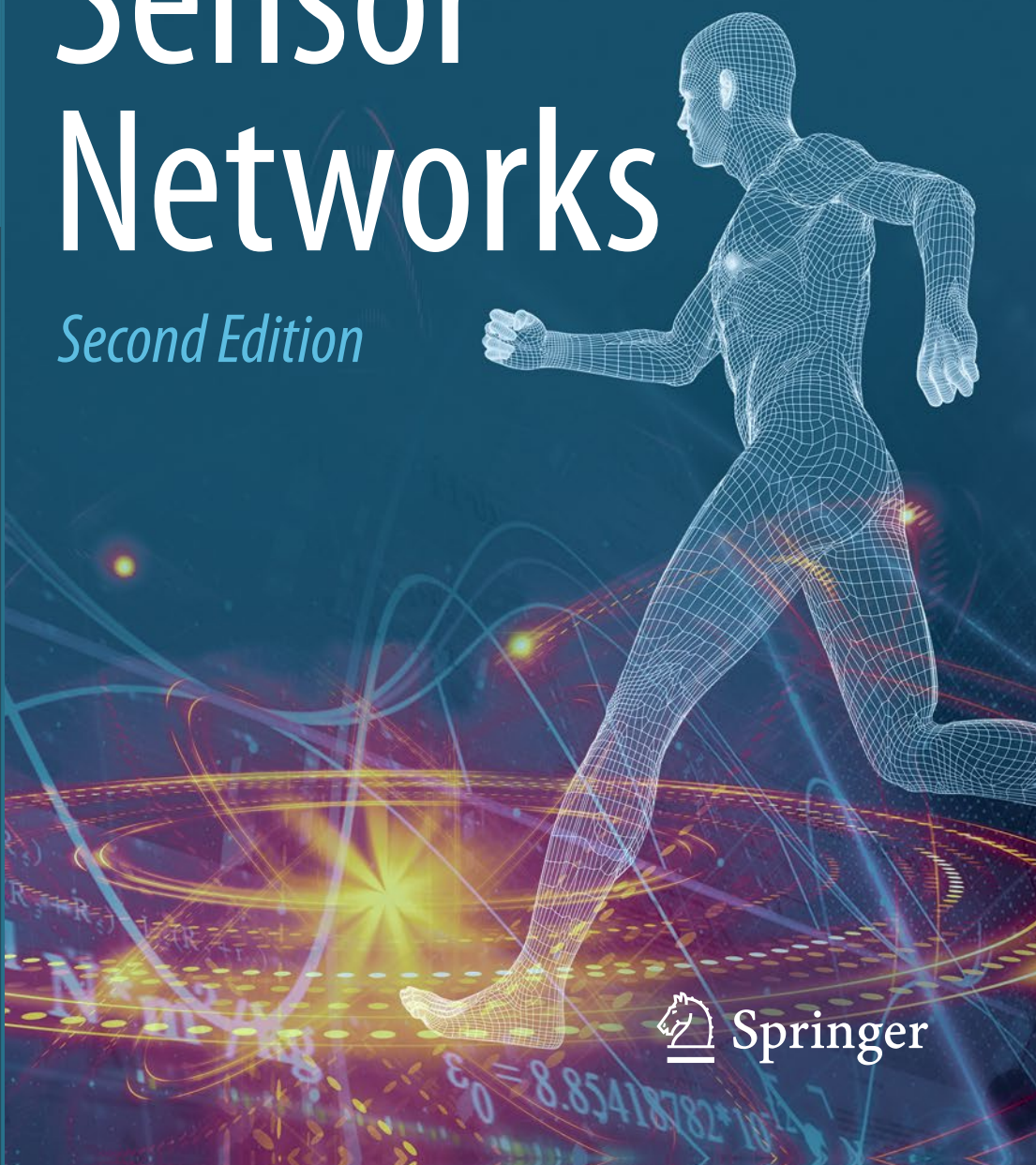
Guang-Zhong Yang *Editor*

Body Sensor Networks

Second Edition



Springer

Body Sensor Networks

Guang-Zhong Yang
Editor

Body Sensor Networks

Second Edition

Springer

Editor
Guang-Zhong Yang, PhD
Director and Co-Founder, The Hamlyn Centre
Imperial College London
London, UK

ISBN 978-1-4471-6373-2 ISBN 978-1-4471-6374-9 (eBook)
DOI 10.1007/978-1-4471-6374-9
Springer London Heidelberg New York Dordrecht

Library of Congress Control Number: 2014937397

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

Use body as the medium, inspiration and a source of energy to provide continuous sensing, monitoring and intervention

Since the first edition of this book, the field of Body Sensor Networks (BSNs) has advanced rapidly. The original motivation of BSN was to harness allied technologies that underpin the development of pervasive sensing for healthcare, wellbeing, sports and other applications that require "ubiquitous" and "pervasive" monitoring of physical, physiological, and biochemical parameters in any environment without activity restriction and behaviour modification. The ultimate aim of BSN is therefore to provide a truly personalised monitoring platform that is pervasive, intelligent and context-aware, yet "invisible", with applications ranging from managing patients with chronic disease and care for the elderly, to general well-being monitoring and performance evaluation in sports. To ensure its widespread use, there are many technical challenges that need to be tackled. These include the need for better sensor design, MEMS integration, biocompatibility, power source miniaturisation, low power wireless transmission, context awareness, secure data transfer and integration with smart therapeutic systems.

In this second edition of the book, we have updated the chapters with the latest developments in the field, addressing sensor design, micro-electronics and information processing aspects of the system. Since its inception, the development of BSN has been focussed on both wearable and implantable sensors. In the last few decades, we have seen rapid advances in both chemical and biosensor developments. The emergence of new biological sensing modalities is fundamentally changing the way we apply biomeasurements in vivo. In terms of implantable sensing, many of the issues associated with the extension of biosensor technology from in vitro to in vivo applications have long been appreciated, and a number of practical issues are addressed in this book. In a BSN with limited bandwidth and power constraints, the conventional method of data acquisition and analogue-to-digital data conversion with signal processing taking place after transmission is no longer optimal. A BSN represents a prime candidate for bio-inspired local processing to take place at the sensor front-end before transmission. This processing

could include spatial and temporal averaging for drift and failure tolerance. The key principle of bio-inspired engineering in this application area is that biology does not often deal in absolute values, but in relative changes from a given norm.

From a sensor data processing and inferencing point of view, the development of the BSN has introduced a whole range of challenging research issues in pattern recognition, behaviour profiling and machine learning. The pursuit of low-power, miniaturised, smart sensing embodied either as a wearable or implantable device has also imposed significant challenges on integrating information from what is often heterogeneous, incomplete and error-prone sensor data. In practice, it is therefore desirable to rely on sensors with redundant or complementary data to maximise the information content and reduce both systematic and random errors.

One important aspect of the book is the introduction of bio-inspired concepts both for hardware design and for developing software components that possess the self-* properties of autonomic sensing. We have discussed the use of artificial neural networks for performing context-aware sensing, and the use of autonomic principles of self-healing, self-organisation and self-protection for developing BSNs with effective fault tolerance and self-protection.

As mentioned in the first edition of this book, advances in science and medicine are intimately linked. In current clinical practice, ranging from prevention to complex intervention, we rely heavily on early, accurate and complete diagnosis followed by close monitoring of the results. Attempts so far, however, are still limited to a series of snapshots of physiological, biomechanical and biochemical data. Transient abnormalities cannot always be reliably captured. The concept of BSN is therefore an important ingredient for the future development of pervasive healthcare because technological developments in sensing and monitoring devices will not only change chronic disease management in a home or community setting, but also reshape the general practice of clinical medicine.

With demographic changes associated with the aging population and the increasing number of people living alone, the social and economic structure of our society is also changing rapidly. In a population consisting of several vulnerable groups, such as those with chronic disease and the elderly, the need for effective individualised health monitoring and delivery is the primary motivation for the development of BSNs. There is little doubt that for the development of the BSN, a panoply of technologies will need to be combined in new and previously unsuspected ways. However, the rewards for success, in terms of the quality and duration of life in the case of many of those suffering from chronic conditions, will be substantial.

There has been tremendous effort from all contributors of this book in making this 2nd edition possible. I would like to express my sincere thanks to all the contributing authors. Without their enthusiasm, support, and flexibility in managing the tight publishing schedule, this book would not have become possible. In particular, I would like to thank Su-Lin Lee, Emily Yang, Benny Lo, and Surapa Thiemjarus, for all their hard work in providing essential editorial support, as well as being actively involved in the preparation of some of the technical chapters. I would also like to thank the editorial staff of Springer, the publisher of

this volume. In particular, I am grateful to Helen Desmond and her colleagues in helping with the editorial matters.

This work would not have been possible without the financial support from all the funding bodies that supported our work, particularly the Engineering and Physical Sciences Research Council (EPSRC), UK. Their generous support has allowed us to establish and promote this exciting field of research – a topic that is so diversified, and yet brings so many challenges and innovations to each of the disciplines involved.

I do hope this book will act as a valuable resource to a very wide spectrum of readers interested in, or inspired by, this multifaceted and exciting topic.

London Guang-Zhong Yang
November 2013

About the Editor

Guang-Zhong Yang (PhD, FREng) is Director and Co-founder of the Hamlyn Centre, Deputy Chairman of the Institute of Global Health Innovation, Imperial College London, UK. Professor Yang also holds a number of key academic positions at Imperial College – he is Director and Founder of the Royal Society/ Wolfson Medical Image Computing Laboratory, co-founder of the Wolfson Surgical Technology Laboratory, Chairman of the Centre for Pervasive Sensing. He is a Fellow of the Royal Academy of Engineering (RAEng), fellow of IEEE, IET, MICCAI, AIMBE, IAMBE, City of Guilds and a recipient of the Royal Society Research Merit Award and The Times Eureka 'Top 100' in British Science.

Professor Yang's main research interests are in medical imaging, sensing and robotics. In imaging, he is credited for a number of novel MR phase contrast velocity imaging and computational modelling techniques that have transformed in vivo blood flow quantification and visualization. These include the development of locally focused imaging combined with real-time navigator echoes for resolving respiratory motion for high-resolution coronary-angiography, as well as MR dynamic flow pressure mapping for which he received the ISMRM I. I. Rabi Award. He pioneered the concept of perceptual docking for robotic control, which represents a paradigm shift of learning and knowledge acquisition of motor and perceptual/cognitive behaviour for robotics, as well as the field of Body Sensor Network (BSN) for providing personalized wireless monitoring platforms that are pervasive, intelligent, and context-aware. Professor Yang is a Distinguished Lecturer for IEEE Engineering in Medicine and Biology Society and Editor-in-Chief, *IEEE Journal of Biomedical and Health Informatics*.

Contents

Contributors

Javier Andreu-Perez The Hamlyn Centre, Imperial College London, London, UK

Omer Aziz The Hamlyn Centre, Imperial College London, London, UK

Anthony E.G. Cass Department of Chemistry and Institute of Biomedical Engineering, Imperial College London, London, UK

Timothy G. Constandinou Department of Electrical and Electronic Engineering, Imperial College London, London, UK

Jonathan M. Cooper School of Engineering, University of Glasgow, Glasgow, UK

David R.S. Cumming School of Engineering, University of Glasgow, Glasgow, UK

Ara Darzi The Hamlyn Centre, Imperial College London, London, UK

Joshua Ellul The Hamlyn Centre, Imperial College London, London, UK

Javier Espina Philips Group Innovation, Eindhoven, The Netherlands

Thomas Falck Philips Group Innovation, Eindhoven, The Netherlands

Paul A. Hammond Dialog Semiconductor, Edinburgh, UK

Henry Higgins Caldicot, Monmouthshire, UK

Xiaopeng Hu School of Computer Science and Technology, Dalian University of Technology, Dalian, China

Kamen Ivanov Institute of Biomedical and Health Engineering, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, PR China

Erik A. Johannessen Førsteamanuensis, Department of Micro- and Nanosystems Technology (IMST), Vestfold University College, Tønsberg, Norway

Christopher J. Johnson Department of Chemistry and Institute of Biomedical Engineering, Imperial College London, London, UK

Richard Kwasnicki The Hamlyn Centre, Imperial College London, London, UK

Thao T. Le Department of Chemistry and Institute of Biomedical Engineering, Imperial College London, London, UK

Benny Lo The Hamlyn Centre, Imperial College London, London, UK

Robert Merrifield The Hamlyn Centre, Imperial College London, London, UK

Paul Mitcheson Department of Electrical and Electronic Engineering, Imperial College London, London, UK

Oliver Mülhens Philips Group Innovation, Eindhoven, The Netherlands

Danny O'Hare Department of Bioengineering, Imperial College London, London, UK

Athanasia Panousopoulou The Hamlyn Centre, Imperial College London, London, UK

Institute of Computer Science, Foundation for Research and Technology – Hellas (FORTH), Heraklion, Crete, Greece

Lars Schmitt Philips Group Innovation, Eindhoven, The Netherlands

Leila Shepherd DNA Electronics Ltd, London, UK

Surapa Thiemjarus National Electronics and Computer Technology Center, Pathumthani, Thailand

Chris Toumazou Department of Electrical and Electronic Engineering, Imperial College London, London, UK

Jakub Trzebinski Department of Chemistry and Institute of Biomedical Engineering, Imperial College London, London, UK

Lei Wang Institute of Biomedical and Health Engineering, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, PR China

Guang-Zhong Yang The Hamlyn Centre, Imperial College London, London, UK

Eric Yeatman Department of Electrical and Electronic Engineering, Imperial College London, London, UK

Zhiqiang Zhang The Hamlyn Centre, Imperial College London, London, UK

Chapter 1
Introduction

Guang-Zhong Yang, Omer Aziz, Richard Kwasnicki, Robert Merrifield, Ara Darzi, and Benny Lo

1.1 Wireless Sensor Networks

Over the past decades, the miniaturisation and cost reduction brought about by the semiconductor industry have made it possible to create computers that are smaller in size than a pin head, powerful enough to carry out the processing required, and affordable enough to be considered disposable. This reduction in size and increase in processing capability will undoubtedly continue in future years, with new classes of computer or smart device emerging in every decade [1]. Similarly, advances in wireless communication, sensor design, and energy storage technologies have meant that the concept of a truly pervasive *Wireless Sensor Network* (WSN) is rapidly becoming a reality [2]. Integrated microsensors no more than a few millimetres in size, with onboard processing and wireless data transfer capability are the basic components of such networks already in existence nearly a decade ago [3, 4]. With rapid expansion of smart devices in recent years, many applications have been proposed for the use of WSNs and they are likely to change every aspect of our daily lives.

The evolution of these wireless devices originated from academic research as shown in Fig. 1.1. There is now a myriad of platforms available from both academic institutions and commercial organisations. One of the first concepts developed to utilise large-scale pervasive wireless sensor networks was "Smart Dust." This was developed at the *University of California* (UC) at Berkeley and funded by the *Defence Advanced Research Projects Agency* (DARPA). The aim of the project was to produce a self-contained, millimetre-scale sensing and communication platform for massively distributed sensor networks [4]. Primarily meant as a military application, the "Smart Dust" concept involved the use of thousands of tiny wireless sensor "motes" that could be spread over a large battlefield area, allowing enemy movements to be monitored in a covert manner. In addition to its

G.-Z. Yang, Ph.D (✉) • O. Aziz • R. Kwasnicki • R. Merrifield • A. Darzi • B. Lo
The Hamlyn Centre, Imperial College London, London SW7 2BY, UK
e-mail: g.z.yang@imperial.ac.uk

G.-Z. Yang (ed.), *Body Sensor Networks*, DOI 10.1007/978-1-4471-6374-9_1,

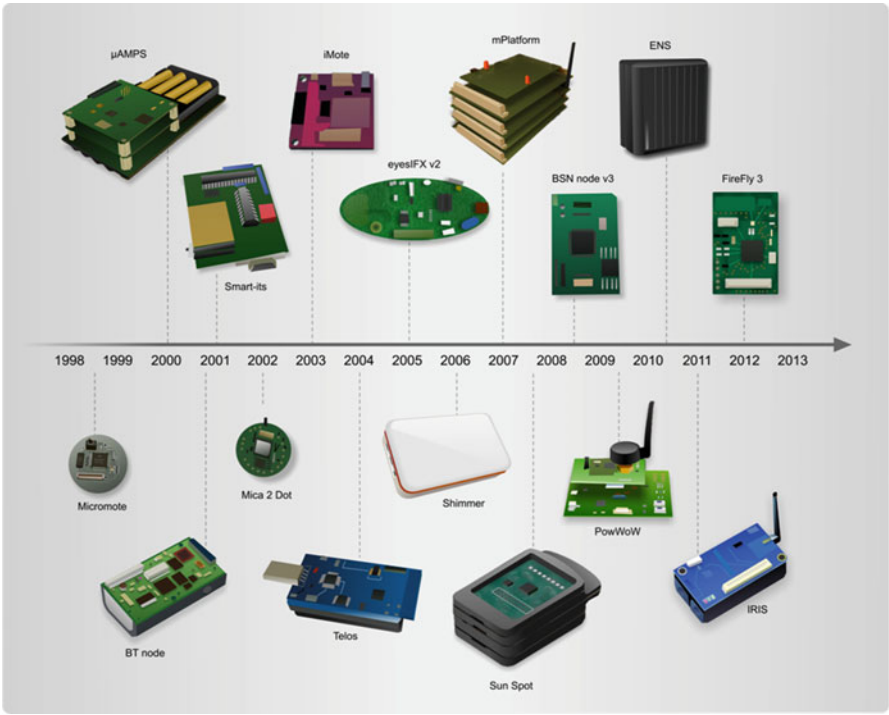


Fig. 1.1 A timeline of the evolution of wireless sensor networks and some of the exemplar platforms that have been developed over the last 15 years. *Details see Appendix A – Wireless Sensor Development Platforms*

military application, *Smart Dust* was intended to be used in a number of different settings ranging from deploying nodes into the atmosphere for weather condition monitoring, to placing them in environments such as factories to monitor their production output. Researchers have also approached *Smart Dust* from a biotechnology perspective to produce motes from chemical compounds rather than electrical circuitry [5].

