fiala *et al.* (1999), the heat transfer coefficient is defined allowing for the heat transfer with insensible water loss. Insensible water loss is not proportional to the difference between body surface temperature and air temperature, as shown by Eq. (3), and therefore should not be represented in the same manner for wide temperature variations. Thus, the equivalent heat transfer coefficient due to insensible water loss was calculated at 28°C. For P_{ins1} as in Table 3, the heat transfer coefficient between the skin and air in the adult female was calculated as 1.7 W/m^2/°C. The heat transfer coefficient from the skin to the air, including the insensible heat loss, was obtained as 5.7 W/m^2/°C. However, the numeric phantom used in this chapter is descretized by voxels, and thus the surface of the phantom is approximately 1.4 times larger than that of an actual human (Samaras *et al.* 2006). Considering the difference in the surface area, the actual heat transfer coefficient with insensible water loss is 7.8 W/m^2/°C, which is well within the uncertain range summarized by Fiala *et al*. (1999).

In Sec. 3, we consider the room temperature of 38°C, in addition to 28°C, in order to allow comparison with the temperatures measured by Tsuzuki et al. (1998). The insensible water loss c assumed to be the same as that at 28°C (Karshlake 1972). The heat transfer coefficient from the skin and air is chosen as 1.4 W/m^2/°C (Fiala *et al*. 1999). Since the air velocity in the lung would be the range of 0.5 and 1.0 m/s, the heat transfer coefficient H_2 can be estimated as 5 – 10 W/m^2/°C (Fiala *et al*. 1999). However, this uncertainty does not influence the computational results in the following discussion, because the difference between the

internal air temperature and the body-core temperature is at most a few degrees, resulting in a marginal contribution to heat transfer between the human and air (see Eq. (3)).

2.5 Thermoregulatory Response in Adult and Child

For a temperature elevation above a certain level, the blood perfusion rate was increased in order to carry away excess heat that was produced. The variation of the blood perfusion rate in the skin through vasodilatation is expressed in terms of the temperature elevation in the hypothalamus and the average temperature increase in the skin. The phantom we used in the present study is the same as that used in our previous study (Hirata *et al*. 2007b). The variation of the blood perfusion rate in all tissues except for the skin is marginal. This is because the threshold for activating blood perfusion is the order of 2°C, while the temperature elevation of interest in the present study is at most 1°C, which is the rationale for human protection from RF exposure (ICNIRP, 1998).

Perspiration for the adult is modeled based on formulas presented by Fiala *et al*. (2001). The perspiration coefficients are assumed to depend on the temperature elevation in the skin and/or hypothalamus. An appropriate choice of coefficients could enable us to discuss the uncertainty in the temperature elevation attributed to individual differences in sweat gland development:

$$SW(\mathbf{r},t) = \{W_S(\mathbf{r},t)\Delta T_S(t) + W_H(\mathbf{r},t)(T_H(t) - T_{Ho})\} / S \times 2^{(T(\mathbf{r})-T_0(\mathbf{r}))/10} \tag{7}$$

$$W_s(\mathbf{r},t) = \alpha_{11}\tanh(\beta_{11}T_s(\mathbf{r},t) - T_{so}(\mathbf{r})) - \beta_{10}) + \alpha_{10} \tag{8}$$

$$W_H(\mathbf{r},t) = \alpha_{21}\tanh(\beta_{21}T_s(\mathbf{r},t) - T_{so}(\mathbf{r})) - \beta_{20}) + \alpha_{20} \tag{9}$$

where S is the surface area of the human body, and W_S and W_H are the weighting coefficients for perspiration rate associated with the temperature elevation in the skin and hypothalamus. Fiala *et al*. (2001) determined the coefficients of α and β for the average perspiration rate based on measurements by Stolowijk (1971). In addition to the set of coefficients in Fiala *et al*. (2001), we determined the coefficients for adults with higher and lower perspiration rates parametrically (Hirata *et al*. 2007b). In this chapter, we used these sets of parameters.

Thermoregulation in children, on the other hand, has not been adequately investigated yet. In particular, perspiration in children remains unclear (Bar-Or 1980, Tsuzuki *et al*. 1995). Therefore, heat stroke and exhaustion in children remain topics of interest in pediatrics (McLaren *et al*. 2005). Tsuzuki *et al*. (1995) and Tsuzuki (1998) found greater water loss in children than in mothers when exposed to heat stress. Tsuzuki *et al*. (1995) attributed the difference in water loss to differences in maturity level in thermophysiology (See also McLaren *et al*. 2005). However, a straightforward comparison cannot be performed due to physical and physiological differences. A number of studies have examined physiological differences among adults, children, and infants (e.g., Fanaroff *et al*. 1972, Stulyok *et al*. 1973). The threshold temperature for activating perspiration in infants (younger than several weeks of age) is somewhat higher than that for adults (at most 0.3°C). On the other hand, the threshold temperature for activating perspiration in children has not yet been investigated. In the present study, we assume that the threshold temperature for activating perspiration is the same in children and adults. Then, we will discuss the applicability of the

present thermal model of an adult to an 8-month-old child by comparing the computed temperature elevations of the present study with those measured by Tsuzuki (1998).

3. Temperature Variation in the Adult and Child Exposed to Hot Room

3.1 Computational Temperature Variation in Adult

Our computational result will be compared with those measured by Tsuzuki (1998). The scenario in Tsuzuki (1998) was as follows: 1) resting in a thermoneutral room with temperature of 28°C and a relative humidity of 50%, 2) exposed to a hot room with temperature of 35°C and a relative humidity of 70% for 30 min., and 3) resting in a themoneutral room.

First, the perspiration model of Eq. (7) with the typical perspiration rate defined in Hirata *et al*. (2007b) is used as a fundamental discussion. Figures 2 and 3 show the time course of the average skin and body-core temperature elevations, respectively, in the adult exposed to a hot room, together with those for an 8-month-old child. As shown in Fig. 2, the computed average temperature elevation of the adult skin was 1.5°C for a heat exposure time of 30 min., which is in excellent agreement with the measured data of 1.5°C. From Fig. 3, the measured and computed body-core temperatures in the adult female were 0.16°C and 0.19°C, respectively, which are well within the standard deviation of 0.05°C obtained in the measurement (Tsuzuki 1998). In this exposure scenario, the total water loss for an adult was 50 g/m² in our computation, whereas it was 60 g/m² in the measurements.

In order to discuss the uncertainty of temperature elevation due to the perspiration, the temperature elevations in the adult female is calculated for different perspiration parameters given in Hirata *et al*. (2007b). From Table 4(a), the set of typical perspiration parameters works better than other sets for determining the skin temperature. However, the body-core temperature for the typical perspiration rate was larger than that measured by

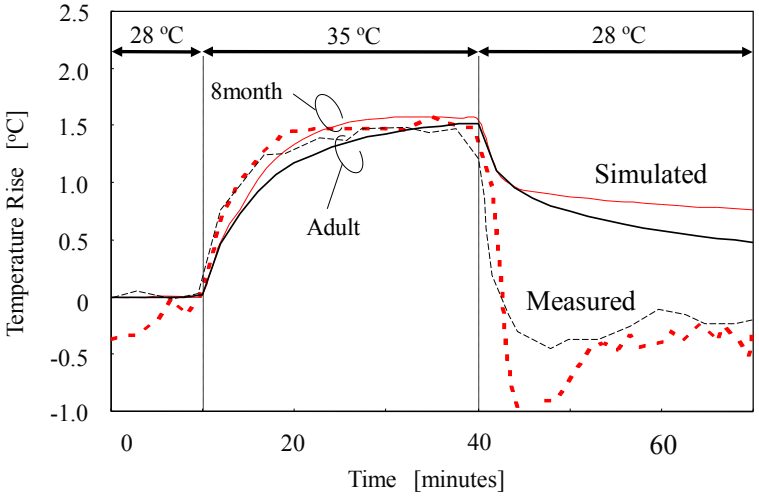


Fig. 2. Time course of average skin temperature elevations in the adult and the 8-month-old child.

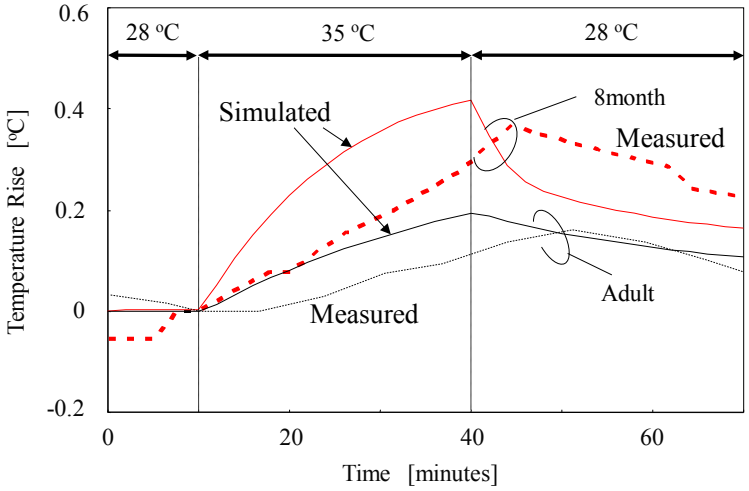


Fig. 3. Time course of body-core temperature elevations in the adult and the 8-month-old child.

child	low	typical	high	measured
skin	2.0	1.5	0.98	1.5
body-core	0.50	0.41	0.32	0.37

(a)

adult	low	typical	high	measured
skin	1.7	1.5	1.2	1.5
body-core	0.21	0.19	0.17	0.16

(b)

Table 4. Temperature elevation in (a) adult and (b) child exposed to a hot room for different perspiration parameters.

Tsuzuki (1998). This is thought to be caused by the decrease in body-core temperature before heat exposure (0-10 min. in Fig. 3).

3.2 Computational Temperature Variation in Adult

Since thermal physiology in children has not been sufficiently clarified, we adapted the thermal model of the adult to the 8-month-old child for the fundamental discussion. The time courses of the average skin and body-core temperature elevations in the 8-month-old child are shown in Figs. 2 and 3, respectively. As shown in Fig. 2, the computed average temperature elevation of the skin of a child at 30 min. of heat exposure was 1.5°C, which is the same as that for an adult as well as the measured data. The measured and computed

body-core temperatures in the child were 0.37°C and 0.41°C, respectively. This difference of 0.04°C is well within the standard deviation of 0.1°C obtained in the measurement (Tsuzuki 1998). In our computation, the total perspiration of the child was 100 g/m², whereas in the measurements, the value was 120 g/m²; the same tendency was observed for the adult.
Table 4(b) lists the temperature elevations in the 8-month-old child for different perspiration parameters which were the same as we did for the adult. As with the adult, the model with the typical perspiration rate works better than the other models.

3.3 Discussion

From Fig. 2, an abrupt temperature decrease in the recovery phase after exposure in a hot room is observed in the measured data but is not observed in the computed data. The reason for this difference is discussed by Tsuzuki (1998), who reported that wet skin is suddenly cooled in a thermoneutral room. This phenomenon cannot be taken into account in our computational modeling or boundary condition (Eqs. (3) and (4)). Such phenomenon would be considered with other boundary conditions, e.g., a formula by Ibrahiem et al (2005). However, this is beyond the scope of the present study, since our concern is on the temperature elevation in the body.
As shown by Fig. 3, the computed body-core temperature increases more quickly than the measured temperature. The time at which the body-core temperature became maximal in the measurement was retarded by 11 min. for the adult female whereas 5 min. for the child. There are two main reasons for this retard. One is caused by our assumption that the blood temperature is spatially constant and varies instantaneously (See Eq. (5)) based on the fact that the blood circulates throughout the body in 1 min. The other reason is that, in the experiment, we consider the blood temperature elevation instead of that in the rectum. The blood temperature in the rectum increases primarily due to blood circulation at an elevated temperature. In Hirata *et al*. (2007b), the temperature elevation in the hypothalamus, which is located in the brain and considerable as body core, was shown to be retarded by a few minutes relative to the blood temperature elevation. The difference of the retard between the adult and the child is attributed to the smaller body dimensions and greater blood perfusion rate of the child compared to those of the adult. The assumption in Eq. (5) was validated for rabbits (Hirata *et al* 2006b), the body dimensions of which are much smaller than those of a human. In addition, the blood perfusion rate of the rabbit is four times greater than that of the human adult, considering the difference in basal metabolic rate (Gordon *et al*. 1976). From this aspect, the thermal computational model developed here works better for the child than for the adult. This retard in the body-core temperature elevation would give a conservative estimation from the standpoint of thermal dosimetry. In the following discussion, we consider not the temperature elevations at a specific time, but rather the peak temperatures for the measured data.
From table 4, we found some difference in total water loss between adult and child. One of the main reasons for this difference is thought to be the difference in race. The volunteers in the study by Tsuzuki (1998) were Japanese, whereas the data used for the computational modeling was based primarily on American individuals (Stolowijk, 1971). Roberts *et al*. (1970) reported that the number of active sweat glands in Korean individuals (similar to Japanese) is 20-30% greater than that in European individuals (similar to American). In addition, the perspiration rate in Japanese individuals is thought to be greater than that in American individuals, which was used to derive the perspiration formula.

Even though we applied a linear scaling when developing the 8-month-old child phantom, its influence on the temperature looks marginal. This is because the body-temperature is mainly determined by the heat balance between the energy produced through metabolic processes, energy exchange with the convection, and the energy storage in the body (Adair and Black 2003, Ebert et al 2005, Hirata et al 2008). Especially, the anatomy of the phantom does not influence from the heat balance equation in the previous studies, suggesting that our approximation of the linear scaling was reasonable.
Tsuzuki et al. (1995) expected a maturity-related difference in themoregulatory response, especially for perspiration, between the adult and the child. The present study revealed two key findings. The first is the difference in the insensible water loss, which was not considered by Tsuzuki et al (1995). The other is the nonlinear perspiration response controlled by the temperature elevations in the skin and body core (Eq. (7)). In addition to these physiological differences, the larger body surface area-to-mass ratio generated more sweat in the child. The computational results of the present study considering these factors are conclusive and are consistent with the measured results.
From the discussion above, the validity of the thermal model for the adult was confirmed. In addition, the thermal model for the 8-month-old child is found to be reasonably the same as that of the adult.

4. Body-core Temperature Elevation in Adult and Child for RF Whole-body Exposures

4.1 Computational Results for Temperature Elevation for RF Exposures

An anatomically based human phantom is located in free space. As a wave source, a vertically polarized plane wave was considered; the plane wave was thus incident to a human phantom from the front. Female adult and 3-year-old child phantoms are considered in this section. The reason for using the 3-year-old child phantom is that this phantom is more anatomically correct than the 8-month-old child phantom, which was developed for comparison purposes in Section 3.1 simply by reducing the adult phantom.
The whole-body-averaged SAR has two peaks for plane-wave exposure at the ICNIRP reference level; more precisely, it becomes maximal at 70 MHz and 2 GHz in the adult female phantom and 130 MHz and 2 GHz in the 3-year-old child phantom. The first peak is caused by whole-body resonance in the human body. The latter peak, on the other hand, is caused by the relaxation of the ICNIRP reference level with the increase in frequency. Note that the power density at the ICNIRP reference level is 2 W/m^2 at 70 MHz and 130 MHz and 10 W/m^2 at 2 GHz. The whole-body-averaged SAR in the adult female phantom was 0.069 W/kg at 70 MHz and 0.077 W/kg at 2 GHz, whereas that in the 3-year-old child phantom was 0.084 W/kg at 130 MHz and 0.108 W/kg at 2 GHz. The uncertainty of whole-body SAR, attributed to the boundary conditions and phantom variability, has been discussed elsewhere (e.g., Findlay and Dimbylow 2006, Wang *et al*. 2006, Conil *et al*. 2008). In order to clarify the effect of frequency or the SAR distribution on the body-core temperature, we normalized the whole-body-averaged SAR as 0.08 W/kg while maintaining the SAR distribution. The normalized SAR distributions at these frequencies are illustrated in Fig. 4. As this figure shows, the SAR distributions at these frequencies are quite different (Hirata *et al*. 2007a). EM absorption occurs over the whole body at the resonance frequency. Compared

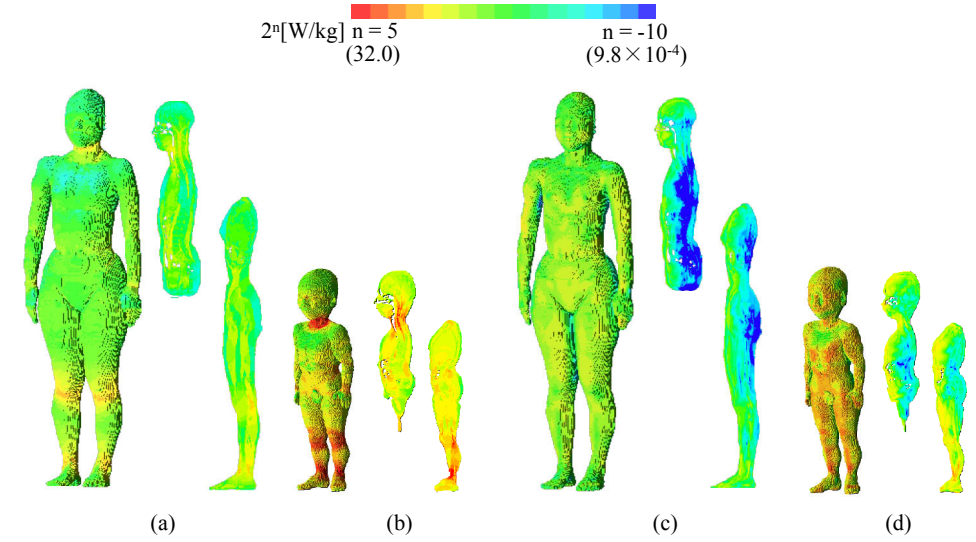


Fig. 4. SAR distributions in the adult female at (a) 70 MHz and (b) 2 GHz and those in the 3-year-old child model at (c) 130 MHz and (d) 2 GHz.

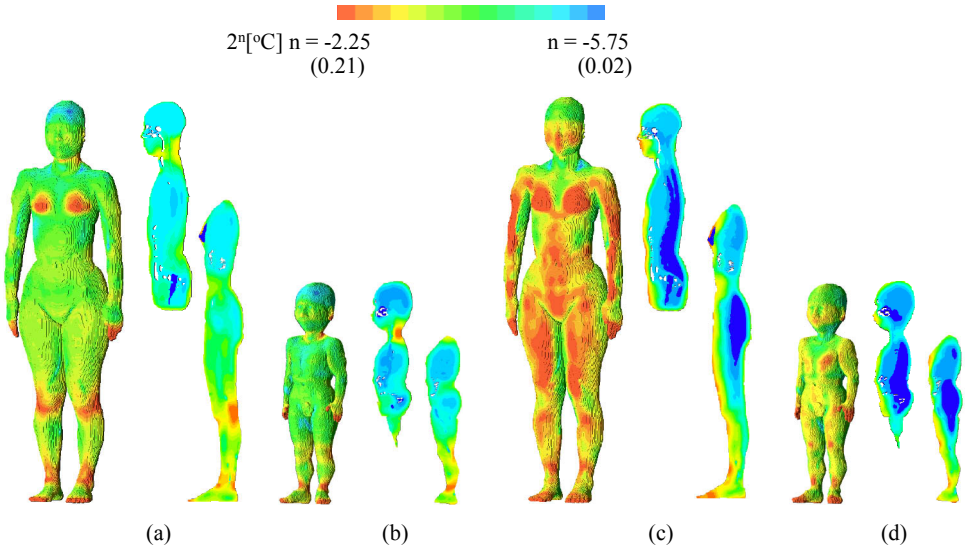


Fig. 5. Temperature elevation distributions in the adult female at (a) 70 MHz and (b) 2 GHz and those in 3-year-old child model at (c) 130 MHz and (d) 2 GHz.

with the distribution at 2 GHz, the absorption around the body core cannot be neglected. In contrast, the SAR distribution is concentrated around the body surface at 2 GHz.

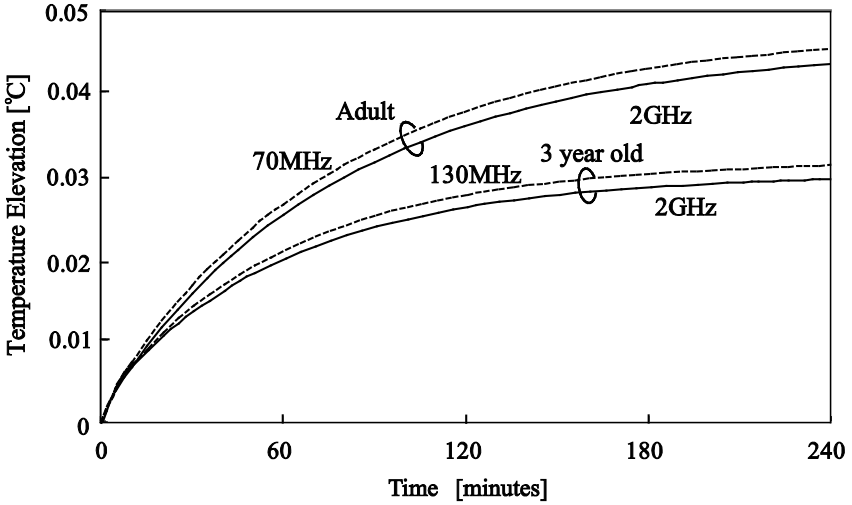


Fig. 6. Temperature elevation in the adult and 3-year-old child at the whole-body averaged SAR of 0.08 W/kg. Exposure duration was 4 hour.

The temperature elevation distributions in a human are illustrated in Fig. 5 for the whole-body-averaged SAR of 0.08 W/kg. The duration of exposure was chosen as 60 min. As shown in Fig. 5, the SAR and temperature elevation distributions are similar. For example, the temperature elevation at the surface becomes larger at 2 GHz. However, the temperature in the body core (e.g., in the brain) is uniform at approximately 0.03°C. This is because the body core is heated mainly due to the circulation of warmed blood (Hirata *et al*. 2007b).
Figure 6 shows the time courses of the temperature elevation in the adult and the child at a whole-body-averaged SAR of 0.08 W/kg. This figure indicates that it took 4 hours to reach the thermally steady state. At 4 hours, the body-core temperature increases by 0.045°C at 65 MHz and 0.041°C at 2 GHz. This confirms the finding in our previous study (Hirata *et al*. 2007b) that whole-body-averaged SAR influences the body-core temperature elevation regardless of the frequency or SAR distribution. On the other hand, the temperature elevation in the child was 0.031°C at 130 MHz and 0.029°C at 2 GHz, which was 35% smaller than that in the adult.

4.2 Discussion

We found in Fig. 6 significant difference of body-core temperature elevation between adult and child. In order to clarify the main factor influencing temperature elevation, let us consider an empirical heat balance equation for the human body as given by Adair et al (1998):

$$M + P_{RF} - P_t = P_s \tag{10}$$

where M is the rate at which thermal energy is produced through metabolic processes, P_{RF} is the RF power absorbed in the body, P_t is the rate of heat transfer at the body surface, and P_S is the rate of heat storage in the body.
More specific expression for (10) is given in the following equation based on on (2) and (3).

$$\begin{aligned}&\int_0^t\int_V \big(A(\boldsymbol{r},t)-A(\boldsymbol{r},0)\big)dVdt+\int_0^t\int_V SAR(\boldsymbol{r})\cdot\rho(\boldsymbol{r})dVdt\\&-\left\{\int_0^t\int_S h(\boldsymbol{r})\big(T(\boldsymbol{r},t)-T(\boldsymbol{r},0)\big)dSdt+\int_0^t\int_S SW(t)dSdt\right\}\\&=\int_V\big(T(\boldsymbol{r},t)-T(\boldsymbol{r},0)\big)\cdot\rho(\boldsymbol{r})\cdot C(\boldsymbol{r})dV\end{aligned}\tag{11}$$

The first term of (11) represents the energy due to the metabolic increment caused by the temperature elevation. In this chapter, this term is ignored for the sake of simplicity, since that energy evolves secondarily via the temperature elevation due to RF energy absorption.
For (11), we apply the following two assumptions: 1) the temperature distribution is assumed to be uniform over the body, and 2) the SAR distribution is assumed to be uniform. Then, we obtained the following equation:

$$\begin{aligned}\big(T(t)-T_0\big)\cdot\rho_{WBave}\cdot V\cdot C_{WBave}&=\int_0^t SAR_{WBave}\cdot\rho_{WBave}\cdot Vdt\\&-\int_0^t\big(T(t)-T_0\big)dt\cdot\left\{\int_S H(\mathbf{r})dS+\int_S sw(t)dS\right\}\end{aligned}\tag{12}$$

where W is the weight of the model [kg], SAR_{WBave} is the WBA-SAR[W/kg], H is the mean value of the heat transfer coefficient between the model and air [W/m^2 ^{o}C], C_{WBave} is the mean value of the specific heat [J/kg ^{o}C]. $sw(t)$ is a coefficient identical to $SW(t)$ except that the temperature is assumed to be uniform; $SW(t)=sw(t)(T(t)-T_0)$.
By differentiating (12), the temperature elevation is obtained as

$$T(t)=T_0+\frac{W\cdot SAR_{WBave}}{\int_S H(\vec{r})dS+\int_S sw(t)dS}\left(1-\exp\left(-\frac{\int_S H(\vec{r})dS+\int_S sw(t)dS}{W\cdot C_{WBave}}t\right)\right).\tag{13}$$

As can be seen from Eq. (13), the ratio of surface area to the weight is considered dominant factor influencing the temperature elevation. The total power deposited in the human is proportional to weight, as we fixed the whole-body-averaged SAR as 0.08 W/kg. On the other hand, the power loss from the human via perspiration is proportional to the surface area, because perspiration of the child can be considered as identical to that of the adult. As listed in Table 1, the ratio of the surface to the weight is 0.029 m^2/kg for the adult, whereas that of the child is 0.043 m^2/kg. This difference of 47% coincides reasonably with the fact that body-core temperature elevation in the child is 35% smaller than that in the adult. Marginal inconsistency in these ratios would be caused by the nonlinear response of the perspiration as given by Eq. (7).
For higher whole-body-averaged SAR(~ 4 W/kg), the ratio of temperature elevations in the adult to that of the child was 42%, which was closer to their body surface area-to-weight

ratio of 47% than that in the case for the whole-body-averaged SAR at 0.08 W/kg. For higher temperature elevation, the effect of body-core temperature elevation on the perspiration rate is much larger than that due to skin temperature elevation. In addition, the perspiration rate becomes almost saturated. Therefore, the thermal response is considered to be linear with respect to the body-core temperature increase.
It is worth commenting on the difference between this scenario and that described in Section 3.1. In Section 3.1, the body-core temperature elevation in the child was larger than that in the adult for the heat stress caused by higher ambient temperature. The thermal energy applied to the body via ambient temperature is proportional to the surface area of the body. On the other hand, in this scenario, the thermal energy moves from the surface area of the body to the air, because the body is cooled via the ambient temperature. For these two cases, the main factor varying the body-core temperature is the same as the body surface area-to-weight ratio. However, the magnitude relation between the body surface and the ambient temperatures was reversed.

5. Conclusion

The temperature elevations in the anatomically-based human phantoms of adult and 3-year-old child were calculated for radio-frequency whole-body exposure. The rationale for this investigation was that further work on thermal dosimetry of children with appropriate thermoregulatory response is listed as one of the high priority researches in the RF research agenda by the WHO (2006). However, systemic work on the difference in the thermoregulation between young child and adult has not been performed, mainly because of ethical reason for experiment and the lack of reliable thermal computational model.
In this chapter, we discussed computational thermal model in the child which is reasonable to simulate body-core temperature elevation in child phantoms by comparing with experimental results of volunteers when exposed to hot ambient temperature. From our computational results, it was found to be reasonable to consider that the thermal response even in the 8-month-old child was almost the same as that in the adult. Based on this finding, we calculated the body-core temperature elevation in the 3-year-old child and adult for plane wave exposure at the ICNIRP basic restriction. The body-core temperature elevation in the 3-year-old child phantom was 40% smaller than that of the adult, which is attributed to the ratio of the body surface area to the mass. This rationale for this difference has been explained by deriving a simple formula for estimating core temperature.

6. References

Adair, E. R.; Kelleher, S. A., Mack, G. W. & Morocco, T. S. (1998). Thermophysiological responses of human volunteers during controlled whole-body radio frequency exposure at 450 MHz, *Bioelectromagnetics*, Vol.19, pp. 232-245

Adair, E. R.; Cobb, B. L., Mylacraine, K. S. & Kelleher, S. A. (1999) Human exposure at two radio frequencies (450 and 2450 MHz): Similarities and differences in physiological response, *Bioelectromagnetics*, Vol.20, pp. 12-20

Adair, E. R. & Black, D. R. (2003). Thermoregulatory responses to RF energy absorption, Bioelectromagnetics, Vol.24 (Suppl. 6), pp.S17-S38

American Conference of Government Industrial Hygienists (ACGIH) (1996) Threshold limit values for chemical substances and physical agents and biological exposure indices (Cincinnati OH)

Bar-Or, O.; Dotan, R.; Inbar, O.; Rotshtein, A. & Zonder, H. (1980) Voluntary hypohydration in 10 to 12 year old boys *J. Appl. Physiol.*, Vol.48, pp.104-108

Bernardi, P; Cavagnaro, M.; Pisa, S. & Piuzzi, E. (2003) Specific absorption rate and temperature elevation in a subject exposed in the far-field of radio-frequency sources operating in the 10-900-MHz range *IEEE Trans. Biomed. Eng*., vol.50, pp. 295-304

Conil, E; Hadjem, A; Lacroux, Wong, M.-F. & Wiart, J. (2008) Variability analysis of SAR from 20 MHz to 2.4 GHz for different adult and child models using finite-difference time-domain *Phys. Med. Biol.* Vol.53, pp. 1511-1525

Cooper, T. E. & Trezek, G. J. (1971) Correlation of thermal properties of some human tissue with water content, *Aerospace Med*., Vol. 50, pp. 24-27

Chatterjee, I. & Gandhi, O. P. (1983) An inhomogeneous thermal block model of man for the electromagnetic environment *IEEE Trans. Biomed. Eng*., Vol.30, pp. 707-715

Dimbylow, P. J. (2002). Fine resolution calculations of SAR in the human body for frequencies up to 3 GHz *Phys. Med. Biol.*, Vol.47,pp. 2835-2846

Dimbylow, P. (2005). Resonance behavior of whole-body averaged specific absorption rate (SAR) in the female voxel model, NAOMI. *Phys. Med. Biol.*, vol.50, pp.4053-4063.

Douglas, H. K. (1977) Handbook of Physiology, Sec. 9, Reactions to environmental agents MD: American Physiological Society

Fanaroff, A. A.; Wald, M.; Gruber, H. S. & Klaus, M. H. (1972) Insensible water loss in low birth weight infants *Pediatrics* , Vol. 50, pp. 236-245

Fiala, D.; Lomas, K. J. & Stohrer, M. (1999) A computer model of human thermoregulation for a wide range of environmental conditions: the passive system, *J Appl Physiol*, Vol. 87, pp. 1957-1972

Fiala, D.; Lomas, K. J. & Stohrer, M. (2001) Computer prediction of human thermoregulation and temperature responses to a wide range of environmental conditions, *Int J Biometeorol*, Vol. 45, pp. 143-159

Findlay, R. P. & Dimbylow, P. J. (2006) Variations in calculated SAR with distance to the perfectly matched layer boundary for a human voxel model, *Phys. Med. Biol.*, Vol. 51, pp. N411-N415

Follow, B. & Neil, E. Eds (1971) Circulation, Oxford Univ. Press (New York USA)

Foster, K. R. & Adair, E. R. (2004) Modeling thermal responses in human subjects following extended exposure to radiofrequency energy, *Biomed. Eng. Online* 3:4

Fujimoto, S.; Watanabe, T.; Sakamoto, A.; Yukawa, K. & Morimoto, K. (1968) Studies on the physical surface area of Japanese., 18. Calculation formulas in three stages over all ages. *Nippon Eiseigaku Zasshi*, Vol. 23, pp. 443-450 (in Japanese).

Gabriel, C. (1996) Compilation of the dielectric properties of body tissues at RF and microwave frequencies. Final Tech Rep Occupational and Environmental Health Directorate. AL/OE-TR-1996-0037 (Brooks Air Force Base, TX: RFR Division)

Gordon, R. G.; Roemer, R. B. & Horvath, S. M. (1976) A mathematical model of the human temperature regulatory system-transient cold exposure response *IEEE Trans Biomed Eng*, Vol. 23, pp. 434-444

Hardy, J. D. & DuBois, E. F. (1938) Basal metabolism, radiation, convection, and vaporization at temperatures of 22-35 °C, *J. Nutr.*, Vol. 15, pp. 477-492

Hirata, A.; Fujiwara, O. & Shiozawa, T. (2006a) Correlation between peak spatial-average SAR and temperature increase due to antennas attached to human trunk, *IEEE Trans. Biomed. Eng.*, Vol. 53, pp. 1658-1664

Hirata, A.; Watanabe, S.; Kojima, M.; Hata, I.; Wake, K.; Taki, M.; Sasaki, K.; Fujiwara, O. & Shiozawa, T. (2006b) Computational verification of anesthesia effect on temperature variations in rabbit eyes exposed to 2.45-GHz microwave energy, *Bioelectromagnetics* , Vol. 27, pp. 602-612

Hirata, A. Kodera, S. Wang, J. & Fujiwara, O. (2007a) Dominant factors influencing whole-body average SAR due to far-field exposure in whole-body resonance frequency and GHz regions *Bioelectromagnetics*, Vol. 28, pp.484-487

Hirata, A.; Asano, T. & Fujiwara, O. (2007b) FDTD analysis of human body-core temperature elevation due to RF far-field energy prescribed in ICNIRP guidelines, *Phys Med Biol* , Vol. 52, pp. 5013-5023

Hirata, A. & Fujiwara, O. (2009) Modeling time variation of blood temperature in a bioheat equation and its application to temperature analysis due to RF exposure, *Phys Med Biol* , Vol. 54, pp. N189-196

Hoque, M. & Gandhi, O. P. (1988) Temperature distribution in the human leg for VLF-VHF exposure at the ANSI recommended safety levels, *IEEE Trans. Biomed. Eng.* Vol. 35, pp. 442-449.

Ibrahiem, A.; Dale, C.; Tabbara, W. & Wiart J. (2005) Analysis of temperature increase linked to the power induced by RF source

International Commission on Radiological Protection (ICRP), (1975) Report of the Task Group on Reference Man, Vol.23, Pergamon Press: Oxford.

International Commission on Non-Ionizing Radiation Protection (ICNIRP). (1998) Guidelines for limiting exposure to time-varying electric, magnetic, and electromagnetic fields (up to 300 GHz)., *Health Phys.*, Vol. 74, pp. 494-522

Karslake D. De. K. (1972) The stress of hot environment. Cambridge Univ. Press, London.

Margaret, W.; Johnston, W. & Newburgh, L. H. (1942) Calculation of heat production from insensible loss of weight, *J Clin Invest.*, Vol.21, pp.357–363

Mason, P. A.; Hurt, W. D.; Walter, T. J.; D'Andrea, A.; Gajsek, P.; Ryan, K. L.; Nelson, D.A.; Smith, K. I.; & Ziriax, J. M. (2000) Effects of frequency, permittivity and voxel size on predicted specific absorption rate values in biological tissue during electromagnetic-field exposure. *IEEE Trans. Microwave Theory Tech.,*Vol.48, No. 11, pp-2050-2058.