In the early stages of the project, the team gained experience by building relatively large motes using *Commercial Off-The-Shelf* (COTS) components. With cooperation from Intel, these motes were created as an open-source hardware and software platform, combining sensors, low power wireless communication, and processing into a single architecture. The motes were also designed to have the ability to "self-organise" (resulting in a self-configuring WSN), and carry out onboard signal processing and distributed inference tasks prior to sending relevant information to a central controller. Whilst tiny, ubiquitous, low-cost, *Smart Dust* motes remain an attractive concept, many reasonably small motes are now commercially available. Existing designs have already integrated a range of sensors monitoring a variety of environmental factors including temperature, humidity, barometric pressure, light intensity, tilt and vibration, and magnetic field with short-distance wireless communication.

One of the key developments for WSNs is the small, open source (freely available) energy-efficient software operating system known as the *Tiny Micro-threading Operating System*, or "TinyOS", which has been developed at UC Berkeley. This operating system provides a basic framework and development environment for WSNs, and it functions well under the constraints of power, size, and cost. TinyOS software runs both the hardware and network, making sensor measurements, routing decisions, and controlling power dissipation. A review of different software development tools for WSN is provided in Appendix B.

Currently, the new applications emerging for WSNs can be categorised into the following three types: those used for monitoring environments (indoor, outdoor, urban or countryside), monitoring objects (such as machines and buildings) and monitoring the interaction of these objects with environments [6]. For example, companies such as *British Petroleum* (BP) have leveraged the huge potential offered by WSN technology and have embarked on the development of its large-scale use. One example of this is the setting up of an experimental WSN for monitoring their refinery equipment in order to measure abnormal vibration and thereby to alert engineers to a potentially malfunctioning piece of equipment before it actually breaks down. Other application areas also include the monitoring of customers' liquefied petroleum gas tank fill levels [2]. Similarly, Shell and HP have initiated on the development of new ultrasensitive, low-power MEMS accelerometers for wireless seismic acquisition system for improved understanding of the earth's subsurface to facilitate oil and gas exploration [7]. In recent years, the use of WSN for civil infrastructure monitoring [8] and smart homes [9, 10] has enjoyed significant growth and maturity, thus realising the vision of truly ubiquitous deployment of these wireless sensors.

WSNs have also been applied to habitat monitoring, with sensor nodes placed on animals or in their surrounding environment. The team at Berkeley illustrated the use of WSNs for monitoring microclimates in and around nesting burrows of the Leach's Storm Petrel on the Great Duck Island in Maine [11]. Other examples include the "Zebranet" project for monitoring the long-range migration, interspecies interactions and nocturnal behaviour of zebras in Africa, and the EU WASP (Wirelessly Accessible Sensor Populations) for human and livestock monitoring.

Whilst WSN technology continues to evolve for the broad range of applications and settings described above, it does not specifically tackle the challenges associated with human body monitoring. The human body consists of a complicated internal environment that responds to and interacts with its external surroundings, but is in a way "separate" and "self-contained". Human body monitoring using a network of wireless sensors may be achieved by attaching these sensors to the body surface as well as implanting them into tissues. In essence, the human body environment is not only on a smaller scale, but also requires a different type and frequency of monitoring, with appreciation of different challenges than those faced by WSN. The realisation that proprietary designed WSNs are not ideally suited to monitoring the human body and its internal environment has led to the development of a wireless *Body Sensor Network* (BSN) platform.

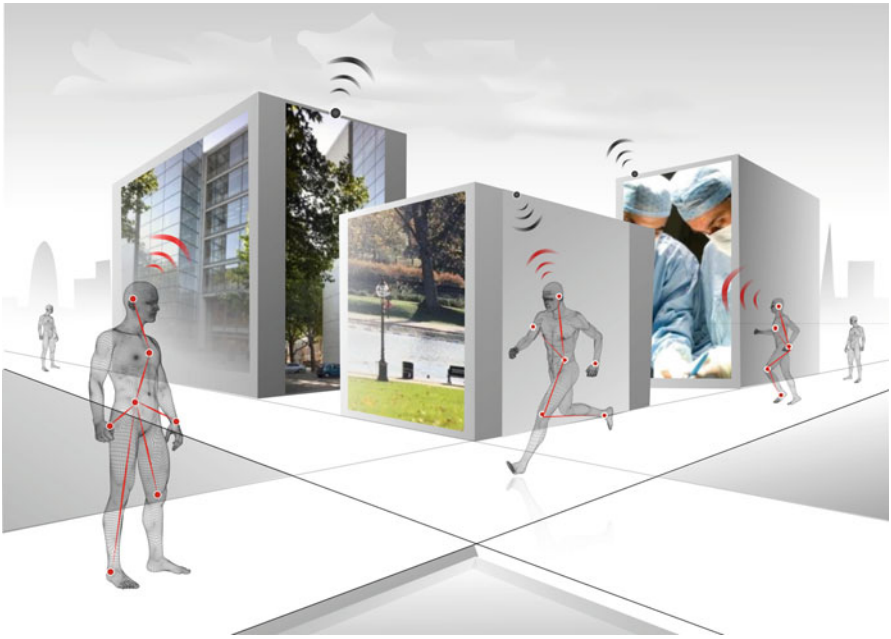


Fig. 1.2 Diagrammatic representation of wearable BSN architecture with wirelessly linked context-aware "on body" (external) sensors and its seamless integration with home, working, and hospital environments

Specifically designed for the wireless networking of implantable and wearable body sensors, the BSN architecture aims to set a standard for the development of a common approach towards pervasive monitoring. Figure 1.2 is a diagram illustrating the basic concept of such architecture. It represents a subject with a number of sensors attached to their body, each sensor also being connected to a small processor, wireless transmitter, and battery, and all together forming a "BSN node complex" capable of seamlessly integrating with home, office, and hospital environments. The BSN node ensures the accurate capture of data from the sensor to which it is connected, carries out low level processing of the data, and then wirelessly transmits this information to a *Local Processing Unit* (LPU). The data from all the sensors is in this way collected by the LPU, processed further, and fused before being wirelessly transmitted to a central monitoring server either via a wireless LAN, Bluetooth, or mobile phone (GPRS or 3G/4G) network [12].

Although the challenges faced by BSNs are in many ways similar to WSNs, there are intrinsic differences between the two, which require special attention. Some of these differences are illustrated in Table 1.1. Other major differences include the development of implantable sensors and new sensing modalities for BSN. The purpose of this chapter is to provide an overview of the development and history of wireless BSNs, highlighting not only the challenges lying ahead but also the future development directions.

Table 1.1 Different challenges faced by WSN and BSN

Challenges	WSN	BSN
Scale	As large as the environment being monitored (m/km)	As large as human body parts (mm/cm)
Node number	Greater number of nodes required for accurate, wide area coverage	Fewer, more accurate sensors nodes required (limited by space)
Node function	Multiple sensors, each perform dedicated tasks	Single sensors, each perform multiple tasks
Node accuracy	Large node number compensates for accuracy and allows result validation	Limited node number with each required to be robust and accurate
Node size	Small size preferable but not a major limitation in many cases	Pervasive monitoring and need for miniaturisation
Dynamics	Exposed to extremes in weather, noise, and asynchrony	More predictable environment but motion artefact is a significant challenge
Event detection	Early adverse event detection desirable; failure often reversible	Early adverse events detection vital; human tissue failure irreversible
Variability	Much more likely to have a fixed or static structure	Biological variation and complexity means a more variable structure
Data protection	Lower level wireless data transfer security required	High level wireless data transfer security required to protect patient information
Power supply	Accessible and likely to be changed more easily and frequently	Inaccessible and difficult to replace in implantable setting
Power demand	Likely to be greater as power is more easily supplied	Likely to be lower as energy is more difficult to supply
Energy scavenging	Solar, and wind power are most likely candidates	Motion (vibration) and thermal (body heat) most likely candidates
Access	Sensors more easily replaceable or even disposable	Implantable sensor replacement difficult and requires biodegradability
Biocompatibility	Not a consideration in most applications	A must for implantable and some external sensors. Likely to increase cost
Context awareness	Not so important with static sensors where environments are well defined	Very important because body physiology is very sensitive to context change
Wireless technology	Bluetooth, Zigbee, GPRS, and wireless LAN, and RF already offer solutions	Low power wireless required, with signal detection more challenging
Data transfer	Loss of data during wireless transfer is likely to be compensated by number of sensors used	Loss of data more significant, and may require additional measures to ensure QoS and real-time data interrogation capabilities

1.2 BSN for Healthcare and Wellbeing

The observations by Hippocrates, the Greek founder of modern medicine, that audible sounds emanating from the chest were produced by the heart, ultimately led to the development of the stethoscope by René Théophile Hyacinthe Laënnec in 1816. The name stethoscope was chosen because it is derived from the Greek stethos (meaning chest), and skopein (meaning to observe). At the time Laënnec describes using a 1 ft long, 1.5 in. wide cedar tube ¼ in. in diameter, with a central channel [13]. Today the acoustic stethoscope looks markedly different, and has evolved into a device has been carefully engineered to accurately relay heart and chest sounds through its diaphragm through to the physician's ear. More recently electronic stethoscopes that convert acoustic sound waves to electrical signals have emerged. These work by amplifying and processing heart and breath sounds thereby optimising auscultation (listening), have the added advantage of being able to record and thereby transmit the information they record. Electronic stethoscopes achieve their function by a number of means including microphones in the chest piece, piezoelectric crystals, or a stethoscope diaphragm with an electrically conductive inner surface forming a capacitive sensor [14]. The emergence of this technology has made remote cardiopulmonary examination into a reality with some groups demonstrating the reliable use of "*telestethoscopy*" to monitor vulnerable patients with cardiovascular diseases such as heart failure, in their home environments [15]. The path is now paved for wireless continuous telestethoscopy for monitoring heart and breath sounds, in a way that Hippocrates himself could never imagine.

Like stethoscopes, a number of other diagnostic tools have continued to evolve, revolutionising medical practice by allowing doctors to extract more and more important information about their patients' physiological states. *Electrocardiography* (ECG) has made the diagnosis of myocardial infarction immediate, respirometry has allowed us to quantify and monitor the progress of chronic pulmonary disease, and pulse oximetry has meant we can acutely monitor a patient's oxygen saturation thereby indirectly monitoring their respiratory function. Whilst these diagnostic tools continue to develop, they still offer information that is nothing more than a "*snapshot in time*". No matter how sophisticated and complex the monitoring equipment gets, the next great challenge for core diagnostic devices lies in being able to monitor a patient's physical parameters, physiology and biochemistry continuously and with the patient in any environment.

The impact of the environment a patient is in on their physiology is well known. Perhaps the best example of this is the "*white coat syndrome*" which results in patients' recorded blood pressures being much higher when visiting their physicians in clinics or in the hospital [16]. For chronic diseases where deterioration is more gradual, being able to monitor a patient in their own environment, be it home, work, or outdoors is therefore key. In the case of hypertension (a raised blood pressure), ambulatory blood pressure monitoring would result not only in more accurate diagnosis, but also more effective monitoring of response to therapy. For acute diseases requiring hospital admission, patients go from being monitored

intensively in the hospital ward, to a completely unmonitored home environment when they have shown a certain degree of recovery and stability in their condition. On many occasions a patient is kept in hospital slightly longer than absolutely necessary just to allow the physician to monitor them for a certain period of time before being confident of their discharge. Having the ability to monitor patients at home would be invaluable in both acute and chronic disease states, allowing safe early discharge to a monitored home environment. The development of wireless BSNs offers a platform to establish such a health monitoring system, and represents the latest evolution of diagnostic tools. BSN patient monitoring systems promise to provide information that is likely to be as important, dramatic and revolutionary as those initial observations made by Hippocrates himself.

With increasing maturity of BSN, the current focus is also shifted towards the general health and wellbeing of the population rather than just the management of disease progression or the efficacy of therapeutic measures. The *World Health Organisation* (WHO) defines health as "*a state of complete physical, mental and social well-being and not merely the absence of disease of infirmity*" [17]. Preventing disease through promotion of healthy lifestyle choice is potentially a cost-effective approach to modern healthcare challenges [18]. Choices such as diet, physical activity, sleep, smoking and alcohol, have all been associated with many medical conditions. One of the most documented conditions associated with lifestyle choices is in fact cardiovascular disease [19]. The development of BSN is therefore important for establishing undesirable lifestyles leading to chronic conditions, which is critical to the promotion of healthy living and help realise the vision of prevention and early intervention of diseases. In the following sections, we will outline the roles of BSN for different sensing scenarios.

1.2.1 Monitoring Patients with Chronic Disease

The scale of the requirement for patient monitoring in healthcare systems can only be appreciated once the magnitude of human disease processes requiring early diagnosis and treatment is considered. Several examples illustrate this need, but none as dramatically as cardiovascular related illnesses.

Abnormalities of heart rhythm (*arrhythmias*) such as *atrial fibrillation* are commonly encountered in clinical practice, occurring in as many as 4 % of the population over the age of 60, increasing with age to almost 9 % in octogenarians [20]. Early symptoms of atrial fibrillation include fatigue and palpitations, and often lead to the patient seeking medical advice. *Electrocardiography* (ECG) is eventually performed along with other investigations, and as soon as the diagnosis is made treatment is begun to try and prevent the longer-term complications of tachycardia (rapid heart rate induced) including cardiomyopathy (enlargement of the heart resulting in pump failure) and stroke. To prevent stroke, the patient is often placed on anticoagulant (blood thinning) medication placing them at risk of potential bleeding complications from the therapy. This results in a twofold

increase in mortality in this elderly patient group, independently of other risk factors [21]. Apart from early detection of this condition using ECG so that prompt treatment can be initiated, regular monitoring is required to ensure control of the heart rate. Whilst the use of 24-h Holter ECG monitoring does help, continuous and pervasive monitoring of heart rate using miniaturised and BSN technology worn by the patient during their day offers the chance to diagnose abnormalities in cardiac rhythm earlier than ever before. Ultimately this guides therapy to control cardiac rate and rhythm, thereby preventing many of the complications of *atrial fibrillation*. For high-risk groups such as elderly patients with *atrial fibrillation* on anticoagulants, BSN's would offer a means of determining the effectiveness of therapy and the need for the patient to remain on the drug.

Hypertension is thought to affect approximately one billion individuals worldwide [22]. The diagnosis of this disease is often made in an otherwise asymptomatic patient who has presented to their doctor for other reasons. This condition can, if untreated, result in end-organ failure and significant morbidity; ranging from visual impairment to coronary artery disease, heart failure, and stroke. *Heart failure* in turn affects up to 10 % of patients above the age of 65 [23], Early diagnosis of high blood pressure is important for both controlling risk factors such as smoking and high cholesterol, but also for early initiation of antihypertensive treatment. The diagnosis is confirmed using serial blood pressure measurements, and once treatment is commenced this is titrated to the required effect by monitoring the patient's blood pressure over a period of weeks or months. Once a patient has been diagnosed with hypertension, they require regular blood pressure monitoring to ensure adequacy of therapy. Indeed over a patient's life, the pharmacotherapy they receive may be altered many times. One can imagine how labour-intensive blood pressure monitoring in these patients can be, often requiring several visits to clinics. Although home blood pressure testing kits have been made available, the limitations of these devices are their dependence on the operator and patient motivation. Recently, a new category termed "*prehypertension*" has been identified and may lead to even earlier initiation of treatment [24]. BSNs would allow doctors to monitor patients with seemingly high blood pressure during their normal daily lives, correlating this to their other physiology in order to better understand not only the disease process but also to decide what therapy to start the patient on, and to monitor their response to this therapy.

Diabetes mellitus is a well-known chronic progressive disease resulting in several end-organ complications. It is a significant independent risk factor for hypertension, peripheral vascular, coronary artery, and renal disease amongst others. The prevalence of diabetes mellitus has increased dramatically over the past four decades, mainly due to the increase in prevalence of obesity [25]. It is estimated that annually 24,000 cases of diabetes induced blindness are diagnosed, and 56,000 limbs are lost from peripheral vascular disease in the United States alone. The diagnosis is often made from measuring fasting blood glucose (which is abnormally raised) either during a routine clinical consultation, or as a result of complications of the condition. Once such acute complication is diabetic ketoacidosis which can be life threatening, and can occur not only in newly diagnosed

diabetics, but also in those with poor blood sugar control due to reduce compliance with medication [26]. Once diagnosed, these patients require the regular administration of insulin at several times during the day, with blood glucose "pinprick" testing used to closely monitor patients' blood sugar in between these injections. This need for repeated drawing of blood is invasive and therefore undesirable for many patients, yet there are at present no clear reliable alternatives. As previously mentioned, variable treatment compliance rates (60–80 % at best) in these patients are made worse the fact that they are on multiple medications [27]. BSN technology used in the monitoring of this group would allow the networking of wireless implantable and attachable glucose sensors not only to monitor patient glucose levels but also to be used in "closed feedback loop" systems for drug (insulin) delivery, as described later on in this chapter.