McLaren, C.; Null, J. & Quinn, J. (2005) Heat stress from enclosed vehicles: moderate ambient temperatures cause significant temperature rise in enclosed vehicles, Vol. 116, pp. e109-e112

Nagaoka, T.; Watanabe, S.; Sakurai, K.; Kunieda, E.; Watanabe, S.; Taki, M. & Yamanaka, Y. (2004) Development of realistic high-resolution whole-body voxel models of Japanese adult males and females of average height and weight, and application of models to radio-frequency electromagnetic-field dosimetry *Phys. Med. Biol.* , Vol. 49, pp. 1-15

Nagaoka, T.; Kunieda, E. & Watanabe, S. (2008) Proportion-corrected scaled voxel models for Japanese children and their application to the numerical dosimetry of specific

absorption rate for frequencies from 30 MHz to 3 GHz, *Phys. Med. Biol.*, Vol. 53, pp.6695-6711

Nakayama, T. & Iriki, M. Ed. (1987) Handbook of Physiological Science vol.18: Physiology of Energy Exchange and Thermoregulation Igaku-Shoin (Tokyo)

Pennes, H. H. (1948) Analysis of tissue and arterial blood temperatures in resting forearm *J. Appl. Physiol.*, Vol. 1, pp. 93-122

Roberts, D.F.; Salzano, F. M. & Willson, J. O. C. (1970) Active sweat gland distribution in caingang Indians *Am. J. Phys. Anthrop.* Vol. 32, pp. 395-400

Samaras, T.; Christ, A. & Kuster, N. (2006) Effects of geometry discretization aspects on the numerical solution of the bioheat transfer equation with the FDTD technique *Phys. Med. Biol.*, Vol. 51, pp. 221-229

Spiegel, R. J. (1984) A review of numerical models for predicting the energy deposition and resultant thermal response of humans exposed to electromagnetic fields *IEEE Trans. Microwave Theory Tech.* Vol.32, pp.730-746

Stolwijk, J. A. J. (1971) A mathematical model of physiological temperature regulation in man. Washington, DC: NASA (CR-1855)

Stulyok, E.; Jequier, E. & Prodhom, L. S. (1973) Respiratory contribution to the thermal balance of the newborn infant under various ambient conditions *Pediatrics*, Vol.51, pp. 641-650

Taflove, A. & Hagness, S. (2003) Computational Electrodynamics: The Finite-Difference Time-Domain Method: 3rd Ed. Norwood. MA: Artech House

Tsuzuki, K.; Tochibara, Y. & Ohnaka, T. (1995) Thermoregulation during heat exposure of young children compared to their mothers, *Eur. J. Appl. Physiol.* Vol. 72, pp. l2-17

Tsuzuki, K. (1998) Thermoregulation during hot and warm exposures of infants compared to their mothers, *Jpn. Soc. Home Economics*, Vol. 49, pp. 409-415

Wang, J; Kodera, S.; Fujiwara, O. & Watanabe, S. (2005) FDTD calculation of whole-body average SAR in adult and child models for frequencies from 30 MHz to 3 GHz *Phys Med Biol*, Vol. 51, pp. 4119-4127

World Health Organization (WHO) (2006) RF research agenda, http://www.who.int/peh-emf/research/rf_research_agenda_2006.pdf

Towards a Robotic System for Minimally Invasive Breast Interventions

Vishnu Mallapragada and Nilanjan Sarkar
Vanderbilt University
USA

1. Introduction

Breast cancer is the most common cancer among American women and is the third leading cause of cancer death in women (American Cancer Society, 2009). The risk of breast carcinomas is approximately 1 in 8 women (Pass et al., 2008). As a result of nationwide implementation of screening for breast cancer, the proportion of small carcinomas found at the time of treatment has increased (Elmore et al., 2005). For small lesions, the standard of care is breast conserving therapy (BCT) with adjuvant radiation therapy (Arriagada et al., 2003). However, the cosmetic outcome of BCT is often disappointing with a satisfaction rate of 81% (Dian et al., 2007). Furthermore, BCT carries a relatively high morbidity rate due to bleeding (up to 11%) and infections (3.63%) (Bakker & Roumen, 2002; El Tamer et al., 2007). Breast asymmetry is also an issue with BCT where 35% of patients experience breast asymmetry after BCT (Dian et al., 2007).

The occurrence of benign epithelial lesions (e.g., fibroadenomas) is also common among women. Approximately 1 in 10 women experience a fibroadenoma in their lifetime (Rewcastle, 2005). As with BCT for carcinoma, surgical resection, the treatment of choice, provides definitive diagnosis, eliminates patient anxiety and reduces the need for follow-up monitoring, but it is expensive, can cause cosmetic or ductal damage and may be unnecessary because of the benign nature of the lesion (Greenberg et al., 1998). Additionally, it has been argued (Greenberg et al., 1998) that surgical resection of every fibroadenoma would place a huge burden on the health care system.

Because of bleeding, infection, cost and cosmetic considerations, several minimally invasive ablative therapies for breast carcinomas as well as benign fibroadenomas are being currently investigated that include cryoablation, radiofrequency ablation (RFA), laser-induced thermal therapy, microwave ablation and focused ultrasound (Van Esser et al., 2007). We present a brief background of cryoablation and RFA in order to motivate the need for development of a robotic breast intervention system.

Cryoablation is a very promising ablation technique to treat small breast tumors. It involves the introduction of a cryoprobe (about 2.4-3 mm in diameter) into the center of a tumor under ultrasound guidance in order to freeze the tissue to temperatures between -160^0 C and -190^0 C to kill the tumor cells (Sabel, 2008). Cryoablation is a simple and safe procedure with minimal discomfort or side effects for the patients (Whitworth & Rewcastle, 2005). After

percutaneous US (ultrasound) guided placement of the probe at the center of the tumor, the procedure involves monitoring the formation of an iceball and occasionally injecting saline between the iceball and the skin to prevent thermal damage. After (generally) two freeze-thaw cycles, the probe is removed and a bandage is placed over the incision. No local or regional anesthesia is needed past the probe placement, since freezing produces the same effect. Several small studies (Sabel et al., 2004; Pfleiderer et al., 2002; Staren et al., 1997; Morin et al., 2004) have demonstrated the safety, feasibility, efficacy and limitations of cryoablation for treating breast cancer. Data from these studies, most of which involve cryoablation followed by surgical resection, demonstrate that cryoablation was successful in destroying 100% of cancers less than 1 cm. For tumors between 1.0 and 1.5 cm, this success rate was achieved in patients with invasive ductal carcinoma without a significant ductal carcinoma-in-situ (DCIS) component. Cryablation was not recommended for tumors greater than 1.5 cm. From these experimental results, it was concluded (Sabel, 2008) that cryoablation can safely and efficiently treat primary breast cancers. It can be performed in an office-based setting with only local anesthesia, and with minimal side effects or discomfort by a skilled practitioner.

Cryoablation has also been applied in treating benign breast tumors such as fibroadenomas (Kaufman et al., 2002; Kaufman et al., 2004). In recent years cryoablation has been an approved treatment by the US Food and Drug Administration for women with biopsy-proven fibroadenomas. There is another important application of cryoablation, which is cryoprobe-assisted lumpectomy (CAL). The gold standard for localization in mammographically-guided breast biopsy is needle wire localization (NWL). While NWL has high degree of diagnostic accuracy, it is less reliable in localizing malignant tumors prior to excision (Fornage & Edeiken, 2005). NWL is often ineffective at assuring a tumor-free margin in the specimen resulting in 20-79% re-excision rate (Rewcastle, 2005). In CAL, a cryoprobe is guided using US into the center of the tumor. The cryoprobe is engaged until an iceball is created that encloses not only the entire tumor but also a margin. The iceball enclosing the tumor is then surgically excised, which may prove to be a superior localization methodology than NWL (Tafra et al., 2003).

US-guided RFA, on the other hand, uses thermal energy to induce thermal tissue necrosis in the target region and is considered one of the most promising ablation techniques in the treatment of breast cancer (Noguchi, 2007). The diameter of RFA probes is similar to those of cryoablation except that the electrodes may have secondary prongs. The goal of RFA of primary breast cancer is to achieve a lethal thermal lesion that encompasses not only the tumor, but also a margin of surrounding normal tissue to destroy possible peripheral microscopic disease. There have been several pilot and feasibility US guided RFA studies conducted in USA, Japan, France and Italy (Jeffrey et al., 1999; Izzo et al., 2001; Elliot et al., 2002; Burak et al., 2003; Hayashi et al., 2003; Fornage et al., 2004; Noguchi et al., 2006; Marcy et al., 2006). It was found that a small T1 carcinoma less than 2 cm was ablated between 95–100%. The ablation rate for T1-T2 tumors less than 3.0 cm was 90%. It is concluded from the above studies that RFA can be effective for well-localized tumors measuring less than 2 cm. The use of real-time sonography has been less effective in RFA because the margins of the tumor may be obscured by a diffuse area of hyperechogenicity. Thus, the role of sonographic guidance in RFA is to ensure that the probe is placed right at the geometric center of the tumor.

Thus, in order to evaluate the clinical efficacy of these ablative procedures, they first need to be performed accurately. In addition, clinicians across the country should be able to perform them otherwise the benefits of these treatments will be limited to a few at specialized cancer centers. Naturally the question arises whether there is variability among the clinicians, which if it exists will adversely affect the use of such ablative procedures. Measuring, monitoring, and regulating the quality of breast surgery are controversial subjects (Schachter et al., 2006). Up to 90% of breast surgery in the US is provided by 25% of surgeons (Pass et al., 2008) and many surgeons perform surgery on fewer than 5 breast cancer patients annually (Neuner et al., 2004). For example, the numbers for lumpectomy per year are (Pass et al., 2008): 50% of surgeons do 2; 25% do 6; 10% do 11; and 1% do 34. Quality discrepancy has been noted in many studies and appears to be related to surgeon volume and/or specific breast surgery or surgical oncology training (Skinner et al., 2003). It is concluded by Pass et al. (2008) that breast-focused surgeons are more competent than the 50% of surgeons who do the occasional case. There is data to suggest that breast-focused surgeons have better outcome. For example, treatment by surgical oncologists resulted in 33% reduction in the risk of death at 5 years (Skinner et al., 2003). Possible reasons for such improved outcome include a volume effect, improved surgical techniques or appropriate use of multidisciplinary approach to cancer treatment (Skinner et al., 2003). However, it is likely that only approximately 10% of patients in the United States are treated by surgeons who perform at least 30 annual operations (Neuner et al., 2004). Variability also exists among radiologists (Woodward et al., 2007). For example, only 10% of all radiologists are breast imaging specialists but 61% of radiologists interpret mammograms, and only 30% of mammograms are interpreted by breast imaging specialists (Lewis et al., 2006). It was found that radiologists in academic medical centers, compared with other radiologists, had a higher sensitivity during interpretations. More training in mammography and more experience in performing breast biopsy were associated with a decreased threshold for recalling patients, resulting in statistically significant increases in sensitivity and false-positive rates (Miglioretti, 2007). The above discussion suggests that the variability among clinicians might adversely impact the clinical usefulness of cryoabalation and RFA because of the requirement of the highest order of eye-hand coordination and may serve as a deterrent for their widespread use, notwithstanding the substantial benefits they can offer in terms of patient comfort, cosmetic results and cost.

Therefore, given the large variability that exists among breast clinicians (Pass et al., 2008; Schachter et al., 2006; Neuner et al., 2004; Skinner et al., 2003; Woodward et al., 2007; Lewis et al., 2006; Miglioretti et al., 2007), an automated system that can help the clinicians access small lesions in a precise manner for treatment without requiring the highest level of eye-hand coordination will likely be clinically useful, especially in community hospitals where a highly skilled radiologist-mammography specialist is less likely to be available. We therefore propose a new robotic technology for precise percutaneous probe placement to enhance the efficacy of minimally invasive ultrasound (US) guided ablative therapies such as cryoablation and radio frequency ablation (RFA) in the breast so that their scope and clinical usefulness can be effectively explored. In addition, this technology will offset clinician variability to a great extent and as a result, will help spread the use of ablative therapies in the breast to communities where such treatment might not otherwise be available.

Robotic technology has become sufficiently mature in recent years to be useful in many medical applications. There are numerous examples of these systems currently in use, such as the da Vinci Surgical System (Intuitive Surgical®). However, robotic aids are noticeably absent for breast surgery due to several unique technical challenges that are yet to be solved. In this work we will address two primary technical challenges to develop a novel robotic system for breast ablative procedures. These are: 1) how to ensure precise placement of a therapy delivery probe at the center of the tumor; and 2) how to minimize difficulty of ablative procedures. A robotic system that addresses these challenges will allow clinicians to precisely and easily access center of the breast lesion for therapeutic purposes.
In this chapter, the following terminology is used: target refers to a suspected lesion/tumor that is to be ablated; probe/needle (used interchangeably since the system development is independent of the specific instrument used to access the targets) refers to a treatment or therapy probe inserted into the target for ablation; US probe is the imaging probe used for acquiring US images.

2. Background and Literature Review

As mentioned before, there are two major problems to be addressed to improve the accuracy and reduce the difficulty of precise probe/needle placement at the center of the target.
Accurate needle placement: During needle insertion, the complex tissue of the breast induces the small target to deflect away from its original location. Figure 1 (DiMaio & Salcudean, 2003) shows a three dimensional (3D) finite element model of needle insertion in soft tissue. We can see from Figure 1 that as the needle is inserted, large tissue deformation causes the target to move away from the line of insertion of the needle. Other factors such as patient movement and breathing also cause inaccuracy in needle-target alignment. Note that accurate needle placement has to be achieved with a single insertion, since multiple insertions cause other risks such as excessive bleeding, tissue damage and significant patient discomfort.

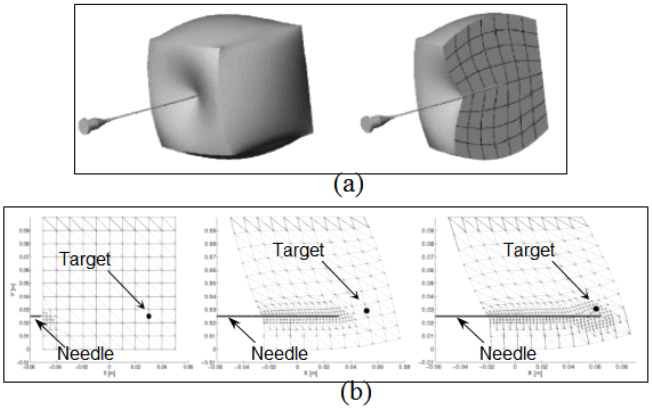


Fig. 1. Finite element model of needle insertion in soft tissue (DiMaio & Salcudean, 2003).
(a) Cross section of 3D model shows the finite element mesh.
(b) Target movement during needle insertion.

Difficulty of operation: Sonography is the widely used imaging technique because of its real-time capability and cost-effectiveness (Liberman et al., 1998). The current state-of-the-art US guided technique is highly dependent on the skill of the clinician (Fornage, 1999). A clinician performs this procedure by holding the US probe with one hand and inserting the needle with the other hand. Since sonography only provides a 2D image, if the target moves out of plane of the transducer, the clinician has to continuously reorient the probe to keep the needle and the target in the imaging plane while inserting the needle. It is critical to orient the imaging plane parallel to the needle. This freehand procedure requires excellent hand-eye coordination. Since stabilization of the breast is problematic (Smith et al., 2001) and steering of the needle inside the breast is extremely difficult, many insertion attempts are required to successfully position the needle at the target. This procedure is very fatiguing for the clinician and uncomfortable for the patient.

To address such challenges, several groups have designed robotic systems to improve the accuracy of needle insertions (Cleary et al., 2006; Tsekos et al., 2001; Stoianovici et al., 1998; Cleary et al., 2001; Patriciu et al., 2001; Stoianovici et al., 2001). The reader is referred to the work of Cleary et al. (2006) for a detailed review of state-of-the-art in interventional robotic systems. Several systems, such as, a device for conditioning of the breast and positioning of the probe (Tsekos et al., 2001), robotic systems for needle insertion (Stoianovici et al., 1998), precise intratumoral placement of therapeutic agents (Cleary et al., 2001), spinal (Patriciu et al., 2001) and renal (Stoianovici et al., 2001) percutaneous procedures have been developed. "Although these innovations greatly improve accuracy by automating needle target alignment, they do not provide active trajectory correction in the likely event that trajectory errors arise" (Okazawa et al., 2005). Needle trajectory errors and target mobility result in multiple insertions at the same site for accurate needle placement.

As a result, significant research effort is being made to investigate techniques that can address the problem of target movement during needle insertion. Okazawa et al. (2005) and other researchers (Glozman et al., 2007; Webster III et al., 2006; Sears & Dupont, 2006) pesented steerable devices are that allow the clinician to steer the tip of the needle towards the target during insertion. With such a device the clinician would have to manoeuvre the needle using one hand with image data from a US monitor while at the same time correctly orienting the US probe and stabilizing the breast with the other hand. As mentioned earlier, such a freehand technique is extremely difficult (due to high level of hand eye coordination required) and fatiguing for the clinician and uncomfortable for the patient. A visually controlled needle-guiding system is developed by Loser & Navab (2000) for automatic or remote controlled percutaneous interventions. In the automated mode, needle insertion path is updated based on image feedback to the needle-guiding system. Azar et al. (2002) and Alterovitz et al. (2003) developed a finite element model of the breast to predict the movement of the target. The needle path is planned based on this prediction to accurately position the needle at the target. To get an accurate prediction of the movement of the target, finite element analysis requires the geometric model and mechanical properties of the breast. In addition, average time for computation is 29 minutes (Azar et al., 2002).

The goal of the current research is to address the problem of ensuring precise placement of the probe/needle at the target for minimally invasive breast procedures leading to the design and development of an innovative robotic breast intervention system to aid the clinician. This system will potentially allow the clinician to solely focus on the detection, decision making, and ablation of the target without being encumbered by the difficulty of

achieving good targeting accuracy. In addition, the system also facilitates breast stabilization, US image acquisition and processing. It is evident from the literature review that there does not currently exist such a system. The robotic system provides a mechanism to compensate for needle–target misalignment for providing access to mobile targets. A novel approach termed, "target manipulation", (Mallapragada et al., 2008) is used to position the target inline with the needle thereby minimizing error in needle–target alignment. In this approach multiple robotic fingers manipulate the tissue externally to position a target inline with the needle during insertion.
In the following sections, basic theoretical framework of target manipulation is presented. Simulations and experimental results on phantoms are used to demonstrate the efficacy of this technique. Some aspects relating to the development of an autonmous US imaging system for minimizing difficulty of breast interventional procedures are also presented. The chapter ends with a discussion on the potential advantages and limitations of this system.

3. Control Framework

3.1. Noncollocated Controller for Target Manipulation

During image guided breast intervention a clinician inserts the needle through an incision in the skin. A schematic of needle insertion in a breast is shown in Figure 2.

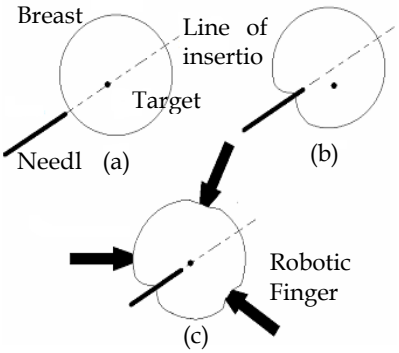


Fig. 2. Needle insertion schematic. (a) and (b) Target movement during needle insertion. (c) Minimizing needle – target misalignment using external robotic fingers.

The two dimensional plane of the figure represents a horizontal plane passing through the target. In the figure, a simplified anatomy for the breast is shown. In reality, breast tissue is inhomogeneous and its biomechanical properties are nonlinear. Hence, if the tip of the needle reaches the interface between two different types of tissue, its further insertion will push the tissue, instead of piercing it, causing unwanted deformations. These deformations move the target away from its original location, as shown in Figure 2b. In this section, we discuss controller design for the external robotic fingers positioned around the breast as illustrated in Figure 2c. These fingers apply forces on the surface of the breast based on the image of the target to guide the target towards the line of insertion of the needle.
During an interventional procedure, the needle is inserted into the breast at a shallow angle (away from the chest wall) to the horizontal plane containing the target. The needle incision

site and the orientation of the needle are chosen by the clinician considering factors such as location of target, location of critical anatomical structures and ease of access to target. The desired target position is the point where the line of insertion (of the needle) intersects the plane containing the target. While one can choose any plane that contains the target and has an intersection with the line of needle insertion, we choose this plane to be the horizontal plane for simplicity. The desired target position is determined by a planner based on the actual target location and needle direction. Note that we only need to control the target position in two dimensions (horizontal plane) to be able to successfully position the target along the line of insertion of the needle. Our goal is to design a controller that acts on the position error to guide the target towards the desired location. The controller is designed such that the effect of needle force (disturbance) on target position is minimized.

Before we discuss the design of the control system, we present a result by Wada et al. (2001) to determine the number of robotic fingers required to position the target at an arbitrary location in the horizontal plane. The following definitions are given according to the convention of Wada et al. (2001).

Manipulation points: Defined as the points that can be manipulated directly by robotic fingers. In our case, the manipulation points are the points where the external manipulators apply forces on the breast.

Positioned points: Defined as the points that should be positioned indirectly by controlling manipulation points appropriately. In our case, the target is the positioned point.

The control law to be designed is non-collocated since sensor feedback is from the positioned points and control force is applied at the manipulation points. The following result is useful in determining the number of fingers required to accurately position the target at the desired location.

Result (Wada et al., 2001): The number of manipulated points must be greater than or equal to that of the positioned points in order to realize any arbitrary displacement.

In our case, the number of positioned points is one, since we are trying to control the position of just the target. Hence, ideally the number of contact points would also be one. But practically there are two constraints: (1) We do not want to apply shear force on the breast to avoid damage to the skin. (2) We can only apply control forces directed into the breast. We cannot pull the skin on the breast since the robotic finger is not attached to the breast. Thus our problem becomes more restrictive since we need to control the position of the target by applying only unidirectional compressive force.

However, there exists a theorem on force direction closure in Mechanics that helps us determine the equivalent number of normal compressive forces that can replace one unconstrained force in a 2D (horizontal) plane.

Theorem (Nguyen, 1986): A set of wrenches W can generate force in any direction if and only if there exists a three-tuple of wrenches $\{w_1, w_2, w_3\}$ whose respective force directions f_1, f_2, f_3 satisfy:

(i) Two of the three directions f_1, f_2, f_3 are independent.

(ii) A strictly positive combination of the three directions is zero,

$$\alpha f_1 + \beta f_2 + \gamma f_3 = 0 . \quad (1)$$

The ramification of this theorem for our problem is this: we need three control forces distributed around the breast (as shown in Figure 1c) such that the end points of their

direction vectors draws a non-zero triangle that includes their common origin point. With such an arrangement we can realize any arbitrary displacement of the target point.

To develop a control law for the robotic fingers, the breast is modeled as a discrete three dimensional network of mass-spring-dampers. Following the notation of McClamroch (1985), dynamics of the system can be written in the form

$$\mathrm{M}\ddot{x}+F(x,\dot{x})=\mathrm{B}f\,. \tag{2}$$

x, $\dot{x}$ and $\ddot{x}$ (variables in italics represent vectors) denote generalized displacement, velocity and acceleration vectors, respectively. Each of these vectors has size 2n x 1 where n is the number of nodes in the discretized model. M is a constant symmetric, positive definite mass matrix. F is a flexible nonlinear restoring force function. B is the influence matrix which links the location of the fingers to the geometry of the structure. f denotes a vector of the external force applied on the system. f is a combination of the control force f_c and the needle force f_n. The output displacement vector is defined as follows:

$$y=\mathrm{C}x\,. \tag{3}$$

Here, y is the position of the target and C is the output matrix. The output position data is obtained through image feedback. Let $p_i\ (i=1,2,3)$ denote displacement of the manipulation point in contact with actuator i along the direction of actuation. We now define the displacement vector of the manipulation points as

$$p=\begin{bmatrix}\mathrm{p}_1 & \mathrm{p}_2 & \mathrm{p}_3\end{bmatrix}^{\mathrm{T}}. \tag{4}$$

Note that the elements of p are also elements of x. We have performed extensive simulations to determine the nature of the control laws that would be appropriate for this control problem. In particular, we investigated three generic class of controllers: adaptive controller, force feedback controller and a proportional-integral (PI) position error-based controller.

Adaptive control:

$$\dot{f}_c=\mathrm{K}_\mathrm{h}(-\dot{p}+\mathrm{g}^\mathrm{s}y_d-\mathrm{g}^\mathrm{d}p-\mathrm{g}^\mathrm{f}f_c)\,. \tag{5}$$

Here, f_c - force applied by the fingers, K_h - compliance matrix, $\dot{p}$ - velocity vector of manipulation points, y_d - desired position of target, g^f, g^d - force and position feedback gain matrices, g^s - feedforward gain.

To drive the steady state error to zero, the feedforward gain g^s is selected based on the stiffness of the system (McClamroch, 1985). Since stiffness of the breast is unknown, feedforward gain is adaptively estimated using MIT rule. Adaptation law is given by

$$\dot{\mathrm{g}}^\mathrm{s}=-\mathrm{v}(y_d-y)y_d\,, \tag{6}$$

where v is the adaptation gain.

Force feedback control:

$$\dot{f}_c = \mathrm{K_h}\{-\dot{p} + \mathrm{g^i}\int (y_d - y)\,\mathrm{dt} - \mathrm{g^d}\ p - \mathrm{g^f}\ f_c\}. \tag{7}$$

$\mathrm{g^i}$ is the integral gain. Instead of adaptively estimating the feedforward gain in (5), an integral term is added to eliminate steady state error.
PI control:

$$f_c = \{K_P(y_d - y) + K_I \int (y_d - y)d\tau\}. \tag{8}$$

$\mathrm{K_P}$ and $\mathrm{K_I}$ are proportional and integral gains. Note that in control laws (5), (7) and (8), geometric or mechanical properties of the breast are not required. Equations (2) and (3) are presented to develop a discretized model of the breast using a network of mass-spring-dampers. These equations (and consequently any model parameters) are not used in developing the control scheme. Equations (2) and (3) are used in simulation to construct a model of the breast for testing the feasibility of such a control methodology.
The implicit assumption in the control laws is that there exists kinematic coupling between the contact points and the target. This means that applying external control force (at the contact point) in a particular direction causes the target to move in a direction that has positive projection along the direction of force. Moreover, this assumes that the internal elastic force around the target can be controlled by applying external force on the surface. This assumption is valid since breast tissue is a continuous medium, however inhomogeneous. Inhomogeneity might cause the target to deflect away from the direction of force application, but continuity of the medium ensures kinematic coupling. Weak coupling (when the target is located away from the line of action of the fingers or due to inhomogeneity in the tissue) may necessitate large external forces to position the target but theoretically this does not undermine the control framework. Large external forces are undesirable so as to prevent patient injury and discomfort. This can be avoided in two ways: (1) appropriate positioning of the external fingers such that their line of action is close to the target; and (2) since breast tissue is not inhomogeneous in all directions, this problem can also be obviated by distributing the actuators around the breast. The theorem and result discussed above inherently address this problem and as an obvious consequence, the actuators are positioned 120^0 apart (in Figure 2c).

3.2. Simulation of Target Manipulation in Soft Tissue

Figure 3 shows a schematic of the control structure. Target position data (y) is obtained through image feedback. The desired target position (y_d) is determined by the planner based on the current target location and needle direction. The desired target position is always along the line of insertion of the needle. The controller acts on the position error and drives the robotic fingers to position the target at the desired location. The force exerted by the needle is the disturbance to the system.

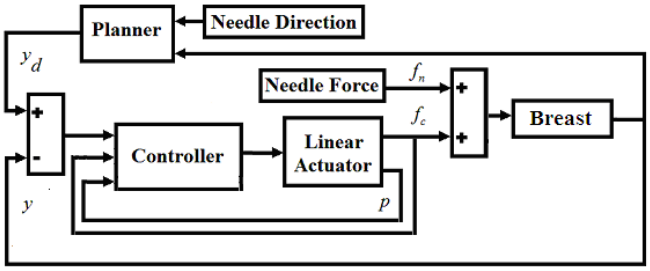


Fig. 3. Control structure.

To test the performance of the controllers, we simulated a needle insertion task using a 3D model of the breast. The target is located inside the breast and is initially at the origin of the coordinate system. A needle is inserted into the breast along the line specified by spherical coordinates: azimuth (θ) 45^0, zenith (φ) 45^0 and passing through the origin. A plot of the needle insertion force is shown in Figure 4. The force profile is based on the elasto-plastic friction model (Yang et al., 2005). The insertion force gradually increases after contact with the breast surface (Point A to Point B). At Point B, the needle punctures the tissue characterized by a sudden drop in force (Point B to Point C). As the insertion continues, needle force steadily rises (Point C to Point D) mainly due to friction between needle and the tissue. Force remains constant once the insertion stops. In reality, there is a slight drop in force at Point D due to expansion of tissue. However, this drop in force is small and does not alter the results of this simulation significantly. In addition, force vibrations caused by the internal structure of the tissue have not been modeled since they do not significantly affect our results.

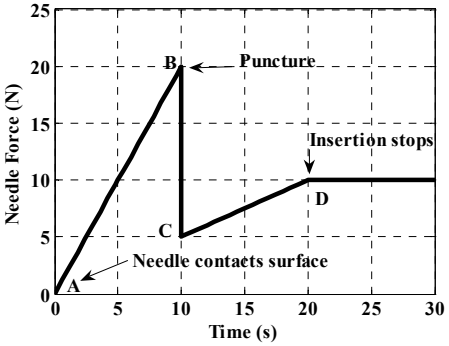


Fig. 4. Needle insertion force.

Figure 5 shows a plot of the trajectory of the target using the three different control laws. In Figure 5, Point A (origin) is the initial location of the target. The needle is inserted along the direction indicated by the broken arrow. As the needle is inserted, inhomogeneity in the tissue causes the target to move away from the needle path. As we can see from Figure 5, all three controllers successfully bring the target back to the needle path. The final target

location is at Point B which lies on the line of needle insertion. The simulation results suggest that adaptive controller and force feedback controller do not provide any significant advantage over the PI controller. Hence, a PI position error-based controller is chosen for this application since it is the cheapest and simplest controller with acceptable performance. Note that Figure 5 only shows the line of insertion of the needle and not the path of the needle. Temporal coordination of needle insertion and target manipulation has to be performed by the clinician through visual feedback from real time ultrasound imaging.

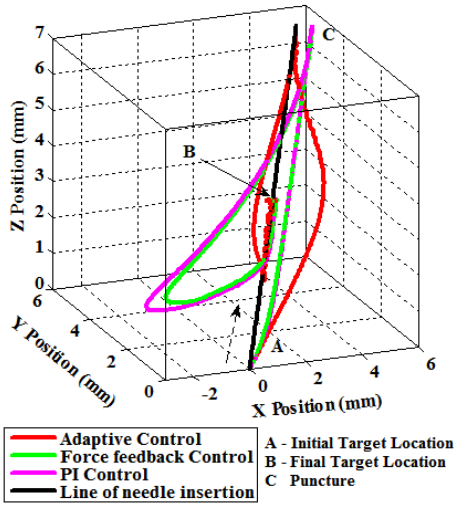


Fig. 5. Target trajectory.

4. Results

4.1. Phantoms

Deformable plastic phantoms with a stiff inclusion (a plastic insert placed in the phantom to mimic a tumor) are created to test the efficacy of the PI controller in positioning the inclusion at a desired location. We make the phantoms in such a manner that their material properties closely resemble breast tissue properties as published in the literature (Wellman, 1999). This step is necessary to ensure the success of the controller when it is applied to real breast tissue. The phantom is a cylindrical structure (radius 60mm, height 35mm) made of PVC (Poly Vinyl Chloride) plastic. The inclusion is a plastic sphere (radius 14 mm) that is much stiffer than the phantom. The inclusion is used as the target in the following experiments. Softening and hardening material is added to this plastic to alter its elastic property. Three phantoms (A, B, and C) are prepared with different mix ratios. Phantom A has plastic to hardener ratio of 4:1. Phantom B does not contain softener or hardener. Phantom C has plastic to softener ratio of 3:1. Procedure for preparing the phantoms is outlined by Mallapragada et al. (2007). The three phantoms are homogeneous.

Uniaxial compression tests are performed on each of the phantoms to determine their elastic properties. Nominal stress – strain values are computed from force – displacement data measured during the compression test. Elastic moduli of the phantoms are determined by

an exponential fit of the stress – strain curve (Wellman, 1999). Using the notation by Wellman (1999), stress – strain relationship is given by

$$\sigma_n^* = \frac{b^*}{m^*}(e^{m^*\varepsilon_n} - 1), \tag{9}$$

where σ_n^* is the nominal stress and ε_n is the nominal strain. b^* and m^* are the exponential fit parameters. The Young's modulus is given by

$$E = b^* e^{m^*\varepsilon_n}. \tag{10}$$

Figure 6 shows a plot of the Young's moduli of the phantoms. Fat tissue data shown in Figure 6 is presented by Wellman (1999). From the figure we can see that the phantoms have Young's moduli similar to that of fat tissue in the breast. Breast tissue properties vary greatly based on factors such as age, presence of tissue abnormality etc. In order to demonstrate feasibility of our technique under significant parameter variation, we prepared phantoms with varying elastic properties to demonstrate that the presented controller can work in realistic scenarios.

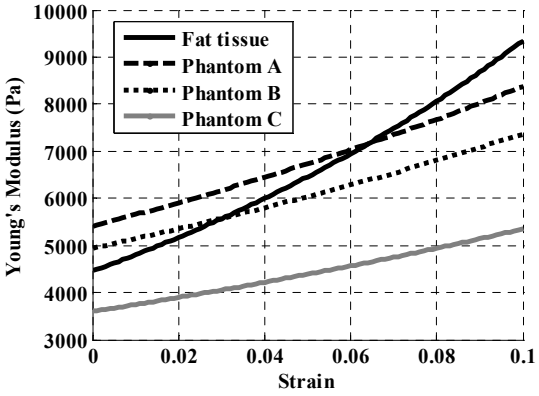


Fig. 6. Comparison of Young's modulus of fat tissue and phantoms.

4.2. Experimental Results

We demonstrate efficacy of this technique through planar manipulation experiments. In these experiments needle is inserted in the plane of target manipulation. Therefore, we only need two control forces (fingers) to position the target inline with the needle. When the needle is inclined to the plane of target manipulation, we will need three control forces (fingers) to position the target on the needle path. This is because when the needle is in the plane of target manipulation, the intersection between the axis of the needle and the plane is a line whereas when the needle is inclined to the plane, the intersection between the axis of the needle and the plane is a point. As discussed in Section 3.1, in either case planar manipulation is sufficient to position the target inline with the needle. Hence, results presented in this Chapter also demonstrate feasibility for out of plane needle insertion since in either case we are manipulating target position in the plane.

We have performed four different experiments to demonstrate different aspects of target manipulation. Needle used in the following experiments is a 10-gauge vacuum assisted device.

Experiment 1: In the first experiment, we demonstrate movement of the target during planar needle insertion. The phantoms used for this experiment are inhomogeneous (as is actual breast tissue) such that during needle insertion the target deflects away from the path of the needle. Figure 7 shows two images grabbed from the video signal during needle insertion. Figure 7a shows the initial target location prior to needle insertion. Figure 7b shows needle inserted along Y axis. As the needle is inserted, the target deflects away from the needle path.

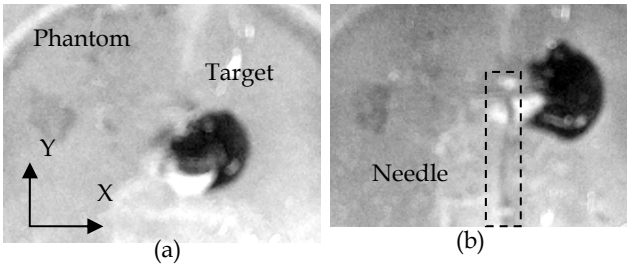


Fig. 7. Images grabbed from the video signal during needle insertion in inhomogeneous phantom. (a) Initial position of target in phantom is shown. (b) Displacement of target from needle path. Needle is identified in the figure with a dotted bounding box.

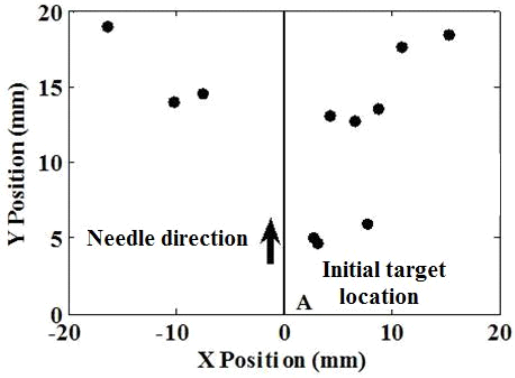


Fig. 8. Movement of target during needle insertion. Needle is inserted along Y axis. Dots indicate maximum displacement of target from the needle path.

Figure 8 shows a scatter plot of the maximum deflection of the target away from the needle path during multiple trials using different phantoms. Point A (origin) is the initial location of the target before the needle is inserted. The arrow indicates the direction of needle insertion. During 11 trials, the average maximum deflection of the target away from the needle path is 8.51 mm. The cone angle for target deflection ranges from 28.80 to 52.30.

These results are consistent with the observation of DiMaio & Salcudean (2003). Thus this experiment demonstrates that the phantoms that we have created behave similar to breast tissue.

Experiment 2: An experimental setup is constructed to test the efficacy of PI controlled robotic fingers in positioning the stiff target at a desired location by applying force on the surface of the phantom. A schematic of the experimental setup is shown in Figure 9.
The phantom is braced against a support on two sides and fingers apply force from the opposite sides. Position feedback is obtained using a Creative Labs video camera (30 Frames per second, 640 X 480 pixels). The target can be viewed using a video camera since the phantom is transparent. Image data from the video camera is converted from RGB to YC_BC_R color space. The phantom is placed against a red background and the target is blue in color. Hence, chrominance (C_B) is used to track the target in real-time. Image processing algorithm consists of the following steps: 1) region segmentation to extract the region of interest, 2) color space conversion to convert from RGB to YC_BC_R, 3) thresholding to differentiate the target from the background, 4) median filtering to remove noise, and 5) blob analysis to extract target centroid coordinates.

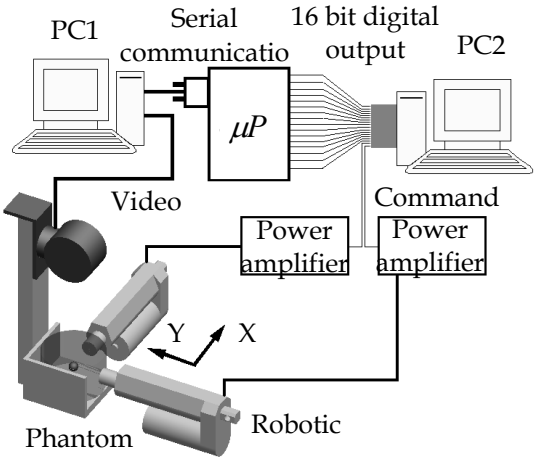


Fig. 9. Experimental setup shows system architecture and the location of actuators for target position control.

During an interventional procedure, image data would be obtained through ultrasound imaging. Image frames from the video camera are sent to a computer (1.6 GHz and 2 GB RAM, shown as PC 1 in Figure 9) running image processing algorithm in Matlab. Image frames are processed to extract position data of the target. Target position data is communicated serially to a microcontroller (Freescale 68HC912B32, 8 MHz clock frequency). The microcontroller outputs this data in a 16 bit parallel format. Each iteration of image processing and data communication requires 0.2 seconds. This is the time delay in the feedback loop of the controller. Medical US systems have frame rates of 5 frames per second

or higher. Hence the time delay in the feedback loop using US systems will be the same or less and the performance of the controller will not be affected. This data is read by another computer (1.6 GHz and 1GB RAM, shown as PC 2 in Figure 9) using a data acquisition card (Measurement computing PCIM DDA06/16). This computer runs the PI control algorithm and outputs control signals to power amplifiers for driving the robotic fingers. The fingers are lead screw driven linear actuators (FA-PO-20-12-2", Firgelli Automation) with inbuilt potentiometers. They have a no load speed of 50 mm/s and a load capacity of 88N at 25% duty cycle. The end-effector of the actuators has a circular surface area of 3.1 cm^2 (2 cm diameter) . Contact between the end-effector of the finger and phantom is frictionless.

In this experiment, we have created a situation that is similar to the target deflection problem due to a needle insertion to demonstrate the feasibility of the concept. In a needle insertion situation, the task is to localize the target so that it remains inline with the needle. Any deviation of the target is seen as an error by the controller and a compensating force is generated to mitigate the error. We assume that the target is already deflected and the task of the controller is to move the target to a desired position by applying an external force. Thus this experiment is conducted to move the target to a desired location within the phantom using the fingers. Initial position of the target is set as the origin. Needle force acting as a disturbance on the system is not included for this experiment. This experiment is designed to move the target along two directions (X and Y axes as shown in Figure 9) using two fingers perpendicular to each other. The goal is to be able to position the target at any point in the horizontal plane (XY plane in Figure 9). Phantom is braced against a support opposite to the fingers.

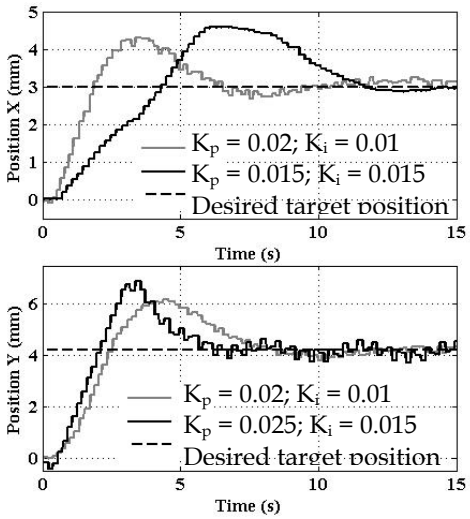


Fig. 10. Target positioning in horizontal plane. Position response of the target for two trials with different control gains. Target reaches the desired position in approximately 12 seconds with K_P = 0.02 and K_I = 0.01.

In PI control law (8), duty cycle of a PWM (Pulse Width Modulation) signal is used as actuator input instead of force. This is chosen to overcome friction based limit cycle behavior seen in DC (Direct Current) voltage controlled actuators. We use a PWM signal with frequency of 4 Hz and amplitude of 2 Volts for driving the fingers. The desired position of the target is 3 mm along X axis and 4.2 mm along Y axis. Two trials are performed with different sets of proportional and integral gains. For trial 1, proportional and integral gains are 0.02 and 0.01 respectively for both the actuators. For trial 2, proportional and integral gains are 0.015 and 0.015 (respectively) for the first finger (acting along X direction); proportional and integral gains are 0.025 and 0.015 (respectively) for the second finger (acting along Y direction). It can be observed from Figure 10 that the position response for trial 1 has less overshoot along both directions. For X direction, target reaches the desired position faster for trial 1. For Y direction, target response has less oscillcation for trial 1. Therefore gains selected for trial 1 are better suited for target position control. It can also be observed from Figure 10 that the target reaches the desired position in approximately 12 seconds for trial 1. Note that any geometric or mechanical properties of the phantom are not used in the control scheme (8).

Experiment 3: For the third experiment, we use the robotic fingers to manipulate target position during a needle insertion task. The experimental setup for this task is shown in Figure 10.

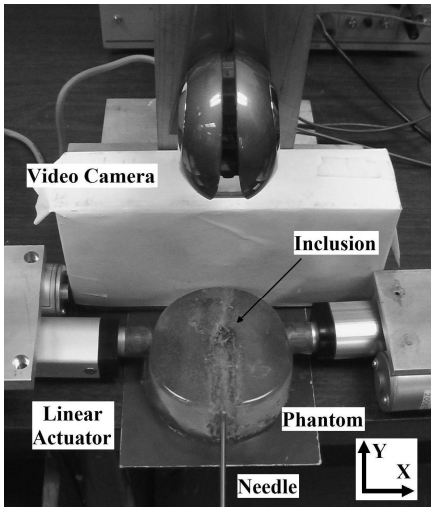


Fig. 11. Experimental setup for needle insertion task. Needle is inserted along Y axis. Robotic fingers control target position along X axis.

Target is initially located at the origin and the needle is inserted along the Y axis. Due to inhomogeneity in the phantom, target moves away from the needle path during insertion. We use fingers positioned along the X axis to steer the target towards the needle path (Y axis). During this experiment, force applied by the needle on the phantom is treated as a disturbance to the system. The task of the controller is to position the target on the Y axis

(needle path). Position of the target along Y axis is not controlled since the needle will intersect with the target no matter where it is located on the Y axis.

Figure 12 shows a plot of the position response of the target (along the X axis) during five trials (shown as Trials 1, 2, 3, 4 and 5) on three different inhomogeneous phantoms. The inhomogeneous phantoms are constructed by distributing the phantom material A, B and C (described in Section 4.1) in an asymmetrical arrangement during the molding process. These phantoms have two kinds of material with elastic moduli similar to fat and glandular tissue.

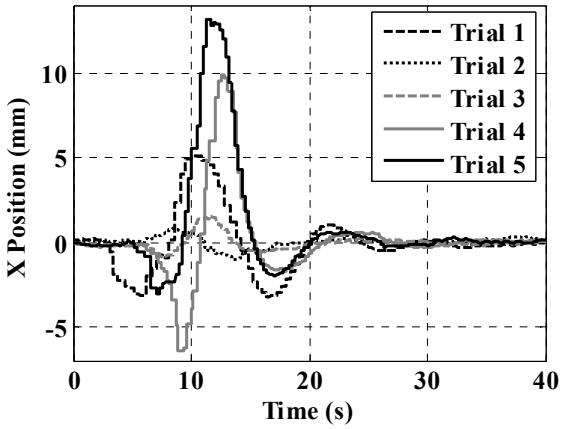


Fig. 12. Target position response during needle insertion. Plot shows that the target reaches the needle path in approximately 28 seconds.

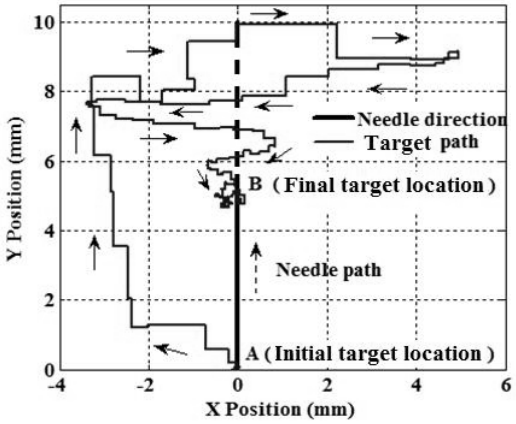


Fig. 13. Locus of target position for Trial 1. Needle is inserted along the Y axis. Point A is the initial target location and point B is the final target location.

From Figure 12, it can be observed that the target is initially located on the Y axis (displacement along X axis is zero). As the needle is inserted, target moves away from the Y axis (non zero displacement along X axis). This initiates a control action by the fingers, which steer and position the target on the Y axis (displacement along X axis is zero) at steady state. As we can see from Figure 12, target is steered back to the needle path in about 30 seconds. In all five trials, we were successful in positioning the target along the needle path. We could steer the target back to the needle path even when the deviation of the target is large (~10 mm).

Figure 13 shows locus of the target position for Trial 1. We can see from Figure 13 that the target is initially located in the path of the needle (Point A), but as the needle is inserted it deviates away from the path and the fingers (under PI control action) steer it back towards the line of insertion. The final location of the target is at Point B on the needle path.

Root mean square error (RMSE) is used to quantify the targeting accuracy of this technique. RMSE is defined as

$$\text{RMSE}=\sqrt{\frac{\sum_{t=t_i}^{t_f}(x_d-x)^2}{n}}, \tag{11}$$

where t_i and t_f represent the limits of the time interval over which RMSE is computed. x_d is the desired target position and x is the actual target position at time t. n is the number of data points in the chosen time interval. Table I shows the RMSE values for the five trials in Figure 12. In these trials the desired target position is along the Y axis, hence $x_d=0$. Targeting accuracy needs to be evaluated when the target is positioned on the needle path, therefore steady state position data (t_i=30 s; t_f =40 s) is used to compute RMSE. The sampling time for the controller is 0.001 s, hence n =10001. As we can see from Table 1, the targeting error in all cases is one order of magnitude less than the diameter of the needle (10-guage, 3.4 mm). Therefore this technique can be used to successfully position the target on the needle path.

Trial #	1	2	3	4	5
RMSE (mm)	0.32	0.18	0.10	0.13	0.12

Table 1. Targeting accuracy with external manipulation

Experiment 4: We present results with a 2D US system to demonstrate applicability of this technique using US image feedback. The experimental setup is shown in Figure 14. Image feedback is obtained using a Toshiba ECCO-CEE (model SSA-340A) US system. US images are acquired using a frame grabber card (Data Translation, DT3120). Figure 15 shows an image grabbed from the video signal of the US system. Dark blob in the image is the target. Region segmentation, thresholding, median filtering and blob analysis are used to extract target position coordinates.

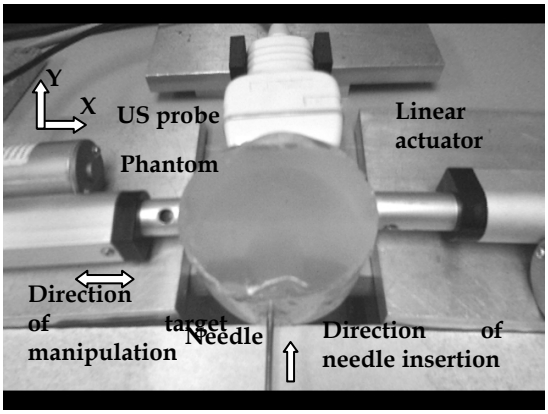


Fig. 14. US experimental setup. Needle is inserted along Y axis and fingers control target position along X axis.

The task description for this experiment is similar to Experiment 3. The target is initially located at the origin and the needle is inserted along the Y axis (Figure 14). We use robotic fingers positioned along the X axis to steer the target towards the needle path (Y axis). The task of the controller is to position the target on the Y axis. XY plane is the imaging plane of the US probe. In this setup, position of the actuators and needle (with respect to the US image space) are hard coded into controller.

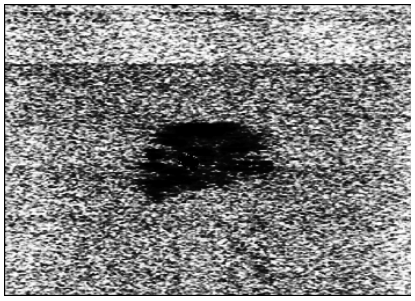


Fig. 15. US image of target in phantom. Dark blob in the image is the target.

Figure 16 shows the locus of target position. We can see from Figure 16 that the target is initially located in the path of the needle (Point A), but as the needle is inserted it deviates away from the path and the actuators steer it back towards the line of insertion. The final location of the target is at Point B on the needle path where it meets the needle. RMSE in this experiment is 0.06 mm.

During needle insertion, real-time video (rendered on US or computer monitor display) provides visual information of the position of the needle and target. In these experiments, operator controls needle insertion depth based on visual feedback so that the needle does not overshoot/cross the target location. If needle insertion is automated, target manipulation has to be coordinated with needle insertion so as to ensure alignment of the target with needle tip.

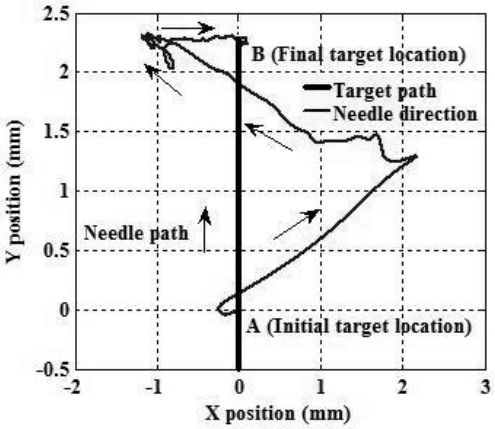


Fig. 16. Locus of target position for Experiment 4. Needle is inserted along the Y axis. Point A is the initial target location and point B is the final target location.

5. Autonomous US Imaging

Researchers have developed robotic systems to alleviate the difficulty associated with acquiring US images during medical procedures (Abolmaesumi et al., 2002; Masuda et al., 2001). Robotic systems with autonomous US imaging minimize the difficulty of breast interventions (Section 2) and also reduce the incidence of muscoskeletal disorders in sonographers. However, despite the potential advantages offered by robotic US imaging systems, there are some technical challenges that need to be overcome in order to realize autonomous US image acquisition.

2D US is the most widely used and cost effective clinical imaging technique. There are two major limitations of 2D US imaging. First, Tracking the location of a target (3D coordinates) using a 2D US probe requires knowledge of the position of the probe and coordinates of the target in the image plane of the probe. Sensors are used to measure the position of the probe. Coordinates of the target are extracted using an image processing algorithm, if the target is in the imaging plane. If the target is out of the imaging plane, there is uncertainty in determining the location of the target with respect to the imaging plane i.e., there is no information regarding the location of the target in a direction perpendicular to the imaging plane. Second, the US probe has to be continuously in contact with the surface of the breast to ensure acoustic coupling. Surface deformation due to target manipulation and needle insertion may result in the US probe losing contact with the breast. Loss of contact has to be detected (robotic systems use force sensors for detecting contact state of the US probe) and the US probe has to be moved to reestablish contact with the surface.

Even though the systems described in literature greatly reduce the difficulty of acquiring US images, the target cannot be tracked in real-time if it moves out of the imaging plane of the probe. Therefore, there is a need for a robust image acquisition system that is capable of automatic search and recovery of the target. Currently, none of the existing systems have this ability.

Real time automated target tracking using 2D US is a challenging task especially since US gives image of a two dimensional cross-section. We are investigating a system for

positioning the US probe guided by an image acquisition algorithm for automatic retrieval of the image of the target. This system continuously orients the transducer such that the 3D coordinates of the target are tracked in real-time. A novel sensorless technique for detecting contact state of US probe is also being investigated. The region in the image close to the edge (where the US probe makes contact with the surface) is used to detect the contact state of the US probe. This region is extracted from the US image and Otsu's method is used to estimate the greyscale threshold for the region. The threshold differential between contact and non contact states is used to infer the contact state of the US probe.
The target manipulation technique presented in this paper will be integrated with the autonomous US imaging system to provide comprehensive robotic assistance during US guided breast interventional procedures.

6. Discussion

PI control architecture has been presented in the paper for guiding a target towards the needle path. The performance of the controller is tested on phantoms with elastic property similar to that of breast tissue. We have demonstrated planar needle insertion as a proof-of-concept of a new control approach to target manipulation. As discussed in Section 4.2, it does not limit the scope of our results. Results show that PI controlled robotic fingers can be used to efficiently position a target inline with needle during insertion. The proposed technique of guiding breast interventions can increase the success rate of the procedure. The entire procedure is predicted to be fast, making it clinically viable. Since the needle is not steered inside the breast, tissue damage is also minimized. Additionally, since multiple insertions are not required, the proposed technique will likely reduce clinician fatigue and patient discomfort and improve the structural integrity of the tissue specimen. Geometric and mechanical properties of the breast are not required for precise positioning of the target.
Target identification in ultrasound is a significant challenge and as such our work does not address this issue. The targets in our experiments are easily identifiable in the video frame. The focus of this paper is on developing a new targeting approach for breast interventional procedures which can be integrated with any real time imaging modality.
The proposed technique is not without limitations. In the experimental setup, position of the needle (with respect to the image space) is hard coded into the controller. However, this would be infeasible in a clinical setting. Therefore, needle position/orientation has to be measured using a 6 degree-of-freedom electromagnetic sensor, or estimated using needle segmentation techniques in US image.
The end-effector of the actuator has a circular cross section with 20 mm diameter. During experimental tests, we observed that maximum target movement in any direction is typically in the range of 5 – 10 mm. Since the end-effector has surface contact (contact are 3.14 cm^2), even if the target moves out-of-plane, the actuators can still control the target motion without causing instability. It is highly unlikely that target movement will be greater than 20 mm. If such an extreme case is observed, the robotic system requires another degree-of-freedom so that the actuators can move in a direction perpendicular to the plane of manipulation. However, issues relating to stability for out-of-plane target movements have not been investigated in this work.

7. Conclusion

We have presented a new paradigm for breast interventions that will allow the clinician to place the treatment probe at the center of a breast tumor for ablative procedures under real-time US image feedback and computer control. However, the clinician will have full control of the treatment probe. The robotic system will control the position of the tumor by careful manipulation of the breast such that the tumor centroid always remains along the line of insertion of the probe. Thus, this unique system will ensure that the probe placement is precise, and as a result, the effect of ablation is optimum.

There are several important benefits that can be derived from this technological innovation. **First**, it will enhance the clinical efficacy of cryoablation and RFA by allowing the clinician to place the probe precisely at the center of the tumor. It will help create an ablative zone that is expected to encompass the whole tumor with a given margin. As a result, one can analyze the clinical outcome of such procedures without the contribution of positioning error. **Second**, given the variability that exists among clinicians treating breast cancers and since ablation procedures require the highest level of eye-hand coordination skills, it is realistic to assume that most patients will not be benefited by this treatment option in its current form. The proposed robotic system will significantly reduce the eye-hand skill requirement on the part of the clinician and thereby will allow a majority of physicians to perform these ablative procedures all over the country. The proposed system will thus help decentralize breast cancer treatment to a certain extent and potentially reduce the load on breast cancer centers. Clinicians in community hospitals and private practice, who may not experience high volume or may not have specialized training, will feel comfortable in administering ablative procedures because the mechanical accuracy will be guaranteed by the robotic system. **Third**, by helping to perform these ablative procedures accurately (in terms of positioning of the probe) and thus allowing many more clinicians to perform these procedures, this robotic system is expected to contribute towards cost savings in healthcare, and improving cosmesis of patients. It has been predicted that by 2010, 50% of newly diagnosed breast cancers may be less than 1 cm (Cady, 2000). That would represent 90,000 patients for whom a lumpectomy will be necessary, with the associated operating room time, anesthesia, cost, and cosmetic and psychological impact (Sabel et al., 2004). If even a fraction of these patients could be helped by this system, the benefits will be substantial and would be a significant improvement to treatment of breast cancer.

8. Acknowledgment

The authors gratefully acknowledge the guidance of Dr. Tarun Podder (Jefferson University Hospital, Philadelphia) during the course of this work. The authors would also like to thank Dr. David Pickens (Radiology Imaging Research Laboratory, Vanderbilt University) for his support.

9. References

Abolmaesumi, P., Salcudean, S.E., Zhu, W., Sirouspour, M.R., DiMaio, S.P. (2002). Image-guided control of a robot for medical ultrasound. *IEEE Transactions on Robotics and Automation*, Vol. 18, No. 1, pp. 11-23

Alterovitz, R., Goldberg, K., Pouliot, J., Taschereau, R. & Hsu, I-C. (2003). Sensorless planning for medical needle insertion procedures. *Proc. of IEEE International Conference on Intelligent Robots and Systems*, pp. 3337 – 3343

American Cancer Society (ACS): Cancer Facts & Figures, http://www.cancer.org/downloads/STT/2008CAFFfinalsecured.pdf, webpage accessed on February 2009

Arriagada, R., Le, M.G., Guinebretiere, J.M., et al. (2003). Late local recurrences in a randomized trial comparing conservative treatment with local mastectomy in early breast cancer patients. *Annals of Oncology,* 14(11):1617-1622

Azar, F.S., Metaxas, D.N. & Schnall, M.D. (2002). Methods for modeling and predicting mechanical deformations of the breast under external perturbations. *Medical image Analysis*, Vol. 6, pp 1-27

Bakker, X.R. & Roumen, R.M. (2002). Bleeding after excision of breast lumps. *Eur J Surg,* 168(7):401-403

Burak, W.E. Jr., Agnese, D.M., Povoski, S.P. et al. (2003). Radiofrequency ablation of invasive breast carcinoma followed by delayed surgical excision. *Cancer*, 98:1369-1376

Cady, B. (2000). Breast cancer in third millennium. *Breast J*, 6:280-287

Cleary, K., Freedman, M., Clifford, M., Lindisch, D., Onda, S. & Jiang, L. (2001). Image-guided robotic delivery system for precise placement of therputic agents. *J. Controlled Release*, Vol. 74, No. 1, pp. 363–368

Cleary, K., Melzer, A., Watson, V., Kronreif, G. & Stoianovici, D. (2006). Interventional robotic systems: Applications and technology state-of-the-art. *Minimally Invasive Therapy and Allied Technologies*, Vol. 15, No. 2, 101-113

Dian, D., Schwenn, K., Mylonas, I., et al. (2007). Aesthetic result among breast cancer patients undergoing autologous breast reconstruction versus breast conserving therapy. *Arch. Gynecol. Obstet*., 275(6):445-450

DiMaio, S.P. & Salcudean, S.E. (2003). Needle insertion modeling and simulation. *IEEE Transactions on Robotics and Automation*, Vol. 19, No. 4

Elliot, R.L., Rice, P.B., Suits, J.A. et al. (2002). Radiofrequency ablation of a steriotactically localized nonpalpable breast carcinoma. *Am Surg,* 68:1-5

Elmore, J.G., Armstrong K., Lehman C.D., et al. (2005). Screening for breast cancer. *JAMA* 293(10):1245-1256

El Tamer, M.B., Ward, B.M., Scifftner, T., et al. (2007). Morbidity and mortality following breast cancer surgery in women: national benchmarks for standard of care. *Ann Surg,* 245(5):665-671

Fornage, B.D. (1999). Sonographically guided needle biopsy of nonpalpable breast lesions. *Journal of Clinical Ultrasound*, Vol. 27, No. 7

Fornage, B.D., Sneige, N., Ross, M.I. et al. (2004). Small (less than equal to 2 cm) breast cancer treated with US-guided radiofrequency ablation: feasibility study. *Radiology*, 231:215-224

Fornage, B.D. & Edeiken, B.S. (2005). Percutaneous ablation of breast tumors. In : *Tumor Ablation: Principles and Practice*, Vansonnenberg, E., Solbiati, L. & Mcmullen, W, 428-439, Springer-Verlag, ISBN 13:978-0387-95539-1, New York

Glozman, D. & Shoham, M. (2007). Image-guided robotic flexible needle steering. *IEEE Transactions on Robotics*, Vol. 23, No. 3, pp 459 – 467

Greenberg. R., Skornick. Y. & Kaplan. O. (1998). Management of breast fibroadenoma. *J Gen Intern Med,* 13:640-645

Hayashi, A.H., Silver, S.F., Van Der Westhuizen, N.G. et al. (2003). Treatment of invasive breast carcinoma with ultrasound-guided radiofrequency ablation. *Am J Surg,* 185:429-435

Izzo, F., Thomas, R., Delrio, P. et al. (2001). Radiofrequency ablation in patients with primary breast carcinoma. A pilot study in 26 patients. *Cancer*, 92:2036-2044

Jeffrey, S.S., Birdwell, R.L., Ikeda, D.M. et al. (1999). Radiofrequency ablation treatment for breast cancer: first report of an emerging technology. *Arch Surg,* 134:1064-1068

Kaufman, C.S., Bachman, B., Littrup, P.J. et al. (2002). Office-based ultrasound-guided cryoablation of breast fibroadenomas. *Am J Surg,* 184:394-400

Kaufman, C.J., Littrup, P.J., Freman-Gibb, L.A. et al. (2004). Office-based cryoablation of breast fibroadenomas: 12-month followup. *J Am Coll Surg,* 198(6):914-923

Lewis, R.S., Sunshine, J.H. & Bhargavan, M. (2006). A portrait of breast imaging specialists and the interpretation of mammography in the U.S. *Am J Roentgenol,* 187:W456-W468

Liberman, L., Feng, T. L., Dershaw, D. D., Morris, E. A. & Abramson, A. F. (1998). US-guided core breast biopsy: use and cost-effectiveness. *Radiology*, Vol. 208, 717-723

Loser, M.H. & Navab, N. (2000). A new robotic system for visually controlled percutaneous interventions under CT fluoroscopy. *Proc. of MICCAI*, Vol. 1935, LNCS, pp.887-896

Mallapragada, V., Sarkar, N., Podder, T. (2007). A Robotic system for real-time tumor manipulation during image guided breast biopsy. *Proc. of IEEE International Conference on Bioinformatics and Bioengineering*, pp. 204-210

Mallapragada, V., Sarkar, N., Podder, T. (2009). Robot assisted real-time tumor manipulation for breast biopsy. *IEEE Transactions on Robotics* (to be published)

Marcy, P.Y., Magne, N., Castadot, P. et al. (2006). Ultrasouind-guided percutaneous radiofrequency ablation in elderly breast cancer patients: preliminary institutional experience. *Br J Radiol*, 80(952):267-273

Masuda, K., Kimura, E., Tateishi, N., Ishihara, K. (2001). Three-dimensional motion mechanism of ultrasound probe and its application for tele-echography system. *Proc. of IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 1112-1116

McClamroch, N.H. (1985). Displacement control of flexible structures using electrohydraulic servo-actuators. *Journal of Dynamic Systems, Measurement and Control*, Vol 107, pp 34-39

Miglioretti. D.L., Smith-Bindman. R., Abraham. L. et al. (2007). Radiologist characteristics associated with interpretive performance of diagnostic mammography. *JNCI* 99(24):1854-1863

Morin, J., Traore', A., Dionne, G., et al. (2004). Magnetic-resonance-guided percutaneous cryosurgery of breast carcinoma: technique and early clinical results. *Can J Surg,* 47:347-351

Neuner, J.M., Gilligan, M.A., Sparapani, R., et al. (2004). Decentralization of breast cancer surgery in the United States. *Cancer*, 101:1323-1329

Nguyen, V.D. (1986). Constructing force-closure grasps. *Proc. of International Conference on Robotics and Automation*, Vol. 3, pp. 1368 – 1373

Noguchi, M., Earashi, M., Fujii, H. et al. (2006). Radiofrequency ablation of small breast cancer followed by surgical resection. *J Sur Oncol*, 93:120-128

Noguchi, M. (2007). Is radiofrequency ablation treatment ready for "prime time"? *Breast Can res Treat,* 106:307-314

Okazawa, S., Ebrahimi, Chuang, J., Salcudean, S.E. & Rohling, R. (2005). Hand-held steerable needle device. *IEEE/ASME Transactions on Mechatronics*, Vol. 10, pp 285-296

Pass, H.A., Klimberg, S.V. & Copeland, E.M. (2008). Are "Breast-Focused" surgeons more competent? *Annals of Surgical Oncology,* 15(4):953-955

Patriciu, A., Solomon, S., Kavoussi, L. & Stoianovici, D. Robotic kidney and spine percutaneous procedures using a new laser-based CT registration method. *Proc. Of MICCAI*, Vol. 2208, pp.249–257

Pfleiderer, S.O., Freesmeyer, M.G., Marx, C. et al. (2002). Cryotherapy of breast cancer under ultrasound guidance: initial results and limitations. *Eur Radiol,* 12:3009-3014

Rewcastle, J.C. (2005). Application of cryoablation in breast, In: *Tumor Ablation: Principles and Practice*, Vansonnenberg, E., Solbiati, L. & Mcmullen, W, 423-427, Springer-Verlag, New York

Sabel, M.S., Kaufman, C.S., Whitworth, P. et al. (2004). Cryoablation of early-stage breast cancer: work-in-progress report of a multi-institutional trial. *Annals of Surgical Oncology,* 11(5):542-549

Sabel, M.S. (2008). Cryoablation for breast cancer: no need to turn a cold shoulder. *J Surg Onco*, 97:485-486

Schachter, H.M., Mamaladze, V., Lewin, G., et al. (2006). Many quality measurements, but few quality measures assessing the quality of breast cancer care in women: a systematic reviw. *BMC Cancer,* 6:291

Sears, P. & Dupont, P. (2006). A Steerable Needle Technology Using Curved Concentric Tubes. *Proc. of IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 2850 – 2856

Skinner, K.A., Helsper, J.T., Deapen, D., et al. (2003). Breast cancer: do specialists make a difference? *Ann Surg Oncol,* 10:606-615

Smith, W.L., Surry, K.J.M., Mills, G.R., Downy, D.B. & Fenster, A. (2001). Three-dimensional ultrasound-guided core needle breast biopsy. *Ultrasound Med Biol*, Vol. 27, No. 8, pp. 1025–1034

Staren, E.D., Sabel, M.S., Gianakakis, L.M. et al. (1997). Cryosurgery of breast cancer. *Arch Surg*, 132:28-33 discussion 34

Stoianovici, D., Whitcomb, L., Anderson, J., Taylor, R. & Kavoussi, L. (1998). A modular surgical robotic system for image guided percutaneous procedures. *Proc. Of MICCAI*, Vol. 1496, LNCS, pp. 404–410

Stoianovici, D., Cadeddu, J.A., Demaree, R.D. et al. (2001). A novel mechanical transmission applied to percutaneous renal access. *Proc. of ASME Dynamic Systems Control Division*, DSC, Vol. 61, pp. 401–406

Tafra, L., Smith, S.J., Woodward, J.E. et al. (2003). Pilot trial of cryoprobe assisted breast-conserving surgery for small ultrasound visible cancers. *Ann Surg Oncol,* 10(9):1018-1024

Tsekos, N.V., Shudy, J., Yacoub, E., Tsekos, P.V. & Koutlas, I.G. (2001). Development of a robotic device for MRI-guided interventions in the breast. *Proc. of Bioinformatics and Bioengineering Conference*, 201-208

Van Esser, S., Maurice, A.A., Van Der Bosch, J., et al. (2007). Minimally invasive ablative therapies for invasive brest carcinomas: an overview of current literature. *World J Surg,* 31:2284-2292

Wada, T., Hirai, S., Kawamura, S. & Kamiji, N. (2001). Robust manipulation of deformable objects by a simple PID feedback. *Proc. of International Conference on Robotics and Automation*, pp. 85-90

Webster III, R.J., Kim, J.S., Cowan, N.J., Chirikjian, G. & Okamura, A.M. (2006). Nonholonomic modeling of needle steering. *International Journal of Robotics Research*, 25(5/6), pp. 509–526

Wellman, P. (1999). *Tactile imaging*. PhD dissertation, Division of Engineering and Applied Sciences, Harvard University, Cambridge, Massachusetts

Whitworth, P.W. & Rewcastle, J.C. (2005). Cryoablation and cryolocalization in the management of breast disease. *J Surg Oncol*, 90:1-9

Woodward, D.B., Gelfand, A.E., Barlow, W.E., et al. (2007). Performance assessment for radiologists interpreting screening mammography. *Statist. Med.,* 26:1532-1551

Yang, H., Liu, P.X., Zhang, J. (2005). Modelling of needle insertion forces for surgical simulation. *Proc. of International Conference on Mechatronics and Automation*, pp 592-595

Spectral Analysis Methods for Spike-Wave Discharges in Rats with Genetic Absence Epilepsy

Elif Derya Übeyli[1,*], Gul Ilbay[2] and Deniz Sahin[2]
*Corresponding author –
[1]TOBB Economics and Technology University; Faculty of Engineering, Department of Electrical and Electronics Engineering; 06530 Söğütözü, Ankara, Turkey;
e-mail: edubeyli@etu.edu.tr
[2]Kocaeli University, Faculty of Medicine, Department of Physiology, Kocaeli, Turkey
e-mail: gulilbay@yahoo.com e-mail: sahindeniztr@yahoo.com

1. Introduction

The WAG/Rij strain is an inbred strain of Wistar rats in which all animals present absence seizures (Coenen et al., 1992). These seizures appear as a sudden interruption of consciousness and are characterized in the cortical electroencephalogram (EEG) by the occurrence of bilateral and synchronous spike-wave discharges (SWDs) (Coenen & Van Luijtelaar, 2003).

SWDs seen in WAG/Rij rats share many clinical characteristics with typical human absence epilepsy and exhibit a similar pharmacological reactivity to drugs (Coenen et al., 1992; Van Luijtelaar & Coenen, 1986; Peeters et al., 1989). Therefore, WAG/Rij strain of rats is considered to be a valid animal model of human absence epilepsy (Ates et al., 1999; Ates et al., 2004). Nowadays this genetic model of absence epilepsy is commonly used for studying the efficacy of new antiepileptic drugs on the occurrence of SWD and the pathogenesis of absence epilepsy (Coenen & Van Luijtelaar, 2003; Bouwman & Van Rijn, 2004). However, the mechanisms underlying SWDs, still remain unclear (Bouwman et al., 2007). Although the analysis of the time–frequency (TF) structure of SWDs may contain important information about the mechanisms of this type of brain paroxysmal activity and can play a significant role in the investigation of antiepileptic drugs, the dynamics of SWDs in rodent models have been poorly investigated (Bosnyakova et al., 2006; Bosnyakova et al., 2007). It is usually indicated that in animals with absence epilepsy the typical SWDs have a mean frequency of 8.7 Hz (Van Luijtelaar & Coenen, 1986). In addition, by means of the Fast Fourier procedure it was shown that the frequency of the SWD is approximately 10-11 Hz at the beginning of and 7-8 Hz at the end of the discharges (Drinkenburg et al., 1993). Bosnyakova et.al (2006) used a modified Morlet wavelet transform to describe significant parameters of the dynamics in the TF domain of the dominant rhythm of SWD. In a recent

paper, analysis of the TF pattern of SWD in patients with absence seizures and WAG/Rij rats revealed that TF dynamics of SWDs had similar properties but in a different frequency range (Bosnyakova et al., 2007).

Spectral analysis methods can be used for representing and/or discriminating the signals recorded from WAG/Rij rats. The basic problem that we consider is the estimation of the power spectral density (PSD) of a signal from the observation of the signal over a finite time interval (Kay & Marple, 1981; Kay, 1988; Proakis & Manolakis, 1996; Stoica & Moses, 1997; Übeyli, 2009). The signals recorded from WAG/Rij rats are conventionally interpreted by analyzing their spectral content. Diagnosis and disease monitoring are assessed by analysis of spectral shape and parameters. In order to determine the variabilities of the signals under study are processed by spectral analysis methods to achieve PSD estimates and to obtain the features representing the signals. In the signal processing stage, numerous different methods can be used so that several diverse features can be extracted from the same raw data. The nonparametric methods (Fast Fourier transform based methods), parametric methods (model-based methods) and TF methods (wavelet transform) are the methods used for spectral analysis. The parameters obtained by these methods characterize the behaviour of the time-varying signals. This feature of using a smaller number of parameters to represent the time-varying signals is particularly important for recognition and diagnostic purposes. The objective of the chapter in the field of detection of changes in the time-varying signals is to extract the representative features of the signals under study. In this chapter, the dynamic parameters in the TF domain of SWD were analysed and results represent good additional tool for discriminating this epileptic event and new perspective for future investigations (Übeyli et al, 2008; Übeyli et al., 2009).