Although the three chronic conditions mentioned above illustrate the need for continuous physiological and biochemical monitoring, there are other examples of disease processes that would also benefit from such monitoring. Table 1.2 lists some of these processes and the parameters that may be used to monitor them.

1.2.2 Monitoring Hospital Patients

In addition to monitoring patients with chronic diseases, there are two other specific areas where BSN applications offer benefit. The first of these is the hospital setting, where a large number of patients with various acute conditions are treated every year. At present, patients in hospital receive monitoring of various levels of intensity ranging from intermittent (four to six times a day in the case of those suffering with stable conditions), to intensive (every hour), and finally to continuous invasive and non-invasive monitoring such as that seen in the intensive care unit. This monitoring is normally in the form of vital signs measurement (blood pressure, heart rate, ECG, respiratory rate, and temperature), visual appearance (assessing their level of consciousness) and verbal response (asking them how much pain they are in).

Patients undergoing surgery are a special group whose level of monitoring ranges from very high during and immediately after operation (under general anaesthesia), to intermittent during the post-operative recovery period. Aside from being restrictive and "wired", hospital ward-based patient vital signs monitoring systems tend to be very labour intensive, requiring manual measurement and documentation, and are prone to human error. Automation of this process along with the ability to pervasively monitor patients wherever they are in the hospital (not just at their bedside), is desirable not only to the healthcare provider, but also to the patient. In the post-operative setting, the use of wearable devices and implantable micro-machined wireless sensors to monitor the site of the operation has already begun. These include direct sensing targets as illustrated in Table 1.2, as well as surrogate signs that can be used to indicate the onset of potential post-surgical complications. The *ear-worn activity recognition* (e-AR) sensor developed by Imperial College London and Sensixa, for example, has been used to quantify

Table 1.2 Disease processes and the parameters commonly used to monitor these diseases. Suggested sensor types for measurement of these parameters are listed in brackets. All of these conditions currently place a heavy administrative and financial burden on healthcare systems, which may be reduced if they are reliably detected

Disease process	Physiological parameter (BSN sensor type)	Biochemical parameter (BSN sensor type)
Hypertension	Blood pressure (*implantable/ wearable mechanoreceptor*)	Adrenocorticosteroids (*implantable biosensor*)
Ischaemic heart disease	*Electrocardiogram* (ECG), cardiac output (*implantable/ wearable ECG sensor*)	Troponin, creatine kinase (*implantable biosensor*)
Cardiac arrhythmias/ heart failure	Heart rate, blood pressure, ECG, cardiac output (*implantable/ wearable mechanoreceptor and ECG sensor*)	Troponin, creatine kinase (*implantable biosensor*)
Cancer (breast, prostate, lung, colon)	Weight loss (body fat sensor) (*implantable/ wearable mechanoreceptor*)	Tumour markers, blood detection (urine, faces, sputum), nutritional albumin (*implantable biosensors*)
Asthma/COPD	Respiration, peak expiratory flow, oxygen saturation (*implantable/ wearable mechanoreceptor*)	Oxygen partial pressure (*implantable/wearable optical sensor, implantable biosensor*)
Parkinson's disease	Gait, tremor, muscle tone, activity (*wearable EEG, accelerometer, gyroscope*)	Brain dopamine level (*implantable biosensor*)
Alzheimer's disease	Activity, memory, orientation, cognition (*wearable accelerometer, gyroscope*)	Amyloid deposits (brain) (*implantable biosensor/EEG*)
Stroke	Gait, muscle tone, activity, impaired speech, memory (*wearable EEG, accelerometer, gyroscope*)	
Diabetes	Visual impairment, sensory disturbance (*wearable accelerometer, gyroscope*)	Blood glucose, glycated haemoglobin (HbA1c) (*implantable biosensor*)
Rheumatoid arthritis	Joint stiffness, reduced function, temperature (*wearable accelerometer, gyroscope, thermistor*)	Rheumatoid factor, inflammatory and autoimmune markers (*implantable biosensor*)
Renal failure	Urine output (*implantable bladder pressure/volume sensor*)	Urea, creatinine, potassium (*implantable biosensor*)
Vascular disease (peripheral vascular and aneurisms)	Peripheral perfusion, blood pressure, aneurism sac pressure (*wearable/ implantable sensor*)	Haemoglobin level (*implantable biosensor*)
Infectious diseases	Body temperature (*wearable thermistor*)	Inflammatory markers, white cell count, pathogen metabolites (*implantable biosensor*)
Post-operative monitoring	Heart rate, blood pressure, ECG, oxygen saturation, temperature (*implantable /wearable and ECG sensor*)	Haemoglobin, blood glucose, monitoring the operative site (*implantable biosensor*)

post-operative home recovery through activity indices and physiology monitoring systems [28]. Similarly an implantable sensor has been used to monitor the pressure in the aneurysm sac following endovascular stenting [29]. The next step for any "hospital of the future" would be to adopt a ubiquitous and pervasive in-patient monitoring system enabling carers to predict, diagnose, and react to adverse events earlier than ever before. Furthermore, in order to improve the efficiency of hospital systems, the movements of patients through its wards, clinics, emergency departments and operating theatres may be tracked to try and understand where workflow is being disrupted and may be streamlined. This would help, for example, to maintain optimal capacity to cater for elective (planned) admissions whilst having the ability to admit patients with acute illnesses. Once a patient is discharged from hospital they are often asked to attend outpatient clinics to monitor their overall clinical progress. Due to the episodic and subjective nature of outpatient clinic review, it is often difficult to objectively determine how a patient is recovering from their diagnosis. The advent of implantable sensors through which information regarding the operative of disease site can be determined may represent a significant step in solving this problem. An example of this is the use of an implantable micro machined wireless sensor to monitor patients undergoing abdominal aortic aneurism surgery has already begun [29]. Here, one of the major abdominal arteries is found to be abnormally dilated, making it prone to rupture and death. A repair with a synthetic graft is required to prevent this, and during the procedure a sensor is placed at the site of surgery, thereby allowing the measurement of pressure around the aorta. In the clinic, the patient is asked to lie down and a radiofrequency reader is 'swiped' over their abdomen, allowing the clinician to measure the pressure in the sack, thereby evaluating the status of the aneurism repair. From examples such as this it can be envisaged that continuous and pervasive BSN patient monitoring has a clear role in providing accurate objective information about a patient's health in the outpatient setting.

1.2.3 Monitoring Elderly Patients

The second scenario where BSNs may prove invaluable is for the regular and non-intrusive monitoring of "at risk" population groups such as the elderly. With people in industrialised nations living longer than ever before and an increase in average life expectancy of more than 25 years, the size of this group is set to increase, along with its potential demand upon healthcare resources [30].

Identifying ways of monitoring this aging population in their home environment is therefore very important, with one key example of the usefulness of this approach being the vulnerable periods during months of non-temperate weather. There is evidence to suggest that at times of the year when weather conditions are at their extremes (either very cold or very hot), elderly patients are at increased risk of requiring hospital admission [31, 32]. They are at risk because they are not able to

seek medical help early enough for simple and treatable conditions, which eventually may lead to significant morbidity. An example of this is an elderly individual who lives alone and acquires a chest infection, which he fails to identify and seek help for until the infection requires hospital admission, or even ventilatory support. This could all be potentially avoided if the infection, or change in patient habits as a result of this infection, was picked up early and antibiotic therapy initiated.

Examples illustrating how people behave differently at the onset of illnesses include a decrease in appetite, a reduction in movement, and propensity to stay indoors. When correlated with physiological vital signs measurement, this system has the potential to clearly identify those most at risk. It also demonstrates an instance in which a WSN linking ambient sensors (set up in the patient's home) with a BSN (on the patient's body) can complement each other. Monitoring elderly patients in their home environment during non-temperate weather will therefore allow earlier detection of any deterioration in their condition, for which prompt treatment may reduce the need for hospital admission, associated morbidity and even mortality.

The concept of an unobtrusive "home sensor network" to monitor an elderly person's social health (giving feedback not only to that person's carers and family members, but also to the elderly individual themselves) is one that is being developed by several companies such as Intel [33]. Whilst such a sensor network attempts to monitor well-being by identifying the individual and the level of activity they are undertaking, it is easy to see how this network could communicate with a body sensor network relaying physiological data about the individual. Combining these two networks would allow for a much better appreciation of the context in which the sensing is taking place.

1.2.4 Life Style and Wellbeing

Thus far, multiple sensing options exist for both monitoring lifestyle behaviours and for intervening when necessary. Physical activity is an excellent example, where the vast number of devices available reflects the abundance of evidence associating inactivity with poor health [34]. Activity monitors have progressed from simple uni-axial pedometers not only through technological advances, but also through user-orientated design. Simple sensors disguised in wristwatches are one popular option for activity monitoring. Devices such as the Nike+ FuelBand, Sensixa e-AR, Fitbit Flex, Jawbone UP, Basis B1 and BodyMedia FIT LINK all allow pervasive monitoring of daily activity, with real-time user feedback, usually through as simple LED display. The value of the sensor is improved through user-friendly software that allows personalized activity goals to be set, and progress towards those goals to be displayed throughout the day or reviewed later. Such sensors have been shown to improve quality of life as much as costly supervised exercise programmes [35]. Similar sensors have also been used for weight

management purposes, with several large-scale randomised clinical trials underway to determine their value [36, 37]. Devices and applications which exploit aspects of human psychology, including incentives to reach goals, or allowing comparisons of activity levels to that of their peer group (norming) are likely to affect the activity-behaviour [38].

Quality of sleep is an often under-looked factor associated with quality of life and wellbeing [39, 40]. Sleep disturbance can start a vicious cycle in which poor sleep results in stress and anxiety, which results in further sleep disturbance, and so on [41]. The conventional 'sleep study' involves a patient visiting a specialised facility for polysomnography, where multiple aspects of sleep quality are measured. This includes brain activity, eye movements, body movement and cardiorespiratory activity. The requirement of specialist facilities, and putting the patient in an unusual environment (usually for only one night), often restricted by wires and bulky devices, limits the feasibility and accessibility of sleep studies. Pervasive BSN options exist, facilitating reliable home-based sleep assessment and assistance. The SenseWear Armband multi-sensor activity monitor (SWA), utilises several sensor inputs to provide valid measurements for sleep duration and efficiency [42]. The SWA combines a two-axis accelerometer with heat flux, skin temperature and galvanic skin response, relationships within which have been shown to correlate with onset of sleep and quiet wakefulness [43]. Smartphones have also been implemented in sleep state detection. Applications utilise either the phone's microphone or accelerometer (placed under the bed sheets) to estimate the sleep phase from the user's movements. Not only does this allow sleep cycle analysis, but it also attempts to improve the waking process by triggering the alarm during the most suitable sleep phases before a user-defined deadline [44].

1.3 The Need for Pervasive Health Monitoring

In most healthcare systems, spiralling costs, inadequate staffing, medical errors, and an inability to get to hospital in time in rural areas are placing a tremendous burden on the provision of care. The concept of "ubiquitous" and "pervasive" human wellbeing monitoring with regards to physical, physiological, and biochemical parameters in any environment and without restriction of activity [45, 46] has now become a reality with the important advances in sensor, miniaturised processor, and wireless data transmission technologies described earlier [47, 48]. Recently, there has been a huge surge in the number of external (wearable) activity tracker devices currently in use to monitor activity levels ranging from wrist-bands to wearable devices to complement clothing in a range of environments. Incorporation of these technologies into mainstream community healthcare programmes for treatment of obesity, ischaemic heart disease, diabetes mellitus, and the promotion of wellbeing is extremely attractive as they off the patient direct feedback on a daily

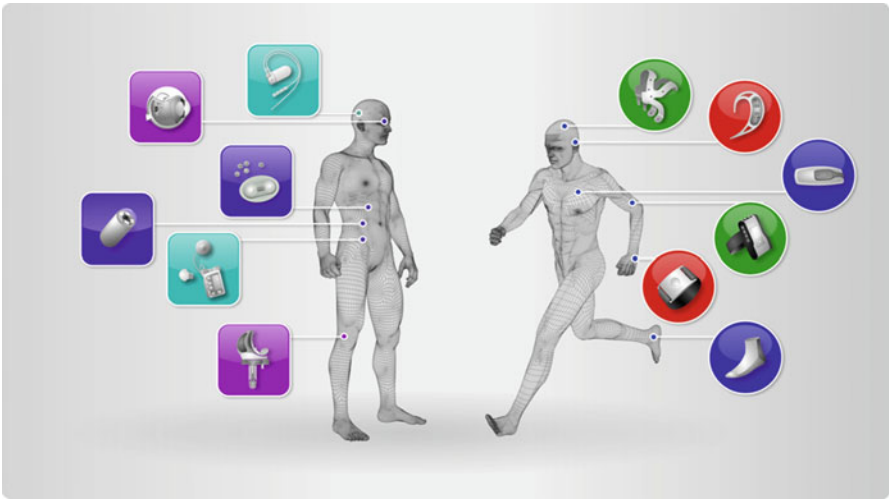


Fig. 1.3 The range of wearable and implantable sensors (ranging from neuro-stimulator to smart knee prosthesis) already in use or being developed for wellbeing and patient monitoring, as discussed in this chapter

basis, and an opportunity for self-management. The result is not only the potential to impact the healthcare of large populations in a way that traditional monitoring cannot, but as the technology becomes more available, an opportunity to do this inexpensively. This is the area of pervasive healthcare that is likely to be the first to enter mainstream clinical practice on a large scale. Figure 1.3 outlines some of the aforementioned technologies in a BSN setting.

As illustrated in Fig. 1.3, another area that has shown promise and generated interest is that of implantable devices and biosensors. Advances in key areas such as power supply miniaturisation, increased battery duration, reduced energy consumption, and power scavenging are essential to such systems [49]. *Micro Electro-Mechanical System* (MEMS) technology is an area which has offered the prospect of sophisticated sensing using a miniaturised sensor device [50]. Pervasive healthcare systems utilising large scale BSN and WSN technology will allow access to accurate medical information at any time and place, ultimately improving the quality of the service provided.

With these advances taking place, clinicians are for the first time able to explore the prospect of not only monitoring patients more closely, but also to do this in an environment where they have never been able to monitor patients before. The chronic conditions of diabetes mellitus and hypertension mentioned previously in this chapter are currently managed on the basis of a series of "snapshots" of information, obtained in a clinical setting which is artificial in comparison to the patient's normal environment.

The long-term management of these conditions would clearly benefit from any technology that could result in a more tailored treatment being offered to the patient. The treatment of atrial fibrillation, which is associated with episodic rather than continuous circulatory abnormalities such as blood pressure surges, paroxysmal arrhythmias or episodes of myocardial ischaemia best illustrates this, as at present much time is wasted in trying to "capture an episode" of these abnormalities. As all of these episodes could be picked up using basic vital signs monitoring but their timing cannot be predicted, a wireless, pervasive, and continuous monitoring system is ideally suited to diagnosing and monitoring the progress of these diseases.

Additionally, better and earlier detection is likely to result in earlier administration of the appropriate treatment, and the prevention of disease-related morbidity. In the more acute hospital setting, the ability to continuously capture data on patient well-being has the potential to facilitate earlier adverse event detection and ultimately treatment. In such a system, the patient would not be required to stay at their bedside for monitoring to take place, increasing their mobility and return to activity in the hospital. In addition to this, a patient's physiological data would be obtained either continuously or at shorter time intervals, picking up said deterioration more quickly. At present, much of the data captured even with the aid of continuous patient monitoring is lost, but in conjunction with an automated pervasive monitoring system all data gathered could be stored for later review and trend analysis.

Historically pervasive monitoring platforms have adopted a number of strategies to deploy external sensors. These may be wearable in clothing, either through integration with a textile platform [51], or by embedding into clothes with integrated electronics resulting in "intelligent biomedical clothes" [52]. The "MIThril" project based at Massachusetts Institute of Technology Media Lab developed a body-worn sensing computation and networking system, integrated in a "tunic" and was one of the first of its kind [52]. Proposed wearable applications include "memory glasses" that aim to provide a context-aware memory aid to the wearer. The highly anticipated launch of Google's 'Glass' concept into a product for mainstream use, shows just how rapidly this is becoming a reality.

All these strategies share a common aim in providing unobtrusive, pervasive human monitoring irrespective of geographical location. In the case of external sensors, whilst embedding these into a garment does provide a more convenient wearable system, it lacks flexibility for the addition and relocation of sensors as dictated by patient size and shape. With implantable sensors, wiring is impractical and to a large extent can limit sensor placement. A wireless platform specifically designed to network external and implantable sensors on the human body is desirable not only because it allows these various sensors to be networked in a less bulky and intrusive way, but also because it allows the potential to add and remove sensors as required. Wireless sensor networking, data acquisition, data capture, and low power transmission all offer the prospect of flexible body sensor networks that are truly pervasive.

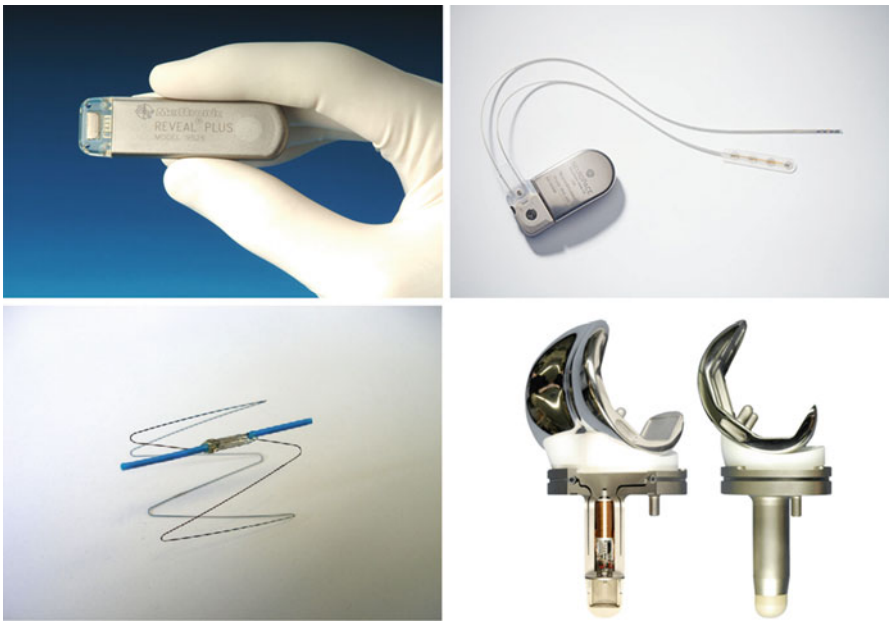


Fig. 1.4 Example implantable sensors in clinical use/development. (*Top Left*) Medtronic "Reveal Insertable Loop Recorder" (*Image courtesy of Medtronic Inc.*). (*Top Right*) NeuroStimulator (RNS) from NeuroPace (*Image courtesy of NeuroPace Inc, Mountain View, CA*). (*Bottom Left*) Implantable cardiovascular pressure sensor, wirelessly powered and read based on a radio-frequency resonator (*Image courtesy Chris McLeod, Imperial College London*). (*Bottom Right*) Smart knee prosthesis with strain gauge-instrumented tibial component for the measurement of forces (*Image courtesy Georg Bergmann and Friedmar Graichen of Julius Wolff Institut für Biomechanik und Muskuloskeletale Regeneration, Germany*)

The development of implantable sensors offers BSN one of its most exciting components. The Bravo pH system, for example, comprises of a radio-telemetric capsule, which is placed by passing and endoscope through the mouth and deployed at the junction between the stomach and oesophagus. It is designed to diagnose *Gastro-Oesophageal Reflux Disease* (GORD), which affects millions of patients worldwide. The current gold standard for pH monitoring involves using a nasogastric catheter-mounted pH electrode passed transnasally into the oesophagus [53] which is physically restricting and socially obtrusive. The implantable pH sensor avoids these problems and the mechanism of deployment involves firing a spring-loaded pin through a small section of mucosa sucked into a trough on the sensor housing. This is mostly frequently performed under direct vision during upper gastro-intestinal endoscopy or nasogastric manometric guidance [54].

Other implantable devices include Medtronic's "Reveal Insertable Loop Recorder" as shown in Fig. 1.4 (top left), which is a fully implantable cardiac

monitor used to record the heart's rate and rhythm during instances of unexplained fainting, dizziness, or palpitations. The device provides the clinician with an ECG that can be used to identify or rule out an abnormal heart rhythm as the cause of these symptoms [55]. CardioMEMS is a company that produces an implantable pressure sensor, which has been developed at Georgia Institute of Technology, that can take pressure readings following implantation into an aneurism sac at the time of endovascular repair [29]. This implanted sensor then provides a means of monitoring the status of the repair during the years following. For medically resistant epilepsy, *NeuroStimulator* (RNS) from NeuroPace (NeuroPace, Mountain View, CA) is offering a solution [56]. The device as shown in Fig. 1.4 (top right) consists of a neurostimulator, connected to subdural leads placed in the brain, which have been identified as causing seizures (seizure foci). The device monitors electrocorticographic activity, and if abnormal, can stimulate the area to stop the seizure.

For continuous in vivo blood pressure monitoring for patients suffering from heart failure and pulmonary hypertension, the team at Imperial College led by Chris McLeod have developed a fully implantable wireless sensor able to provide continuous, real-time pressure measurements based on *Surface Acoustic Wave* (SAW) technology. It is used to deposit resonators on crystalline quartz wafers, which are then assembled to produce a pressure sensitive device. The device as shown in Fig. 1.4 (bottom left) is wirelessly powered and read, thus forgoing the need of batteries.

Finally Bergmann et al. have designed and tested a total knee replacement prosthesis with strain gauge-instrumented tibial component for the measurement of forces. Over 80,000 knee replacements occur in the UK every year, and the prevalence is increasing [57]. Although the procedure is considered to be a successful and durable surgery, multiple revisions occur, forming 8 % of all knee arthroplasties performed [58]. The smart implant as shown in Fig. 1.4 (bottom right) provides a novel insight into the mechanisms of prosthesis failure, as well as human knee kinetics and pathology. As demonstrated in a small group of patients, the prosthesis provided force measurements throughout the gait cycle, and highlighted how the parts of the prosthesis under the most load.

1.4 Technical Challenges Facing BSN

Although the BSN platform aims to provide the ideal wireless setting for the networking of human body sensors and the setting up of pervasive health monitoring systems, there are a number of technical challenges that lie ahead. These include the need for better sensor design, MEMS integration, biocompatibility, power source miniaturisation, low power wireless transmission, context awareness,

secure data transfer, and integration with therapeutic systems, each of which are mentioned briefly below, and covered in more detail throughout this book.

1.4.1 Improved Sensor Design

Advances in biological, chemical, electrical, and mechanical sensor technologies have led to a host of new sensors becoming available for wearable and implantable use. Although the scope of these sensors is very wide, the following examples highlight the potential they offer to pervasive patient monitoring. In the case of patients with diabetes mellitus, trials of implantable glucose sensors are underway in an attempt to rid this patient population of the need for regular invasive blood glucose pinprick testing [59]. In addition, the ability to determine tissue and blood glucose levels using an implantable wireless glucose sensor may also form the sensing part of a "closed feedback loop" system. The other half of this loop would consist of a drug delivery pump which would continuously infuse a variable amount of fast-acting insulin based upon the patient's glucose level [60]. This concept effectively results in the closed feedback loop system acting as an "artificial pancreas", which maintains blood glucose within a closely defined reference range. As a result, diabetics may be able to avoid not only the complications of an acutely uncontrolled blood sugar (hypo- or hyper-glycaemia), but also much of the end-organ damage associated with the condition (retinopathy, cardiac, renal, and peripheral vascular disease).

Reliability is a very important requirement for sensors in closed feedback loop systems such as this, because they ultimately guide treatment delivery. It may be therefore that in this case an implantable biosensor array offers a more accurate result than one isolated sensor [61].

Improvements in sensor manufacturing and nano-engineering techniques, along with parallel advances in MEMS technology offer the potential for producing even smaller implantable and attachable sensors than were previously possible. An example of one such miniaturised nano-engineered sensor currently is a fluorescent hydrogel alginate microsphere optical glucose sensor [62]. Physiological sensors benefiting from MEMS technology integration include the microneedle array and the implantable blood pressure sensor [63]. Although much of this technology is still experimental, it is not inconceivable that over the next decade these sensors will guide therapy for chronic conditions such as hypertension and congestive cardiac failure. MEMS devices in particular may prove pivotal in the drug delivery component of any closed feedback loop [64]. In addition, when mass-produced, MEMS technology offers the prospect of delivering efficient and precise sensors for no more than a few dollars. This is well illustrated in the case of accelerometers, which have been used by the automobile industry to efficiently and reliably trigger car airbag releases during simulated accidents.

1.4.2 MEMS and BioMEMS

Although mentioned several times the role of MEMS in integrated sensor system, it is useful to discuss its role for both physical and biochemical sensing. The general term of MEMS refers to a wide range of microfabricated non-integrated circuit devices. These include but not limited to physical and biochemical sensors as well as systems that offers complete assays from sample preparation. The original MEMs evolved from investigations in the 1950s on harnessing the technology behind silicon integrated circuit fabrication such as photolithography and ion implantation, to create miniaturised, mass-producible low unit cost sensors [65]. Microfabrication processes specific to MEMS that are not used in conventional IC manufacturing are also developed [66]. Amongst these are electroplating, spin casting, soft lithography, metal sputtering and vapour deposition of polymers. Some of these facilitate MEMS devices fabrication with non-silicon substrates such as SU-8 photoresist and *Polydimethylsiloxane* (PDMS) and other biocompatible polymers. This eventually led to a wide range of miniaturised and robust MEMS sensor chips for various physical and biochemical substances. Common microfabricated MEMS building blocks include beams, cantilevers, bearing and hinges, thin film diaphragms and modified substrates to form deep wells/pits, channels and vias. Metal patterning on these micro structures allows signal transduction for readout and actuation of the MEMS devices. The integration of MEMS device sensors with readout electronics allows further system miniaturisation as well as introducing benefits such as high SNR [67]. The MEMS manufacturing process however has evolved away from IC processing steps such that the integration of the two in the same foundry will incur additional costs for process control and modifications as well as carry extra risks of lower yield compared to those of independent IC or MEMS processes. The assembly of MEMS device and IC with system-in-package solutions creates other tradeoffs such as tracking parasitics which impedes performance.

Common physical sensors realised with MEMS technology include strain gauges, pressure sensors, and inertia sensors such as accelerometers and gyros [68–70]. Pressure sensors are commonly constructed with a thin film diaphragm (often made of silicon) attached to sensors sensitive to the diaphragm deformation. Piezoresistive sensors rely on the fact that silicon can alter its resistivity upon the application of mechanical stress. When piezoresistive sensors are attached to a thin film diaphragm, any deformation of the diaphragm due to a change in pressure results in a change in resistance of the corresponding piezoresistive elements. This change in resistance is subsequently detected by readout electronics [71]. In the most basic form, inertia sensors such as accelerometers and gyros are constructed with a proof mass attached to the end of a micro cantilever. The combination of designer specified geometry and process dependent physical parameters of the construction determine the behaviour and performance of the inertia sensor. The mechanical part of the sensor is often hermetically sealed to prevent contamination. Signal transduction can be carried out via piezoresistive or capacitive (charge) sensing [72]. For capacitive sensing, one or more electrodes on the proof mass act as one of the

plates of a capacitor whereas the corresponding conductive patterns on a fixed frame of reference act as the other plate. Displacement of the proof mass alters the distance between the two capacitor plates and hence creates a change in capacitance. In practice the change in capacitance is detected by a change in stored charge while a known voltage is applied across the two plates. Complex geometries of multiple proof masses and capacitive transducer constructs allow the realisation of vibratory rate gyroscopes. The gyros operate according to the principle of Coriolis acceleration sensing to recover angular velocity information. To accomplish this, the proof mass is set into vibration along one axis with the reverse process of capacitive sensing. Rotation of the proof mass due to an external moment is translated into displacement orthogonal to the vibration axis. This displacement is sensed capacitively and converted to angular velocity in readout. Besides piezoresistive and capacitive readout, piezoelectric *Surface Acoustic Wave* (SAW) devices are also used to transduce the sensor signal [73]. Physical qualities of the SAW such as amplitude, phase, and frequency are modulated by the sensing element through signal dependent changes in mass and sensor finger spacing.

The term BioMEMS broadly covers a range of microfabricated devices geared towards biomedical applications such as drug discovery and tissue engineering [74]. Non-electrical/mechanical systems such as DNA/protein microarrays for high throughput assays also fall in the category of BioMEMS. Integrated systems such as *lab-on-chip and micro-total analysis systems* (LoC/uTAS) [75, 76] carry out biological assays from sample preparation to readout by utilising microfluidic constructs and biochemical sensors. BioMEMS fabrication on glass, ceramic or polymer substrates follow similar work flow to their silicon counter parts but with processing steps such as soft lithography and plasma polymerisation to cater for non-silicon material processing [76]. In the context of sensing, species are targeted by combining a biological probe with microfabricated transducers for electrical, mechanical, or optical readout. Target species include proteins, cells, DNAs and chemical markers with corresponding probes such as antibodies, extracellular matrix proteins, cells, oligonucleotides and enzymes (e.g. see [77–79]). The probes are typically attached to the MEMS transducers via processes such as covalent immobilisation. This improves specificity of surface absorption (fouling reduction) and minimise denaturation of the bio/molecular probes. In the case of binding proteins and especially antibodies, an appropriate orientation and confirmation of the proteins attached to the sensor surfaces are ensured.

Different substrates can be functionalised with specific processes for the covalent attachment of bio-probes [79]. For example, plasma treatments are routinely used to mediate the formation of hydroxyl, amine and Carboxyl groups on substrates such as PDMS, *Polymethylmethacrylate* (PMMA), and SU8. Gold surfaces can be functionalised through *Self-Assembled Monolayers* (SAMs), where the high affinity of sulphur towards gold allows alkanethiol molecules with interface groups of interests to be immobilised [80]. Antibodies are commonly formed as SAMs on sensor surfaces. *Layer-by-Layer* (LBL) is a process that takes advantage of electrostatic attractions between oppositely charged polyelectrolytes for sensor surface functionalization. Surface oxidation resulting in charged surfaces on metals

or glass can be used as a substrate for LBL deposition of charged molecules such as some proteins and RNA. Multilayers of deposition of a chain of molecules of alternating net charge polarity is also possible with LBL.

For sensor signal transduction and readout, electrochemical detection with enzyme-based biosensors are perhaps the most mature with a wide range of work published in literature and commercially available sensors [81]. For example, Glucose sensors are usually formed by the immobilisation of glucose oxidase on platinum or carbon electrodes with readout via amperometric amplifiers. Field effect transistors with their gate oxide exposed to solution are used to measure ion concentrations as gate surface charges are varied according to solution concentration [82]. These transistors are known as *Ion-Sensitive Field Effect Transistors* (ISFETs). Enzyme monolayers can be used to further functionalise gates of ISFETS, extending the range of sensing targets. Readouts for ISFETs rely on potentiometric electronics. Advanced electrochemical methods for surface characterisation such as cyclic voltammetry and chronopotentiometry have also been used to detect specific target binding of immobilised nucleotides on gold surfaces. Apart from mechanical and electrical readout, optical readout is also demanded by a number of bioMEMS devices. Wavelengths used span from ultraviolet to infrared and bio-probes are often attached to optical fibres to aid measurements. Species absorption onto gold/silver as well as change in refractive index can be measured through *Surface Plasmon Resonance* (SPR) [83]. Phase shift caused by the targeted specimen can be detected with interferometry. Colorimetric readout is possible for targeted species with reaction products with specific absorption/emission spectra. Data readout of immobilised microbial biosensors are carried out via measuring microbial respiration products upon exposure to targeted species. Optical and electrochemical measurements are common in these cases [84].

1.4.3 Biocompatibility, Integratability and Resorbability

Implantable sensors and stimulators have had to overcome the problems of long-term stability and biocompatibility, with perhaps one of the most successful examples of this being the cardiac pacemaker and the *Implantable Cardioverter-Defibrillator* (ICD) [85]. The scale of implantable ICD use is best demonstrated by the fact that in 2001, a total of 26,151 were implanted at 171 centres in the UK and Ireland [86]. One of the main indications for an ICD is sudden cardiac death, which affects approximately 100,000 people annually in the UK, demonstrating the size of the patient population that may benefit from this device [87]. Other implantable devices currently used in clinical practice include implantable drug delivery systems for chronic pain [88], sacral nerve stimulators for anal incontinence [89], and high frequency brain (thalamic) stimulation for neurological conditions such as Parkinson's disease [90] and refractory epilepsy [91]. Figure 1.5 illustrates some of the example implantable devices for sacral nerve stimulation, drug delivery for chronic pain management, and deep brain stimulation.

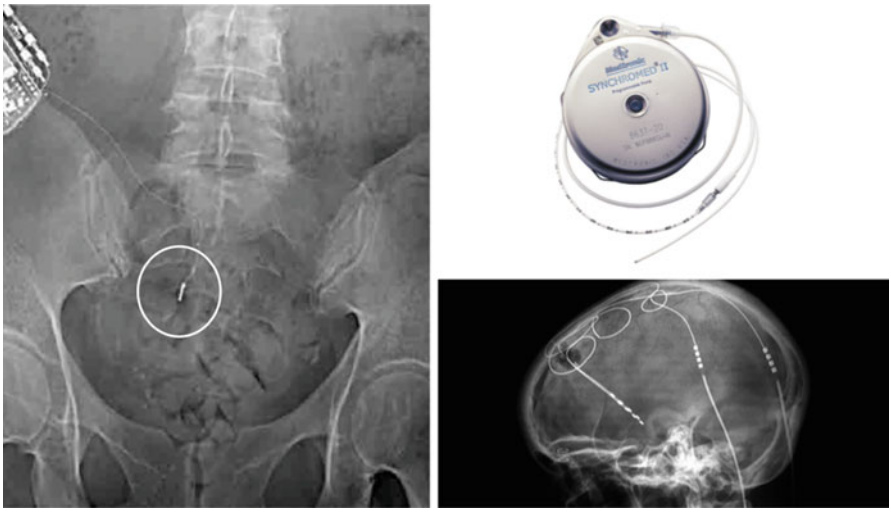


Fig. 1.5 (*Left*) An example of an implantable device (sacral nerve stimulator) in current clinical use. (*Top Right*) Implantable Synchromed® II drug delivery system for chronic pain (*Courtesy of Medtronic Inc.*). (*Bottom Right*) Deep brain stimulators (*Courtesy of the Radiological Society of North America*)

The fact that large groups of the patients already carry implanted devices such as those mentioned above, means that many of the lessons learnt from their use can be extended to any proposed implantable biosensor research. In addition to this, the integration of these already implanted sensors and effectors into a larger wireless BSN is something that deserves consideration. With regards to this last statement, power consumption is obviously an important issue, and until this is addressed it is unlikely that, for example, pacemakers would be used to monitor the cardiovascular status as part of a body sensor network. Security and interference of these devices with each other, as well as with day-to-day technologies used by patients such as mobile phones, is a concern that has been noted and must be addressed [92]. The concern here is that interference may result in not only sensor malfunction, but also might affect implanted drug delivery systems and stimulators. This supports the call for a new industrial standard for the wireless transmission frequency used in BSNs.