2. Human Absence Epilepsy: The Wag/Rij Rat as a Genetic Model

Epilepsy is a sudden and recurrent brain malfunction and is a disorder that reflects an excessive and hypersynchronous activity of the neurons within the brain. It encompasses a number of different syndromes, the cardinal feature of which is a predisposition to recurrent unprovoked seizures. It can occur at all ages, and is characterized by a variety of presentations and causes. The estimated incidence is one case per 2000 persons in the Western population per year, whereas the prevalence of active epilepsy with recent seizures is around 5–10 per 1000. For unknown reasons, the incidence of epilepsy is highest in the first year of life and increases again for those over 60 years of age. The cumulative incidence, that is, the chance of having epilepsy during a lifetime of 80 years, is about 3% (Engel et al., 2007; Fisher et al., 2005; Kwan & Sander, 2004)

By convention, the diagnosis of epilepsy requires that the patient has had at least two unprovoked seizures. Seizures are sudden, brief attacks of altered consciousness; motor, sensory, cognitive, psychic, or autonomic disturbances; or inappropriate behavior caused by abnormal excessive or synchronous neuronal activity in the brain (Fisher et al., 2005). The phenotype of each seizure is determined by the point of origin of the hyperexcitability and its degree of spread in the brain. If given a sufficient stimulus (e.g., hypoxia, hypoglycemia), even the normal brain can discharge excessively, producing a seizure. However, a person with isolated nonrecurrent, externally provoked seizures that are also caused by excessive

discharge of cerebral neurons is not thought to have epilepsy as long as the seizures are not recurrent and each seizure is preceded by a provocation (e.g., substance abuse, fever, exposure to alcohol combined with lack of sleep) (Elger & Schmidt, 2008).
Epilepsy can also occur as a syndrome. An epileptic syndrome is an epileptic disorder characterized by a cluster of signs and symptoms occurring together; these include items such as the type of seizure, etiology, anatomy, precipitating factors, age of onset, severity, chronicity, diurnal and circadian cycling, and sometimes prognosis (ILAE, 1989; Valentin et al., 2007).

Recently, epileptic seizures and epileptic syndromes is broadly divided into partial (focal or localization-related), beginning in a part of one hemisphere, and generalized, which were bilaterally symmetrical without local onset (Engel & Pedley, 2007; Weber & Lerche, 2008; Hauser et al., 1996).Partial seizures, are those in which, in general, the first clinical and electroencephalographic changes indicate initial activation of a system of neurons limited to part of one cerebral hemisphere. A partial seizure is classified primarily on the basis of whether consciousness is impaired during the attack. When consciousness is not impaired, the seizure is classified as a simple partial seizure. When consciousness is impaired, the seizure is classified as a complex partial seizure where the patient is unable to respond normally to exogenous stimuli. A partial seizure may progress to a generalized motor seizure. Generalized seizures however, are those in which the first clinical changes indicate initial involvement of both hemispheres. Consciousness may be impaired, and this impairment may be the initial manifestation. In addition, motor manifestations are bilateral. The ictal electroencephalographic patterns initially are bilateral and presumably reflect neuronal discharge, which is widespread in both hemispheres. Generalized seizures can be classified into one of the following six fundamental groups: absence seizures, myoclonic seizures, tonic-clonic (or clonic-tonic-clonic) seizures, atonic seizures, tonic seizures, and clonic seizures. Further subdivision separates epilepsies of known etiology (symptomatic or secondary epilepsies) from those that are idiopathic (primary) and those that are cryptogenic (Weber & Lerche, 2008; Hauser et al., 1996).

Much less is known about idiopathic epilepsies. This type has no specific cause, but is assumed to have a genetic etiology. The epilepsies that are idiopathic almost always have onset childhood or adolescence, although there are exceptions: Some patients develop these kinds of epilepsies after the second decade of life, and in some rare cases, the onset may be even later. Absence, myoclonic, and tonic-clonic seizures are associated with idiopathic generalized epilepsies (IGE) (ILAE, 1989). The symptomatic epilepsies differ from the idiopathic epilepsies by having a known cause, such as head trauma, or there may be some evidence of brain or neurological dysfunction. In cryptogenic epilepsies, the etiology or side of abnormality may not be detectable or is hidden (Mattson, 2003).

One of the distinct groups of the generalized seizures is the absence seizures. Absence seizures were first described by Poupart in 1705 and, subsequently, in 1770, were called "petits accés" (petit access) by Tissot. Introduced by Calmeil (Temkin, 1994; van Luijtelaar & Sitnikova, 2006), the name "absence seizures" dates from 1824. Absence seizures are brief epileptic seizures characterized by generalized spike wave discharges (SWDs) in the cortical EEG accompanied by short falls in consciousness (absence) without the tonic-clonic

manifestations. Ongoing activity is interrupted during the seizure, responsiveness is decreased, and mental functioning is impaired. There are only minimal myoclonic jerks of the eyes and perioral automatisms. The beginning and end of the seizure are abrupt, and there is no aura or postictal state (Snead, 1995; Lockman, 1989). Depending on EEG findings absence seizures are divided into typical and atypical absence seizures (Posner et al., 2005).

The typical absence seizures are the most common and are characterized by a loss of consciousness that is time-locked with bursts of bilaterally synchronous 3 cycles/second SWDs. They are generally associated with minimal or no cognitive impairment (Snead, 1995). However, atypical absence seizures are less common, but are often associated with severe neurological impairment. Although the pharmacological profiles of the two absence types are the same, a number of features may be used to distinguish typical from atypical absence seizures (Holmes et al., 1987). One characteristic concerns voluntary behavior during the ictus (Carmant et al., 1996; Bare et al., 1998; Snead et al., 1999; Cortez et al., 2001; Nolan et al., 2005); another major difference relates to the frequency of the SWDs, both in human patients and in rodent experimental models of atypical and typical absence epilepsy (Snead, 1995; Cortez et al., 2001; Nolan et al., 2005; Velazquez et al., 2007). In addition, typical absences are brief generalized epileptic seizures with a sudden onset and termination. However, in the atypical absences seizures, the onset and termination are not as abrupt as in typical absences, and changes in tone are more pronounced (Panayiotopoulos, 2008). The usual frequency of the bilaterally synchronous SWD in typical absence seizures is 3 cycles/second, while that seen in atypical absence seizures are faster or slower than 3 cycles/second (Lockman, 1989; Snead, 1995). A third distinguishing feature is that there is a major difference in outcome between children with typical versus atypical absence seizures; atypical absence seizures are associated with a severely abnormal cognitive and neurodevelopmental outcome in children (Pavone et al., 2001; Høie et al., 2005; Henkin et al., 2005; Markand, 2003). The final distinguishing characteristic involves the neural circuitry involved in the SWD. In typical absence seizures, the epileptiform activity is constrained within thalamocortical circuitry. In addition, some evidence existing evidence indicate that while no SWD can be recorded from any circuitry other than thalamus and cortex during typical absence seizures. In contrast, there are data for the involvement of both thalamocortical and limbic circuitry in atypical absence seizures (Snead et al., 1999; Velazquez et al., 2007).

In human absence seizures, it is important to note that, typical absence seizures may be the only seizure type experienced by a child and this then constitutes either an epileptic syndrome called childhood absence epilepsy or juvenile absence epilepsy. However; absence seizures may also be only one of multiple types of seizures, as in the case of juvenile myoclonic epilepsy where myoclonic and tonic-clonic seizures occur as well as absence seizures. The Commission on Classification and Terminology of the International League Against Epilepsy recognizes four epileptic syndromes with typical AS: childhood absence epilepsy; juvenile absence epilepsy; juvenile myoclonic epilepsy and myoclonic absence epilepsy. (Posner et al., 2005; ILAE, 1989).

The annual incidence rate of absence seizures has been estimated to be 1/10,000. The estimation of the prevalence of absence seizures varies from 2.3% to 37.7%. Absence seizures

are more often seen in childhood epilepsy, and predominantly occur in children of school age; however, absence seizures also are seen in adults, albeit less commonly. A family history of epilepsy is found in 15% to 44% of patients with generalized absence seizures, and an inherited factor in human absence seizures was recognized. In addition, genetic factors play a predominant role in the etiology of IGE with typical absence seizures. Because the concordance for absence seizures in monozygotic twins does not reach 100%, acquired factors probably also play a role in this type of seizure. The mean age of onset is 7 years (range: 9 months to 12 years). Recently, the γ-aminobutyric acid (GABA) A receptor γ-2 subunit mutation has been reported to be an autosomal dominant mutation associated with childhood absence epilepsy. This mutation appears to impair GABAA receptor function (Stefan et al., 2007). The drug of choice for typical absence seizures is valproic acid or ethosuximide. In addition, absence seizures also respond well to the newer medications like lamotrigine and topiramate. However, in contrast to all other types of seizures, GABA-mimetic anti-epileptic drugs such as vigabatrin and tiagabine exacerbate absence seizures and aggravate SWDs, as predicted from outcomes in rodent models (Coenen, 1995; Cocito and Primavera, 1998; Knake et al., 1999; Depaulis & van Luijtelaar, 2006). Therefore, it is suggested that absence epilepsy is a disturbance in inhibition. (van Luijtelaar & Sitnikova, 2006). Nevertheless, pathophysiology of idiopathic generalized absence epilepsy is not fully understood (Tolmacheva et al., 2004). The benign nature of absence epilepsy precludes invasive investigation in humans; therefore, emphasis is placed on experimental animal models (acute, chronic, pharmacological, genetic, etc.) that allow the investigation of the mechanisms of pathogenesis and propose better diagnostic and therapeutic procedures of this disease.

An animal model of generalized absence seizures should, in addition to reflecting the clinical and pharmacological characteristics of this disorder, fulfill other requirements. These criteria include reproducibility and predictability, as well as the ability to standardize and quantitate the model. In addition, animal models of absence should reflect the fact that both clinical and experimental absence seizures are exacerbated by both direct and indirect GABA agonists. Finally, if an animal model of absence seizures is to be considered as a valid one, involvement of thalamocortical circuitry and specific noninvolvement of hippocampal circuitry should be demonstrated (Snead, 1995).

Although a number of pharmacological and genetic rat models meet these criteria, genetic models are the pertinent choice among the models available for human absence epilepsy. The WAG/Rij (Wistar Albino Glaxo Rijswijk) strain of rats is such a model (Snead, 1995; Coenen & van Luijtelaar, 2003).

The WAG/Rij strain is an inbred strain of rats in which brother-sister cross breeding has taken place for more than 100 generations, implying that the rats are homozygous. Therefore, rats from this strain offer an eminent possibility to study the genetic background and heredity of absence epilepsy. Furthermore, the rats are fertile and show no sings of behavioral abnormalities (Coenen et al., 1992). All individuals of this strain develop spontaneous SWDs in their EEG (van Luijtelaar & Coenen, 1986). SWDs in WAG/Rij rats are accompanied by behavioral arrest and immobility, minimal facial myoclonic jerks, twitching of eyes and vibrissae, altogether mimicking clinical manifestation of absence epilepsy in humans (WAG/Rij model has face validity). Likewise, the pharmacological profile of seizures in WAG/Rij rats and in humans is similar. This enables predictions from the model

to the patient (predictive validity of WAG/Rij model). Finally, absence seizures in WAG/Rij rats and in humans are based on the same theoretical grounds (WAG/Rij model has a construct validity). Consequently, WAG/Rij rats fulfill all the necessary criteria for a valid and reliable animal model of human absence epilepsy. (Coenen & van Luijtelaar, 2003, Sitnikova & van Luijtelaar 2006).

3. Electroencephalograms and Spike-Wave Discharges

Rhythmic electrical activity can be recorded from the cerebral cortex. This activity is known as the EEG when the activity is recorded from the surface of the skull. An EEG signal is a measurement of currents that flow during synaptic excitations of the dendrites of many pyramidal neurons in the cerebral cortex. When brain cells (neurons) are activated, the synaptic currents are produced within the dendrites. This current generates a magnetic field measurable by a secondary electrical field over the scalp measurable by EEG systems. Differences of electrical potentials are caused by summed postsynaptic graded potentials from pyramidal cells that create electrical dipoles between the soma (body of a neuron) and apical dendrites, which branch from neurons. Since the human head consists of different layers including the scalp, skull, brain, and many other thin layers in between, the skull attenuates the signals approximately one hundred times more than the soft tissue. On the other hand, most of the noise is generated either within the brain or over the scalp. Therefore, only large populations of active neurons can generate enough potential to be recordable using the scalp electrodes. These signals are later amplified greatly for display purposes (Berne & Levy, 1993; Sanei & Chambers, 2007).

The EEG is an important diagnostic tool in clinical neurology and is particularly useful in patients with epilepsy. Diagnosis of epilepsy is made primarily on patient history and neurological exam, but clinical criteria alone may not be sufficient for characterization of its type, or may not be infallible. The EEG verifies that the event in question is an epileptic seizure and not something else. It helps to classify the type of seizure and the underlying epileptic syndrome. Often, the EEG helps to pinpoint where seizures arise in the brain. The frequency of seizure occurrence and the effectiveness of therapy can be evaluated with EEG (Koutroumanidis & Smith, 2005; Berne & Levy, 1993).

Seizures characterized by paroxysmal changes of the brain's electrical activity that produce impairment in consciousness or bilateral movements; such as tonic–clonic, tonic and complex partial seizures almost always appear in the scalp EEG.

Ictal (seizure) and interictal (between seizure) EEG patterns correspond to specific seizure types and types of epilepsy. Several distinct EEG patterns appear in generalized seizures that are related to the underlying cause of the epilepsy. Seizures arising in patients with idiopathic generalized epilepsy usually begin with generalized SWDs or a burst of generalized polyspikes. Patients with symptomatic generalized epilepsy have a broader variety of ictal onset patterns, although a generalized spike and slow wave, generalized slow wave, generalized fast activity, or generalized attenuations are usually seen. Subsequent ictal EEG findings depend on the seizure type. During absence seizures, the spike-wave (S&W) pattern continues; during tonic seizures, the generalized fast activity persists. Tonic–clonic seizures display an evolving spike pattern consonant with the muscular contractions. On the other hand, absence seizures have varying characteristics. Typical absence seizures (Fig. 1) usually begin with a generalized, frontally predominant

SWD at 3.5 to 4 Hz that gradually slows to 2.5 Hz by the end of the seizure. Absence seizures often show increasing spike amplitude in the first two or three discharges. Seizure offset, while relatively quick, also often reflects a "build down," with one to three rhythmic slow waves of diminishing amplitude following the last spike discharge. There is no significant postictal slowing of the background after absence seizures. Atypical absence seizures may show somewhat more irregular S&W or polyspike and-wave discharges, sometimes at initial higher frequencies. Patients with symptomatic generalized epilepsy have atonic or tonic seizures. Atonic seizures may begin with a generalized spike or sharp wave (with or without a slow wave), a generalized slow wave, or simply a diffuse attenuation of the background EEG with or without low amplitude fast activity. In contrast, a characteristic feature of tonic seizures is the progressive buildup of generalized fast activity (15–30 Hz), which may be preceded by a S&W or a slow wave or may appear without antecedent. The seizure is usually followed by a brief period of postictal background slowing. On the other hand, during and between focal seizures, scalp recordings may reveal EEG spikes. (Sperling & Clancy, 2007).

It should be noted that seizures are infrequent events in the majority of patients, and a prolonged EEG recording session may be required to capture a seizure. Fortunately for diagnosis, 50-80% of patients with epilepsy display interictal discharges in their routine interictal EEG test. Interictal epileptiform discharges (IEDs) are more prevalent and more persistent in some epilepsy syndromes such as childhood-onset absence. In adults with partial epilepsy, IEDs are more common when seizures originate in the temporal lobes than when seizures originate elsewhere. Interictal discharges may be divided morphologically into sharp waves, spikes, S&W complexes (also called spike-and-slow-wave complexes), and polyspike-wave complexes (also called multiple-spike-and-slow-wave-complexes) (Fisher & Leppik, 2008; Worrell et al., 2002).

IEDs are difficult to describe precisely. IED must be paroxysmal and clearly distinguished from background activity. There must be an abrupt change in polarity lasting several milliseconds. Duration must be <200 milliseconds. The Committee on Terminology distinguishes between spikes, which have a duration <70 milliseconds, and sharp waves, which have a duration between 70 and 200 milliseconds. In addition, the IED must have a physiologic field. Practically, this means that the IED is recorded by more than one electrode and has a voltage gradient across the scalp. This requirement helps distinguish IEDs from artifacts. Also, focal IEDs suggest localization-related epilepsies, whereas generalized IEDs suggest generalized epilepsies (Walczak et al., 2007**)**.

The prototype of generalized epileptiform abnormalities is the sudden onset of bilaterally synchronous 2.5-4 cycles/s S&W activity. Although described as S&W activity, the morphology of the complexes is somewhat more complicated. Two spikes, a positive transient and a slow wave, make up the complex. Spike I is negative and low in amplitude (25-50 μV*)*, short in duration, and usually not seen in the first few complexes of the burst. A positive transient lasting 100-150 ms follows spike I; this is followed by spike II which is high in amplitude (three times spike I) and lasts 30-90 ms. The slow wave following spike II is a surface negative wave, and, if one can distinguish it from the positive transient one, its duration is in the 150-250 ms range. In addition to sudden near simultaneous diffuse onset, the S&W activity phenomenon usually stops simultaneously over both hemispheres and is followed by an abrupt return of normal background activity. Potentially, the S&W activity phenomenon can be considered an aberrant age-dependent thalamocortical oscillatory

rhythm. A common assumption holds that the thalamocortical relay cells involved in spindle generation have special Ca^{2+} channels called transient channels which provide them the ability to burst fire when stimulated. Also, the nucleus reticularis thalami neurons impose oscillatory behavior on the thalamocortical relay cells. Alterations seen in the nucleus reticularis thalami - thalamocortical relay -cortical neuron loop are responsible for S&W bursts (Gloor & Fariello, 1988; Snead, 1995; Schaul 1998).

WAG/Rij rats show spontaneously occurring bilaterally synchronised SWDs in the cortical EEG with a frequency of 7–11 Hz and an amplitude of 100 to 450 μV, and a mean duration of 5 seconds (1-30 seconds; Fig 2). This feature of the WAG/Rij rats allows them to be considered as a genetic animal model for typical absence seizures. SWDs in WAG/Rij rats are age-dependent. The first EEG symptoms of absence epilepsy appear in 2–3 month-old WAG/Rij rats. Later, with age, increases in the number and duration of SWDs are observed. By the age of 6 months, virtually all WAG/Rij rats show SWDs on their cortical EEGs (Coenen & Van Luijtelaar, 2003; Ilbay et al., 2001;Ates et al., 2004). SWDs do not appear randomly in time; rather, they tend to appear in clusters in rats as in humans. The amplitude of SWDs is largest in the frontal midline region and gradually decreases in the lateral and posterior directions. Onset and termination are abrupt; the attacks may be preceded and immediately followed by normal EEG activity, especially when recorded in the waking (resting) state (Kellaway, 1985; Midzyanovskaya et al., 2006; Rodin & Ancheta, 1987; van Luijtelaar & Sitnikova, 2006). Recent studies have shown that, SWDs are generated in a neuronal network involving cortical and thalamic areas in both hemispheres. The somatosensory cortex is assumed to contain the site of SWD initiation whereas, the rostral part of the reticular thalamic nucleus probably maintains SWD activity by acting as a pacemaker (Avanzini et al., 2000; Meeren et al., 2005).

4. Advanced Methods in Epileptology

Most traditional epilepsy analysis methods, based on the EEG, are focused on the detection and classification of epileptic seizures. Among these, the best method of analysis is still the visual inspection of the EEG by a highly skilled EEG specialist. Visual analysis of long term EEG records is, however, time consuming and laborious. Therefore, with the advent of new signal processing methodologies, several computerized techniques have been proposed to detect and localize epileptic seizures. Consequently, various automated spike detection approaches have been developed. Artificial neural networks (ANNs) have been used for seizure detection by many researchers. The Kohonen self-organizing feature map ANN was used for spike detection. The major problem with these methods is that the epileptic seizure signals do not follow similar patterns. Presenting all types of seizure patterns to the ANN, on the other hand, reduces the sensitivity of the overall detection system. Therefore, a clever feature detection followed by a robust classifier often provides an improved result (Sanei & Chambers, 2007).

Among recent works, TF approaches effectively use the fact that the seizure sources are localized in the TF domain. Most of these methods are mainly for detection of neural spikes of different types. Different TF methods following different classification strategies have been proposed by many researchers in this area. The methods are especially useful since the EEG signals are statistically nonstationary. The discrete wavelet transform (DWT) obtains a better TF representation than the TF based on the short-term Fourier transform due to its

multiscale (multilevel) characteristics; i.e. it can model the signal according to its coarseness. The DWT analyses the signal over different frequency bands, with different resolutions, by decomposing the signal into a coarse approximation and detail information. In a recent approach, a DWT-based TF method followed by an ANN has been suggested. The ANN classifies the energy of various resolution (detail) levels. Using this technique, it is possible to detect more than 80% of adult epileptic seizures. Other TF distributions such as the pseudo-Wigner–Ville can also be used for the same purpose (Sanei & Chambers, 2007).

Epilepsy is considered to be a dynamic disease. For this reason, it is characterized by qualitative changes from normal behavior to abnormal dynamics of some variables. Epileptic subjects display long periods of normal EEG activity intermingled occasionally with epileptiform paroxysmal activity. Thus, some measures of dynamical change have also been used for seizure detection. These measures significantly change in the transition between the preictal and ictal states or even in the transition between the interictal and ictal states (Suffczynski et al., 2006; Sanei & Chambers, 2007).

The mechanisms of epileptogenesis -the transformation of a naive network to one that generates seizures- are poorly understood (Khalilov et al., 2005). One of the classical ways to study epileptogenesis is by studying clinical and electroencephalographic characteristics. Modern approaches utilize advanced methods such as computational models that are based on neuroanatomical and electrophysiological properties of the circuitry that is involved in the development of SWDs (van Luijtelaar, et al., 2004). Wladimir Yakhno and colleagues demonstrated that normal and pathological oscillations may emerge in the same sensory network and that changes in the degree of interdependence among the distributed neuron ensembles, constituting of the reticular thalamic nucleus, the thalamic relay nuclei and the cortex, gives rise to various models of operandi of neural firing patterns similar to normal firing as well as to SWDs (Yakhno et al., 2004).

Nowadays, genetic absence epilepsy rodent model is commonly used for studying spectral and TF analysis of SWD patterns under various physiological and pharmacological drug conditions, automatic seizure detection, and seizure prediction. Inna Midzyanovskaya from the Institute of Higher Nervous Activity in close collaboration with Vasely Strelkov analysed long-term EEG records of WAG/Rij rats. They reported in their article the general parameters of S&W activity and its spectral characteristics. They found various types of regularities: a strong modulation with a period length of 24-hours, as well as modulations with period lengths ranging from a few minutes to several hours. Their analyses of the intervals between two successive periods with SWDs also revealed a weak rhythm, with an interval of 3-6 seconds, corresponding to a frequency of 0.17 to 0.33 Hz (Midzyanovskaya, et al., 2006).

Wavelet transforms provide information about the time and the frequency structures of a signal simultaneously. For the investigation of dynamic characteristics of SWDs characterizing absence epilepsy during physiological conditions and after the administration of a drug, Bosnyakova et al., used the wavelet transform. This allowed the researcher to observe that the periodical SWD amplitude changes in the range from tenths of a second to one second, as well as the SWDs frequency from the beginning to the end of the discharge. The results of this work also demonstrated the usefulness of applying wavelet transforms for TF analysis of SWD patterns under various pharmacological drug conditions, addressing different brain mediator systems (Bosnyakova, et al., 2006).

In an established work by Tel'nykh et al., an algorithm has been proposed for detecting SWDs, sleep spindles and other characteristic phasic events in the EEG recorded from the cortical and subcortical structures in WAG/Rij rats. The program is capable for recording, analyzing, and automatic finding of characteristic features in the EEG (Tel'nykh et al., 2004).
In addition, based on the mathematical theory of nonlinear dynamics, there has been an increased interest in the analysis of the EEG for the prediction of epileptic seizures. It has been shown that epileptic sources gradually tend to be less chaotic from a few minutes before the seizure onset. This finding is clinically very important since it indicates that the patients do not need to be under anticonvulsant administration permanently, but from just a few minutes before seizure (Sanei & Chambers, 2007).
In terms of genetic absence epilepsy rats and human patients, SWDs emerge suddenly from a normal background EEG and do not seem to be anticipated by any peculiar EEG changes. This gives the impression that S&W seizures are "suddenly generalized". The common definition of S&W seizures, given as suddenly generalized and bilaterally synchronous activities, may be valid at the macroscopic EEG level. However; forerunners of S&W seizures are almost invisible on macroscopic EEG level, although neuronal activity has definitely been changed before the seizure onset. In fact, cortical neurons display time lags between their rhythmic spike trains and progressively increased synchrony. These neuronal processes may involve subtle EEG changes that cannot be easily seen in EEG, but can be detected and validated with methods of mathematical analysis of digitally recorded EEG signal. Unfortunately, only few studies have reported EEG changes during transitional state between background activity and S&W seizures (Pinault et al., 2001; Inouye et al., 1990; Steriade & Amzica, 1994; van Luijtelaar & Sitnikova, 2006).

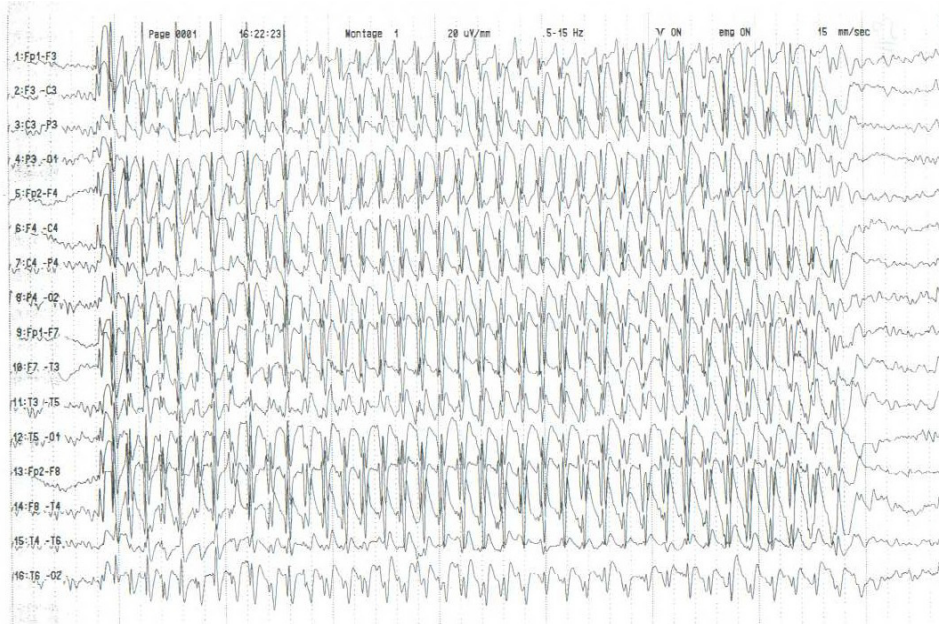


Fig. 1. EEG recording of an absence seizure showing the spike-wave discharges.

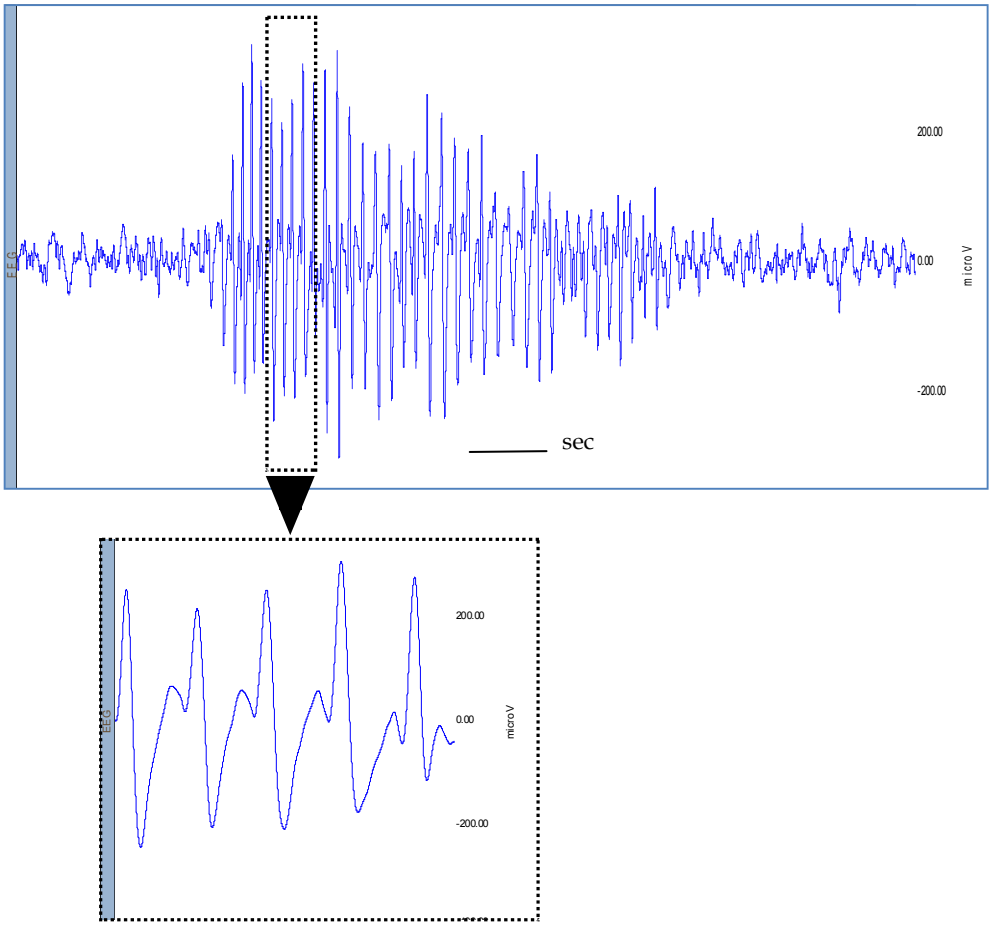


Fig 2. Representative spike-wave discharges in the EEG of a WAG/Rij rat.

5. Spectral Analysis Methods for Spike-Wave Discharges

In order to achieve PSD estimates which represent the changes in frequency with respect to time and to obtain the features, the classical methods (nonparametric or fast Fourier transform-based methods), model-based methods (autoregressive, moving average, and autoregressive moving average methods), TF methods (wavelet transform) are presented in the following.

5.1 Nonparametric methods

The nonparametric methods of spectral estimation rely entirely on the definitions of the equations (1) and (2) of PSD to provide spectral estimates. These methods constitute the "classical means" for PSD estimation. We first introduce two common spectral estimators,

the periodogram and the correlogram derived directly from equations (1) and (2), respectively.

$$P(f) = \lim_{N\to\infty} E\left\{\frac{1}{N}\left|\sum_{n=1}^{N} x(n)e^{-j2\pi fn}\right|^2\right\} \tag{1}$$

$$P(f) = \sum_{k=-\infty}^{\infty} r(k)e^{-j2\pi fk} \tag{2}$$

where $P(f)$ is power spectral density and $r(k)$ is autocorrelation function of the signal under study.

These methods are equivalent under weak conditions. The periodogram and correlogram methods provide reasonably high resolution for sufficiently long data lengths, but are poor spectral estimators because their variance is high and does not decrease with increasing data length. The high variance of the periodogram and correlogram methods motivates the development of modified methods that have lower variance, at a cost of reduced resolution. The modified power spectrum estimation methods described in this section are developed by Bartlett (1948), Blackman and Tukey (1958), and Welch (1967) (Kay & Marple, 1981; Kay, 1988; Proakis & Manolakis, 2007; Stoica & Moses, 1997). These methods make no assumption about how the data were generated and hence are called nonparametric. The spectral estimates are expressed as a function of the continuous frequency variable f, in practice, the estimates are computed at discrete frequencies via the fast Fourier transform (FFT) algorithm.

5.2 Parametric methods

The parametric or model-based methods of spectral estimation assume that the signal satisfies a generating model with known functional form, and then proceed by estimating the parameters in the assumed model. The signal's spectral characteristics of interest are then derived from the estimated model. The models to be discussed are the time series or rational transfer function models. They are the autoregressive (AR) model, the moving average (MA) model, and the autoregressive-moving average (ARMA) model. The AR model is suitable for representing spectra with narrow peaks. The MA model provides a good approximation for those spectra which are characterized by broad peaks and sharp nulls. Such spectra are encountered less frequently in applications than narrowband spectra, so there is a somewhat limited interest in using the MA model for spectral estimation. For this reason, our discussion of the MA spectral estimation will be brief. Spectra with both sharp peaks and deep nulls can be modeled by ARMA model. However, the great initial promise of ARMA spectral estimation diminishes to some extent because there is yet no well-established algorithm, from both theoretical and practical standpoints, for ARMA parameter estimation. The theoretically optimal ARMA estimators are based on iterative procedures whose global convergence is not guaranteed. The practical ARMA estimators are computationally simple and often quite reliable, but their statistical accuracy may be poor in some cases (Kay & Marple, 1981; Kay, 1988; Proakis & Manolakis, 2007; Stoica & Moses, 1997).

5.2.1 AR method

AR method is the most frequently used parametric method because estimation of the AR parameters can be done easily by solving linear equations. In the AR method, data can be modeled as output of a causal, all-pole, discrete filter whose input is white noise. The AR method of order p is expressed as the following equation:

$$x(n) = -\sum_{k=1}^{p} a(k)x(n-k) + w(n), \tag{3}$$

where $a(k)$ are the AR coefficients and $w(n)$ is white noise of variance equal to σ^2. The AR (p) model can be characterized by the AR parameters $\{a[1], a[2], \ldots, a[p], \sigma^2\}$. The PSD is

$$P_{AR}(f) = \frac{\sigma^2}{|A(f)|^2}, \tag{4}$$

where $A(f) = 1 + a_1 e^{-j2\pi f} + \ldots + a_p e^{-j2\pi fp}$.

To obtain stable and high performance AR method, some factors must be taken into consideration such as selection of the optimum estimation method, selection of the model order, the length of the signal which will be modeled, and the level of stationary of the data (Kay & Marple, 1981; Kay, 1988; Proakis & Manolakis, 2007; Stoica & Moses, 1997).

Because of the good performance of the AR spectral estimation methods as well as the computational efficiency, many of the estimation methods to be described are widely used in practice. The AR spectral estimation methods are based on estimation of either the AR parameters or the reflection coefficients. Except the maximum likelihood estimation, the techniques estimate the parameters by minimizing an estimate of the prediction error power. The maximum likelihood estimation method is based on maximizing the likelihood function (Kay & Marple, 1981; Kay, 1988; Proakis & Manolakis, 2007; Stoica & Moses, 1997).

5.2.2 MA method

The MA method is one of the model-based methods in which the signal is obtained by filtering white noise with an all-zero filter. Estimation of the MA spectrum can be done by the reparameterization of the PSD in terms of the autocorrelation function. The q th-order MA PSD estimation is (Kay & Marple, 1981; Kay, 1988; Proakis & Manolakis, 2007; Stoica & Moses, 1997)

$$\hat{P}_{MA}(f) = \sum_{k=-q}^{q} \hat{r}(k)e^{-j2\pi fk}. \tag{5}$$

5.2.3 ARMA method

The spectral factorization problem associated with a rational PSD has multiple solutions, with the stable and minimum phase ARMA model being one of the model-based methods. A reliable method is to construct a set of linear equations and to use the method of least squares on the set of equations. Suppose that for an ARMA of order p, q the autocorrelation

sequence can be accurately estimated up to lag M, where $M > p+q$. Then the following set of linear equations can be written:

$$\begin{bmatrix} r(q) & r(q-1) & \cdots & r(q-p+1) \\ r(q+1) & r(q) & \cdots & r(q-p+2) \\ \vdots & \vdots & & \\ r(M-1) & r(M-2) & & r(M-p) \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{bmatrix} = - \begin{bmatrix} r(q+1) \\ r(q+2) \\ \vdots \\ r(M) \end{bmatrix}, \tag{6}$$

or equivalently,

$$Ra = -r. \tag{7}$$

Since dimension of R is $(M-q)xp$ and $M-q > p$ the least squares criterion can be used to solve for the parameter vector a. The result of this minimization is

$$\hat{a} = -\left(R^* R\right)^{-1}\left(R^* r\right). \tag{8}$$

Finally the estimated ARMA power spectrum is (Kay & Marple, 1981; Kay, 1988; Proakis & Manolakis, 2007; Stoica & Moses, 1997)

$$\hat{P}_{ARMA}(f) = \frac{\hat{P}_{MA}(f)}{\left|1 + \sum_{k=1}^{p} \hat{a}(k) e^{-j2\pi fk}\right|^2}, \tag{9}$$

where $\hat{P}_{MA}(f)$ is estimate of the MA PSD and is given in equation (5).

5.2.4 Selection of AR, MA, ARMA model orders

One of the most important aspects of the use in model-based methods is the selection of the model order. Much work has been done by various investigators on this problem and many experimental results have been given in the literature (Kay & Marple, 1981; Kay, 1988; Proakis & Manolakis, 2007; Stoica & Moses, 1997). One of the better known criteria for selecting the model order has been proposed by Akaike (1974), called the Akaike information criterion (AIC), is based on selecting the order that minimizes equation (10) for the AR method, equation (11) for the MA method, and equation (12) for the ARMA method:

$$AIC(p) = \ln \hat{\sigma}^2 + 2p/N, \tag{10}$$

$$AIC(q) = \ln \hat{\sigma}^2 + 2q/N, \tag{11}$$

$$AIC(p,q) = \ln \hat{\sigma}^2 + 2(p+q)/N, \tag{12}$$

where $\hat{\sigma}^2$ is the estimated variance of the linear prediction error.

5.3 Wavelet Transform

The WT is designed to address the problem of nonstationary signals. It involves representing a time function in terms of simple, fixed building blocks, termed wavelets. These building blocks are actually a family of functions which are derived from a single generating function called the mother wavelet by translation and dilation operations. Dilation, also known as scaling, compresses or stretches the mother wavelet and translation

shifts it along the time axis (Akay, 1998; Daubechies, 1990; Unser & Aldroubi, 1996; Mallat, 1998; Soltani, 2002; Übeyli, 2008).

The WT can be categorized into continuous and discrete. Continuous wavelet transform (CWT) is defined by

$$\mathrm{CWT}(a,b) = \int_{-\infty}^{+\infty} x(t)\psi_{a,b}^{*}(t)dt \,, \tag{13}$$

where $x(t)$ represents the analyzed signal, a and b represent the scaling factor (dilatation/compression coefficient) and translation along the time axis (shifting coefficient), respectively, and the superscript asterisk denotes the complex conjugation. $\psi_{a,b}(\cdot)$ is obtained by scaling the wavelet at time b and scale a :

$$\psi_{a,b}(t) = \frac{1}{\sqrt{|a|}}\psi\left(\frac{t-b}{a}\right) , \tag{14}$$

where $\psi(t)$ represents the wavelet.

Continuous, in the context of the WT, implies that the scaling and translation parameters a and b change continuously. However, calculating wavelet coefficients for every possible scale can represent a considerable effort and result in a vast amount of data. Therefore DWT is often used. The WT can be thought of as an extension of the classic Fourier transform, except that, instead of working on a single scale (time or frequency), it works on a multi-scale basis. This multi-scale feature of the WT allows the decomposition of a signal into a number of scales, each scale representing a particular coarseness of the signal under study. In the procedure of multiresolution decomposition of a signal $x[n]$, each stage consists of two digital filters and two downsamplers by 2. The first filter, $g[\cdot]$ is the discrete mother wavelet, high-pass in nature, and the second, $h[\cdot]$ is its mirror version, low-pass in nature. The downsampled outputs of first high-pass and low-pass filters provide the detail, D_1 and the approximation, A_1, respectively. The first approximation, A_1 is further decomposed and this process is continued (Akay, 1998; Daubechies, 1990; Unser & Aldroubi, 1996; Mallat, 1998; Soltani, 2002; Übeyli, 2008).

6. Results of Analysis

The PSDs describe the distribution of power with frequency. In this study, the PSDs of the SWDs of WAG/Rij rats were obtained by using the FFT, Burg AR, MA, and least squares modified Yule-Walker ARMA methods. The sample PSDs of the SWD records of WAG/Rij rats are presented in Figures 3 and 4. When the PSDs are examined, it is seen that classical method (FFT) has large variance (Figures 3 and 4). The FFT method is based on a finite record of data, the frequency resolution of these methods equal to the spectral width of the window length N, which is approximately $1/N$. The principal effect of windowing that occurs when processing with the FFT is to smear or smooth the estimated spectrum. This method suffers from spectral leakage effects, due to windowing that are inherent in finite-length data records. Often, the spectral leakage masks weak signals that are present in the

data. Smearing and spectral leakage are particularly critical for spectra with large amplitude ranges, such as peaky spectra.

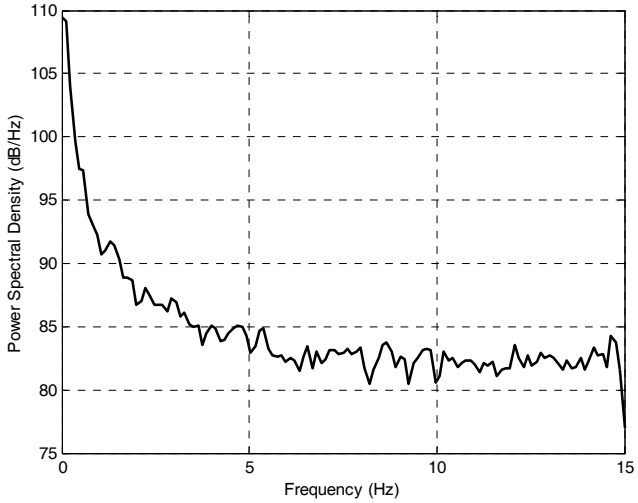


Fig. 3. PSD of sample SWD recods obtained by FFT method

A model for the signal generation can be constructed with a number of parameters that can be estimated from the observed data. From the model and the estimated parameters, the PSD can be computed. The modeling approach eliminates the need for window functions and the assumption that the autocorrelation sequence is zero outside the window. Spectra with both sharp peaks and deep nulls cannot be modeled by either AR or MA methods. In these cases, the ARMA spectral estimation provides an opportunity to improve on the AR and MA spectral estimations. By combining poles and zeros, the ARMA method provides a more efficient representation, from the viewpoint of the number of model parameters, of the spectrum of a random process. When the PSDs are examined, the Burg AR and the least squares modified Yule-Walker ARMA methods' performance characteristics have been found to be superior to the FFT and the MA methods.

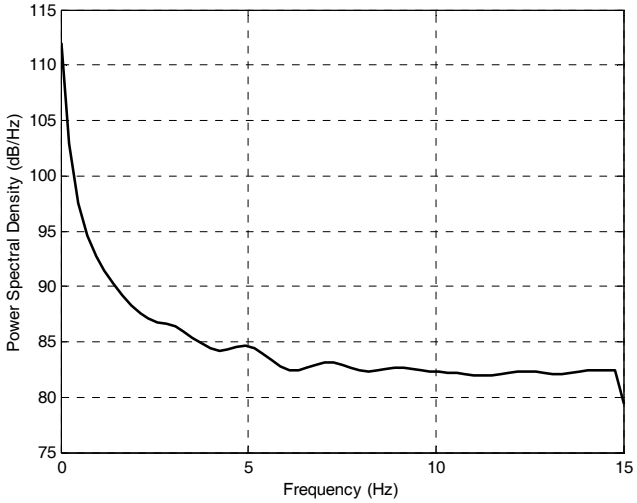


Fig. 4. PSD of sample SWD recods obtained by ARMA method

The selection of the model orders in the AR, MA, and ARMA spectral estimators is a critical subject. Too low order results in a smoothed estimate, while too large order causes spurious peaks and general statistical instability. In the case of the dimension of autocorrelation matrix is inappropriate and the model orders chosen incorrect, poor spectral estimates are obtained by the AR, MA, and ARMA spectral estimators. Heavy biases and/or large variabilities may be exhibited. In this study, Akaike Information Criteria (Akaike, 1974) was taken as the base for choosing the model order. According to the equations (10), (11), and (12) model order p was taken as 10 for the AR method, model order q was taken as 10 for the MA method and model orders p and q were taken as 10 for the ARMA method.

The mean values of the peak frequencies and power levels of the PSDs of all SWD records of WAG/Rij rats are given in Table 1. According to the values presented in Table 1, the PSD of the FFT method has spurious peaks and does not produce accurate spectral estimates due to limits on resolution. Since the PSD of the SWD records obtained by the MA method is smooth, the MA method has been found inappropriate for the SWDs. The peak frequencies and power levels of the AR and ARMA methods are similar for the PSDs of the SWDs of WAG/Rij rats. The AR and the ARMA methods produce frequency estimates which are unbiased and nearly attain the Cramer-Rao bound. From Table 1, one can see that the AR and the ARMA methods produce the true frequencies as the peaks of the spectral estimates for the SWDs. The obtained results demonstrated that the peak frequencies and the power levels of the AR and ARMA PSDs can be used as the features representing the SWD records of WAG/Rij rats.

Method	P_1 / f_1	P_2 / f_2	P_3 / f_3	P_4 / f_4
FFT	Spurious peaks			
AR	113.1706/0	85.0981/4.9219	82.8657/7.2656	83.0971/14.2969
MA	110.8564/0	–	–	81.4569/14.7656
ARMA	111.8974/0	85.5421/4.9219	82.8964/7.2656	81.8633/14.2969

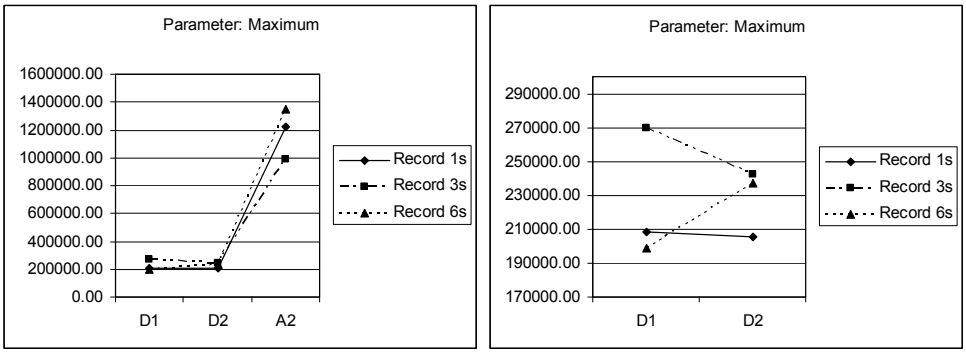
Table 1. Mean values of peak frequencies and power levels of PSDs of all SWD recods

The spectral analysis of the SWDs of WAG/Rij rats was performed using the DWT. The selection of appropriate wavelet and the number of decomposition levels is very important in analysis of signals using the DWT. The number of decomposition levels is chosen based on the dominant frequency components of the signal. The levels are chosen such that those parts of the signal that correlate well with the frequencies required for classification of the signal are retained in the wavelet coefficients. In the present study, the number of decomposition levels was chosen to be 2. Thus, the SWD records were decomposed into the details $D_1 - D_2$ and one final approximation, A_2. Usually, tests are performed with different types of wavelets and the one which gives maximum efficiency is selected for the particular application. The smoothing feature of the Daubechies wavelet of order 2 (db2) made it more suitable to detect changes of the signals under study. Therefore, the wavelet coefficients were computed using the db2 in the present study. The frequency bands corresponding to different levels of decomposition for db2 with a sampling frequency of 60 Hz are: D_1 (7.5-15Hz); D_2 (3.75-7.5Hz); and A_2 (0-3.75Hz). The wavelet coefficients were computed using the MATLAB software tool (Übeyli et al., 2008; Übeyli et al., 2009).

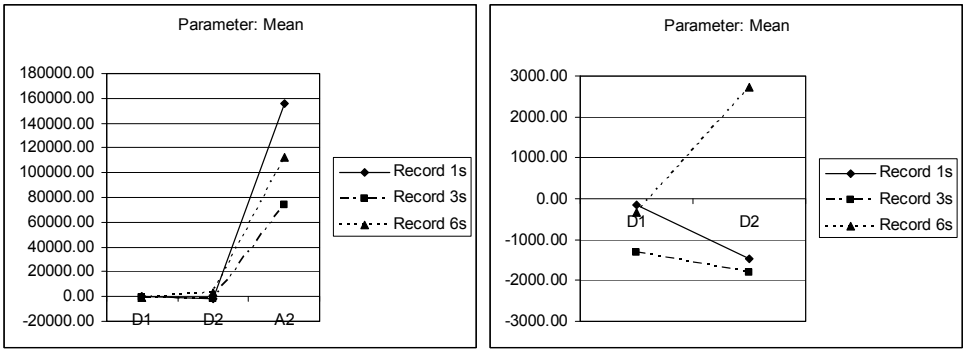
The computed wavelet coefficients provide a compact representation that shows the energy distribution of the signal in time and frequency. Therefore, the computed wavelet coefficients of the SWD records for each WAG/Rij rats were used as the feature vectors representing the signals. In order to reduce the dimensionality of the extracted feature vectors, statistics over the set of the wavelet coefficients was used. The following statistical features were used to represent the TF distribution of the signals under study:

1. Maximum of the wavelet coefficients in each subband.
2. Mean of the wavelet coefficients in each subband.
3. Minimum of the wavelet coefficients in each subband.
4. Standard deviation of the wavelet coefficients in each subband.

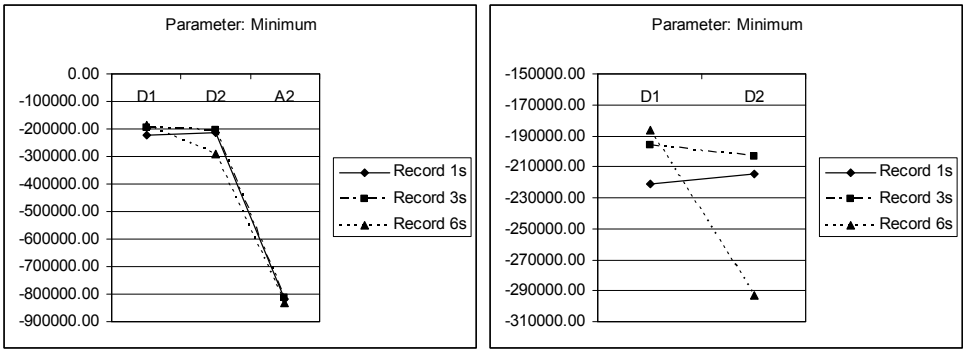
68 typical SWDs (length 4-8s) obtained from 8 WAG/Rij rats were analyzed. The wavelet coefficients of 1s, 3s and 6s were computed for each SWD. Figure 5 demonstrates the maximum, mean, minimum and standard deviation of the wavelet coefficients in each subband (D_1, D_2, A_2) of the SWD records obtained from the 1st WAG/Rij rat for 1s, 3s and 6s. From Figure 5, one can see that the computed features (wavelet coefficients in each subband) of the SWD records of WAG/Rij rats in various seconds are different from each other. This figure indicated that the wavelet coefficients can be used to identify characteristics of the SWD records of WAG/Rij rats that were not apparent from the original time domain signal (Übeyli et al., 2008; Übeyli et al., 2009).


Parameter: Maximum
1600000.00
1400000.00
1200000.00
1000000.00
800000.00
600000.00
400000.00
200000.00
0.00
D1
D2
A2
Record 1s
Record 3s
Record 6s
Parameter: Maximum
290000.00
270000.00
250000.00
230000.00
210000.00
190000.00
170000.00
D1
D2
Record 1s
Record 3s
Record 6s

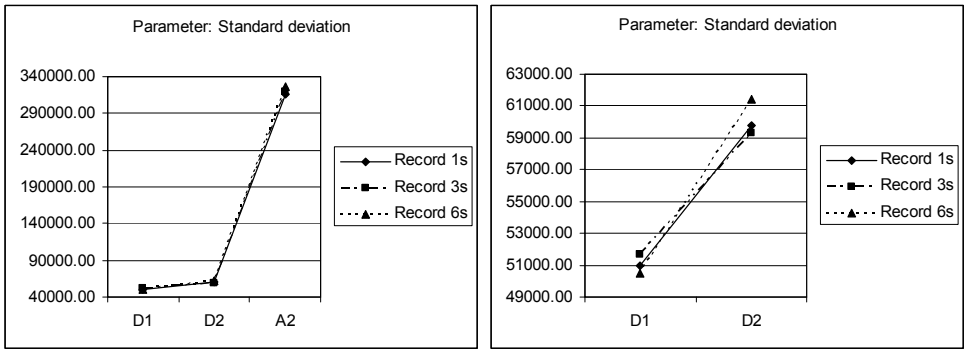
(a)


Parameter: Mean
180000.00
160000.00
140000.00
120000.00
100000.00
80000.00
60000.00
40000.00
20000.00
0.00
-20000.00
D1
D2
A2
Record 1s
Record 3s
Record 6s
Parameter: Mean
3000.00
2000.00
1000.00
0.00
-1000.00
-2000.00
-3000.00
D1
D2
Record 1s
Record 3s
Record 6s

(b)


Parameter: Minimum
0.00
-100000.00
-200000.00
-300000.00
-400000.00
-500000.00
-600000.00
-700000.00
-800000.00
-900000.00
D1
D2
A2
Record 1s
Record 3s
Record 6s
Parameter: Minimum
-150000.00
-170000.00
-190000.00
-210000.00
-230000.00
-250000.00
-270000.00
-290000.00
-310000.00
D1
D2
Record 1s
Record 3s
Record 6s

(c)



(d)

Fig. 5. Analysis results of SWD records of WAG/Rij rat 1

7. Conclusion

In this chapter, promising results in detecting the changes in the SWD records of WAG/Rij rats were presented. The SWD records of WAG/Rij rats were processed using the FFT, Burg AR, MA, and least squares modified Yule-Walker ARMA methods. Performance of these methods were compared in terms of their frequency resolution and the effects in clinical applications. Since the FFT and the MA methods have low spectral resolution, these two methods have not been found appropriate for evaluating the PSDs of the SWD records of WAG/Rij rats. The performance characteristics of the Burg AR and the least squares modified Yule-Walker ARMA methods have been found extremely valuable for analysis of the SWD records, because of their clear spectra. In conclusion, it should be emphasized that the AR and ARMA methods were found extremely valuable for extraction of the features representing the SWD records of WAG/Rij rats.

The features from the SWD records of WAG/Rij rats were obtained by usage of the DWT. The computed wavelet coefficients can be used as features representing and/or discriminating the SWD records of WAG/Rij rats in various seconds. The results showed that the DWT can be useful to analyze TF dynamics of SWDs both in physiological conditions and after pharmacological interventions for future investigations.

Acknowledgement
The authors would like to thank Dr Hikmet Aloğlu for providing the patient EEG sample from his sleep laboratory recordings.

8. References

Akaike, H. (1974). A new look at the statistical model identification, *IEEE Transactions on Automatic Control*, AC-19, 716-723, ISSN: 0018-9286.

Akay, M. (1998). *Time Frequency and Wavelets in Biomedical Signal Processing*, Institute of Electrical and Electronics Engineers, Inc., New York, ISBN: 0-7803-1147-7.

Ates, N.; Esen, N. & Ilbay, G. (1999). Absence epilepsy and regional blood-brain barrier permeability: the effects of pentylenetetrazole-induced convulsions. *Pharmacological Research*, Vol. 39, No. 4, 305-310, ISSN: 1043-6618.

Ates, N.; Sahin, D. & Ilbay, G. (2004). Theophylline, a methylxanthine derivative, suppresses absence epileptic seizures in WAG/Rij rats. *Epilepsy Behavior*, Vol. 5, No. 5, 645-648, ISSN: 1525-5050.

Avanzini, G.; Panzica, F. & de Curtis, M. (2000). The role of the thalamus in vigilance and epileptogenic mechanisms. *Clinical Neurophysiology,* Vol. 111, Supp. l 2, 19-26, ISSN: 1388-2457.

Bare, M.A.; Glauser, T.A. & Strawsburg, R.H. (1998). Need for electroencephalogram video confirmation of atypical absence seizures in children with Lennox-Gastaut syndrome. *Journal of Child Neurology,* Vol. 13, 498–500, ISSN: 0883-0738.

Berne, R.M. & Levy, M.N.(1993). *Physiology*. Mosby-Year Book, USA, ISBN: 0-8016-6465-9.

Bosnyakova, D.; Gabova, A.; Kuznetsova, G.; Obukhov, Y.; Midzyanovskaya, I.; Salonin, D.; van Rijn, C.; Coenen, A.; Tuomisto, L. & van Luijtelaar G. (2006). Time-frequency analysis of spike-wave discharges using a modified wavelet transform. *Journal of Neuroscience Methods*, Vol. 154, 80-88, ISSN: 0165-0270.

Bosnyakova, D.; Gabova, A.; Zharikova, A.; Gnezditski, V., Kuznetsova, G. & van Luijtelaar, G. (2007). Some peculiarities of time-frequency dynamics of spike-wave discharges in humans and rats. *Clinical Neurophysiology*, Vol. 118, 1736-1743, ISSN: 1388-2457.

Bouwman, B.M. & Van Rijn, C.M. (2004). Effects of levetiracetam on spike and wave discharges in WAG/Rij rats. *Seizure*, Vol. 13, No. 8, 591-594, ISSN: 1059-1311.

Bouwman, B.M.; Suffczynski, P.; Lopes da Silva, F.H.; Maris, E. & van Rijn, C.M. (2007). GABAergic mechanisms in absence epilepsy: a computational model of absence epilepsy simulating spike and wave discharges after vigabatrin in WAG/Rij rats. *European Journal of Neuroscience*, Vol. 25, No. 9, 2783-2790, ISSN: 0953-816X.

Carmant, L.; Kramer, U.; Holmes, G.L.; Mikati, M.A.; Riviello, J.J. & Helmers, S.L. (1996). Differential diagnosis of staring spells in children: a video-EEG study. *Pediatric Neurology,* Vol. 14, 199–202, ISSN: 0887-8994.

Cocito, L. & Primavera, A. (1998). Vigabatrin aggravates absences and absence status. N*eurology,*Vol. 51, 1519–1520, ISSN: 0028-3878.

Coenen, A.M.; Drinkenburg, W.H.; Inoue, M. & van Luijtelaar, E.L.. (1992). Genetic models of absence epilepsy, with emphasis on the WAG/Rij strain of rats. *Epilepsy Research*, Vol. 12, No. 2, 75-86, ISSN: 0920-1211.

Coenen, A.M.L. (1995). Neuronal activities underlying the electroencephalogram and evoked potentials of sleeping and waking: implications for information processing. *Neuroscience & Biobehavioral Reviews* Vol. 19, 447–463, ISSN: 0149-7634.

Coenen, A.M. & Van Luijtelaar, E.L. (2003). Genetic animal models for absence epilepsy: a review of the WAG/Rij strain of rats. *Behavior Genetics*, Vol. 33, No. 6, 635-55, ISSN: 0001-8244.

Cortez, M.A.; McKerlie, C. & Snead, O.C. III. (2001). A model of atypical absence seizures: EEG, pharmacology, and developmental characterization. *Neurology,* Vol. 56**,** 341–349, ISSN: 0028-3878.

Daubechies, I. (1990). The wavelet transform, time-frequency localization and signal analysis. *IEEE Transactions on Information Theory*, Vol. 36, No. 5, 961-1005, ISSN: 0018-9448.

Depaulis, A. & van Luijtelaar, G. (2006). Genetic models of absence epilepsy in the rat. In: Pitkanen A.,. Schwartzkroin P and Moshe S., Editors, *Models of Seizures and Epilepsy*, Elsevier, ISBN: 978-0-12-088554-1.

Drinkenburg, W.H.; van Luijtelaar, E.L.; van Schaijk, W.J. & Coenen, A.M. (1993). Aberrant transients in the EEG of epileptic rats: a spectral analytical approach. *Physiology & Behavior*, Vol. 54, No. 4, 779-83, ISSN: 0031-9384.

Elger, C.E. & Schmidt, D. (2008). Modern management of epilepsy: a practical approach. *Epilepsy Behavior,* Vol. 12, No. 4, 501-39, ISSN: 1043-6618.

Engel, J.Jr. & Pedley, T. (2007). Introduction: What Is Epilepsy? In: Engel J Jr, Pedley T, editors. *Epilepsy: a comprehensive textbook*. Philadelphia: Lippincott Williams & Wilkins, ISBN: 978-0-7817-5777-5.

Engel, J.; Williamson, P.D.; Berg, A.T. & Wolf, P. (2007). Classification of Epileptic Seizures. In: Engel J Jr, Pedley T, editors. *Epilepsy: a comprehensive textbook*. Philadelphia: Lippincott Williams & Wilkins, ISBN: 978-0-7817-5777-5.

Fisher, R.S.; van Emde, B.W. & Blume, W.*et al*., (2005). Epileptic seizures and epilepsy: definitions proposed by the International League Against Epilepsy (ILAE) and the International Bureau for Epilepsy (IBE), *Epilepsia,* Vol. 46, 470–472, ISSN: 0013-9580.

Fisher, R.S. & Leppik, I. (2008). Debate: When does a seizure imply epilepsy? *Epilepsia,* Vol. 49, Suppl 9, 7-12, ISSN: 0013-9580.

Gloor, P. & Fariello, R.G. (1988). Generalized epilepsy: some of its cellular mechanisms differ from those of focal epilepsy. *Trends in Neuroscience,* Vol. 11**,** 63–68, ISSN: 0166-2236.

Hauser, W.A.; Annegers, J.F. & Rocca, W.A. (1996). Descriptive epidemiology of epilepsy: contributions of population-based studies from Rochester, Minnesota. *Mayo Clinic Proceedings*; Vol. 71, 576–86, ISSN: 0025-6196.

Henkin, Y.; Sadeh, M.; Kivity, S.; Shabtai, E.; Kishon-Rabin, L. & Gadoth, N. (2005). Cognitive function in idiopathic generalized epilepsy of childhood. *Developmental Medicine & Child Neurology,* Vol. 47, 126–132, ISSN: 0012-1622.

Holmes, G.L.; McKeever, M. & Adamson, M. (1987). Absence seizures in children: clinical and electroencephalographic features. *Annals of Neurology,* Vol. 21, 268–273, ISSN: 0364-5134.

Høie, B.; Mykletun, A.; Sommerfelt, K.; Bjornaes, H.; Skeidsvoll, H. & Waaler, P.E. (2005). Seizure-related factors and non-verbal intelligence in children with epilepsy. A population-based study from Western Norway. *Seizure,* Vol. 14, 223–231, ISSN: 1059-1311.

ILAE. (1989). Proposal for revised classification of epilepsies and epileptic syndromes. Commission on Classification and Terminology of the International League Against Epilepsy. *Epilepsia,* Vol. 30, 389–399, ISSN: 0013-9580.

Ilbay, G.; Sahin, D.; Karson, A. & Ates, N. (2001). Effects of adenosine administration on spike-wave discharge frequency in genetically epileptic rats. *Clinic & Experimental Pharmacology & Physiology,* Vol. 28, No. 8, 643-6, ISSN: 0305-1870.

Inouye, T.; Sakamoto, H.; Shinosaki, K.; Toi, S. & Ukai, S. (1990). Analysis of rapidly changing EEGs before generalized spike and wave complexes. *Electroencephalography & Clinical Neurophysiology,* Vol. 76, No. 3, 205-21, ISSN: 0013-4694.

Kay, S.M. & Marple, S.L. (1981). Spectrum analysis – A modern perspective. *Proceedings of the IEEE*, Vol. 69, No. 11, 1380-1419, ISSN: 0018-9219.

Kay, S.M. (1988). *Modern Spectral Estimation: Theory and Application*, Prentice Hall, New Jersey, ISBN: 0-13-598582-X 025.

Kellaway, P. (1985). Childhood seizures. *Electroencephalography & Clinical Neurophysiology,* Vol. 37, 267-83, ISSN: 0013-4694.

Khalilov, I.; Le Van Quyen, M.; Gozlan, H. & Ben-Ari, Y (2005). Epileptogenic actions of GABA and fast oscillations in the developing hippocampus. *Neuron,* Vol. 48, No.5, 787-96, ISSN: 0896-6273.

Knake, S.; Hamer, H.M.; Schomburg, U.; Oertel, W.H. & Rosenow, F. (1999)Tiagabine-induced absence status in idiopathic generalized epilepsy. *Seizure,* Vol. 8**,** 314–317, ISSN: 1059-1311.

Koutroumanidis, M. & Smith, S. (2005). Use and Abuse of EEG in the Diagnosis of Idiopathic Generalized Epilepsies. *Epilepsia,* Vol. 46, Suppl. 9, 96–107, ISSN: 0013-9580.

Kwan, P. & Sander, J.W. (2004). The natural history of epilepsy: an epidemiological view, *J Neurology Neurosurgery, Psychiatry,* Vol. 75, 1376–1381, ISSN: 0022-3050

Lockman, L.A. (1989). Absence, myoclonic, and atonic seizures. *Pediatric Clinics North America,* Vol. 36, No. 2, 331-41, ISSN: 0031-3955.

Mallat, S. (1998). *A Wavelet Tour of Signal Processing*, Academic Press, San Diego, USA, ISBN: 0-12-466605-1.

Markand, O.N. (2003). Lennox-Gastaut syndrome (childhood epileptic encephalopathy). *Journal of Clinical Neurophysiology, Vol.* 20, 426–441, ISSN: 0736-0258.

Mattson, R.H. (2003). Overview: idiopathic generalized epilepsies. *Epilepsia,* Vol. 44, Suppl 2, 2-6, ISSN: 0013-9580.

Meeren, H.; van Luijtelaar, G.; Lopes da Silva, F. & Coenen, A. (2005). Evolving concepts on the pathophysiology of absence seizures: the cortical focus theory. *Archives of Neurology,* Vol. 62, No. 3, 371-6, ISSN: 0003-9942.

Midzyanovskaya, I.; Strelkov, V.; Rijn, C.; Budziszewska, B.; van Luijtelaar, E. & Kuznetsova, G. (2006). Measuring clusters of spontaneous spike-wave discharges in absence epileptic rats. *J Neuroscience Methods.* Vol. 30, No. 154(1-2), 183-9, ISSN: 0165- 0270.

Nolan, M.; Bergazar, M.; Chu, B.; Cortez, M.A. & Snead O.C. III. (2005). Clinical and neurophysiologic spectrum associated with atypical absence seizures in children with intractable epilepsy. *Journal of Child Neurology,* Vol. 20, 404–410, ISSN: 0883-0738.

Panayiotopoulos, C.P. (2008). Typical absence seizures and related epileptic syndromes: assessment of current state and directions for future research. *Epilepsia,* Vol. 49, No. 12, 2131-9. ISSN: 0013-9580.

Pavone, P.; Bianchini, R.; Trifiletti, R.R.; Incorpora, G.; Pavone, A. & Parano, E. (2001). Neuropsychological assessment in children with absence epilepsy. *Neurology,* Vol. 56, 1047–1051, ISSN: 0028-3878.

Peeters, B.W.; Van Rijn, C.M.; Van Luijtelaar, E.L. & Coenen, A.M.. (1989). Antiepileptic and behavioural actions of MK-801 in an animal model of spontaneous absence epilepsy. *Epilepsy Research*, Vol. 3, No. 2, 178-81, ISSN: 0920-1211.

Pinault, D.; Vergnes, M. & Marescaux C. (2001). Medium-voltage 5-9-Hz oscillations give rise to spike-and-wave discharges in a genetic model of absence epilepsy: in vivo dual extracellular recording of thalamic relay and reticular neurons. *Neuroscience*, Vol. 105, No. 1, 181-201, ISSN: 0306-4522.

Posner, E.B.; Mohamed, K. & Marson, A.G. (2005). Ethosuximide, sodium valproate or lamotrigine for absence seizures in children and adolescents. *Cochrane Database Of Systematic Reviews,* Vol. 19, No: 4, ISSN: 1469-493X.

Proakis, J.G. & Manolakis, D.G. (2007). *Digital Signal Processing Principles, Algorithms, and Applications*, Prentice Hall, New Jersey, ISBN: 0-13-187374-1.

Rodin, E. & Ancheta, O. (1987). Cerebral electrical fields during petit mal absences. *Electroencephalography & Clinical Neurophysiology,* Vol. 66, No. 6, 457-66, ISSN: 0013-4694.

Sanei, S. & Chambers, J.A. (2007). *EEG Signal Processing*. England: John Wiley & Sons Ltd, ISBN: 978-0-470-02581-9.

Schaul, N. (1998). The fundamental neural mechanisms of electroencephalography. . *Electroencephalography & Clinical Neurophysiology,* Vol. 106, No. 2, 101-7, ISSN: 0013-4694.

Sitnikova, E. & van Luijtelaar, G. (2006). Cortical and thalamic coherence during spike-wave seizures in WAG/Rij rats. *Epilepsy Research*, Vol. 71, No. 2-3, 159-80, ISSN: 0920-1211.

Snead, O.C. 3rd. (1995). Basic mechanisms of generalized absence seizures. *Annals Neurology,* Vol. 37, No. 2, 146-57, ISSN: 0364-5134.

Snead, O.C.; Depaulis, A.; Vergnes, M. & Marescaux, C. (1999). Absence epilepsy: advances in experimental animal models. *Advances in Neurology,* Vol. 79**,** 253–278, ISSN: 0091-3952.

Soltani, S. (2002). On the use of the wavelet decomposition for time series prediction. *Neurocomputing*, Vol. 48, 267-277, ISSN: 0925-2312.

Sperling, M.R. & Clancy, R.R. (2007). Ictal Electroencephalogram. In: Engel J Jr, Pedley T, eds. *Epilepsy: a comprehensive textbook*. Philadelphia: Lippincott Williams & Wilkins, 850-853, ISBN: 978-0-7817-5777-5.

Stefan, H.; Snead, O.C. III. & Eeg-Olofsson, O. (2007). Typical and Atypical Absence Seizures, Myoclonic Absences, and Eyelid Myoclonia. In: Engel J Jr, Pedley T, eds. *Epilepsy: a comprehensive textbook*. Philadelphia: Lippincott Williams & Wilkins, 573-574, ISBN: 978-0-7817-5777-5.

Steriade, M.& Amzica, F. (1994). Dynamic coupling among neocortical neurons during evoked and spontaneous spike-wave seizure activity. *J Neurophysiology*, Vol. 72, No. 5, 2051-69. ISSN: 0022-3077.

Stoica, P. & Moses, R. (1997). *Introduction to Spectral Analysis*, Prentice Hall, New Jersey, ISBN: 0-13-258419-0.

Suffczynski, P.; Lopes da Silva, F.H.; Parra, J.; Velis, D.N.; Bouwman, B.M.; van Rijn C.M.; van Hese, P.; Boon, P.; Khosravani, H.; Derchansky, M.; Carlen, P. & Kalitzin, S. (2006). Dynamics of epileptic phenomena determined from statistics of ictal transitions. *IEEE Transaction Biomedical Engineering,* Vol. 53, No. 3, 524-32, ISSN: 0018-9294.

Tel'nykh, O.S.; Tel'nykh, A.A.; Shilin, S.G. (2004). Program for recording, analyzing, and automatic finding of characteristic features in the electroencephalograms. In: Van

Luijtelaar, G.; Kutznetsova, G.D.; Coenen, A. & Chepurnov, S.A. *The WAG/Rij Model of Absence Epilepsy: The Nijmegen-Russion Federation Papers,* NICI, Nijmegen, ISBN: 90-808599-1-5.

Temkin, O. (1994). The Falling Sickness. *A History of Epilepsy from the Greeks to the Beginnings of Modern Neurology*, Johns Hopkins Press, Baltimore, ISBN: 0-8018-4849-0.

Tolmacheva, E.A.; van Luijtelaar, G.; Chepurnov, S.A.; Kaminskij, Y. & Mares, P. (2004). Cortical and limbic excitability in rats with absence epilepsy. *Epilepsy Research*, Vol. 62, No. 2-3, 189-98, ISSN: 0920-1211.

Unser, M. & Aldroubi, A. (1996). A review of wavelets in biomedical applications. *Proceedings of the IEEE*, Vol. 84, No. 4, 626-638, ISSN: 0018-9219.

Übeyli, E.D. (2008). Wavelet/Mixture of experts network structure for EEG signals classification. *Expert Systems with Applications*, Vol. 34, No. 3, 1954-1962, ISSN: 0957-4174.

Übeyli, E.D. (2009). Features for analysis of electrocardiographic changes in partial epileptic patients. *Expert Systems with Applications*, Vol. 36, No. 3, Part 2, 6780-6789, ISSN: 0957-4174.

Übeyli, E.D.; Ilbay, G.; Sahin, D. & Ates, N. (2008). Discrete wavelet transform for analysis of spike-wave discharges in rats, *30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society EMBC08*, pp. 4680-4683, ISBN: 978-1-4244-1814-5, Vancouver, Canada, August 2008.

Übeyli, E.D.; Ilbay, G.; Sahin, D. & Ates, N. (2009). Analysis of spike-wave discharges in rats using discrete wavelet transform. *Computers in Biology and Medicine*, Vol. 39, No. 3, 294-300, ISSN: 0010-4825.

Valentin, A.; Hindocha, N.; Osei-Lah, A.; Fisniku, L.; McCormick, D.; Asherson, P.; Moran, N.; Makoff, A. & Nashef, L. (2007). Idiopathic generalized epilepsy with absences: syndrome classification. *Epilepsia,* Vol. 48, No. 11, 2187-90, ISSN: 0013-9580.

Van Luijtelaar, E.L. & Coenen, A.M. (1986). Two types of electrocortical paroxysms in an inbred strain of rats. *Neuroscience Letters*, Vol. 70, No. 3, 393-7, ISSN: 0304-3940.

Van Luijtelaar, G.; Kutznetsova, G.D.; Coenen, A. & Chepurnov, S.A. (2004). *The WAG/Rij Model of Absence Epilepsy: The Nijmegen-Russion Federation Papers,* NICI, Nijmegen, ISBN: 90-808599-1-5.

Van Luijtelaar, G. & Sitnikova, E. (2006). Global and focal aspects of absence epilepsy: the contribution of genetic models. *Neuroscience & Biobehavioral Reviews*, Vol. 30, No. 7, 983-1003, ISSN: 0149-7634.