In his recent review on advances in implantable sensors, Robert Black [93] examined the translational pathways of different implants ranging from glucose sensing, cardiac pressure, neural prosthetics, to devices for monitoring radiation therapy. It outlines the complex environment that these sensors need to operate, as well as the regulatory and reimbursement hurdles that must be overcome. The paper provides both success stories, as well as failures in this challenging, yet increasingly important research and clinical translational space.

For implants, there has also been significant attention to the use of biodegradable materials. The use of these transient implants has many advantages, particularly for

the avoidance of addressing their long-term effect on the body. The challenge here is how the device is absorbed, fragmented and secreted safely. In fact in surgery, particularly orthopaedic surgery, there is a large variety of biodegradable implants including sutures, staples, anchors and interference screws. Biodegradable implants with different polymeric raw materials can have different material properties and tissue response. In orthopaedic implants, these materials include Polyglycolide (PGA) and copolymers such as poly-glycolide-co-trimethylene carbonate; Poly-(L-lactide) (PLLA), poly-(DL-lactide) (PDLLA) and their stereocopolymers; Polydioxanone (PDS) [94]. The reason for the use of biodegradable materials for orthopaedic surgery is that a secondary operation for removal is not necessary. This is in contrast to metallic implants, which need to be removed due to the long-term issue of osteopenia, corrosion and irritation of adjacent tissues. Furthermore, as biodegradable implants degrade, they lose strength and this puts pressure to bone, strengthening it and therefore preventing bone resorption [95]. However, certain long-term effects still need further investigation as clinical results have shown remnants of high molecular PLLA implants could be found several years after implantation [94].

Most of the existing biodegradable devices, either used clinically or in laboratory development, are passive and without the sensing or actuation capabilities. To address this limitation, it is necessary to develop bio-fragmentable or resorbable electronics to fulfil the transient sensing capabilities. Recent work by John Rogers and team have demonstrated a physically transient form of silicon electronics for this purpose [96]. They have experimented with a set of materials and manufacturing schemes. Figure 1.6 shows an example of using magnesium oxide for the dielectrics, monocrystalline silicon nanomembranes for the semiconductors and silk for the substrate and packaging material. The fabrication process involves a combination of transfer printing of the monocrystalline silicon nanomembranes, physical vapour deposition and solution-casting for the silk substrate. All components disintegrate and dissolve when immersed in deionised water [96].

Recently, they have further demonstrated bioresorbable radio frequency electronics [97] as shown in Fig. 1.7, in which it illustrates a wireless powering system with a magnesium receiving antenna, full wave rectifying circuit for powering up a standard LED. It also shows how the circuit dissolve in deionised water over time. Although these are only demonstrated in laboratory settings and the long-term in vivo effects need to be carefully assessed, particularly in terms of toxicity and inflammatory reaction, these developments opens up exciting new opportunities for the development of a new generation of smart transient implants that can potentially reshape the future of surgery.

1.4.4 Energy Supply and Demand

One of the key considerations for BSNs is power consumption. This is because power consumption determines not only the size of the battery required but also the length of time that the sensors can be left in situ. The size of battery used to store

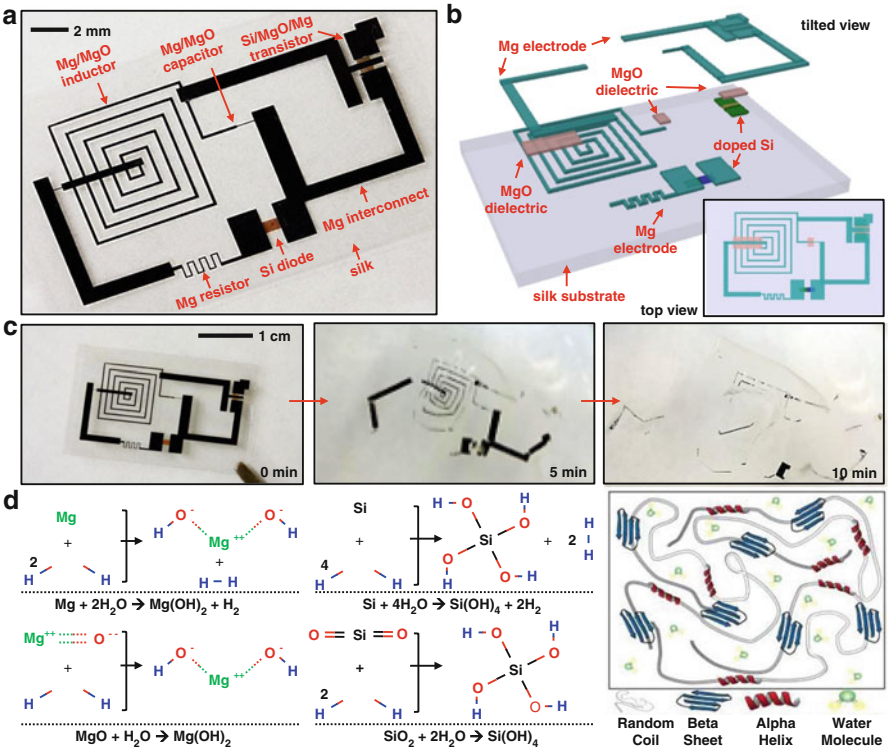


Fig. 1.6 Transient silicon electronics developed by John Rogers and his group [96], showing (**a, b**) the device and its schematic component overview, (**c**) time sequence of dissolution in deionised water, and (**d**) chemical reactions for its constituent materials (*Courtesy of Dr John Rogers, University of Illinois at Urbana-Champaign with permission from publisher*)

the required energy is in most cases the largest single contributor to the size of the sensor apparatus in term of both dimensions and weight. These factors are important not only in the implantable but also the external sensor settings because they determine how "hidden" and "pervasive" the sensors are. Ultra wideband radio has been suggested as a mode of short-range wireless communication with relatively high data rates and low power consumption [98]. Self-configuring networks carry the advantage of reducing energy consumption from unnecessary nodes classified as "redundant", thereby increasing the system's lifespan. Whilst the use of small, high-density batteries for both wearable and implantable sensors remains the most convenient solution, methods that can 'harvest' the required power have been an active research topic.

To this end, reducing battery consumption through the increased use of power scavenging from on-body sources such as vibration and temperature is a strategy being developed to enhance battery life; especially in the case of implantable

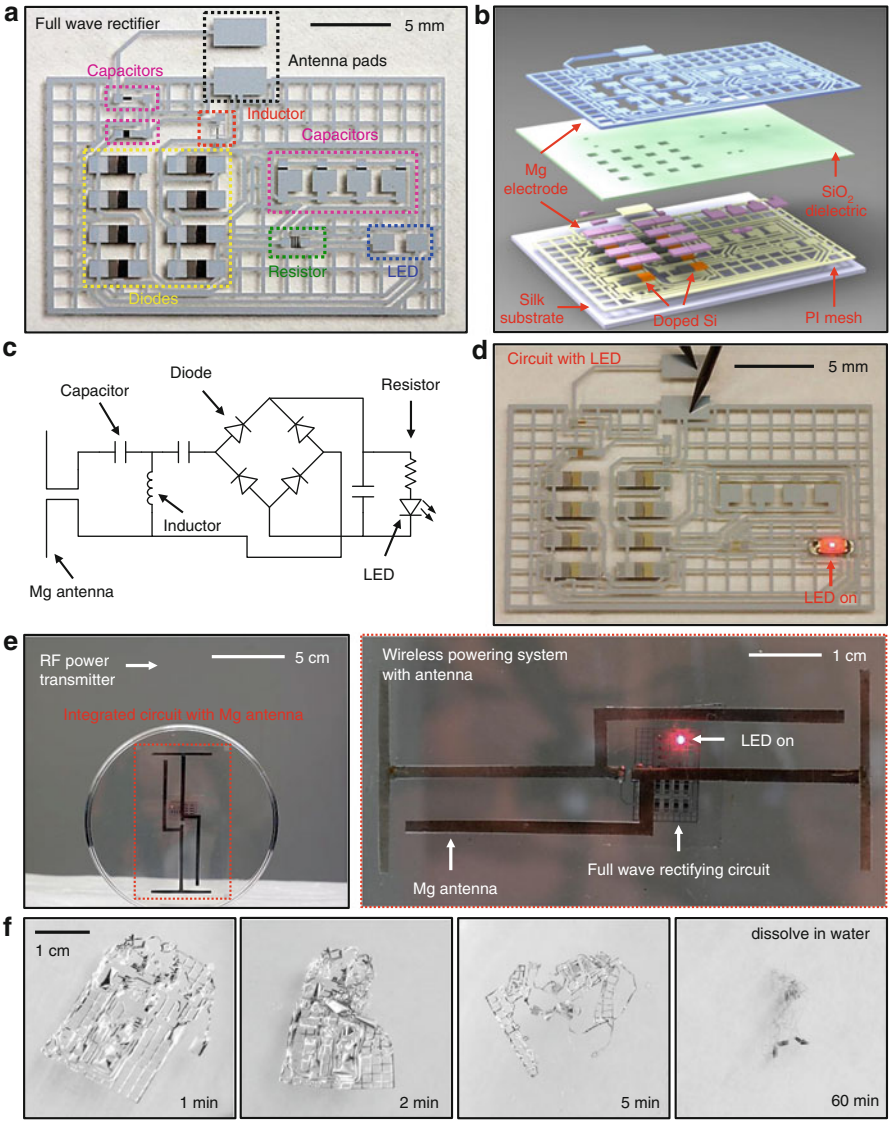


Fig. 1.7 Transient RF power scavenging circuits developed by John Rogers and his group, showing (**a**) the transient full-wave rectifying circuit, (**b**) a schematic overview of the device, (**c**) the circuit diagram, (**d**) operational demonstration with a commercial LED, (**e**) powered with an RF transmitter and a Mg receiving antenna (with a magnified view on the right), and (**f**) a time sequence of dissolution in deionised water after 1–60 min (*Courtesy of Dr John Rogers, University of Illinois at Urbana-Champaign with permission from publisher* [97])

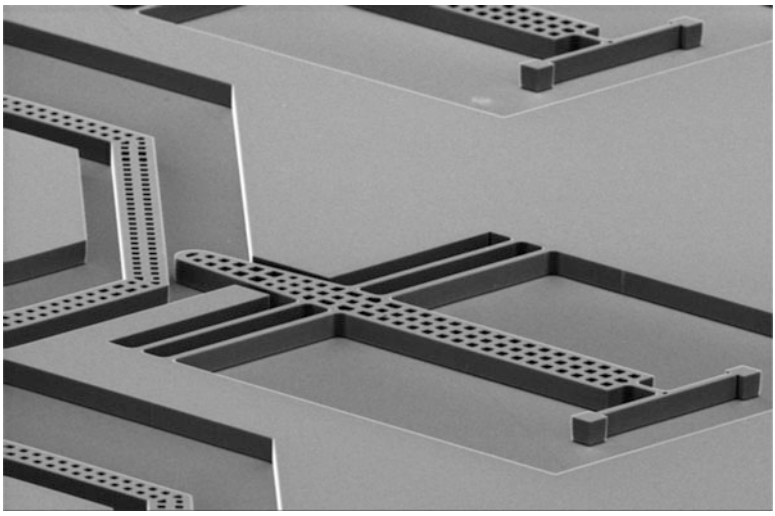


Fig. 1.8 An ultrasonic power transfer MEMS device with stepwise microactuators (*Courtesy of Prof Eric Yeatman, Imperial College London, UK*)

sensors [99, 100]. At present, a large proportion of the electrical power consumed by biosensors goes towards the measurement circuit. In a wireless BSN system, the wireless communication link is likely to be the greatest consumer of power. The development of low-power wireless data paths is therefore a key to the successful development of wireless BSN systems. Reducing the power consumption of the radio transceiver is crucial to the practical deployment of BSNs [101, 102].

Advances in miniature devices for biomedical applications are creating ever-increasing requirements for their continuous, long lasting and reliable energy supply, particularly for implanted devices. Hitherto, the use of inductive coupling remains a popular technique for implants. However, many existing biomedical implants operate in the low-MHz range, necessitating large antennas and limiting data transmission rates. Recent work at Stanford has shown millimetre-sized antennas in vivo operating with optimal power transfer efficiency in the low-GHz range, thus simplifying the design of data link as well as the antenna and matching network [103, 104].

As to be discussed in Chap. 6, potential sources of power for energy harvesting include solar, vibration, heat or pressure gradient, air or fluid flow, chemical/radioactive reactions, as well as a wide variety of biological sources. For implants, for example, it is possible to harvest power from low-frequency body motion. Karami and Inman have shown the potential of powering pacemakers from heat-beat vibrations using linear and nonlinear energy harvesters [105]. These are designed according to the characteristics of heartbeats and robust to heart rate variations. Yeatman et al have also demonstrated a number of low-frequency body motion power harvesters. More recently, they have further demonstrated the use of ultrasonic transfer for powering stepwise microactuators [106] as shown in Fig. 1.8.

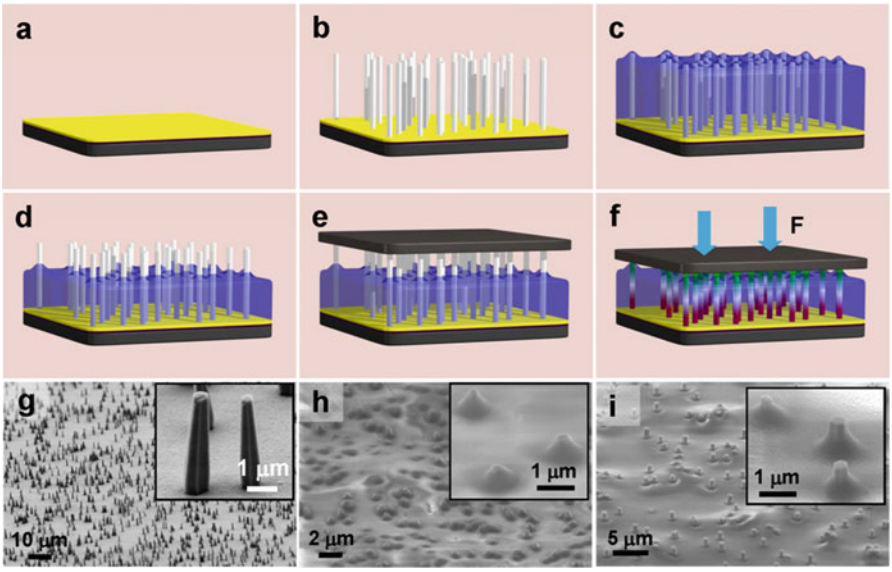


Fig. 1.9 Fabrication steps of vertical nano-wire array integrated nanogenerator (**a**) gold-coated silicon wafer, (**b**) ZnO nanowire arrays grown by low-temperature hydrothermal decomposition, (**c**) a layer of polymethyl-methacrylate by spin coating, (**d**) exposed tips of the nanowires after oxygen plasma etching, (**e**) placement of a platinum-coated flat electrode to form Schottky contact, (**f**) uniaxial stress applied at the top of the electrode and the nanowires compressed and a macroscopic piezoelectric potential generated, (**g**) scanning electron microscopy images of the ZnO nanowire array, (**h**) after spin-coating with polymethyl-methacrylate, and (**i**) after oxygen plasma etching (*Courtesy of Dr Zhong Lin Wang, Georgia Institute of Technology, Atlanta, Georgia, USA with permission from publisher* [107])

The unique feature is that ultrasound is used to power actuation directly, rather than via piezoelectric conversion, for applications such as drug release or mechanical adjustment of prosthetic devices. A coupled mechanical system remotely excited by ultrasound at 200 kHz provides conversion of acoustic energy into motion of a MEMS silicon mechanism, using a receiving membrane coupled to a discrete oscillator. The oscillator motion in turn is converted to continuous motion of the actuator. Frequency multiplexing can be used to allow reversible motion, or the use of multiple actuators in one system.

Another area in power harvesting is in the use of nano-materials for self-powered nano-systems, owing to the pioneering work of Zhong Lin Wang and his team. An example of their work is shown in Fig. 1.9, demonstrate the integration of ZnO nanowires into arrays for power generation [107]. The so-called vertical nano-wire array integrated nanogenerator (VING) can generate output voltage up to 0.243 V and sufficient for powering real low-power devices.

Several other strategies have also been employed for powering implantable or ingestible sensors. For example, the Proteus Smart Pill [108] (Proteus Digital Health, Inc.) as shown in Fig. 1.10 is a novel technological solution to managing

Fig. 1.10 Proteus Smart Pill integrated with sensors harvest power from the gastric fluid (*Courtesy of Proteus Digital Health, Inc. Redwood City, California, USA*)

non-compliance/poor adherence with/to medication, which works by detecting when a patient takes a pill. The system comprises of a cutaneous patch sensor, which wirelessly detects small ingestible sensors that are incorporated into medication. The ingestible sensors harvest power from the gastric fluid, causing transient activation (7 min).