Velazquez, J.L.; Huo, J.Z.; Dominguez, L.G.; Leshchenko, Y. and Snead, O.C. 3rd. (2007). Typical versus atypical absence seizures: network mechanisms of the spread of paroxysms. *Epilepsia,* Vol. 48, No. 8, 1585-93, ISSN: 0013-9580.

Walczak, T.S.; Jayakar, P. & Mizrahi, E.M. (2007). Interictal Electroencephalography: In: Engel J Jr, Pedley T. *Epilepsy: a comprehensive textbook*. Philadelphia: Lippincott Williams & Wilkins, 809-810, ISBN: 978-0-7817-5777-5.

Weber, Y.G. & Lerche, H. (2008). Genetic mechanisms in idiopathic epilepsies. *Developmental Medicine & Child Neurology,* Vol. 50, No. 9, 648-54, ISSN: 0012-1622.

Worrell, G.A.; Lagerlund, T.D. & Buchhalter, J.R. (2002). Role and limitations of routine and ambulatory scalp electroencephalography in diagnosing and managing seizures. *Mayo Clinic Proceedings,* Vol. 77, No. 9, 991-8, ISSN: 0025-6196.

Yakhno, V.G.; Ivanov, A.E.; Nuidel, I.V.; Khurlapov, P.G.; Coenen, A.M.L.; Luijtelaar, E.L.J.M. & Kuznetsova, G.D. (2004). A computational model of transitions between normal and pathological brain states. In:Van Luijtelaar, G.; Kutznetsova, G.D.; Coenen, A. & Chepurnov, S.A. *The WAG/Rij Model of Absence Epilepsy: The Nijmegen-Russion Federation Papers,* NICI, Nijmegen, ISBN: 90-808599-1-5.

A 3D Graph-Cut based Algorithm for Evaluating Carotid Plaque Echogenicity and Texture

José C. R. Seabra and João M. R. Sanches
Instituto de Sistemas e Robótica, Instituto Superior Técnico
Universidade Técnica de Lisboa
Portugal

1. Introduction

Carotid atherosclerosis is the main cause of brain stroke, which is one of the most common neurological diseases in western countries. A distinguishing feature of this disease is plaque formation owing to progressive sub-endotelial accumulation of lipid, protein, cellular waste products, calcium and other substances in medium and large-sized blood vessel walls.
Plaques can grow large enough to significantly reduce the blood's flow through particular arteries such as the carotid arteries, leading to what is commonly designated as *stenosis*. However, most of the damage occurs when they become fragile and rupture. Plaques that rupture cause blood clots to form and travel along circulation possibly blocking smaller vessel ahead. If it blocks a blood vessel that feeds the brain, it causes a *brain stroke*.
The degree of stenosis is often pointed as the most significant clinical indicator of stroke risk and it was until recently the most critical criterion used to determine a surgical intervention (Consensus Group, 1995), together with age, health and patient's clinical history.
The benefit of surgery for plaque removal, termed *endarterectomy*, is clearly demonstrated for patients presenting high degree of *stenosis* (more than 60%) (Consensus Group, 1995). It was also demonstrated, however, that patients on medical treatment remained free of symptoms for a long-time period despite the presence of considerable stenotic lesions. This advises for investigating and employing other clinical features for evaluating the risk of plaque rupture (Barnett et al, 2002; Elatrozy et al, 1998).
Currently, there are accurate methods to assess the disease severity based on (Wintermark et al, 2008) or MRI (Saam et al, 2007). However their application is expensive, time consuming and requires equipment which is not yet available and accessible in most clinical facilities. On the other hand, 2D ultrasound is widely available and provides real time data-acquisition and visualization, so it has been so far the preferred technique in the diagnosis and monitoring of the disease.
Plaque echo-morphology (appearance in grey-scale intensity)assessed through B-mode 2D ultrasound plays nowadays an important role in the evaluation of stroke risk in the scope of the atherosclerotic disease.
Several methods were described in the literature (Kyriacou et al, 2009; Mougiakakou et al, 2007; Sztajzel et al, 2005) aiming at assessing the risk of plaque rupture based on the average

echogenicity and texture characterization extracted from 2D ultrasound images of carotid plaques. Current diagnostic methods present, however, two main flaws. First, the computation of echo-morphological features is usually affected by inaccuracy and subjectivity associated with data acquisition and operator dependent image selection. On the other hand, the analysis of overall echogenicity and texture of the carotid plaque, encoded by means of an averaged measure of pixel intensity values and variance, respectively, may not be accurate enough in many cases, namely when plaques are heterogeneous or present significant hypoechogenic regions. An overall measure of echogenicity or texture is incomplete and does not reveal unstable foci inside the plaque which should be considered for a correct evaluation of the risk of plaque rupture. Moreover, most methods for characterization of plaque echo-morphology do not usually take into account the *speckle* noise which can significantly corrupt the ultrasound data.

Fundamentals of plaque echo-morphology, including the most noticeable clinical indicators and methods for echo-morphological analysis are addressed in Section II.

Section III delineates a more robust, objective and complete framework proposed by the authors. The proposed methodology is based on 3D ultrasound which considers all the information contained on the carotids and plaque structures without depending on a subjective selection of a particular cut or cross-sectional image for diagnosis. To accomplish this task, a semi-automatic 3D reconstruction (de- speckling) of the plaque is performed and features related to echogenicity are computed. Moreover, a procedure for investigating different textural contents and patterns is here investigated by extracting features from estimated speckle fields.

The proposed technique provides diagnostic views not usually accessible via conventional techniques and a new set of features useful for accurately study the plaque morphology. To complement this clinical information, it is suggested that the presence of vulnerable regions inside the plaques, together with their extension and location, may play an important role in the early assessment of stroke risk. These vulnerable regions are defined by low echogenicity (intensity) and high heterogeneity, quantified by reference values mentioned in the literature (El-Barghouty et al, 1996; Elatrozy et al, 1998; Sztajzel et al, 2005).

For this purpose, a 3D Graph-Cuts based algorithm is described to segment/label the most vulnerable (unstable) regions inside the carotid plaque. The algorithm allows to binary segment the data by minimizing an energy function that uses spatial correlation among neighboring pixels/voxels to remove small misclassified regions. The algorithm, highly efficient from a computational point of view, is able to find the global minimizer of this huge combinatorial optimization problem.

The work described by the authors presents an improvement of the 2D based classical approach in the analysis of carotid plaques. This 3D based methodology, together with new risk indicators and a local labelling of unstable foci within the plaque volume is more robust and reduces the subjectivity and operator dependency of the classical method.

Experimental results presented in Section IV show that this labelling procedure is less noisier and favours clustering, being more meaningful from a clinical point of view than the one obtained with simple thresholding.

Section V concludes the chapter and points some potentially meaningful research lines in the context of characterization of carotid plaques.

2. Fundamentals of Plaque Echo-Morphology

Besides the classical indicators such as the degree of *stenosis*, plaque morphology has been pointed as a relevant indicator for assessing the risk of plaque rupture.
2D *B-mode* (Brightness mode) ultrasound is a widely used imaging tool to assess the degree of *stenosis* as well as the plaque morphology. The two most important parameters to characterize the plaque morphology are its echogenicity and texture (Elatrozy et al, 1998).
The echogenicity is evaluated from the image intensities. A region is called *hypoechogenic* if it appears dark in the image and *hyperechogenic* if it appears bright. Concerning the texture it is usually classified as homogeneous or heterogeneous.
Several studies were pursued to statistically characterize the morphology of carotid plaques in 2D ultrasound (Sztajzel et al, 2005). The *Gray-Scale Median* (GSM) is one of the most important indicators considered on plaque diagnosis and is generally used to classify plaques as *hypoechogenic* (GSM < 32) or *hyperechogenic* (GSM > 32) (Pedro et al, 2002). The total percentage of *hypoechogenic* pixels (P40), defined as the percentage of pixels with gray levels below 40, is also an important measure for the characterization of plaque echogenicity. In fact, multiple regression analysis (Sztajzel et al, 2005) has revealed that the GSM and the P40 are the most significant variables related with the presence of symptoms. Additionally, an activity index aiming at quantifying the risk of stroke (Pedro et al, 2002) has been proposed. This overall index merges several indicators, such as plaque overall texture, degree of *stenosis*, global *echogenicity* and location of *hypoechogenic* sites across the plaque.
A vulnerable plaque is associated with thinning of the fibrous cap and infiltration of inflammatory cells that lead to plaque rupture. Studies which established a correlation between quantitative analysis based on ultrasound *B-mode* images and histology (Baroncini et al, 2006) have suggested that *hypoechogenic* regions have more lipid and hemorrhage, indicating inflammatory activity and therefore instability. Conversely, *hyperechogenic* regions are associated with the presence of stable components. Therefore, the location and extension of vulnerable regions throughout the carotid plaque could be a sensitive and relevant marker of stroke risk. Analysis of global information about plaque morphology, despite its unquestionable usefulness, may not be accurate enough in many cases, namely, when plaques are *heterogeneous* or present significant *hypoechogenic* regions. An overall measure of *echogenicity* or texture is incomplete and does not reveal unstable *foci* inside the plaque which may be threatening.

3. Methods

Three-dimensional reconstruction of tissues and organs are generally performed by use of two rendering techniques: surface and volume rendering. Surface rendering (Meairs & Hennerici, 1999; Schminke et al, 2000) can be used to extract the bifurcation walls of carotid arteries and quantify the degree of *stenosis* together with plaque volume. A volume rendering approach (Bullitt & Aylward, 2002) is often used to reconstruct a particular *volume of interest* (VOI) from which echo-morphological or textural analysis can be further performed. Here, a combination of the two approaches is used. First, volume renderings are obtained from sets of 2D ultrasound images of the carotid artery. The regions corresponding to atherosclerotic plaques are visually detected in several cross-sections taken from the obtained volume and then segmented using semi-automatic segmentation methods, e.g. active contours (Xu & Prince, 1998) guided by experienced physicians. Surface rendering

employs the interpolation of these contours providing a three-dimensional representation of the plaque.
The data used in this study consists of 3 sets of n = 100 nearly parallel cross-sections of the carotid artery, from 3 distinct patients, acquired near the bifurcation where plaque formation is more frequent.
The reconstruction algorithm proposed by the authors in (Seabra et al, 2009) is used. The acquisition protocol proposed in the referred publication does not require any additional equipment but the conventional 2D ultrasound scanner available in most medical facilities.
Ultrasound images are corrupted by *speckle*, which makes data visualization and interpretation often a challenging task. The method used here to compute the 3D morphology of the *region of interest* (ROI) containing the carotid plaque is composed of two main steps: i) de-noising and ii) reconstruction. In the first step the ultrasound images, assumed parallel and evenly spaced, are de- speckled with the method described in (Seabra & Sanches, 2008). In the second step, these de- speckled images are interpolated to estimate the 3D ROI containing the plaque.

3.1 De-speckling

The de-noising algorithm, designed in a Bayesian framework, takes into account the RF signal compression performed by the ultrasound equipment which changes the distribution of the raw data provided by the ultrasound probe known to be, under particular conditions, *Rayleigh* distributed (Eltoft, 2006; Michailovich & Tannenbaum, 2006).
Here, the following *Log-compression* model is considered

$$z_{i,j} = a\log(y_{i,j} + 1) + b\,, \tag{1}$$

where a and b are the parameters to be estimated and account for the contrast and brightness adjustments, respectively, performed by the physician during the medical exam. In (1) $z_{i,j}$ are the noiseless pixel intensities of the compressed image and $y_{i,j}$ are the pixels of the corresponding de-compressed image that are assumed to be *Rayleigh* distributed. The parameters (a, b), estimated directly from the observed noisy images as described in (Seabra & Sanches, 2008), are used to invert the transformation (1) and obtain an estimation of the envelope of the RF data (*Rayleigh*),

$$y_{i,j} = e^{\frac{z_{i,j}-b}{a}}\,. \tag{2}$$

In Fig. 1 it is illustrated the effect of the parameter x on the shape of the *Rayleigh* probability density function (PDF). It is also shown how the *Log-Compression Law*, which is an approximation of the ultrasound processing block, changes the *Rayleigh* distribution of the envelope of the RF data.
The de- speckling algorithm, designed in a Bayesian framework with the *Maximum a Posteriori* (MAP) criterion may be formulated as folllows

$$\hat{X} = \arg\min_{X} E(X, Y)\,, \tag{3}$$

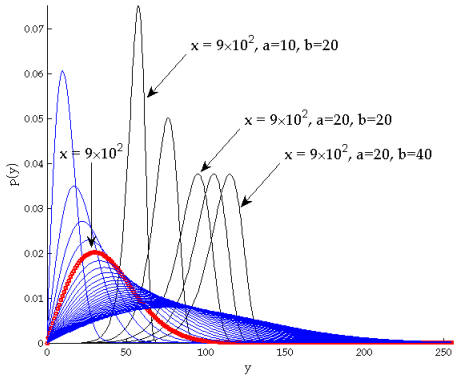


Fig. 1. *Rayleigh* probability density functions (PDF's) for parameters x from 10^2 to 6×10^3 in steps of 10^2, generated according to (6). PDF's using $x = 9\times10^2$ and different values of a and b illustrate the significant change occurring on the original *Rayleigh* PDF.

where $X = \{x_{i,j}\}$ is the denoised image and $Y = \{y_{i,j}\}$ is the estimated envelope of the RF signal obtained after inverting (1). The energy function is

$$E(X,Y) = E_Y(X,Y) + E_X(X), \tag{4}$$

where $E_Y(X,Y)$, called *data fidelity term*, pushes the solution toward the data Y and $E_X(X)$, called *prior term*, regularizes the solution. By assuming statistical independence of the observations the data *fidelity* term is given by

$$E_Y(X,Y) = \log \prod_{i,j} p(y_{i,j|} \mid x_{i,j|}), \tag{5}$$

where

$$p(y_{i,j} \mid x_{i,j}) = (y_{i,j}/x_{i,j}) e^{-y_{i,j}/(2x_{i,j})} \tag{6}$$

is the Rayleigh distribution with parameter $x_{i,j}$. Additionally, the *prior term* is given by

$$E_X(X) = -\log p(X), \tag{7}$$

By assuming that X is a *Markov Random Field* (MRF),

$$p(X) = \frac{1}{\Omega} e^{-\alpha \sum_{i,j} V_{i,j}}, \tag{8}$$

is a Gibbs distribution where Ω is the partition function, α is the prior parameter and $V_{i,j}$ are the potential functions associated with the *cliques* of the MRF, X (Geman & Geman, 1984). Therefore, the *prior term* $E_X(X) = \alpha \sum_{i,j} V_{i,j} + C$ is the so-called Gibbs energy that is here proportional to the *Total Variation* (TV) of X (Vogel & Oman, 1998) where

$$V_{i,j} = \sqrt{(x_{i,j} - x_{i-1,j})^2 + (x_{i,j} - x_{i,j-1})^2} \tag{9}$$

are the magnitudes of the discrete gradient of X at the locations (i, j). These functions are edge preserving because, in relative terms, they penalize more small differences between neighbouring pixels due to noise than large differences due to anatomical transitions.
Details on the optimization of (3) are given in (Seabra et al, 2008; Seabra et al, 2009) and an example to illustrate de application of the algorithm is displayed in Fig. 2.
In this step the speckle field is also estimated from the estimated envelope of the RF and de-speckled images. The speckle noise corrupting the ultrasound images is multiplicative in the sense that its variance depends on the underlying signal $x_{i,j}$. The image formation model may be formulated as follows

$$y_{i,j} = \eta_{i,j} \sqrt{x_{i,j}} \ , \tag{10}$$

where $x_{i,j}$ is the de-speckled pixel intensity at location (i, j) and $y_{i,j}$ and $\eta_{i,j}$ the corresponding pixel intensities of the RF image and the speckle field, respectively.
In this model, the speckle field $\eta = \{\eta_{i,j}\}$ is independent of the signal as occurs in a common *Additive White Gaussian Noise* (AWGN) model where the noisy pixels, $Y = X + \eta$, are the sum of two independent terms, X and η.
In the case of *multiplicative* noise the operation is not additive but multiplicative as shown in (10). The distribution of η is

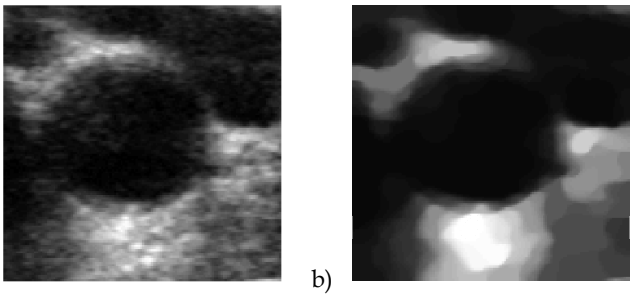
$$p(\eta) = \left| \frac{dy}{d\eta} \right| p(y) = \eta e^{-\frac{\eta^2}{2}}, \qquad \eta \geq 0 \, , \tag{11}$$

which is a unit parameter *Rayleigh* distribution independent of x. The computation of the speckle field, η, is performed from the estimated/observed RF noisy image, Y, and from the de-speckled field, X,

$$\eta_{ij} = \frac{y_{ij}}{\sqrt{x_{ij}}} \, . \tag{12}$$

Given this, the authors argue that the de- speckling procedure may have a two-fold purpose: first, it can provide clearer ultrasound images for better visualization of the structures in the image and and second it allows to extract a speckle field for further texture

analysis. This speckle field can be used to differentiate between tissues having different densities and patterns.



a) b)

Fig. 2. Original ultrasound (a) and de-noised (b) images.

3.2 Reconstruction

The 3D reconstruction of the ROI containing the plaque aims at estimating a 3D field, $F = \{f_p\}$ from the de-speckled images computed in the previous step, $X^t = \{x^t_{i,j}\}$, where t denotes the t^{th} image in the sequence of evenly spaced parallel ultrasound cross sections and $p = (i, j, k)$ represents a node index in the 3D matrix F. The node locations of F, f_p, and the locations of the pixels of X^t, $x^t_{i,j}$ do not necessarily match and therefore an interpolation process is required. Furthermore, some nodes may not be observed which poses a problem of missing data. The following approach is proposed.

Let us consider each voxel as the cubic region, centred at the location of the node f_p, μ_p with dimensions $(\Delta_1, \Delta_2, \Delta_3)$ and the locations of the denoised pixels $x^t_{i,j}$, $\tau^t_{i,j}$. Let $\Sigma_p = \{x^t_{r,l} : \tau^t_{r,l} \in S_{\Delta1,\Delta2,\Delta3}(\mu_p)\}$ be the set of all pixels inside the neighbourhood (voxel) $S_{\Delta1,\Delta2,\Delta3}(\mu_p)$ of the node f_p displayed in Fig. 3.

In a first step a 3D volume, $Z = \{z_p\}$ with the same dimensions of X, is computed where each element $z_p = \langle \Sigma_p \rangle$ is the weighted mean of the set Σ_p where the weights are the normalized distances of the pixel locations to the centre of the corresponding voxel, $z_p = (\sum_r \|\tau_r - \mu_p\| x_r) / (\sum_r \|\tau_r - \mu_p\|)$ where τ_r denotes the location of the r^{th} pixel within the set Σ_p. Therefore, each element of Z contains the average intensity of the pixels within the voxel. However, some elements of Z may be undefined when there are no observations (pixels) inside the corresponding voxel. In this case an interpolation is needed. This missing data problem may be solved by minimizing the following energy function

$$E(F, Z) = \sum_p \left[n_p (f_p - z_p)^2 + \alpha g_p^2 \right], \tag{13}$$

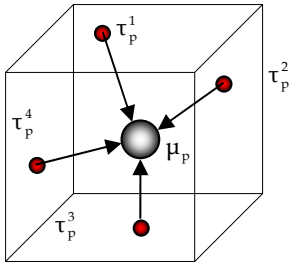


Fig. 3. Voxel representation associated with the node f_p (grey), located at μ_p, and several image pixels, at locations τ_p^r in its neighbourhood (red).

where $g_p = \sqrt{(f_{i,j,k} - f_{i-1,j,k})^2 + (f_{i,j,k} - f_{i,j-1,k})^2 + (f_{i,j,k} - f_{i,j,k-1})^2}$ is the gradient magnitude of F at the voxel p and n_p is the number of observations (pixels) associated with the voxel p. The minimization of (13) may be iteratively solved by optimizing with respect to one unknown at a time, which leads to the following recursion

$$f_p^t = \beta(n_p) z_p + (1 - \beta(n_p)) \bar{f}_p^{t-1}, \tag{14}$$

where $()^t$ denotes the iteration t, $\bar{f}_p^{t-1}$ denotes the mean value of the neighbours of f_p computed in the iteration $t-1$, n_p is the number of image pixels inside the voxel and $\beta(n_p) = \frac{n_p}{n_p + \alpha N_v}$ is a regularization parameter function, displayed in Fig. 4, depending on n_p, on the number of neighbours of f_p, $N_v = 6$, and on the tuning parameter α.

The equation (14) reveals the underlying interpolation mechanism performed during the minimization of the energy function (13). Each new estimate of F, F^t, depends on the previous one, F^{t-1}, and on the field of the mean pixel intensities, Z, computed in the previous step. As large is the number of pixels associated with a given voxel, n_p, as close to one is the parameter β, which means, in the limit, $f_p^t \approx z_p$. Conversely, a small number of observed pixels leads to small values of β, which means that the new estimates of f_p^t result from a combination of z_p and the neighbours, $\bar{f}_p^{t-1}$ computed in the previous iteration. In the limit, when no observations are available, $\beta = 0$, and the new estimate is $f_p^t = \bar{f}_p^{t-1}$, that is, it is equal to the mean intensity of its neighbours. This method adopts the following strategy: when a large number of observed pixels are available for a given voxel its value is mainly computed from Z and when the number of observations is small or even zero the estimates of f_p is obtained mainly from the neighbourhood.

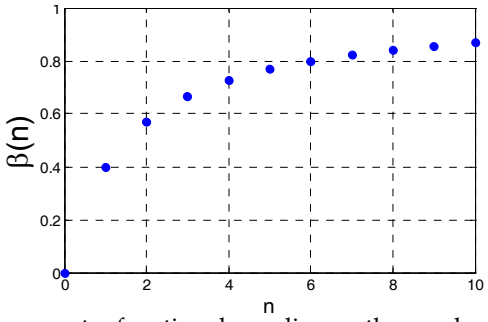


Fig. 4. Regularization parameter function depending on the number of observed pixels.

3.3 Feature Extraction

The volume field F describes the value of the *Rayleigh* parameter across the VOI containing the carotid plaque and it may be used to compute local intensity and textural indicators that characterize the components and tissues of the plaque.

Traditionally, plaque characterization is based on statistics computed from the observed noisy pixels. Here, instead of computing these indicators from the noisy data, the characterization is based on statistical estimators depending on F containing the *Rayleigh* echogenicity (intensity) parameters throughout the whole plaque.

The echogenicity and texture measures are derived from statistical estimators for the mean $f_\mu(p)$, median (GSM) $f_v(p)$, standard deviation $f_\sigma(p)$ and percentile 40 $f_{P40}(p)$ depending on f_p where $p = (i, j, k)$. These statistics, derived from the *Rayleigh* distribution (Abramowitz et al., 1972), are given by

$$\begin{cases} f_\mu(p) = \sqrt{\dfrac{f_p \pi}{2}} \\ f_v(p) = \sqrt{2\log(2) f_p} \\ f_\sigma(x) = \sqrt{\dfrac{4-\pi}{2}} f_p \\ f_{P40}(x) = 1 - e^{-\frac{40^2}{2f_p}} \end{cases} \tag{15}$$

For each 3D reconstructed volume F, four 3D matrices F_μ, F_v, F_σ and F_{P40} were built containing the values of $f_\mu(p)$, $f_v(p)$, $f_\sigma(p)$ and $f_{P40}(p)$, respectively, at each location p. By averaging the elements of these 3D matrices, global indicators about the plaque can be computed. Thus, the global echogenicity of the plaque is given by the mean of F_μ, $\langle F_\mu \rangle$ and the percentage of pixels with grey-scale lower than 40, which is known to be a relevant feature, is given by $< F_{P40} >$. Additionally, the intensity variability is given by $<F_v>$ and $<F_\sigma>$.

3.4 Local Characterization by use of Graph-Cuts

The global characterization of carotid plaques, described in the previous section by averaging the statistical estimators in (15), despite its unquestionable usefulness may not be

enough for a correct assessment of plaque vulnerability, especially in cases where the plaque is significantly heterogeneous or is plagued by artefacts.

Here a local based characterization approach is proposed. The goal is to use the statistical estimators (15) to assess the risk of plaque rupture on a local basis. The method allows to identify sites within the plaque whose features (hypoechogenicity and heterogeneity) point towards potential foci of vulnerability.

The classification of the plaque at each location $p = (i, j, k)$ can be made by thresholding where the elements of the matrices F_μ, F_v, F_σ and F_{P40} are binarized according to a threshold defined by the physician (El-Barghouty et al, 1996; Pedro et al, 2002). The resulting 3D maps characterize locally the plaque with respect to each indicator.

This thresholding algorithm is simple but does not take into account spatial correlation between neighbouring nodes because the process is performed in a voxel-by-voxel basis. Here, a more sophisticated and accurate method is used where the labelling procedure considers the intensity value of the statistical function at location p and also the values of its neighbouring nodes. The spatial correlation is introduced in order to reduce the misclassification rate by assuming that the plaque is composed of homogeneous regions separated by abrupt transitions. This assumption is acceptable from an anatomical perspective and is usually adopted in de-noising and de-blurring algorithms on medical imaging.

Let f_p be the estimated value of F at the p^{th} node. The labelled maps, L_τ with $\tau = \{\mu, v, \sigma, P_{40}\}$, are defined on a plane-by-plane basis, *i.e.* each plane is labelled independently of the others. The segmentation is binary, which means $L(p) \in \{0,1\}$ where $L(p)$ is the p^{th} node of L.

The labelling procedure of the whole volume is performed in three steps, as depicted in Fig. 5: (i) all stacked planes along the vertical direction are independently labelled, (ii) all stacked planes along the horizontal direction are independently labelled, and (iii) both volumes obtained in the previous steps, L^v and L^h, are fused by making $L = L^v \otimes L^h$ where $\otimes$ denotes the element wise Boolean product.

The labelling process of each plane is performed by solving the following optimization problem

$$L_\tau = \arg\min_L E(F_\tau, L), \tag{16}$$

with the energy function being given by

$$E_\tau(F_\tau, L) = \sum_p (\lambda_\tau - f_p)(2L(p) - 1) + \theta \sum_p \frac{V(L(p), L(p_v)) + V(L(p), L(p_h))}{\tilde{g}_p}, \tag{17}$$

where λ_τ is the threshold associated with the indicator τ and θ is a parameter to tune the strength of smoothness, $\tilde{g}_p$ is the normalized gradient at location p, and p_v and p_h are the locations of the causal vertical and horizontal neighbours of the p^{th} node. $V(l_1, l_2)$ is a penalization function defined as follows

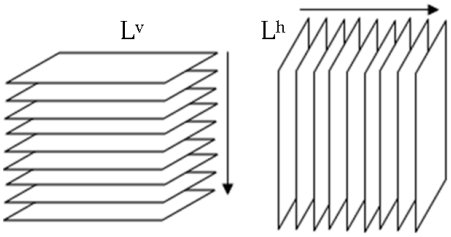


Fig. 5. Labelling procedure performed on a plane by plane basis in two steps, providing L^v and L^h.

$$V(l_1,l_2)=\begin{cases}0 & l_1=l_2\\ 1 & l_1\neq l_2\end{cases}. \tag{18}$$

Similarly to the de-speckling algorithm described in 3.1, the energy function is composed of two terms: the first called *data term* and the second called *regularization term*. The first forces the classification to be $L(p)=1$ when $f_p \geq \lambda_\tau$ because the corresponding term in the energy function (17) is smaller when compared with the one if $L(p)=0$. The opposite occurs if $f_p < \lambda_\tau$. The second term forces the uniformity of the solution because the cost associated with uniform labels is smaller than with nonuniform ones (see (18)). In order to preserve the transitions the terms are divided by the normalized gradient magnitude of f_p, $\tilde{g}_p$ .Therefore, when the gradient magnitude increases the regularization strength is reduced at that particular location.

The minimization of (17) formulated in (16) is a huge optimization task performed in the Ω^{NM} high dimensional space where $\Omega=\{0,1\}$ is the set of labels and N and M are the dimensions of the image.

In (Kolmogorov & Zabih, 2004) it is shown that several energy minimization problems in high dimensional discrete spaces can be efficiently solved by use of *Graph-Cuts* (Boykov, 2001) based algorithms. Opportunely, the energy function given in (17) belongs to this class of problems. The authors have designed a very fast and efficient algorithm to compute the global minimum of the energy function.

4. Experimental Results

In this section results from synthetic and real data are presented. Fig. 6 shows an example of application of the de-speckling /reconstruction procedure, where a log-compressed ultrasound image and its corresponding estimated envelope of the RF signal and de-speckled images are displayed. The log-compressed image was created by use of an echogenicity model of a synthetic plaque (regarded as a matrix of *Rayleigh* parameters) which was corrupted with *Rayleigh* noise and then processed according to the *Log-Compression Law* in (1). Given this result, the de-speckled image is regarded as a "clean" image, which is suitable for visualizing the underlying structures of interest. However, it

may have a second interpretation: in fact, it represents a matrix of localwise estimated Rayleigh parameters.

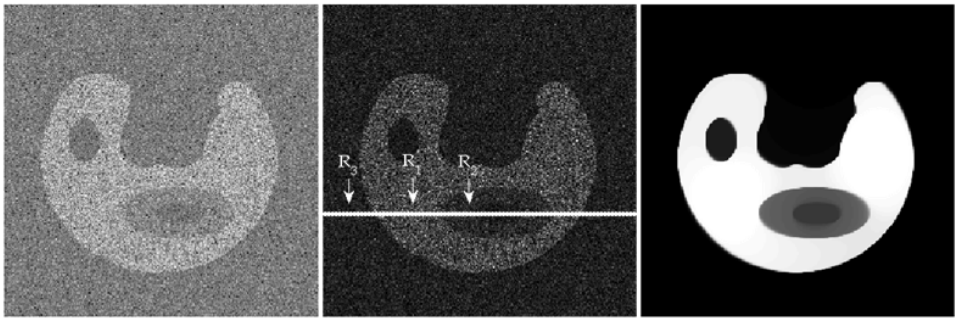


Fig. 6. Reconstruction procedure applied to a synthetic plaque. Log-compressed ultrasound image was used to estimate the RF image (middle) and its de-speckled version (right).

Fig. 7 displays the signal (pixel intensities) along a particular path (see Fig. 6). The signal is overlapped with the real *Rayleigh* parameters f_O which are part of the echogenicity model and the estimated ones, f_p, taken from the de-speckled image. This result clearly shows the ability of the de-speckling algorithm to reduce speckle while preserving the important edges of the image. The PDF's originated from the averaged values of the estimated *Rayleigh* parameters computed within portions R_1, R_2 and R_3 (see Fig. 6) were also computed. The histograms taken from each portion were overlapped with the PDF's to show the efficiency of the de-speckling algorithm on estimating correctly the Rayleigh parameters which model the data. Given this observation, the authors argue that the parameter of the Rayleigh distribution can be used to perform echo-morphological analysis of the tissue because it is a feature which clearly varies from one region to another in the image.

Although texture characterization is beyond the scope of this work, an example to illustrate the potential usefulness of the proposed methodology for texture characterization of the tissues, in particular of the carotid plaque, is here shown.

Fig. 8 shows results of the extraction of texture information from the speckle field of a synthetic image of a plaque presenting distinct pattern regions. Each image quadrant was convolved with a Gaussian mask of different size to simulate distinct smoothness levels. First, a de-speckled image is estimated where clearly the texture pattern was removed and only echogenicity information from the plaque is visible. To retrieve potential significant information encoded in the speckle field a wavelet decomposition of the speckle image is computed to obtain the relative importance of approximation $E_a = E_a/(E_a + E_d)$ and detail $E_d = E_d/(E_a + E_d)$ energies. Fig. 8 (right) depicts the corresponding speckle field together with the computed approximation and detail energies. From the speckle field it is clearly visible that this image is not affected by the information encoded in the echogenicity (intensity) of the synthetic plaque, providing a suitable field for textural analysis. Higher percentages of approximation energies are obtained for the 2nd and 3rd image quadrants, which visibly correspond to smoother regions. On the other hand, detail values are particularly higher for the apparently noisier 4th quadrant.

In the remainder of this section, we present a set of four examples of application of the labelling method based on *Graph-Cuts* to characterize the echo-morphology of carotid

plaques on a localwise basis, where two of them use synthetic data while the remaining two use real ultrasound data.

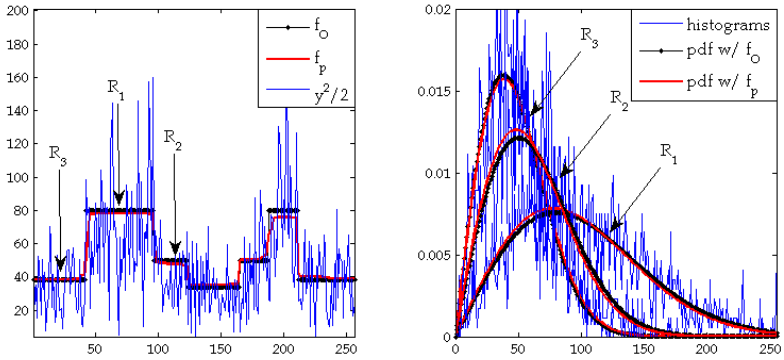


Fig. 7. Representation of a noisy signal (path extracted from Fig. 6) and the corresponding real f_O and estimated rayleigh parameters f_p. Histograms and PDF's originated with averaged values of rayleigh parameters extracted from R_1, R_2 and R_3.

Fig. 8. Extraction of texture information from the speckle field. (Left) Synthetic plaque, presenting regions with distinct texture patterns. (Middle) De-speckled image, where only anatomical information (echogenicity) is visible. (Right) Speckle field, from where approximation and detail energies were computed within each four regions by use of wavelet decomposition.

In the first example depicted in Fig. 9, the previously referred synthetic plaque was used to illustrate the *Graph-cuts* based labelling procedure. Here, the number of classes to detect given as input to the algorithm was *3* and the *data term* associated with the energy function to minimize was simply the image intensities. This simple example provides an illustration of how the *Graph-cuts* labelling algorithm is able to correctly label the different parts of the image.

The performance of the reconstruction and labelling algorithms was also evaluated (Fig. 10a) with a "stack" of synthetically generated echogenicity models of the plaque. This time, the distinction between the different regions in the volume was made harder and almost impossible after corruption with *Rayleigh* noise (Fig. 10b). Not only is the algorithm able to attenuate the speckle, providing a clearer reconstructed volume than the original one

(Fig. 10c), but also allows us to identify the synthetically generated vulnerable hypoechogenic (dark) regions across the plaque. Moreover, the true *Rayleigh* parameters associated with these regions ($f_p = 20$)

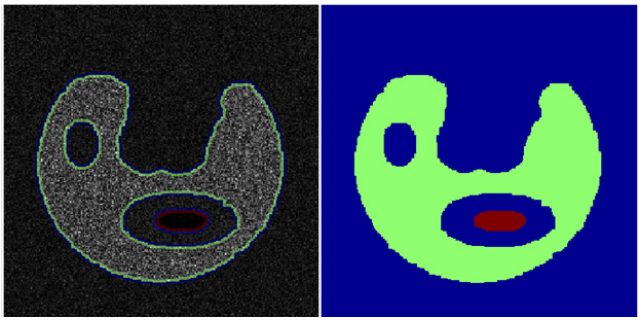


Fig. 9. Illustration of labelling based on *Graph-cuts* using a synthetic plaque corrupted with Rayleigh noise. The different regions are correctly labelled

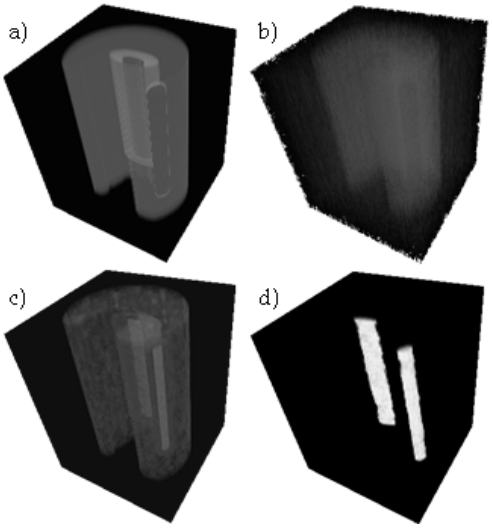


Fig. 10. Reconstruction and labelling using a 3D synthetic carotid plaque (a). Potentially vulnerable regions (two hypoechogenic ($f_p = 20$) and one mid-gray ($f_p = 35$) were created. (b) Carotid plaque, after being corrupted with Rayleigh noise. (c) De-speckled volume of the plaque, and (d) vulnerable sites were correctly labelled.

were correctly recovered by use of the labelling procedure based on *Graph-Cuts* where $f_\lambda = 20$ in (17) (see Fig. 10d).