Another strategy is the development of micro-fuel cells that could be used in the case of implantable sensors, reducing the size of the power supply whilst increasing the lifetime of the battery and therefore the sensor. Characteristics that render fuel cells highly attractive for portable power generation include their high energy efficiency and density, combined with the ability to rapidly refuel [109, 110]. Polymer-electrolyte direct methanol [111] and solid-oxide fuel cells [112] are examples of two such technologies that have been suggested as alternatives to Lithium ion batteries in portable settings.

An alternative approach is to use biocatalytic fuel cells consisting of immobilised micro organisms or enzymes acting as catalysts, with glucose as a fuel to produce electricity [113]. This concept of an "enzymatic microbattery" is attractive because it offers the prospect of dramatically reducing the size of the sensor apparatus and is ideally suited to implantable devices to the point that it has even been suggested as a power supply for a proposed "microsurgery robot" [113]. Miniaturising the packaging required to hold a battery's chemical constituents, which currently consists of

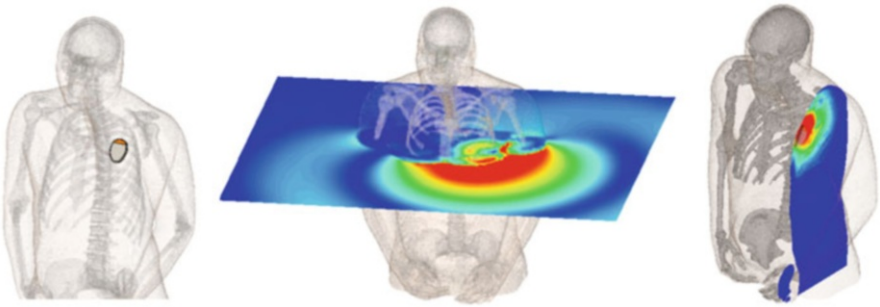


Fig. 1.11 Images showing the field radiated from the loop antenna of an implanted pacemaker at 400 MHz. In the horizontal cutplane, the radiated electrical field is shown in isolines and the vertical cutplane, the power absorbed by the human tissue is displayed. The fields were simulated using the time domain solver of CST MICROWAVE STUDIO® (www.cst.com), which is based on the Finite Integration Technique (FIT). The human phantom includes 32 different tissues at a 1 mm resolution, and was obtained from the male model of the Visible Human Project® – http://www.nlm.nih.gov/research/visible/visible_human.html (*Courtesy Dr Tilmann Wittig, CST – Computer Simulation Technology*)

strong metals such as steel because the battery houses highly corrosive chemicals, is also important. Replacing these chemicals with a substance that is much less corrosive (for example oxygen and water) would therefore require less bulky packaging [114]. As mentioned earlier, an alternative strategy is the use of acoustic (ultrasound) power transmission into an implantable device, with piezoelectric discs as power transducers [115]. Challenges faced in this case include increasing power storage capability as well as controlling the acoustic beam to achieve maximal efficiency.

1.4.5 Wireless Data-Paths, Antenna Design, System Security and Reliability

For both wearable and implantable sensors, efficient wireless data-paths and antenna design are essential to their practical deployment. Different antenna designs have been proposed, whilst wireless communication through the air has been extensively studied, communication from implanted devices through the human body is a new area of study as human body is an uninviting and often hostile environment for a wireless signal. Wireless implants are restricted to a compact antenna that needs to be fully characterised and effectively coupled to the transceiver, also conforming to the low power constraints imposed by implantable devices. In order to design power efficient in-body communication schemes, understanding the mechanism of wave propagation and attenuation inside human body is important and recently both whole body statistical shape models [116] and visible human, as shown in Fig. 1.11 have been used for radiofrequency simulation. Accurate modelling of induced

Fig. 1.12 A picture from the 1918 influenza epidemic is an extreme example of the number of hospital patients that can be present close proximity (*Courtesy of the National Museum of Health and Medicine, Armed Forces Institute of Pathology, Washington, D.C., NCP 1603*)

electromagnetic fields and propagation in the body is a prerequisite to the design of wearable and implantable wireless sensors.

One emerging yet important area of research and development is system security and reliability for BSN. This is of paramount importance for implantable sensors. Vulnerabilities have been found in several existing types of implantable medical device and simultaneously protecting the device, confidentiality and integrity of the data and the privacy of the user remain important research topics. The challenges are often associated with the scarcity of power and computational resources, as well as the conflicting goals of availability (e.g. in emergency) and security, and the large range of contexts in which the devices need to operate.

Figure 1.12 is a historic photograph showing mass crowding in the healthcare environment. Although this is not representative of the clinical setting in modern hospitals, it highlights the importance of secure and reliable data transfer for BSNs. Security and reliability of the network are two of the crucial elements of the BSN design, as sensitive patient information is being transmitted through the wireless network. Unlike typical wired or wireless network architectures in which the network configuration is mostly static and there is limited constraint on resources, the architecture for BSN is highly dynamic, placing more rigorous constraints on power supply, communication bandwidth, storage and computational resources.

In terms of security, BSN data must be secured with strong cryptography to protect the patient's privacy. However, strong cryptography requires extensive computation and resources. Considering the limited resources that a BSN node can have, new approaches need to be taken to maximise security whilst minimising

resource utilisation [117, 118]. Furthermore, the highly dynamic nature of the BSN means that typically static network authentication methods will not be applicable. Even methods proposed for ad hoc networks such as the asymmetric cryptography technique would be computationally too expensive for BSN applications [119]. As such, a robust, efficient and lightweight security infrastructure is required for the practical deployment of BSN applications.

The reliability of the network, on the other hand, directly affects the quality of patient monitoring, and in a worst-case scenario, it can be fatal when a life threatening event has gone undetected. However, due to the constraints on communication bandwidth and power consumption, traditional network reliability techniques such as the retransmission mechanism for TCP protocol, may not be practical for BSN applications. With the similar constraints on WSNs, researchers have proposed several methods for improving its reliability. One simple approach is to use limited retransmission where packets are retransmitted for a fixed number of times until the reception of the acknowledgement; however, retransmission often induces significant overhead to the network. Another approach is to form a multi-path network and exploit the multiple routes to avoid disrupted links [120]. It is expected that this will be an area that will attract significant research interest in the coming years, particularly in exploring the autonomic sensing paradigm for developing self-protecting, self-healing, self-optimising, and self-configuring BSNs.

Thus far, most security and network reliability techniques aim to provide maximum security and reliability for generic wireless sensor network applications. However, in the case of BSN, instead of relying solely on the low-level network infrastructure, high level context information can also be used to reinforce the security and reliability of the system. An example of this is the use of biometric information for enhancing the inherent security of the network [121]. Furthermore, as multiple sensors are often used in a BSN application to measure or infer the same physiological parameters, the use of intelligent multi-sensor data fusion techniques can significantly enhance the reliability of the system.

Other implications of deploying BSNs include the appreciation of the long-term consequences of their effect on the body, particularly in the case of implantable sensors. This is likely to govern the materials and manufacturing process used to construct BSN nodes, their battery supply, and the type of wireless data transfer used. In the case of WSN, where the effect of large numbers of redundant energy-depleted nodes is likely to be detrimental on the environment [2], these nodes must be re-usable or in some way biodegradable. For BSN, both biodegradability and inertness of materials offer potential solutions, but finding the right material for manufacturing the nodes is likely to pose an important challenge.

1.4.6 Context Awareness

In addition to being able to monitor physiological parameters, research on BSNs has identified the importance of the context (environment) the person being

monitored is in when interpreting these parameters. Simple activities such as "sleeping" and "walking" have an effect on not only vital signs such as heart rate and blood pressure, but also on any measure of activity and mobility that is being used. With BSNs, adverse event such as falls in elderlies can be detected along with users' location so assistance can be provided in a timely manner [122]. Long-term monitoring of emotional status [123] has led to several promising applications, particularly those that involves mental disabilities. This "context awareness" can also help account for motion artefacts and errors detected by the sensors. Under normal conditions, visual monitoring provides this contextual information most effectively, but in the pervasive healthcare monitoring setting this is not possible. It is therefore important to identify methods of "inferring" context using techniques such as "dynamic Bayesian networks" [124, 125], "hierarchical hidden semi-Markov models" [126] and "meta-level classifiers" [127] for activity recognition and tracking daily activities. Fusion of data from multiple sensors may provide this contextual information, with selected classifiers designed to yield optimal results based on fusing all sensor readings.

Several types of sensors have been used to develop this context awareness. Accelerometers have been suggested as appropriate candidates for determining activity state (driving, sleeping, exercising), posture (lying, sitting, standing), fall and fall recovery [122, 128, 129]. Other widely used motion sensors are gyroscopes [130] and magnetometers [131]. While accelerometers provide motion measured relative to the gravitational force, magnetometers provide additional reference information along the horizontal axis allowing more accurate range-of-motion analysis. Flex sensors, widely used in glove sensors, have been successfully used for hand gesture recognition [132] and surgical skill assessment [133]. Audio sensors that act by determining either the level of environmental noise or verbal communication, have also showed potential [134]. Together with a wearable camera, a food-intake monitoring system has been developed [135]. Changes in temperature and heat flux may be able to not only determine whether the subject is active or at rest, but also when the subject moves from a warmer indoor to a colder outdoor environment [129]. Skin conductance (affected by sweat gland activity) may be measured using galvanic skin response, with an electrode in contact with the skin surface [123, 136]. Integration of several of these context-sensing modalities is a strategy used by SenseWear^{TM} (BodyMedia Inc.), who have produced a device consisting of a multi-sensor array that is worn on the upper arm and includes a two-axis accelerometer, heat flux sensor, galvanic skin response sensor, skin temperature sensor, and near-body ambient sensor [136]. Over a decade ago, researchers at MIT already developed a modified ring sensor that uses photoplethysmography to detect a person's pulse rate and blood-oxidation level [137]. They have built "context awareness" into this ring using photocells to detect ambient light, thermistors to detect temperature, and accelerometers to detect motion, all of which may interfere with readings and must therefore be accounted for. Currently active research topics on context awareness include behavioural profiling, detecting transitions, detecting concurrent and interleaving context, and distributed inferencing.

Fig. 1.13 An integrated insulin drug-delivery system consisting of two communicating skin patches, one monitoring blood subcutaneous glucose, and the other delivering insulin (*Courtesy of Animas Corporation*)

1.4.7 Integrated Therapeutic Systems

Integrating sensors and therapeutic systems and thereby "closing the feedback loop", is likely to play a major part in defining the role for BSN in clinical practice [138]. This is particularly well illustrated in the delivery of pharmacotherapy where currently drugs are administered at doses and frequencies that are based on average sizes, and metabolic rate. When considering the fact that for an individual patient who has an individual size and metabolism, this optimal dosage is likely to vary considerably from the "average", it is clear that individualised dosing is preferable. In addition to this, a patient's drug requirement may temporarily change during an illness, or for example when they are on other medications such as antibiotics. Whilst underdosing in these situations will result in inadequate treatment (for example seizures in patients on anti-epileptic medication), overdosing will result in an increased risk of the patient suffering unwanted side-effects. Drug-delivering medical feedback loops consist of miniaturised sensors that continuously monitor a drug's effect and through medical control algorithms, adjust its delivery from miniaturised drug pumps. The dosage of the drug would therefore be individualised to the patient.

An example of integrated drug-delivering therapeutic systems for fast-acting insulin in diabetics is shown in Fig. 1.13. This proposed system consists of two

patches on the skin. One is an implanted sensor-amplifier-transmitter which may be replaced by the user every few days. The other patch would be an insulin-delivering system comprising of a calibrator, *Radio Frequency* (RF) receiver, drug reservoir, pump, battery, and miniature subcutaneously inserted drug inlet. The RF signal from the "sensor" patch would be received by the "insulin delivery" patch and translated through a medical algorithm, to a series of micro-doses of insulin. Like the "sensing" patch, the "insulin delivery" patch would also be replaced by the user every few days.

The other disease processes (some of which have been mentioned earlier in this chapter) that would benefit from a similar integrated drug delivery process include epilepsy, hypertension, ischemic heart disease, and conditions requiring blood anticoagulation. All these disease processes currently require "average dosing" and their efficacy is currently monitored by measuring the drug level in the patient's blood, or in the case of hypertensives, their blood pressure when attending their health clinic. It is important to remember that drug delivery may not be the only stimulus that is delivered by such integrated feedback systems, with electrical stimuli, for example to brain, nerves, and muscle being other important examples. Whatever the application, it is clear that the pace at which these integrated feedback systems develop is dependent largely on the development of suitable components and medical control algorithms to construct the miniature subsystems.

1.5 From Wellbeing to Personalised Healthcare

In a population consisting of several vulnerable groups such as those with chronic disease and the elderly, the need for effective individualised health monitoring and delivery has resulted in the concept of "personalised healthcare". Such a system is expected to be 'dynamic' and customised to specifically address the health needs of individuals. In essence, personalised healthcare systems should take into account an individual's chronic (long-term) and episodic (short-term) healthcare needs, and have clear healthcare objectives. They should also account for the cognitive level of the patient, and for both social and community factors. BSNs offer perhaps the greatest chance of developing a personalised healthcare system where treatment may be tailored to the patient at several levels.

At the monitoring level, this system would have to reliably observe the patient's physiology, activity, and context, detecting adverse changes in their wellbeing early. At the delivery of care level, data processing and decision-making algorithms must prompt the appropriate action to deliver correct therapy. Drug delivery, which as previously mentioned is at present dosed according to population averages, could be tailored exactly to an individual's needs, perhaps by infusion rather than by tablet. The cost-effectiveness of such a personalised healthcare system over existing technology solutions is also important and is likely to drive its development. New bionic technologies such as neuromodulation and neurostimulation devices are likely to enable BSNs to interact with and control a patient's

physiological systems themselves. Ultimately, these devices could use information from BSN sensors to control the human body's musculoskeletal system itself. Perhaps one of the most successful examples of a bionic device in clinical use is the cochlear implant, which has had tremendous impact on patients' lives [139].

At the research level, pervasive healthcare systems will allow doctors to learn much more about the disease processes they commonly see in clinical practice. Finally at the information delivery level, giving the patient personalised information (according to their healthcare needs) is likely to help them understand and self-manage their conditions more appropriately [140]. The ultimate aim of all this is the early detection of disease leading to an early intervention, both of which are attributes that may make personalised healthcare-based treatment the next best thing after prevention itself.

In order to deliver truly personalised healthcare, BSN sensors have to become invisible to the patient, thereby avoiding activity restriction or behaviour modification. Whilst sensor miniaturisation and implantability are potential solutions to this, another option being explored is the integration of the sensor into non-clothing items that patients already wear. The ring sensor developed at MIT over 10 years ago, for example, can act as an ambulatory telemetric continuous health monitoring device [137]. It uses advanced photoplethysmographic techniques to acquire data on the patient's heart rate, heart rate variability and oxygen saturation. This ring sensor contains an optical sensor unit, an RF transmitter, and a battery, connected to a microcomputer in the ring itself. This ensures onsite low-level signal processing, data acquisition, filtering, and bidirectional RF communication with a cellular phone which can access a website for data acquisition and clinical diagnosis.

There are, of course, other areas aside from clinical practice where wireless BSN surveillance may be useful in monitoring people and their activity. Professional groups such as fire-fighters and paramedics, who commonly face hazardous situations, as well as policemen and soldiers, may all be monitored with such a system. Fitsense is a company that has developed a system known as "The BodyLAN™". This uses low power wireless body monitoring sensors for a variety of physiological and environmental parameters, which via a proprietary wireless personal area network collect and send data to its users. They can then assess their fitness and performance [141].

One of the challenges of a personalised healthcare system is the wealth of information that the system is going to generate for the healthcare provider above and beyond what is currently available. How this information will be accumulated, stored, and interpreted, and how healthcare systems will respond to adverse events are all questions to be considered. With cheaper, faster and smarter devices such as the iPad become mass market, the use of these devices will have significant influence on personal health, well-being, social integration and rehabilitation. When used in conjunction with a smart tablet as shown in Fig. 1.14, devices such as the e-AR by Sensixa can be transformed into a mobile gait lab for monitoring a whole range of gait and biomechanical indices otherwise have to be obtained through instrumented gait laboratories.

Fig. 1.14 The e-AR device integrated with a commercial tablet as a mobile gait lab for rehabilitation after knee replacement and monitoring the recovery of lower-limb trauma patients

It is important to appreciate that at present whilst much patient information is collected by continuous monitoring, for example during hospital admission, most of this information is lost. Although personalised pervasive healthcare systems will collect a vast amount of information, separating this into "important" and "non-important" is going to require very accurate context sensing. This echoes the general trend of "*quantified-self*" with increasing availability of sensor enriched smart, wearable devices. The myriad of data generated continuously in real-time represents one of the major challenges in the current hype of *big data*. Mining this data and representing it to a user is yet another important task. Finally, as previously mentioned, reacting to this information is going to require major process automation and structural change to existing healthcare systems.

1.6 Finding the Ideal Architecture for BSN

The human body houses what is perhaps the most sophisticated and well-developed example of a network of body sensors in existence. Innervated by small neurones, the *Autonomic Nervous System* (ANS) comprises of autonomic ganglia and nerves. Also known as the involuntary nervous system, it is concerned primarily with the control of the body's internal environment. It is ironic that in developing a wireless body sensor network to monitor a person's physiological state, we are in essence trying to monitor and act on the reactions of the body's own nerves, sensors, and effectors, to both external and internal environments. Looking towards this

advanced sensor network is likely to set the standard for what BSN aims to achieve. The complexity of the human nervous system and its components is clearly much greater than that of any proposed pervasive BSN for patient monitoring. When looking for ideas and solutions to the problems faced by BSN however, the human body may itself hold the key. In order to understand the way in which the autonomic nervous system fulfils the requirement of the ideal body sensor network, we must understand how it overcomes each of the technical challenges mentioned previously in this chapter. We may then be able to translate some of these lessons into design concepts for a truly pervasive patient monitoring system, and look to the challenge of developing feedback loop between wireless sensors as well as effectors. By understanding the principles of human body sensor network, better designs of BSN can be derived. Sensor design, as mentioned previously is a very important part of any such network, and in the case of the human body, this consists of a range of sensor types that are both very sensitive and very accurate. For example, chemoreceptors (chemical receptors) respond to changes in oxygen and carbon dioxide in the blood, and are located either peripherally (in carotid and aortic bodies) or centrally (in the brain). Based on the concentration of these solutes the receptors are able to regulate respiratory rate and cardiac activity, to maintain adequate perfusion of tissues and vital organs. Alternatively baroreceptors (pressure sensors) found in the aortic arch and carotid sinus, are sensitive to the rate of blood pressure change as well as to the steady or mean blood pressure, and are thus able to communicate this information to higher centres in the brain, thereby regulating the blood pressure by altering both the heart's output, as well as the diameter of blood vessels.

Other types of mechanical receptors (mechanoreceptors) that the body possesses include "muscle spindles, which are found between skeletal muscle fibres. Arranged in a parallel distribution with these fibres, the spindles respond to the passive stretch of the muscle, but cease to discharge if the muscle contracts isotonically, thus signalling muscle length. They are therefore the receptors responsible for the stretch reflex such as that elicited by tapping at the knee. Similar mechanical receptors also exist within the cardiac musculature that when overstretched due to increased filling of the heart chambers, result in a compensatory increased strength of cardiac contraction.

Although biocompatibility is not an issue within the body's own sensors because they are self-manufactured, power source miniaturisation is an impressive feature of this system. In general, the body utilises glucose as a substrate using either anaerobic or aerobic respiration to turn this fuel into packets of energy from which is carries in the form of specialised molecules such as *Adenosine Triphosphate* (ATP). Nerves forming the wiring of the system transmit the information in the form of action potentials, with each action potential of approximately 110 mV lasting 5–10 ms. The conduction velocity of nerves can be as high as 100 m per second, making this a very efficient system. Multi-sensory data fusion occurs in the human brain, where each nerve is connected to approximately 10,000 other nerves through special dendritic connections.

Finally, nonvisual context awareness is another important feature of the human nervous system, and is under higher brain centre control. A very good example of this is the body's use of *proprioception* (position sense), particularly with regard to body extremities such as the limbs. To achieve this, the human body uses a number of receptors which signal the position and movement of a limb. These include joint afferents (located in the joints), sensitive to extremes in joint angle, muscle spindles located in the muscle sensitive to position and movement (velocity), Golgi tendon organs located in the muscle tendon sensitive to tension (force), and touch receptors in muscle and overlying skin. Input from all these sensors is processed by the brain and allows the body to know exactly where in space and in what position its different components (limbs) are without the need to look at them. Once the input coming into the nervous system is processed, depending on the state of the muscle, commands are sent back to either maintain or change position.

1.7 The Future: Going from "Micro" to "Nano"

Until now, applications for the use of BSNs in clinical practice have focussed around external and implantable sensors that lie relatively static within the body. However, it is the luminal organs such as blood vessels, gastrointestinal tract, urinary tract, ventricles of the brain, spinal canal, lymphatic, and venous systems that offer the greatest opportunity to sense acute disease processes and monitor chronic illnesses quickly and efficiently. These cavities are essentially the "high-ways" filled with body fluids, inflammatory mediators, cells, and pathogens, forming what are the "battlefields" where disease processes are fought. As the accuracy of our sensors increases and their size decreases, it is in these domains that we would like to have the maximal effect on any disease process. Recent advances in nano-technology have meant that delivering sensors within these luminal cavities is for the first time a real possibility.

Miniaturisation of many sensors to the "micro" scale is already taking place. Nanoscale particles are likely to be the smallest component of sensing systems and may be deployed in a number of ways. For example, nanoscale particles may themselves be coated to form a biomechanical sensing system. An existing example of this is a protein-encapsulated single-walled carbon nanotube sensor that alters its fluorescence depending on exposure to glucose in the surrounding tissues [55]. In fact nanoparticles have even been attached to antigen amyloid-derived diffusible ligands in order to develop a nanoscale optical biosensor for Alzheimer's disease [142]. Finally, nanoparticles may act as sensors themselves. The scenario of injecting nanoscale biosensors into luminal cavities, where the sensor comes in contact with its substrate, binds to it, and is carried to the site of maximal disease activity is no longer unrealistic. Such targeted sensor delivery and binding may allow extremely targeted disease process monitoring, and therapy.

An example of nanoscale particles already in use is *Ultra-Small Particles of Iron Oxide* (USPIO) for pathologic tissue characterisation. The long blood circulating time and the progressive macrophage uptake in inflammatory tissues of USPIO particles are properties that can be interrogated by imaging techniques such as *Magnetic Resonance Imaging* (MRI). The USPIO signal alterations observed in ischemic areas of stroke patients can be related to the visualisation of inflammatory macrophage recruitment into human brain infarction, whereas in brain tumours, USPIO particles which do not pass the ruptured blood-brain barrier fairly soon after injection can be used to assess tumour microvascular heterogeneity. Existing research has also shown that USPIO contrast agents can reveal the presence of inflammatory multiple sclerosis lesions. They can also help pick up the spread of rectal cancers to the lymphatic system by increasing the diagnostic accuracy of rectal imaging [143].

Whilst miniaturisation means that deploying microscale sensors is going to become easier, getting the information out of these sensors will be a significant challenge [55]. Optical imaging techniques have been suggested as a solution to the problem of data extraction from such a small sensor, and would allow, for example, dynamic investigation of the signalling processes that go on inside the cell itself. Self-configuration, self-assembly and actuation are both essential requirements of such sensing systems, as we find we are able to place them in increasingly inaccessible locations. Optical technologies such as fluorescence resonance energy transfer may provide on solution to this requirement for self-assembly [144].

The team led by Joe Wang at UCSD (*University of California-San Diego, USA*) have been at the forefront of developing nanoscale motors capable of converting energy into movement and force, with selective binding and controlled manoeuvres [145]. One example is their Lectin-modified self-propelled microengines for bacterial isolation as shown in Fig. 1.15 [146]. The ability to selectively scour, interact, and isolate pathogenic bacteria intrigues a whole range of possible applications of these nanomachines, representing the future directions of BSN, which does not always need to be embodied as silicon devices.

In his recent article of the Journey to the Centre of a Tumour [147], Sylvain Martel and his team of École Polytechnique de Montréal illustrated the use of a microcarrier-bacteria hybrid based on MC-1 bacterium, which has a spherical body about 2 μm across and a pair of spinning, whiplike tails that propel it at speeds up to 150 times its body length per second. Anticancer drugs can be "cocooned" in liposomes and attached to the bacterium using antibodies. They are then loaded into fat-like bubbles, such as vesicles or micelles, with magnetic nanoparticles and navigated using an MRI machine through large blood vessels. When they reach narrower vessels blocking their path, they would burst and dispatch the bacteria as illustrated in Fig. 1.16. These armed microbes would swim to the tumour and releasing their drug cargo. "*Almost 50 years after Hollywood dreamed up the tale, a real-life fantastic voyage is just beginning*."

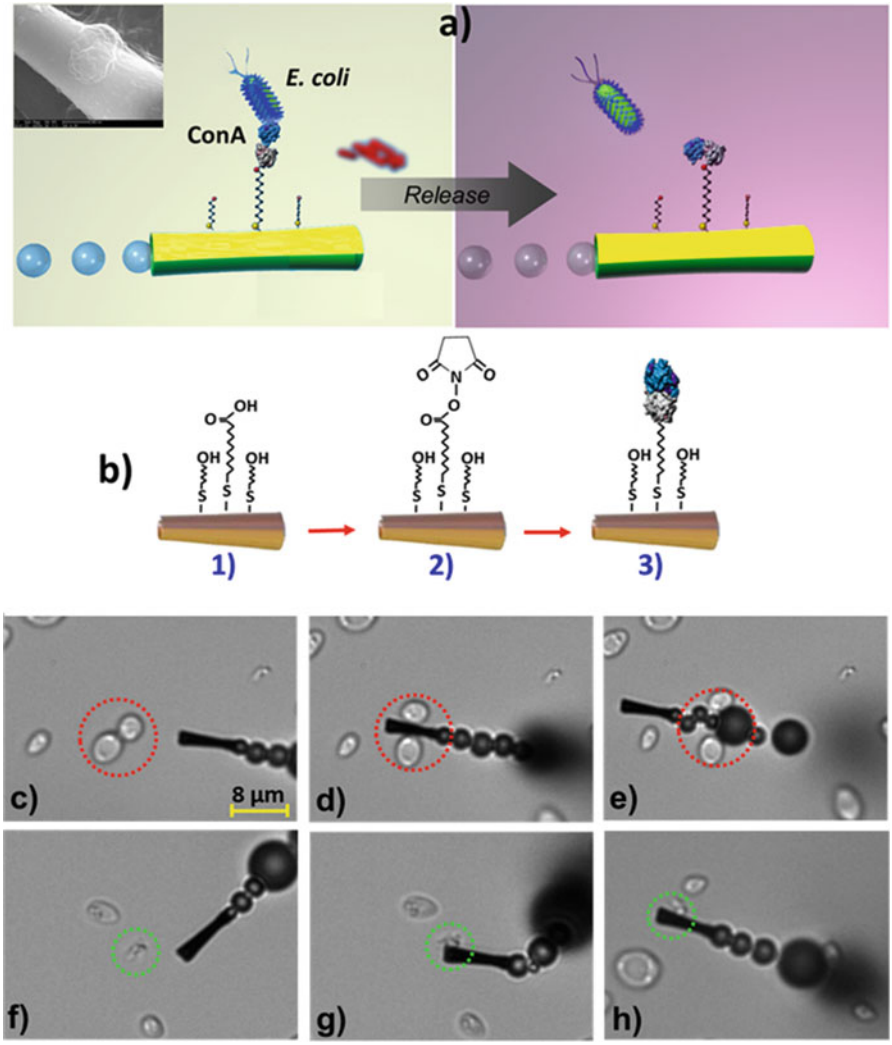


Fig. 1.15 Lectin-modified microengines for bacteria isolation (**a**) schematic diagram illustrating the pick-up, transport and release of the target bacteria, (**b**) the underlying surface chemistry involved on the microengine function with the lectin receptor, (**c–h**) selective binding and transport of the rod-shaped gram-negative E. coli bacteria (~2 µm, in *green dotted circles*) whilst ignoring the S. cerevisiae cells (~5 µm, in *red dotted circles*), (*Courtesy of Dr Joe Wang, Department of Nanoengineering, University of California-San Diego, La Jolla, California, USA, with permission from publisher* [146])

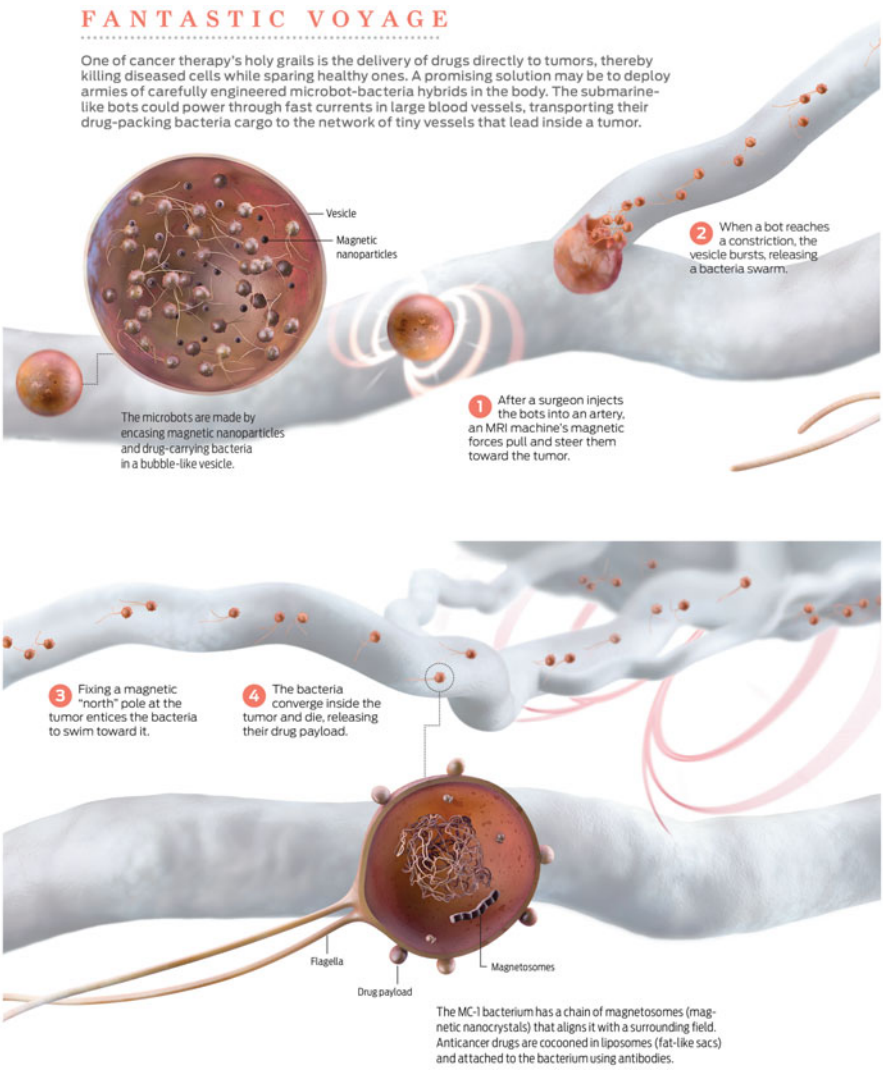


Fig. 1.16 Swimming microbes targeted for tumour with controlled release of drug cargo conceptualised by Sylvain Martel and colleagues. The microcarrier-bacteria hybrid has a spherical body and a pair of spinning, whiplike tails that propel it at speeds up to 150 times its body length per second. Anticancer drugs can be "cocooned" in liposomes and attached to the bacterium using antibodies. They are then loaded into fat-like bubbles, such as vesicles or micelles, with magnetic nanoparticles and navigated using an MRI machine through large blood vessels. When they reach narrower vessels blocking their path, they would burst and dispatch the bacteria (*Courtesy of Dr Sylvain Martel, École Polytechnique de Montréal, Canada, with permission from publisher* [147])

1.8 The Scope of the Book

In this chapter, we have highlighted the scope of BSN for monitoring human disease processes. Although proprietary designed wireless sensor networks provide some solutions to the problems of health monitoring, it is clear that a specialised BSN platform offers the opportunity to monitor human beings in a way that has not previously been possible. This truly ubiquitous and pervasive patient monitoring system will, in the first instance, allow us to identify these disease processes, and will later on allow us to accurately monitor their progress and devise effective therapeutic measures. The challenges of improving sensor design, biocompatibility, energy supply, power scavenging, secure information transfer, and context awareness must all be overcome before such a system is effective. Closing the feedback loop to deliver targeted therapy, and thus reducing unnecessary drug side-effects is also a real possibility.

The purpose of this book is to address both basic issues and emerging trends concerning biosensor design and interfacing, protein engineering for biosensors, wireless communication, network topology, communication protocols and standards, energy scavenging, bio-inspired sensor processing, multi-sensor fusion, and context-aware and autonomic sensing. Figure 1.17 illustrates the structure of the materials covered.