In the third example, the labelling is performed in two longitudinal ultrasound images showing carotid plaques (see Fig. 11a). The segmented carotid plaques (see Fig. 11b) were characterized according to the labelling methods based on thresholding and *Graph-Cuts* (see Figs. 11c-d). In this example the original noisy images were used and a threshold value of 32 (gray-scale: 0−255) was employed to locally characterize the plaque echogenicity.

It is observed that a characterization based on a simple thresholding of the pixel intensities leads to labeled images with a great amount of noise, which are not realistic from a clinical point of view. In fact, physicians aim at identifying particular regions of vulnerability across

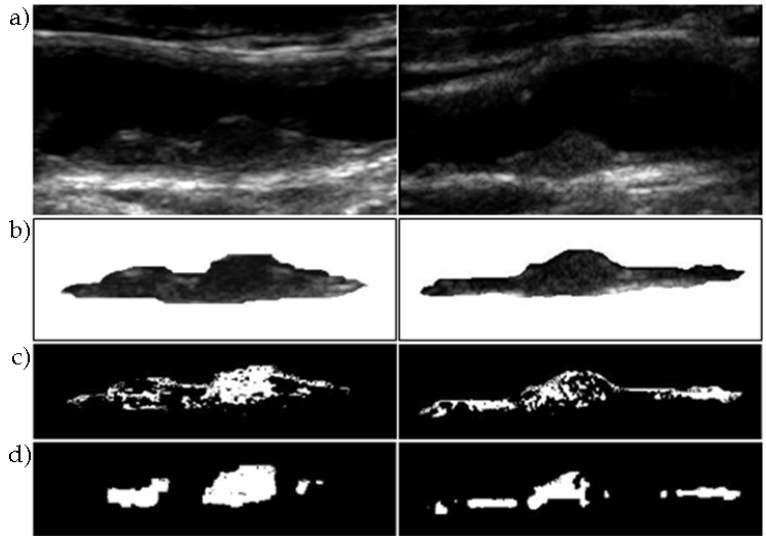


Fig. 11. Original ultrasound images (a) and segmented plaques (b). Labeling of plaques, by use of thresholding (c) and Graph-Cuts based methods (d).

the carotid plaques, where isolated or dispersed pixels seen in Fig. 11c (here termed outliers) are not expected to occur. It is verified (see Fig. 11(d)) that the labelling based on *Graph-Cuts* is less noisier and favors clustering, being more clinical meaningful than the one obtained with simple thresholding.

In the last example, three reconstructed real carotid plaques were locally characterized according to the previously described statistical indicators depending on the estimated parameters f_K of the *Rayleigh* distribution. The statistical estimators considered in this example are f_σ and f_{P40}.

Published studies (Pedro, 2002) in the carotid plaque characterization field suggest that hypoechogenic regions, corresponding to instability *foci*, have GSM < 32 and P40 > 43%.

Fig. 12 displays the labelling of potentially dangerous sites inside the plaque, using the two labelling methods described in this paper (thresholding and *Graph-Cuts*). It is again observed that the labelling using *Graph-Cuts* allows to better discriminate the regions of interest across the carotid plaques.

Volumes labelled with *Graph-Cuts* appear less noisier than when the threshold method is used. This suggests that the use of graph-cuts may improve the characterization of carotid plaques, namely by providing a more appropriate identification and definition of unstable regions across the plaque. As it is pointed out (Pedro, 2002) the degree of extension of these unstable regions as well as their location throughout the plaque should be considered and used as markers of risk of plaque rupture and thus of stroke risk.

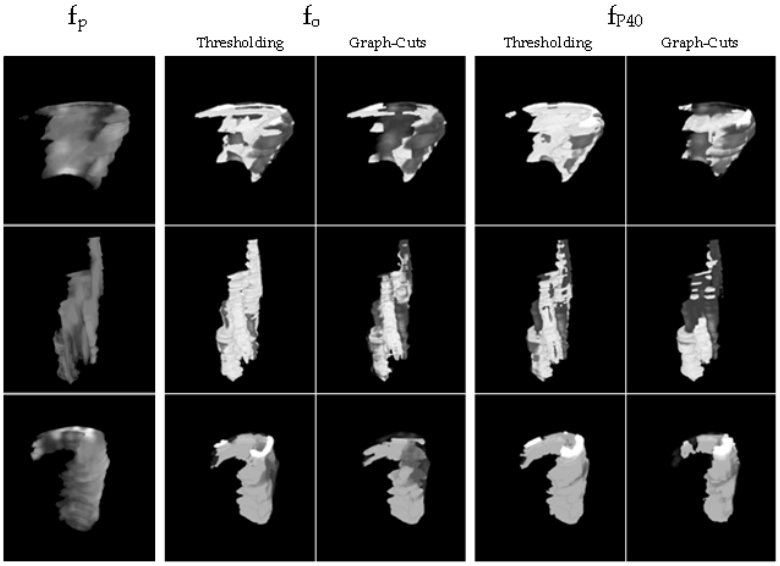


Fig. 12. Echo-morphology of three different plaques encoded in f_p. Comparison of two labeling methods based on - thresholding and *Graph-Cuts* - computed with the local *Rayleigh* estimators of median $f_o(x)$ and $f_{P40}(x)$.

5. Conclusions

Echo-morphology of atherosclerotic plaques as described by its echogenicity and texture assessed through ultrasound is nowadays considered a powerful indicator of stroke risk.
The authors present a robust, objective and complete methodology which consists of de-speckling, 3D reconstruction of the plaque as well as extraction of features related to plaque echogenicity and texture. It is shown that the Bayesian estimation method, employed for reconstruction/ de-noising is able to correctly estimate the parameters of the *Rayleigh* distribution and that these vary significantly from one tissue/ component to another.
An example showing the extraction of texture features from estimated speckle fields suggest its usefulness toward the texture characterization of the plaque tissues/ components.
Moreover, it is proposed a new approach to overcome the major limitations of an averaged characterization of the plaque composition by computing local indicators derived from *Rayleigh* statistics obtained from the noiseless reconstructed images/volumes containing the plaque. This method uses a fast computational labelling algorithm based on *Graph-cuts* to improve the segmentation of potential foci of instability across the plaques. Results show that the labelling method is less noisier and favors clustering, being more clinical meaningful than using simple thresholding.

6. References

Abramowitz, Milton; Stegun, Irene A., eds. (1972). *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, New York: Dover Publications, ISBN 978-0-486-61272-0

Barnett, H.; Meldrum, H. & Eliasziw M. (2002). The appropriate use of carotid endarterectomy. *Canadian Medical Association*, No. 166, 1169–1179

Baroncini, L.; Filho, A.; Junior, L., A. Martins, and S. Ramos. (2006). Ultrasonic tissue characterization of vulnerable carotid plaque: correlation between videodensitometric method and histological examination. *Cardiovascular Ultrasound*, 4-32. Comparative Study

Boykov, Y.; Veksler, O. and Zabih, R. (2001). Fast approximate energy minimization via graph cuts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. Vol. 23, No. 11, 1222-1239

Bullitt, E. & Aylward, S. (2002). Volume rendering of segmented image objects. IEEE Transactions on Medical Imaging, Vol. 21, No. 8, 998-1002

Consensus Group (1995). Consensus statement on the management of patients with asymptomatic atherosclerotic carotid bifurcation lesions: international angiology. *International Angiology*, No. 14, 5–17

Elatrozy, T.; Nicolaides, A.; Tegos, T. & Griffin, M. (1998) The objective characterization of ultrasonic carotid plaque features. *European Journal of Vascular and Endovascular Surgery*, No. 16, 223–230

El-Barghouty, N.; Levine, T.; Ladva, S.; Flanagan, A. & Nicolaides, A. (1996). Histological verification of computerised carotid plaque characterisation. *European Journal of Endovascular Surgery*. Vol. 11, No. 4, 414–416

Eltoft, T. (2006). Modeling the amplitude statistics of ultrasonic images. *IEEE Transactions on Medical Imaging*, Vol. 25, No. 2, 229–240. Comparative Study

Geman, S. & Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 6, No. 6, 721–741

Kolmogorov, V.; and Zabih, R. (2004). What energy functions can be minimized via graph cuts?. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. Vol. 26, No. 2, 147–159

Kyriacou, E.; Pattichis, M.; Pattichis, C.; Mavrommatis, A.; Christodoulou, C.; Kakkos, S. & Nicolaides, A. (2009). Classification of atherosclerotic carotid plaques using morphological analysis on ultrasound images. *Applied Intelligence*, Vol. 30, No. 1, 3-23

Meairs, S. & Hennerici, M. (1999). Four-dimensional ultrasonographic characterization of plaque surface motion in patients with symptomatic and asymptomatic carotid artery stenosis. *Stroke*, Vol. 30, 1807–1813

Michailovich, O. & Tannenbaum, A. (2006). Despeckling of medical ultrasound images. *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*. Vol. 53, No. 1, 64-78

Mougiakakou, S.; Golemati, S.; Gousias, I.; Nicolaides, A. & Nikita, K. (2007). Computer-aided diagnosis of carotid atherosclerosis based on ultrasound image statistics, laws' texture and neural networks. *Ultrasound in Medicine and Biology*, Vol. 33, No. 1, 26-36

Pedro, L.; Fernandes, J.; Pedro, M.; Gonçalves, I. & Dias, N. (2002). Ultrasonographic risk score of carotid plaques. *European Journal of Vascular and Endovascular Surgery*, Vol. 24, 492–498

Saam, T.; Yuan, C.; Chu, B.; Takaya, N.; Underhill, H.; Cai, J.; Tran, N.; Polissar, N.; Neradilek, B.; Jarvik, G.; Isaac, C.; Garden, G.; Maravilla, K.; Hashimoto, B. & Hatsukami, T. (2007) Predictors of carotid atherosclerotic plaque progression as measured by noninvasive magnetic resonance imaging. *Atherosclerosis*, Vol. 194, No. 2, Pages 34-42

Schminke, U.; Motsch, L.; Hilker, L. & Kessler, C. (2000). Three-dimensional ultrasound observation of carotid artery plaque ulceration. *Stroke*, Vol. 31, No. 7, 1651–1655

Seabra, J. & Sanches, J. (2008) Modeling Log-Compressed Ultrasound Images for Radio Frequency Signal Recovery, *Proceedings of the 30th International Conference of the IEEE Engineering in Medicine and Biology Society,* 426-429, Vancouver, Canada, Aug 2008

Seabra, J.; Xavier, J. & Sanches, J. (2008). Convex ultrasound image reconstruction with log-Euclidean priors, *Proceedings of the 30th International Conference of the IEEE Engineering in Medicine and Biology Society,* 435-438, Vancouver, Canada, Aug 2008

Seabra, J.; Pedro, L.; Fernandes, J. & Sanches, J. (2009) A 3D Ultrasound-Based Framework to Characterize the Echo-Morphology of Carotid Plaques. *IEEE Transactions on Biomedical Engineering* Vol. 56, No. 5, 1442-1453

Sztajzel, R.; Momjian, S.; Momjian-Mayor, I.; Murith, N. & Djebaili, K. (2005). Stratified gray-scale median analysis and color mapping of the carotid plaque: correlation with endarterectomy specimen histology of 28 patients. *Stroke*, Vol. 36, No. 4, 741–745

Vogel, C. and & Oman, M. (1998). Fast, robust total variation-based reconstruction of noisy, blurred images. *IEEE Transactions on Image Processing,* Vol.7, No. 6, 813-824

Wintermark, M.; Jawadi, S.; Rapp, J.; Tihan, T.; Tong, E.; Glidden, D.; Abedin, S.; Schaeffer, S.; Acevedo-Bolton, G.; Boudignon, B.; Orwoll, B.; Pan, X. & Saloner, D. (2008) High-Resolution CT Imaging of Carotid Artery Atherosclerotic Plaques. *American Journal of Neuroradiology*, Vol. 29, 875-882

Xu, C. & Prince, J. (1998). Snakes, shapes, and gradient vector flow. *IEEE Transactions on Image Processing*, Vol. 7, No. 3, 359-369

Specular surface reconstruction method for multi-camera corneal topographer arrangements

A. Soumelidis, Z. Fazekas, A. Bódis-Szomorú, F. Schipp,
B. Csákány and J. Németh
Computer and Automation Research Institute,
Budapest University of Technology and Economics,
Eötvös Loránd University &
Semmelweis University, Hungary

1. Introduction

In a recent research and development project, a mathematically profound and technologically viable approach was developed to ensure a precise and reliable measurement and geometrical reconstruction of the human corneal surface. The intention was to get rid of the shortcomings of known corneal topographers and to produce reliable measurement results even for irregular corneas. With an eye on these expectations, a multi-view – i.e., a multi-camera – measurement and reconstruction approach was proposed and used in the experiments. This approach and the lessons learnt from the implementation and testing of the topographer arrangement are presented in this chapter.

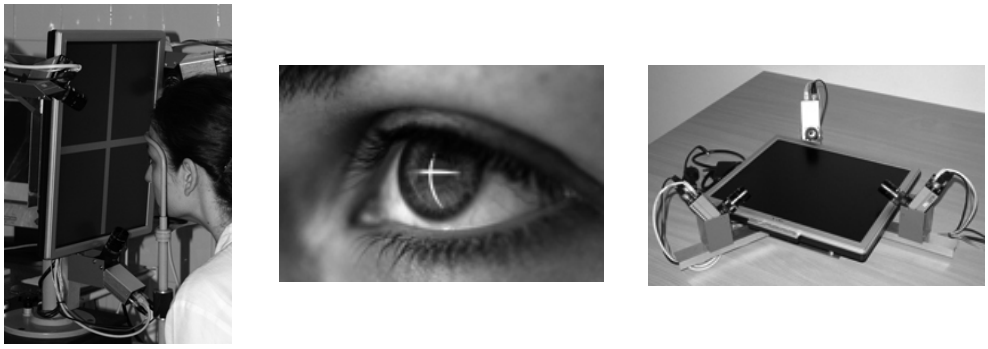


Fig. 1. The three-camera corneal topographer arrangement used in the measurements described in this paper (left). Virtual image of a bright cross-shaped test pattern (middle). Aside from the test pattern, reflections of the patient's eye-lashes and nose are also discernible in the image. The optical subsystem of the topographer arrangement (right).

Firstly, however, some background material on the measured organ (the human eye), and particularly, on its frontal section (the cornea), on its measurement methodology (corneal

topography) and on the principle of the corneal measurement (specular surface reconstruction) are given. It is followed by a detailed presentation of the proposed measurement method. Finally, conclusions are drawn and future work is outlined.

2. The human cornea and its geometrical measurement

In technical terms, the **human eye** can be considered an imaging sensor with its frontal section responsible for focusing the incoming light rays and its rear section responsible for converting the image – formed on its internal surface – into electrical signals available for further processing. The **cornea** – located in its frontal section – is the primary optical structure of the human eye. The corneal tissue is transparent. The human cornea is an optical structure, which generates about the 70% of the total refractive power of the eye, while other structures in the light-path – including also the crystalline lens – contribute significantly less the total refractive power.

The **outer corneal surface** itself is not an optical surface. It is the coating of the corneal surface, i.e., the pre-corneal tear film that provides a smooth optical surface. The tear film is replenished with every blink and after its build-up; it provides a smooth optical surface over the microscopically irregular corneal surface (Németh et al., 2002). The frontal and the inner surface of the human cornea together contribute about 43 dioptres to the eye's total refractive power, which is about 65 dioptres. Physiologically, the shape of the frontal corneal surface is close to spherical.

Primarily, it is because of the cornea's mentioned high share in the total refractive power that the detailed description of the corneal surface and of its refractive power map is of considerable value ophthalmic diagnosis. For this reason, it is not surprising that the **corneal measurement devices**, such as keratometers, corneal topographers, and methods – aiming at measuring and determining the corneal shape and using it for further optical calculations – have a relatively long history; bibliographical details of some milestone papers written by Helmholtz, Purkinje, Placido and Gullstrand are given in (Jongsma et al., 1999).

Purkinje images are virtual images reflected from the consecutive surfaces of the light-path, namely from those of the cornea and crystalline lens. These reflections are known since the first half of the nineteenth century. There are four Purkinje images corresponding to the anterior (1st Purkinje image) and posterior (2nd) surfaces of the cornea, the anterior (3rd) and posterior (4th) surfaces of the crystalline lens. The first Purkinje image is the brightest. As the size and shape of this image depends on the geometry of the frontal corneal surface, it can be used for measurement and surface reconstruction purposes. Most of the keratometers and corneal topographers use this principle. Another important application field of the Purkinje images are the eye trackers.

Presently, **corneal topographers** are used in a range of examinations in ophthalmic diagnosis, see e.g., (Corbett et al., 1999) for a good overview of the topic. A selection of corneal topographers is reviewed in (Jongsma, 1999). The majority of these measurement devices and methods rely on the specularity of the cornea, or more precisely, that of the pre-corneal tear-film and use the 1st Purkinje images. These methods are referred as reflection-based methods, and the topographers using these methods as **reflection-based topographers**. In case of such topographers, a bright measurement-pattern of known and well-defined geometry – e.g., concentric rings called in this context Placido-disk, see Figure 5 – is generated and displayed in front of the patient's cornea. The virtual image of this pattern is photo-

graphed by one or more camera. The distorted virtual image, or images are then analyzed, and the corneal surface is mathematically reconstructed. Based on this reconstruction, height- and refractive power-maps are produced and displayed for inspection by the ophthalmologist.

Cornea topography is an important field of ophthalmologic diagnosis. The purpose of a cornea topographic examination is to determine and display the shape and the refractive power map – that is, the (local) measure of the focusing effect caused by an optical structure (e.g., a lens, or a refractive surface) – of the living human cornea. This contributes to the total refractive power of the eye that determines whether the light-rays entering the eye are properly focused on the fovea, or not. The refractive power of the cornea varies slightly over the corneal surface. The cornea topographers usually display these location-dependent refractive power values – i.e., the **corneal refractive power map** – as pseudo-coloured maps, see Figure 3. From this map, optical aberration features are calculated for the measured cornea by the topographers, or by some standalone programs. After a corneal measurement, the ophthalmologist visually inspects the corneal map, considers the optical aberrations mentioned above, other available relevant measurement data (e.g., the optical aberrations of the whole eye computed from measurement data produced by a Shack-Hartmann wavefront-sensor) and circumstances. Then she suggests an appropriate treatment – if necessary – for the patient's eye. Corneal topography plays a significant role in the diagnosis of corneal diseases, in contact lens selection and fitting, in planning sight-correcting refractive surgical operations, and in their post-operative check-ups.

In case of healthy and regular corneal surfaces, the presently available corneal topographers generally produce good quality corneal snapshots, and based on these, precise and reliable maps are generated. However, even for healthy regular surfaces, small impurities and tiny discontinuities in the pre-corneal tear-film may produce extensive measurement errors. This is because the **simplistic measurement patterns**, such as the aforementioned Placido-disk that are still widely used in corneal topographers, do not provide the necessary information to correct such seemingly local problems. The formation of such a measurement anomaly can be followed in Figure 6. The resulting refractive power map exhibits a serious artifact. It arises as a result of the poorly tracked Placido-rings that give rise to some highly curved arcs, and these arcs in turn result in significant and extensive – i.e., non-local – errors in the generated optical power map.

The **measurement anomalies** could be even greater, if callused irregular corneal surfaces are examined using a Placido-based topographer. In such cases, many of the reflected ring-points may disappear and it may be difficult, if at all possible, to determine the exact radial order of the ring-arcs. Consequently, only a partial, rough and unsatisfactory surface reconstruction can be achieved that is of no use for any diagnostic purposes. To this end several more informative measurement patterns were suggested, e.g., (Vos et al., 1997). Some of the proposed patterns facilitate the identification of point-correspondences by using position-coding colour patterns, e.g., (Griffin et al., 1992; Soumelidis et al., 2008).

Another source of measurement anomalies is the dynamic behaviour of the pre-corneal tear film. It should be taken into consideration even when carrying out a single (i.e., non-dynamic) corneal topographic measurement. The pre-corneal tear film coating the corneal surface keeps changing between blinks. In the first few seconds after a blink, the tear film spreads over and builds up on the corneal surface. Then, it begins to evaporate and after a while it gradually breaks up. The optical smoothness of the corneal surface is guaranteed

only during the evaporation phase of the pre-corneal tear film, not during its break up. Snapshots taken at other times exhibit various degrees of measurement anomaly. Furthermore, weird and incorrect optical power maps are produced, if the measurement pattern has not been properly centred onto the cornea by the ophthalmologist, or if the image taken was blurred because of the improper camera-eye distance setting used.

There are some **alternative image-based examination methods of the cornea**. For example, the slit-scanning method is based on the physiologic Tyndall phenomenon of the corneal stroma. Using a narrow slit-like light beam, images can be obtained from a line of the surface. This method may be combined with the reflective one. Another example is the Scheimpflug imaging, which is a photographical technique. Both the front and rear surfaces of the cornea can be examined with this method. Optical coherence tomography examination of the anterior segment is based on low coherence interferometry (Swartz et al., 2007). All the mentioned alternative imaging methods are line-based, i.e., they process one scanline in a moment. This way of operation, however, constitutes a major drawback when compared with the reflective imaging method.

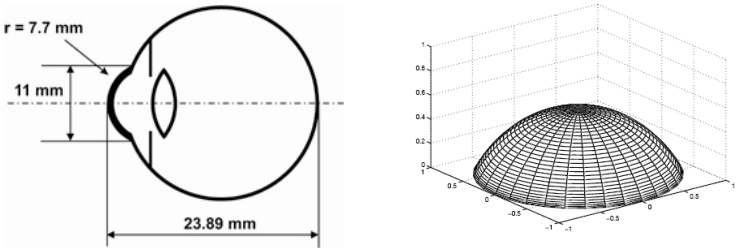


Fig. 2. In Gullstrand's eye model (left), the corneal surface is represented by a spherical calotte surface (right).

3. Eye models, cornea models

The **Gullstrand eye model** – since its creation in 1911 – had been the basis of optical modelling and calculations concerning the human eye for a long period of time. In this early eye model, like in many others proposed later, the refractive power of the cornea is represented solely by its anterior surface. This approximation describes the optical behaviour of living corneas – except for those having certain special ophthalmic conditions – reasonably accurately. This is because, in case of the typical corneas, one can estimate the whole corneal refractive power fairly precisely from the radii of curvature of the anterior surface.

Recent eye models – such as those proposed by Schwiegerling in 1995 and by Liou and Brennan in 1997 – are more realistic and more precise models with more tuneable parameters involved. One of the reasons for using more realistic eye models is to estimate the retinal image quality achieved. The latter model, for example, incorporates also the continuous change of the refractive index within the crystalline lens. The inclusion of such fine details into the model, however, does not prejudice the optical importance of the anterior surface (de Almeida & Carvalho , 2007).

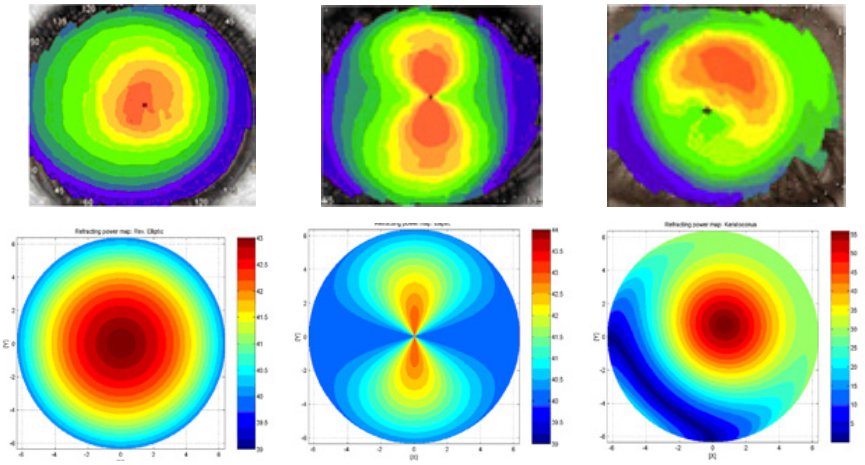


Fig. 3. Refractive power maps of healthy and abnormal human corneas (top row) and their corresponding test models (bottom row). Regions marked with red have the highest dioptre values in the above maps, while the dark blue regions have the lowest.

A **test cornea** is an optical device that serves as a physical model – a proxy – of the (human) cornea. It is primarily used for testing and calibration purposes in conjunction with corneal topographers and other corneal measurement devices. If used with reflective corneal topographers, the test cornea must have a reflective optically smooth surface. The shape and the size of the test cornea should match those of the "average" human cornea. A test cornea is shown in the left image of Figure 4.

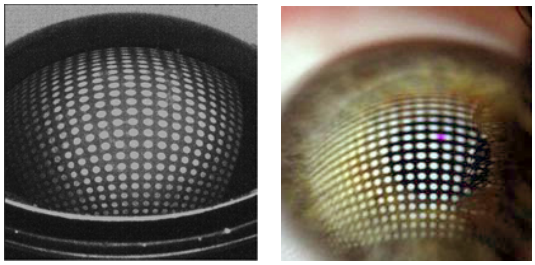


Fig. 4. A virtual image produced by a spherical test cornea (left) and that produced by a living cornea (right).

A test cornea looks like plano-convex lens; except for it is not transparent. Usually, it is fixed onto a planar holder with its planar side. The elaborate part of the test cornea is its convex side; it must be a high precision optical surface. The precision of the planar side is less important, though it still plays a significant role in the positioning of the test cornea. Test corneas are usually made of some opaque machinable material (e.g., dark glass, ceramics, metals, plastic) appropriate for producing high-quality optical surfaces.
In case of glass and ceramics test corneas, these can be manufactured using the so-called single-point diamond turning process. In case of metals (e.g., stainless steel), a customary high-precision lathe can be used for turning. Among the plastic materials used in optics, the

polymethyl-methacrylate (PMMA) is probably the best candidate material for a test cornea. (As in case of the glass used for this purpose, also the PMMA material has to be dark.) There is considerable manufacturing experience with respect to PMMA, as spectacles, and contact lenses are known to be manufactured from this plastic.

Test corneas that are most commonly used in conjunction with corneal topographic measurements have **spherical calotte surface**. Such a test cornea is shown in Figure 4 together with a living cornea photographed from a similar angle for comparison. The radius of such a test cornea is chosen to be equal to the radius of curvature of the typical human cornea. However, different eye models specify and use slightly different "typical" dimensions. It should be noted here that the surface of a living human cornea is not a perfectly spherical surface. This fact is taken into consideration by a number of eye models; as these regard the "typical" corneal surface as ellipsoid surface. Even if the spherical test corneas are considered a simplistic physical model of the cornea, they serve as the most common calibration devices for corneal measurement instruments in respect of general calibration. Clearly, such test corneas are of no use for more complex and more realistic sensitivity analysis purposes, such as sensitivity of the measurement regards to optical aberrations, see e.g., (Wang, 2006); furthermore, they do not facilitate testing of surface reconstruction algorithms.

Three test surfaces were considered essential for testing purposes in the mentioned project. Note that further corneal test surfaces are specified in (ANSI, 2008). Two of these three surfaces are surfaces of revolution; therefore, they can be manufactured by precision turning. However, the third one is a free-form surface which cannot be manufactured by turning.

Surface modelling a healthy cornea is a slightly more realistic model than a spherical calotte model. The surface is a surface of revolution, characterized by two radii of curvature; one for the central region, i.e., near the axis of revolution measured in an axial plane, and one for the periphery, i.e., near the brim of the test surface also measured in an axial plane. Between these two extremes, the local radius of curvature should change in a linear fashion, or – as it would results in a function that cannot be given explicitly – according to some smooth continuous function. In practice, the test surface can be chosen to be the calotte of a prolate spheroid. The refractive power map of such a test surface is shown on the left-hand-side of Figure 3. In this case, the central refractive power was set to 43 dioptres, while the peripheral refractive power was set to 40 dioptres. These are typical values for healthy subjects.

Regular astigmatic corneal surface is a typical degenerative deformation of corneal surface. Plainly speaking, the corneal surface becomes somewhat cylindrical. It is characterised by the direction of the magnitude of the astigmatism. In case of an astigmatic corneal surface, the local curvature and consequently the local refracting power vary along and perpendicular to – and for that matter, skew to – the direction of astigmatism differently. A simple astigmatic surface is the calotte of a prolate spheroid (cut from a prolate spheroid by a plane parallel to the axis of revolution). The refractive power map of such a test surface is shown in the middle image in lower row of the Figure 3. There, the direction of the astigmatism is chosen to be horizontal, furthermore, the horizontal and vertical central refractive powers were set to 42 dioptres and 45 dioptres, respectively; while the peripheral refractive power was set to 39 dioptres, respectively.

Keratoconic surface is a characteristic degenerative deformation of corneal surface. It is characterised by a relatively steep conic upthrust in the corneal surface. For modelling puposes, the base surface can be chosen from the aforementioned test surfaces, while the kera-

toconic upthrust can be modelled as the base surface being multiplied with a unimodal function. This will result in a free-form surface that cannot be manufactured by turning, a but milling. The refractive power map of a keratoconic test surface is shown on the right-hand-side of in lower row of the Figure 3.

4. Specular surface reconstruction approaches

The mathematical **reconstruction of a smooth specular surface** from the distortions it brings about to some known, well-defined measurement pattern, called fiducial, when this measurement pattern is viewed and photographed in the virtual images, is an active research area. The reconstruction can be local or global. In case of **local reconstruction**, sufficient conditions are given for the uniqueness of the reconstruction in (Savarese & Perona, 2001). The formulae to determine the actual surface-patch are given in (Savarese & Perona, 2002a). Multi-view methods for the practically more important **global reconstruction** were published recently. These methods rely on the smoothness of the surface and use several views to make the unique reconstruction possible. For an unknown smooth, convex specular surface – viewed by several cameras – those points are located on – or near to – the specular surface for which the unit normal vectors – calculated from the reflections at these points – are approximately the same for different views. This observation is the basis of the voxel-carving method suggested by (Bonfort & Sturm, 2003). Clearly, this method has an important limitation; it can be used only for those surface portions that reflect the measurement-pattern into more than one camera. An elegant surface reconstruction method was proposed by (Kickingereder & Donner, 2004). There the light-reflection at the surface is described with a total differential equation. The mathematical reconstruction of the surface is achieved by the numerical solution of this total differential equation.

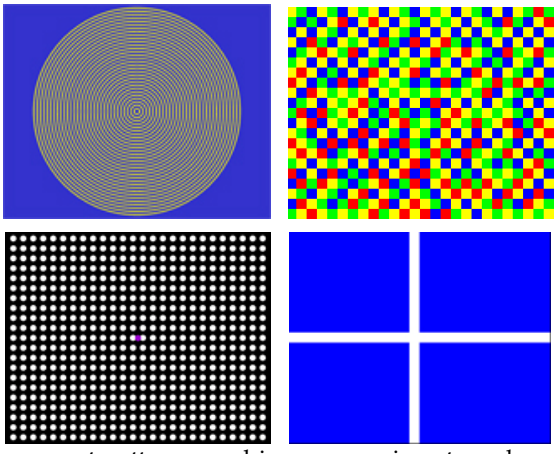


Fig. 5. Various measurement patterns used in our experiments and measurements. Placido disk consisting of concentric rings (top left). A position-coding colour checkerboard (top right). A square grid of circular spots (bottom left) with its centre marked. A cross-shaped pattern (bottom right) used only for basic checks.

The purpose of the **measurement patterns** is to facilitate the identification of the mathematical mappings – produced by the specular corneal surfaces – by providing spatial correspondences between the elements of the measurement pattern (e.g., circles, characteristic points) on one hand and their images on the other. A great variety of measurement patterns is used in the reflection-based corneal topographers. Here, we consider planar patterns only; though, measurement patterns do appear in various topographers on spherical, paraboloid of revolution, and other surfaces.
In Figure 5, a selection of **planar measurement patterns** are shown. These were used either for measurement, or test purposes in the experiments reported herein. In the top left image, Placido's disk is shown. It consists of bright concentric rings. This measurement pattern supports the circle to curve correspondence, but does not support a point-to-point correspondence; and as explained in (Sicam, 2007), for this reason the Placido-based corneal topography gives rise to skew ray ambiguity and as a consequence theoretically inadequate for surface reconstruction. Though, to be fair with the Placido-based topographers, the effect of skew ray ambiguity is insignificant in most of the corneal surfaces.
A position-coding colour checkerboard is shown in the top right image of Figure 5. A somewhat similar colour-coding approach was proposed earlier for another application field in (Griffin, 1992). The generation of the position-coding colour checkerboard shown in the figure is outlined in (Soumelidis et al., 2008). A square grid of circular spots with its centre marked with a coloured spot is shown in the bottom left image of the figure; while a cross-shaped pattern – used only for basic settings and checks – is show in the bottom right image of the same figure.

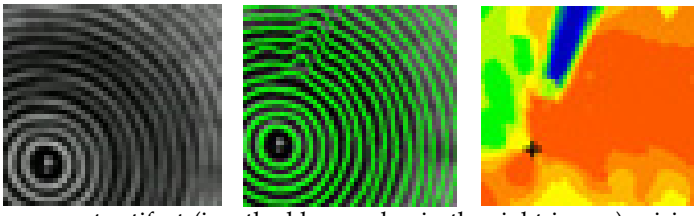


Fig. 6. A measurement artifact (i.e., the blue wedge in the right image) arising after a poor ring-tracking step (middle). The distorted rings of the Placido-disk (left). These rings do not facilitate their local identification.

5. An experimental multi-view corneal topographer arrangement

5.1 Arrangement with up to four cameras

The surface reconstruction method proposed in (Fazekas et al., 2008) – and described herein in more detail – was originally developed for the experimental corneal topographer arrangement described in (Soumelidis et al., 2008). In the following, this arrangement will be referred as the target-arrangement. As it will be apparent later on, the proposed method is not limited to the target-arrangement; however, up to now, it was tested only on that test-bed. The camera-system of the target-arrangement may comprise at most of four cameras due to some internal networking limitation. Presently, however, a **three-camera topographer arrangement** is used for the experiments. The camera-system is mounted rigidly on a high-quality TFT-display which serves as measurement-pattern generator of the topographer arrangement. The target-arrangement is shown in Figure 1, and schematically in

Figure 9. Clearly, this arrangement is very delicate and needs proper calibration for any meaningful measurements. The individual and joint calibration steps of its cameras are outlined below.

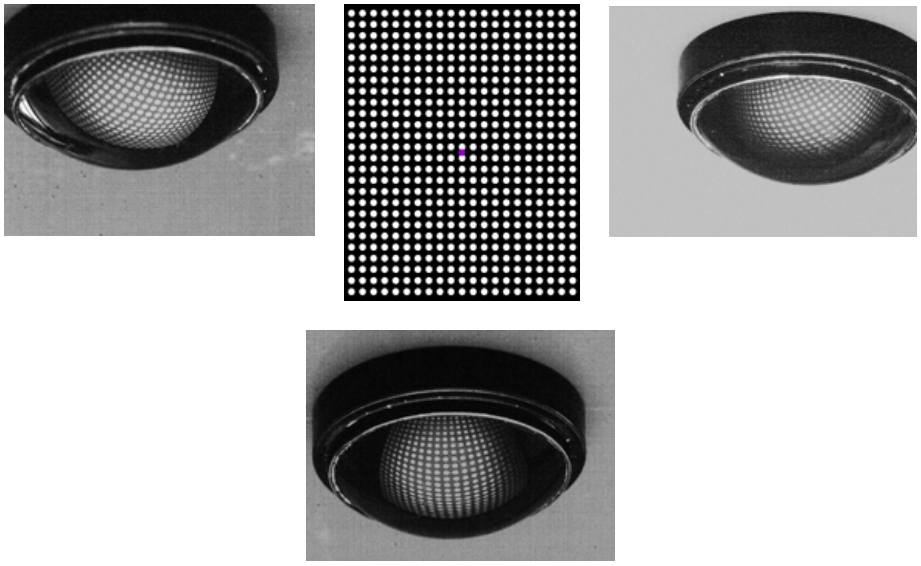


Fig. 7. Distorted virtual images (left, right, bottom) of the measurement pattern (centre) – reflected by the specular spherical surface of a test cornea – as "seen" by the three cameras of the topographer arrangement.

5.2 Individual calibration of the cameras in the arrangement

In the recent years, **camera calibration** and position estimation methods based on planar fiducials have gained popularity in the stereo vision community in respect of spatial reconstruction problems. One of the reasons behind this popularity is the fact that these fiducials are easy to produce; they can be printed on a desktop laser printer in acceptable quality; therefore their production is fairly inexpensive.

A general camera calibration algorithm based on such fiducials is given in (Hartley & Zissermann, 2004), while a concrete **practical implementation** – based on images of a checkerboard placed in different orientations and positions, and photographed with the camera to be calibrated – was proposed in (Zhang, 2000).

Firstly, the **calibration of a single camera** is outlined below. For a more detailed description of this algorithm – used in a different context – see (Bódis-Szomorú et al., 2008); the notation used herein without a detailed explanation is similar to that of appearing in the cited paper. The position of a checkerboard in the 3D space has 6 degree-of-freedom. With m images taken, one has to determine $(9+6m)$ camera parameters in total.

– Localisation of characteristic points in the image (i.e. the localisation of the image points corresponding to the corners of the black and white squares of the checkerboard) using an appropriate corner detector.

– Estimation of the homography $\boldsymbol{H} = [\boldsymbol{h}_1 \quad \boldsymbol{h}_2 \quad \boldsymbol{h}_3]$ – for each checkerboard position photographed – existing between the characteristic points of the planar checkerboard and their corresponding image points.
– Estimation of the intrinsic camera parameters based on the homographies. These parameters can be derived from the condition that image-points of the plane's complex-valued circular points lie on the image of the absolute conic ω. See details in (Hartley & Zisserman, 2004). ω can be estimated with a least-square method from the homographies from conditions $\boldsymbol{h}_1^T \omega \, \boldsymbol{h}_2 = 0$ and $\boldsymbol{h}_1^T \omega \, \boldsymbol{h}_1 = \boldsymbol{h}_2^T \omega \, \boldsymbol{h}_2$. Since $\omega = \boldsymbol{K}^{-T}\boldsymbol{K}^{-1}$, therefore the camera calibration matrix $\boldsymbol{K}$ can be computed via Cholesky-decomposition. See details in (Zhang, 2000).

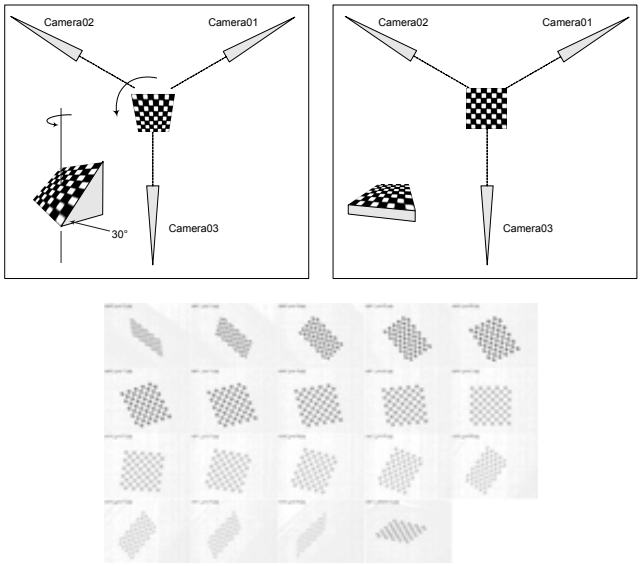


Fig. 8. A 10 mm by 10 mm checkerboard was used for the calibration of the cameras in the topographer arrangement. Firstly, the checkerboard was placed in a horizontal position and photographed with the cameras (top right). Then the checkerboard was fixed to wedge, turned around and photographed in a number of positions with the cameras (top left). The checkerboard images taken with one of the cameras (bottom).

– Computing the extrinsic camera parameters – for each image – based on $\boldsymbol{K}$ and the appropriate $\boldsymbol{H}$ from $\lambda[\boldsymbol{h}_1 \quad \boldsymbol{h}_2 \quad \boldsymbol{h}_3] = \boldsymbol{K}[\boldsymbol{r}_1 \quad \boldsymbol{r}_2 \quad \boldsymbol{t}_W]$ condition.
– At this stage, estimates are available for all the parameters of the following linear camera model $\lambda\widetilde{\boldsymbol{C}} = \boldsymbol{P}\widetilde{\boldsymbol{W}} = \boldsymbol{K}[\boldsymbol{R} \quad -\boldsymbol{R}\boldsymbol{t}]\widetilde{\boldsymbol{W}} = \boldsymbol{K}[\boldsymbol{R} \quad \boldsymbol{t}_W]\widetilde{\boldsymbol{W}}$ for each of the images.
– Parameter optimisation using the Levenberg-Marquardt method with the following cost function to be minimised

$$f(\mathbf{p}_{in}, \mathbf{p}_{ex}^{(1)}, \mathbf{p}_{ex}^{(2)}, \ldots, \mathbf{p}_{ex}^{(m)}) = \sum_{i=1}^{m}\sum_{j=1}^{n}\left\| I_{ij} - \varphi(\mathbf{p}_{in}, \mathbf{p}_{ex}^{i}, \mathbf{W}_{ij}) \right\|_2^2 = \sum_{i=1}^{m}\sum_{j=1}^{n} d^2(I_{ij}, \breve{I}_{ij}) .$$

In order to **calibrate a multi-camera arrangement** based on the camera calibration method outlined above, one needs choose a **reference checkerboard position** and orientation in such a way that the checkerboard placed in that position and orientation can be properly "seen" by the cameras; furthermore, all the images taken by the cameras of the checkerboard in this position and orientation are appropriate for feature extraction. Among the series of checkerboard images shown in Figure 8, the checkerboard position appearing in the right-most sub-image in the bottom row was this common reference position.
If, for instance, the camera parameter vector for *Camera01* – derived via the above calibration process – is in the following form

$$\mathbf{p}_{Camera01}^{all} = (\mathbf{p}_{in}^{T}, \mathbf{p}_{ex}^{(1)T}, \mathbf{p}_{ex}^{(2)T}, \ldots, \mathbf{p}_{ex}^{(m)T})^{T} ,$$

and the *k*-th checkerboard-image is the one that was taken of the checkerboard in its reference position, then the parameter vector characterising the position and orientation of *Camera01* with respect to the reference checkerboard-position is as follows.

$$\mathbf{p}_{Camera01} = (\mathbf{p}_{in}^{T}, \mathbf{p}_{ex}^{(k)T})^{T} .$$

The parameter vector for the other cameras, i.e., for *Camera02* and *Camera03*, can be derived similarly. As shown in Figure 8, a checkerboard – with a size comparable to that of a typical human cornea – was used for the calibration of the cameras of the topographer arrangement. Firstly, this small checkerboard was placed in a horizontal position and photographed with each of the cameras.
Then the **checkerboard was fixed onto wedge** and was turned around and photographed in a number of positions with each of the cameras. Then with the method outlined above the camera parameter vectors were derived. The resulting geometrical positions and orientations of the cameras (of the calibrated three-camera arrangement) are shown in graphic reconstruction of Figure 9. The reconstructed arrangement is qualitatively acceptable.

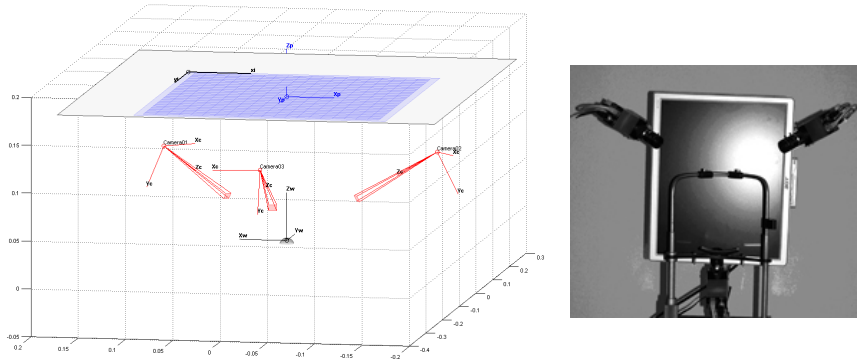


Fig. 9. The graphical representation of the calibration results (left). The pyramids represent the cameras. The three-camera arrangement – with the head-rest – for reference (right).

6. Determining the discrete and continuous distortions

6.1 Segmentation, feature extraction, filtering

In our experiments, a rectangular array of bright circular spots – shown in Figure 7 – was used as measurement-pattern. The **blobs** in the reflected images – corresponding to the circular spots in the rectangular array of the measurement-pattern – were detected at a number of thresholds, and the centre-point and a few shape descriptors were calculated for each blob. The detected blobs were then checked for their areas. Blobs within **blob-hierarchies**, – i.e., blobs that were more or less intactly preserved at a number of thresholds, see examples in Figure 10 top inset – and blobs that greatly overlapped with ones gained at other thresholds were merged with each other; as these were assumed to be the images of the same circular spots. Blobs with much bigger areas than those of the blobs in their neighbourhood were removed from the list of acceptable blobs, as these were considered be segmentation artifacts, rather than reliable segmentation results.

A distorted virtual image – formed by a test cornea – is shown in Figure 10. The process of segmentation, feature extraction, **blob merging and filtering** outlined above can be followed through images and insets of the figure. The area of each detected spot is indicated by that of a coloured disk.

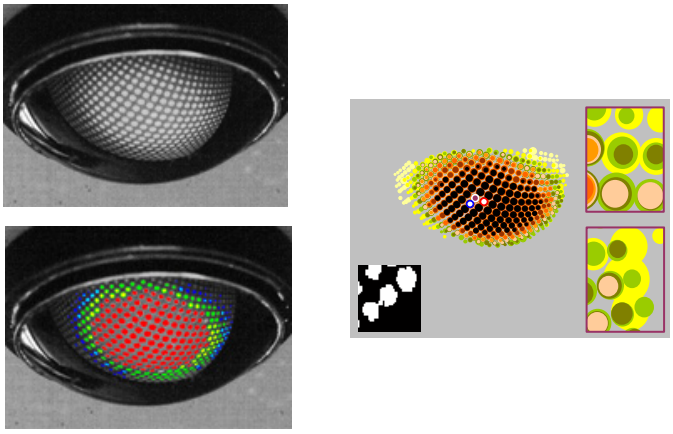


Fig. 10. The virtual image of a square grid of circular spots as being reflected by a test cornea (top). The areas of the spots detected at different intensity thresholds are indicated by that of the coloured disks (right). Blob hierarchies (see main text for details). (top inset). Blobs – shown as yellow disks – split into two separate blobs each – shown as light green disks – at higher thresholds (right inset). A compound white blob – corresponding to two separate circular spots of the measurement pattern – appearing in one of the thresholded images (left inset). Consolidated – i.e., filtered – blobs (left) with colours indicating the highest threshold that produced a shape-wise and size-wise acceptable blob at that location.

6.2 Tracking blobs within curved rows and columns

A coloured circular spot, shown in the central image of Figure 7, was used for marking the **geometric centre of the rectangular array**. Blobs corresponding to these spots in the three reflection images were identified, and their centres were used as starting points for the sub-

sequent blob-identification process. Each of the three 2D-point-sets – formed by the blob-centres of the blobs detected in the reflection images – was submitted to **Delaunnay-triangulation**, see Figure 11. A similar approach appears in (Trichet & Merialdo, 2007) in the context of a different application.
The sides of the resulting triangles were then tracked – starting from the mentioned starting points – in **"row" and "column" directions**. As an example, the images of a living cornea taken by the three-camera topographer arrangement are shown in Figure 16. The results of the tracking process are shown in Figure 17.
For the 2D point-set shown in right-hand-side image of Figure 10, these directions are identified by the white ring to blue ring and white ring to red ring directions, respectively. Based on the tracking results, correspondences between the original circular spots and the detected blobs were established; in other words, a mapping from the original rectangular grid to each of the three point-sets was established.
For the purpose of the tracking, the side-lengths of the Delaunnay-triangles, the directions of sides and the area-ratio of the blobs – connected by the side concerned – were used.

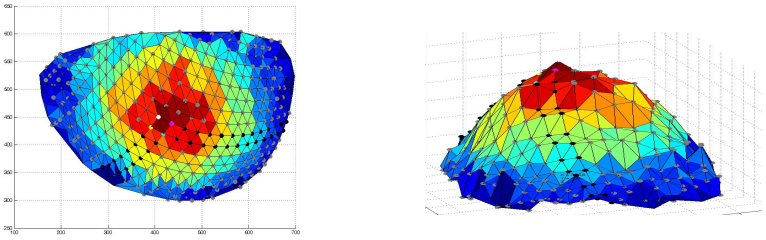


Fig. 11. Delaunnay-triangulation of the detected blobs' centre-points (left). The area of the corresponding blobs indicated in the z-coordinates and by the colour of the triangles.

6.3 Approximation of the apparent distortions using splines

Spline interpolation was used to convert the **three discrete mappings** to continuous mappings, thereby approximating the mappings (optical distortions, in this case) between the points of the measurement-pattern and those of the reflection images. A spline-based approximation of a known mapping is shown as an example in Figure 12.
Each mapping is approximated using two collections of splines, one parameterised with the "row" variable (x_1), and the other one parameterised with the "column" variable (x_2) of the measurement pattern's plane. With these two splines, the location of any intermediate image point can be calculated in two steps by creating interpolated image points first in "row" direction, and then fitting a spline onto these in "column" direction.
Having approximated the three **continuous planar mappings**, that is, one for each camera, now, we can turn our attention – once again – to spatial issues, namely, the surface reconstruction problem.

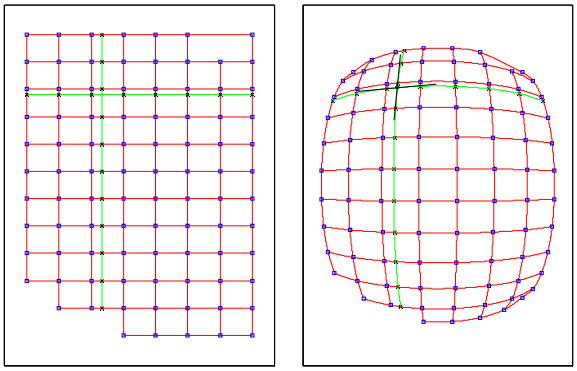


Fig. 12. A rectangular grid with some missing grid-points (left). The distorted image of the grid. A two interpolation lines and their corresponding interpolation curves are shown in green.

7. Surface reconstruction from the continuous approximation of distortions

7.1 Mathematical model of light-reflection

Mathematically, the **smooth specular convex surface** F is described and sought in preferably chosen spatial coordinate systems. Each of these coordinate systems corresponds to one of the cameras of the topographer arrangement. For notational simplicity, only one of these coordinate systems is considered in the following formulae. The origin of the camera coordinate system B is placed in the optical centre of the particular camera and the z-axis of this coordinate system is the optical axis of the camera. The specular surface F, i.e., the surface of an artificial or a living cornea is described in the following form:

$$F(x_1, x_2) = S(x_1, x_2)\hat{\boldsymbol{x}} \quad (\ \hat{\boldsymbol{x}} = (x_1, x_2, 1)^T \ )$$

Here, $S(x)$ ($x = (x_1, x_2)$) is a scalar-factor describing the inverse distance ratio, measured from B, of the 3D point P_x – corresponding to $\hat{\boldsymbol{x}}$ – and the surface-point appearing in the same direction as P_x from B. The propagation of light from the points of the measurement pattern to those of the distorted image, i.e., P_yPP_x,, is described in the coordinate system.
By doing so, a mapping is identified between the points P_y of the measurement pattern and the points P_x of the image. This mapping is $P_y \rightarrow P_x$.
It follows from the conditions prescribed for the mathematical surface that mapping $P_y \rightarrow P_x$ is one-to-one. It follows from the physical law of light reflection that the two-variable function $S(x)$ describing surface F satisfies the following **first-order partial differential equation** (PDE).

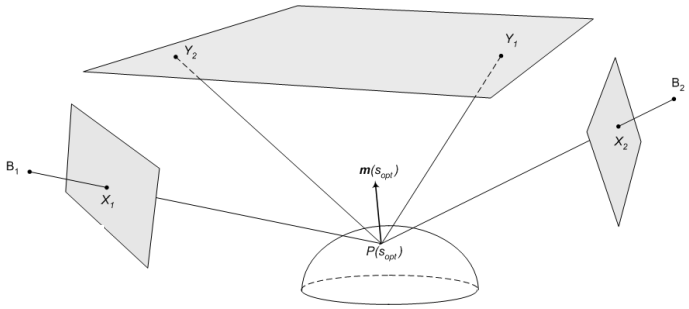


Fig. 13. Two points (Y_1 and Y_2) being reflected by the same patch of the specular surface into the two cameras; m is the normal of the surface patch.

$$\frac{1}{S(x)}\frac{\partial S(x)}{\partial x_j}=\frac{v_j(x)-x_j}{\left\langle \hat{\boldsymbol{x}},\hat{\boldsymbol{x}}-v_j(x)\right\rangle} \quad (j=1,\ 2), \quad v(x)=|\hat{\boldsymbol{x}}|\frac{k+f(x)-S(x)\hat{\boldsymbol{x}}}{\left|k+f(x)-S(x)\hat{\boldsymbol{x}}\right|}\ .$$

Function $f(x)$ can be expressed with mapping $P_x \rightarrow P_y$. Here, k is a vector pointing to a reference point in the plane of the measurement pattern; while $\langle .,. \rangle$ denotes the scalar product of the 3D space. It follows from the mathematical model described above that surface F can be determined uniquely under the starting condition of $S(0, 0) = s_0$, if the $P_y \rightarrow P_x$ mapping is known.

In Figure 13, the same surface-patch – with unit normal $\boldsymbol{m}$ – of smooth specular surface F reflects into the two cameras according to the physical law of light reflection.

7.2 Joint camera calibrations using a test-cornea of known dimensions

In the model outlined above, it is implicitly assumed that the cameras involved are calibrated. An indeed, the chessboard-based camera calibration approach outlined in Subsection 5.2 was used for this purpose. The resulting camera positions and orientations were as expected. However, it soon turned out that the calibration of the multi-camera arrangement – based on these individual camera calibrations only – does not ensure the precision required for the specular surface reconstruction. Therefore, a second – fine-tuning – calibration step proved to be necessary. A **joint and specularity-based calibration** of the cameras in the topographer arrangement was accomplished. The simulation programs that were developed for modelling reflections of smooth specular surfaces were used for this purpose. A spherical artificial cornea with know radius was used as a secondary fiducial in this case.

Again, the Levenberg-Marquardt method – mentioned earlier in Subsection 5.2 – was used to fine-tune the vector of parameters that included, in this case, all the extrinsic camera parameters of all the three cameras and the position and orientation of the test cornea. The aim was to minimize the total sum of quadratic error between the blob-centres in the simulated reflection images and the detected ones.

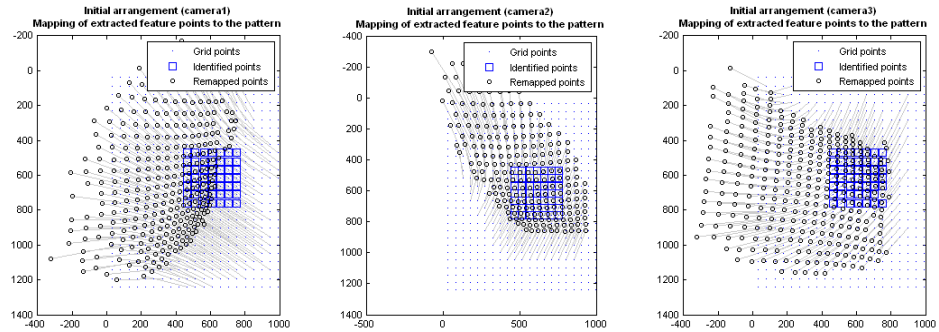


Fig. 14. The importance of the joint camera calibration is indicated by the considerable change in the back-projected grid point locations (i.e., locations derived by back-projecting the detected points onto the measurement pattern's plane). The ○'s mark the grid-point locations according to the individual camera calibration results (i.e., *before* the joint camera calibration), while the •'s indicate the grid point locations *after* the joint camera calibration.

The optimization resulted in the correction of the camera parameters. The need for this fine-tuning step is underlined by the significant (measurement pattern) point location-changes shown in the final plane of measurement pattern. Such point location-changes are shown for each of the cameras in Figure 14.

7.3 Surface reconstruction via solving the partial differential equation. Choosing an advantageous parameterization

A numerical procedure was devised to solve the above PDE's. The solution of these PDE's in effect conveys the reconstructed corneal surface. The solution involves transcribing the PDE's for individual 'rows' and the 'columns. In this manner, **first-order ordinary non-linear differential equations** were generated. These were then numerically integrated – using the well-known Runge-Kutta method – along 'row' and 'column' directions, as necessary.

Figure 15 illustrates a conceptual processing direction – e.g., from left to right in a particular row of the measurement pattern – and corresponding curves on the specular surface F and in the image plane of one of the cameras, as an ordinary differential equation derived from the PDE is solved.

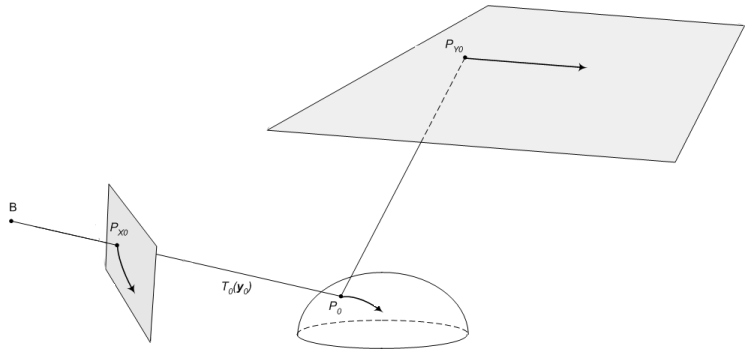


Fig. 15. Directions followed explicitly – in this case on the measurement pattern's plane – or implicitly – on the specular surface and in the image plane – during the numerical integration of the partial differential equation describing the surface reconstruction problem.

Starting conditions are necessary to obtain realistic solution for the PDE given in Subsection 5.1. These were computed with the method outlined in the next subsection. An **efficient computational scheme** was devised to re-use the points computed earlier to improve reconstruction speed. The reconstruction paths used in this scheme is shown in Figure 19 for one of the cameras.

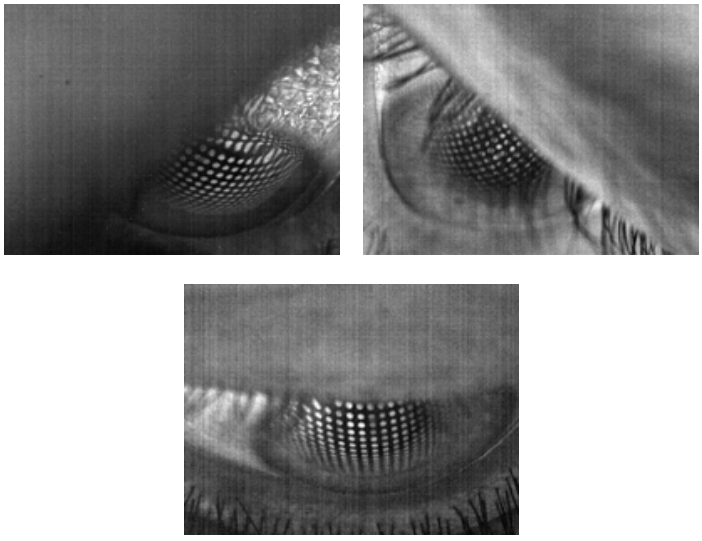


Fig. 16. Purkinje-images of the planar measurement pattern as "seen" by the three cameras of the topographer arrangement.

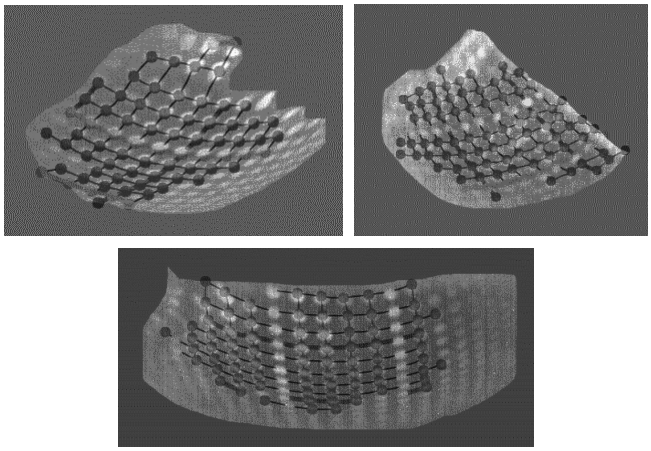


Fig. 17. The grid of the detected blobs overlaid on the corresponding image segments.

7.4 Setting of starting condition

The exact value of the above starting condition could be found out by including two laser-lights for each camera (as it is frequently done in other topographers). However, this distance setting mechanism is unnecessary here, as **precise stereo information can be gathered** with two calibrated cameras looking at the same patch of specular surface (a situation illustrated in Figure 13). An assumed surface-point and the unit normal vector at this point – determined from the first reflection image – are used to determine the point in the measurement pattern that should be seen in the second image. See Figure 18.

The Euclidean distance between the 'should-be-seen' point and the point really seen by the second camera (i.e., the length of the segment marked red in Figure 18) is calculated. If this distance (error) is considerable for an assumed surface-point, then it is – in reality – off the specular surface *F*.

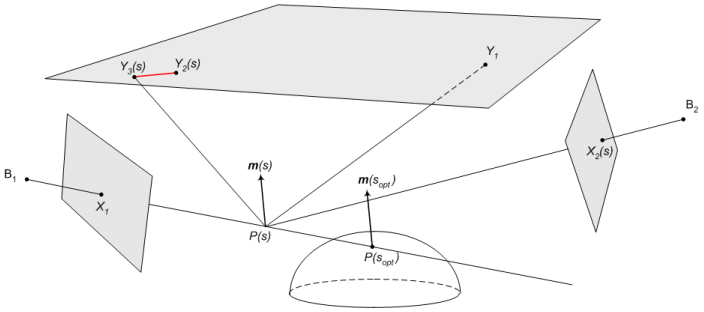


Fig. 18. Locating a point of a specular surface in the space based on two views.

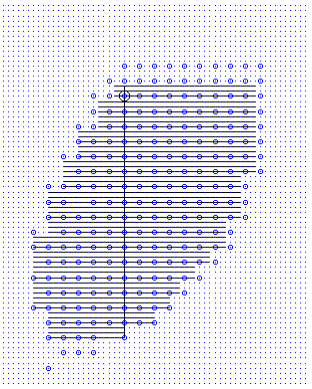


Fig. 19. Surface reconstruction paths – in the plane of the measurement pattern – for the surface region "seen" by one of the three cameras.

By moving along possible surface-points and considering also their unit normal vectors that are for the first camera, the aforementioned distance must be minimized.
The near-zero valued minima mark the spatial positions of surface-points that can be used as stating points for the reconstruction.

7.5 Reconstruction results

In Figure 21, the reconstruction precision achieved in case of a test cornea is shown. Clearly, the joint and specularity-based calibration resulted in usable positions for the cameras. In Figure 20, the distances of the reconstructed surface points from the spherical surface of the test cornea is plotted. The mentioned error distance values in peripheral areas – far from the starting point of the processing – seem to be relatively high (in the order of 10μm).

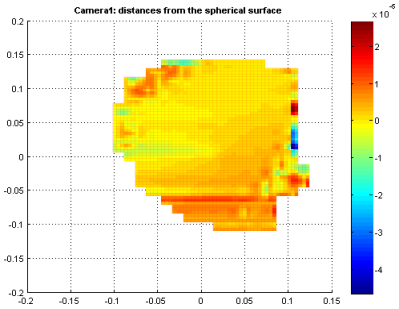


Fig. 20. Reconstruction errors encountered in case of the test cornea shown before for the surface region "seen" by one of the cameras.

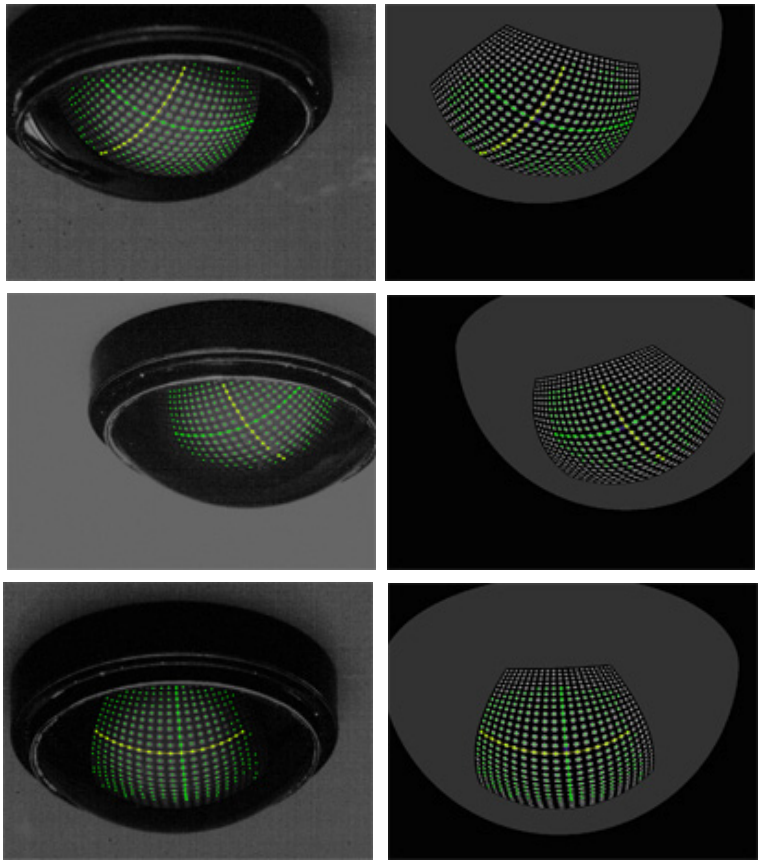


Fig. 21. Virtual images – formed by a test cornea – taken with the three cameras of the topographer arrangement (left). The detected blobs are marked with their centroids. Corresponding syntetic images (right) created via ray-tracing.

8. Conclusions and further work

The majority of the topographers in use today, rely on one view only, which is theoretically insufficient for the unique reconstruction of the corneal surface. To overcome this essential measurement deficiency, a multi-camera arrangement was proposed by the authors and mathematical methods were devised for the specular surface reconstruction. However, further experiments and simulations are required to improve the surface coverage of the multi-camera arrangement.

9. Acknowledgements

This research has been partially supported by the National Office for Research and Technoogy (NORT), Hungary, in the frame of the NKFP-2/020/04 research contract.

10. References

ANSI (2008). *Corneal Topography Systems – Standard Terminology, Requirements,* Z80.23-2008, American National Standards Institute, Washington, DC, USA

Bódis-Szomorú, A.; Dabóczi, T. & Fazekas, Z. (2008). Calibration and sensitivity analysis of a stereo vision-based driver assistance system, In: *Stereo Vision,* Bhatti, A. (Ed.), pp. 1-26, InTech Education and Publishing, ISBN 978-953-7619-22-0, Vienna, Austria

Bonfort, T. & Sturm, P. (2003). Voxel carving for specular surfaces, *Proceedings of the 9th IEEE International Conference on Computer Vision*, pp. 691-696, ISBN 0-7695-1950-4, Nice, France, October 2003, IEEE Computer Society, Washington, DC

Corbett, M. C.; Rosen, E. S.; & O'Brart, D. P. S. (1999). *Corneal Topography: Principles and Applications,* BMJ Publishing, ISBN 0-7279-1226-7, London, UK

de Almeida, M. S. & Carvalho, L. A. (2007). Different schematic eyes and their accuracy to the *in vivo* eye: a quantitative comparison study. *Brazilian Journal of Physics,* Vol. 37, No. 2A, pp. 378-387, ISSN 0103-9733

Fazekas, Z.; Soumelidis, A.; Bódis-Szomorú, A. & Schipp, F. (2008). Specular surface reconstruction for multi-camera corneal topographer arrangements, *Proceedings of the 30th Annual International IEEE EMBS Conference*, pp. 2254-2257, ISBN 978-1-4244-1814-5, Vancouver, British Columbia, Canada, August 2008, IEEE EMBS, Piscataway, NJ, USA

Fleming, R. W.; Torralba, A. & Adelson, E. H. (2004). Specular reflections and the perception of shape. *Journal of Vision,* Vol. 4, No. 9, pp. 798-820, ISSN 1534-7362

Griffin, P. M.; Narasimhan, S. & Yee, S. R. (1992). Generation of uniquely encoded light patterns for range data acquisition. *Pattern Recognition,* Vol. 25, No. 6, pp. 609–616, ISSN 0031-3203

Halstead, M. A.; Barsky, B. A.; Klein, S. A. & Mandell, R. B. (1996). Reconstructing curved surfaces from specular reflection patterns using spline surface fitting to normals, *Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques*, pp. 335-342, ISBN 0-89791-746-4, New Orleans, Louisiana, USA, August 1996, ACM, New York

Hartley, R. & Zissermann, A. (2004). *Multiple View Geometry in Computer Vision,* (2nd edition) Cambridge University Press, ISBN 0-5215-4051-8, Cambridge, UK

Iskander, D.R.; Collins, M. J. & Davis, B. (2000). Optimal modeling of corneal surfaces with Zernike polynomials. *IEEE Transactions on Biomedical Engineering,* Vol. 48, No. 1, pp. 87-95, ISSN 0018-9294

Jongsma, F.; de Brabander, J. & Hendrikse, F. (1999). Review and classification of corneal topographers. *Lasers in Medical Science,* Vol. 14, No. 1, pp. 2-19, ISSN 0268-8921

Kickingereder, R. & Donner, K, (2004). Stereo vision on specular surfaces, *Proceedings the 4th IASTED International Conference on Visualization, Imaging, and Image Processing*, pp. 335-339, ISBN 0-88986-454-3, Marbella, Spain, September 2004, IASTED, Calgary, Alberta, Canada

Lellmann, J.; Balzer, J.; Rieder, A. & Beyerer, J. (2008). Shape from Specular Reflection and Optical Flow. *International Journal of Computer Vision,* Vol. 80, No. 2, pp. 226-241, ISSN 0920-5691

Németh, J.; Erdélyi, B.; Csákány, B.; Gáspár, P.; Soumelidis, A.; Kahlesz, F. & Lang, Zs. (2002). High-speed videotopographic measurement of tear film build-up time. *Investigative Ophthalmology and Visual Science,* Vol. 43, pp. 1783-1790, ISSN 0146-0404

Savarese, S. & Perona, P. (2001). Local analysis for 3D reconstruction of specular surfaces , *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Vol. 2, pp. 738-745, ISBN 0-7695-1272-0, Kauai, Hawaii, USA, December 2001, IEEE Computer Society, Washington, DC

Savarese, S. & Perona, P. (2002a). Local analysis for 3D reconstruction of specular surfaces – Part II, *Proceedings of the 7th European Conference on Computer Vision*, Lecture Notes in Computer Science, Vol. 2351, pp. 1148-1158, ISBN 3-540-43744-4, Copenhagen, Denmark, May 2002, Springer, Heidelberg

Savarese, S.; Chen, M. & Perona, P. (2002b). Second order local analysis for 3D reconstruction of specular surfaces, *Proceedings of the First International Symposium on 3D Data Processing Visualisation and Transmission*, pp. 356-361, ISBN 0-7695-1521-4, Padova, Italy, June 2002, IEEE Computer Society, Washington, DC

Savarese, S.; Chen, M. & Perona, P. (2004). What do reflections tell us about the shape of a mirror? *Proceedings of the 1st Symposium on Applied Perception in Graphics and Visualization*, pp. 115-118, ISBN 1-58113-914-4, Los Angeles, California, USA, August 2004, ACM, New York

Sicam, V. A; Snellenburg, J. J.; van der Heide, R. G. L. & van Stokkum, I. H. M. (2007). Pseudo Forward Ray-Tracing: A new method for surface validation in cornea topography. *Optometry & Vision Science,* Vol. 84, No. 9, pp. E915-E923, ISSN 1040-5488

Soumelidis, A.; Fazekas, Z.; Schipp, F.; Edelmayer, A.; Németh, J. & Csákány, B. (2008). Development of a multi-camera corneal topographer using an embedded computing approach, *Proceedings of the 1st International Conference on Biomedical Electronics and Devices*, pp. 126-129, ISBN 978-989-8111-17-3, Funchal, Madeira, Portugal, Jan. 2008, INSTICC, Setúbal, Portugal

Swartz, T.; Marten, L. & Wang, M. (2007). Measuring the cornea: the latest developments in corneal topography. *Current Opinion in Ophthalmology.,* Vol. 18, No. 4, pp. 325-333, ISSN 1040-8738

Trichet, R.. & Merialdo, B. (2007). Probabilistic matching algorithm for keypoint based object tracking using a Delaunay triangulation, *Proceedings of the Eight International Workshop on Image Analysis for Multimedia Interactive Services*, pp. 17-20, ISBN 0-7695-2818-X, Santorini, Greece, June 2007, IEEE Computer Society, Washington, DC, USA

Vos, F. M.; van der Heijde, G.L.; Spoelder, H.J.W.; van Stokkum, I.H.M. & Groen, F.C.A. (1997). A new instrument to measure the shape of the cornea based on pseudorandom color coding. *IEEE Transactions on Instrumentation and Measurement,* Vol. 46, No. 4, pp. 794-797, ISSN 0018-9456

Wang, W.; Wang, Z.; Wang, Y. & Zuo, T. (2006). Optical aberrations of the cornea and the crystalline lens. *Optik - International Journal for Light and Electron Optics,* Vol. 117, No. 9, pp. 399-404, ISSN 0030-4026

Zhang, Z. (2000). A flexible new technique for camera calibration. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* Vol. 22, No. 11, pp. 1330-1334, ISSN 0018-9340