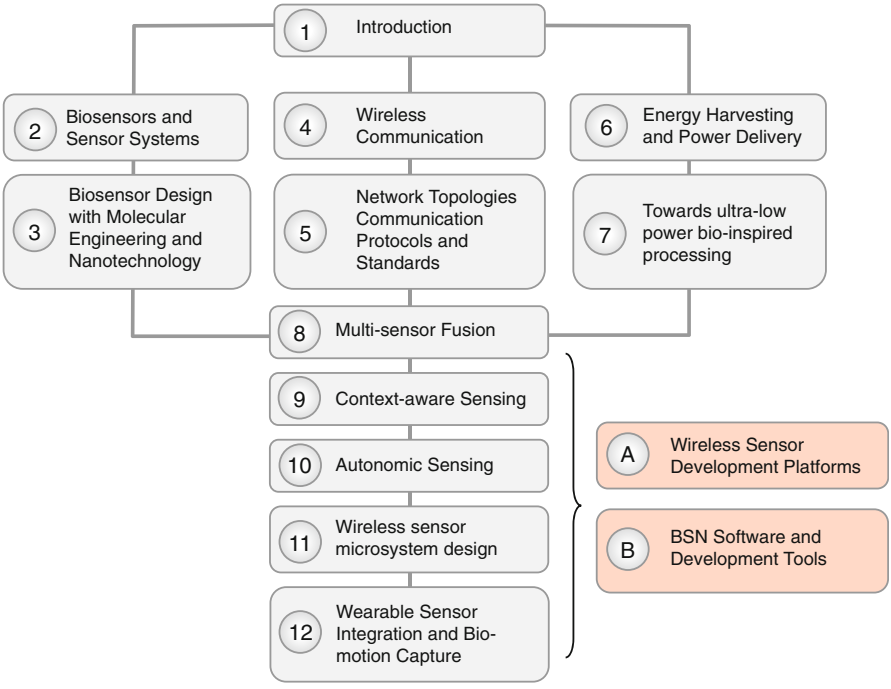


Fig. 1.17 The structure of this book and the interdependency of the chapters

In Chap. 2, we will introduce the basic concept of electrochemical sensors and biosensors. This chapter covers the basic principles of electrochemical devices based on potentiometry, amperometry and voltammetry. The chapter also outlines the basic instrumentation requirements for these devices and looks at issues surrounding biocompatibility and sensor data handling. In Chap. 3, we will focus on the biological aspects of biosensors in two important regards; the first is the biological molecules involved in the molecular recognition processes that give the biosensors their specificity and sensitivity and the other is concerned with biocompatibility. We will discuss how these proteins can be engineered to improve sensor performance and address the mutual interaction between the sensor and the tissue within which it is located. Although progress has been made in making implantable biosensors reliable and robust for a period of a few days, there are still significant technical issues associated with long-term implantation. This reflects in part the response of the tissue to trauma and the inherent robustness of the biological molecules used in the sensor. This implies that the solution to long-term implantation will come from a combination of factors including minimally invasive implantation, understanding and modulating tissue response to implantation and modifying the properties of the biomolecules.

Chapters 4 and 5 will address wireless communication, network topologies, communication protocols and standards for BSNs. In Chap. 4, we will discuss two types of communication links: the inductive loop and radio frequency communication. The inductive loop is widely used today for transferring small packets of data without requiring an implanted power source. Whilst an RF system does require an implanted battery source, it is capable of transferring larger packets of data within a shorter time period and over greater distances. For this reason, RF-based communication will be the main topic of this chapter. Whilst wireless communication through the air has been extensively documented, communication from implanted devices through the human body is on-going research topic. This chapter will discuss body properties and their effect on radio propagation. The human body is an uninviting and often hostile environment for a wireless signal. One of the most important considerations for implanted devices is physical size, meaning in-body communication system designs are restricted to an extremely small antenna that needs to be characterised to enable it to be effectively coupled to the transceiver. A significant portion of this chapter is devoted to antenna measurement and coupling circuit design, as it is critical to the success of an implanted RF system.

Based on the contents described in previous chapters, Chap. 5 illustrates the use of different network topologies for practical deployment of BSNs. Although the initial use of BSNs will only consist of limited sensor nodes, more complex network topology will be adopted as the devices get smaller and more ubiquitous. This will allow more effective use of sensor message hopping, distributed inferencing for improved system robustness and noise resilience with built-in redundancies. Particular emphasis will be made in the chapter touching upon the integration of body sensing with ambient sensing. In such cases, the joining and rejoining of a shifting series of different networks (home, public, and hospital) and the addition or removal of differing sensor nodes under different context requirements can pose

significant challenges. A detailed overview of the current and emerging communication protocols and standards for implantable, wearable, and ambient wireless sensing will be provided, and issues concerning standards for overall healthcare system integration with pervasive sensing will also be discussed.

With increasing miniaturisation and cost reduction of sensors, circuits and wireless communication components come new possibilities for networks of wireless sensors, in wearable and other applications. However, for sensors to be wireless, or untethered, requires not only wireless communication to and from the nodes, but also wireless powering. Batteries, of course, provide this capability in the great majority of portable electronic devices, and thus are the obvious solution for the wireless sensor node application. However, the need for replacement or recharging of batteries introduces a cost and convenience penalty which is already undesirable in larger devices, and is likely to become unacceptable for sensor nodes as their ubiquity grows. As an alternative, sources which scavenge energy from the environment are highly desirable. With decreasing power demands for sensing, processing, and wireless communication for BSNs due to improved electronic design and miniaturisation, alternatives to battery power based on energy scavenging techniques are becomingly increasingly realistic. In Chap. 6, we will discuss the basic power requirements for BSN nodes and possible architectures for different energy scavenging techniques. Issues concerning fabrication, module design, power electronics and system effectiveness will be discussed.

The natural world is analogue and yet the modern microelectronic world to which we are exposed represents real world data using discrete quantities manipulated by logic. The new trend set by BSNs is beginning to see the processing interface move back to using continuous quantities, which are more or less in line with the biological processes. This computational paradigm we label "bio-inspired" because of the ability of silicon chip technology to enable use of inherent device physics, allowing us to approach the computational efficiencies of biology. In contrast to the digital approach, where each operation is performed through a network of devices operated in switched fashion, the physics of the elementary device itself, either electrical, chemical or electrochemical, can be exploited to perform the same operation in an analogue way. Therefore both the energy per unit computation and, silicon real-estate are reduced, resulting in significantly increased overall resource efficiency. In Chap. 7 we will first look at the motivation for bio-inspired signal processing and discuss the relative merits of analogue and digital signal processing, and the need for hybrid architectures. The concept of applying bio-inspired design methodologies to CMOS-based biosensors will then be introduced. *Field-Effect Transistor* (FET) based sensors will be presented, including a detailed example of the application of analogue processing techniques to these devices. Finally, future directions and applications for biochemically inspired design will be discussed.

The pursuit for low power, miniaturised, distributed sensing under natural physiological conditions of the patient has also imposed significant challenges on integrating information from what is often heterogeneous, incomplete, and error prone sensor data. For BSNs, the nature of errors can be attributed to a number of

sources, but motion artefact, inherent limitation and malfunctions of the sensors, and communication errors are the main causes of concern. In practice, it is desirable to rely on sensors with redundant or complementary data to maximise the information content and reduce both systematic and random errors. This, in essence, is the main drive for multi-sensor fusion described in Chap. 8, which is concerned with the synergistic use of multiple sources of information. In this chapter, we will discuss the basic concept of multi-sensor fusion and the methods related to data, feature and decision levels of data fusion techniques. The key emphasis of the chapter is the introduction of optimal averaging techniques for sensor array data and a feature selection algorithm based on Bayesian theory and receiver operating characteristic analysis.

With BSNs, effective sensor fusion and statistical feature reduction is crucial to the use of wireless sensor networks incorporating both built-in redundancies and tissue heterogeneity. For the monitoring of patients under normal physiological status, the contextual information is important to the capture of clinically relevant episodes. In Chap. 9, we will investigate issues concerning context aware sensing for the practical deployment of BSNs. One of the main topics covered by this chapter is the introduction of a novel framework, called *Spatio-Temporal Self-Organising Map* (STSOM), by incorporating the spatio-temporal behaviour of self-organising neural networks for reliable context detection. The significance of the contents provided in this chapter is twofold. First, it provides an effective framework for reliably extracting contextual information based on the feature selection framework described in Chap. 8. Secondly, it illustrates the possibility of complete analogue implementation of the framework based on STSOM for low power operation. The chapter has also covered the use of HMMs and factor graphs for context-aware sensing, as well as emerging techniques for behaviour profiling, transient activity detection, and the handling of concurrent and interleaving events. A distributed inferencing model has been introduced by combining the use of multi-objective feature selection, dependency analysis and factor graph representation. In this way, multiple context recognition problems can be concurrently solved with minimal use of computational resources and the ease of logical-to-physical model mapping.

The use of BSNs can be influenced by a wide range of limitations including processing power, storage, network connectivity, and available power sources. In most cases, it is not possible to guarantee or accurately predict the characteristics of these resources in advance. In particular, wireless network communication is subject to many unpredictable environmental effects, even over short-range connections such as between body and wearable devices. Issues related to *Quality of Service* (QoS), variable resources monitoring, adaptation techniques applied to data including compression, data security, authentication and privacy will need to be addressed. In Chap. 10, we will introduce the basic concepts involved in autonomic sensing and describe a number of other approaches that are inspired by biological systems. The use of the autonomic sensing paradigm for BSNs is relatively new, and in this chapter we will discuss the general issues and new opportunities involved in the development of self-protecting, self-healing, self-optimising, and

self-configuring BSNs. The main technical details of the chapter will cover the use of multidimensional scaling for self-discovery of sensor co-locations for automatic configuration of routing structures, and the use of Bayesian belief propagation for efficient, distributed inferencing and fault tolerant sensing.

In the final two chapters of this book, practical examples of system design and software integration are provided. In Chap. 11, we will provide a practical perspective of designing wireless sensor microsystems. Wireless sensor microsystems offer very diverse functionality, and this brings about a range of technical design problems. Design skills include sensors, ASICs, wireless, low power, packaging, software, networking and power sources. These problems become more challenging the smaller the final device must be. In this chapter, we will cover a range of design topics that are of relevance to wireless BSN microsystem designs and the general design principles involved will be elaborated in the context of developing an ingestible lab-in-a-pill device. This offers a particularly challenging case study echoing many of the issues that have been discussed in other chapters.

In Chap. 12, the key processing steps involved in human motion capture based on BSN are provided. The chapter covers the basics of quaternion based orientation representation and a Bayesian fusion framework incorporating quaternion-based unscented Kalman filter for 3D body posture and movement reconstruction. Detailed algorithm design has been presented and with reference to Chap. 5, issues related to network QoS have been highlighted.

To help readers entering into the field and starting up some of the practical experiments involving BSNs, we have provided two appendices outlining the wireless sensor development platforms, as well as detailed technical and programming issues for the use of the BSN development kit and operating systems.

The inherent diversity of the materials covered for the effective development of BSNs means it is not essential for the readers to go through the chapters provided in a rigid sequence, and Fig. 1.17 outlines the interdependency of all the chapters. It is suggested that readers can take an appropriate route depending on your technical background to follow the materials presented, either used as a reference for BSN research and development or as a graduate level text book. A dedicated web site (http://www.bsn-web.info) that accompanies the contents of the book has also been created. Interested readers can use this site to find out the latest updates and access BSN related resources, particularly those related to the BSN development kit as described in the appendices of this book.

References

1. Bell G. Bell's law for the birth and death of computer classes. Communications of the ACM 2008; 51(1):86–94.
2. Bulusu N, Jha S. Wireless sensor network systems: a systems perspective. Artech House Publishers, 2005.
3. Warneke B, Last M, Liebowitz B, Pister KSJ. Smart dust: communicating with a cubic-millimeter computer. Computer 2001; 34(1):44–51.

4. Kahn JM, Katz RH, Pister KSJ. Next century challenges: mobile networking for smart dust. In: Proceedings of the International Conference on Mobile Computing and Networking, Boston, MA, 2000.
5. Link JR, Sailor MJ. Smart dust: self assembling, self-organizing photonic crystals of porous Si. Proceedings of the National Academy of Sciences of the United States of America 2003; 100(19):10607–10610.
6. Culler S, Estrin D, Srivastava M. Overview of sensor networks. Computer 2004; 37(8):41–49.
7. HP, "Shell and HP advance seismic sensing capabilities", http://www8.hp.com/us/en/hp-news/press-release.html?id=908010#.Uil03z_hH8p, 2011.
8. Spencer BF, Jr, Cho S, Sim SH. Wireless monitoring of civil infrastructure comes of age. Structure Magazine 2011:12–16.
9. Chan M, Campo E, Esteve D, Fourniols J-Y. Smart homes – current features and future perspectives. Maturitas 2009; 64:90–97.
10. Zhang S, McClean SI, Scotney BW. Probabilistic learning from incomplete data for recognition of activities of daily living in smart homes. IEEE Transactions on Information Technology in Biomedicine 2012; 16(3):252–462.
11. Mainwaring A, Polastre J, Szewczyk R, Culler D, Anderson J. Wireless sensor networks for habitat monitoring. In: Proceedings of the First ACM International Workshop on Wireless Sensor Networks and Applications, Atlanta, Georgia, USA, 2002; 88–97.
12. Lo BPL, Yang GZ. Key technical challenges and current Implementations of body sensor networks. In: Proceedings of the Second International Workshop on Wearable and Implantable Body Sensor Networks, London, UK, 2005; 1–5.
13. Markel H. The stethoscope and the art of listening. New England Journal of Medicine 2006; 354(6):551–553.
14. Tavel ME. Cardiac auscultation: a glorious past–and it does have a future! Circulation 2006; 113(9):1255–1259.
15. Fragasso G, Cuko A, Spoladore R, Montano C, Palloshi A, Silipigni C, et al. Validation of remote cardiopulmonary examination in patients with heart failure with a videophone-based system. 13 2007; 4(281–286):281–286.
16. Hond DE, Celis H, Vandenhoven G, O'Brien E, Staessen JA. Determinants of white-coat syndrome assessed by ambulatory blood pressure or self-measured home blood pressure. Blood Pressure Monitoring 2003; 8(1):37–40.
17. WHO. International Health Conference, New York 1948.
18. Charakida M, Masi S, Deanfield JE. The year in Cardiology 2012: focus on cardiovascular disease prevention. European Heart Journal 2013; 34(4):314–317.
19. Berry JD, Dyer A, Cai X, Garside DB, Ning H, Thomas A, et al. Lifetime risks of cardiovascular disease. New England Journal of Medicine 2012; 366(4):321–329.
20. Go AS, Hylek EM, Phillips KA, Chang Y, Henault LE, Selby JV, et al. Prevalence of diagnosed atrial fibrillation in adults: national implications for rhythm management and stroke prevention: the AnTicoagulation and Risk Factors in Atrial Fibrillation (ATRIA) study. Journal of the American Medical Association 2001; 285(18):2370–2375.
21. Benjamin EJ, Wolf PA, D'Agostino RB, Silbershatz H, Kannel WB, Levy D. Impact of atrial fibrillation on the risk of death: the Framingham heart study. Circulation 1998; 98(10):946–952.
22. Chobanian AV, Bakris GL, Black HR, Cushman WC, Green LA, Izzo JLJ, et al. Seventh report of the joint national committee on prevention, detection, evaluation, and treatment of high blood pressure. Hypertension 2003; 42(6):1206–1252.
23. Hunt SA, et al. ACC/AHA guidelines for the evaluation and management of chronic heart failure in the adult: executive summary a report of the American college of cardiology/American Heart Association task force on practice guidelines (committee to revise the 1995 guidelines for the evaluation and management of heart failure). Circulation 2001; 104(24):2996–3007.
24. Qureshi AI, Suri MF, Kirmani JF, Divani AA, Mohammad Y. Is prehypertension a risk factor for cardiovascular diseases? Stroke 2005; 36(9):1859–1863.

25. Brown AS. Lipid management in patients with diabetes mellitus. American Journal of Cardiology 2005; 96(4A):26–32.
26. Maldonado M, D'Amico S, Otiniano M, Balasubramanyam A, Rodriguez L, Cuevas E. Predictors of glycaemic control in indigent patients presenting with diabetic ketoacidosis. Diabetes, Obesity and Metabolism 2005; 7(3):282–289.
27. Rubin RR. Adherence to pharmacologic therapy in patients with type 2 diabetes mellitus. American Journal of Medicine 2005; 118(Suppl 5A):27S–34S.
28. Aziz O, Atallah L, Lo B, Gray E, Athanasiou T, Darzi A, et al. Ear-worn body sensor network device: an objective tool for functional postoperative home recovery monitoring. Journal of the American Medical Informatics Association 2011; 18(2):156–159.
29. Allen MG. Implantable micromachined wireless pressure sensors: Approach and clinical demonstration. In: Proceedings of the Second International Workshop on Wearable and Implantable Body Sensor Networks, London, UK, 2005; 40–43.
30. Butler RN. Population aging and health. British Medical Journal 1997; 315(7115):1082–1084.
31. Aronow WS, Ahn C. Elderly nursing home patients with congestive heart failure after myocardial infarction living in New York City have a higher prevalence of mortality in cold weather and warm weather months. Journals of Gerontology Series A: Biological Sciences and Medical Sciences 2004; 59(2):146–147.
32. Koken PJ, Piver WT, Ye F, Elixhauser A, Olsen LM, Portier CJ. Temperature, air pollution, and hospitalization for cardiovascular diseases among elderly people in Denver. Environmental Health Perspectives 2003; 111(10):1312–1317.
33. Dishongh T, Rhodes K, Needham B. Room to room location using wearable sensors for tracking social health of elders. In: Proceedings of the Second International Workshop on Wearable and Implantable Body Sensor Networks, London, UK, 2005; 18–20.
34. Kohl HW, Craig CL, Lambert EV, Inoue S, Alkandari JR, Leetongin G, et al. The pandemic of physical inactivity: global action for public health. Lancet 2012; 380(9838):294–305.
35. Gardner AW, Parker DE, Montgomery PS, Scott KJ, Blevins SM. Efficacy of quantified home-based exercise and supervised exercise in patients with intermittent claudication: a randomized controlled trial. Circulation 2011; 123(5):491–498.
36. Barry VW, McClain AC, Shuger S, Sui X, Hardin JW, Hand GA, et al. Using a technology-based intervention to promote weight loss in sedentary overweight or obese adults: a randomized controlled trial study design. Diabetes, Metabolic Syndrome and Obesity 2011; 4:67–77.
37. Pellegrini CA, Duncan JM, Moller AC, Buscemi J, Sularz A, DeMott A, et al. A smartphone-supported weight loss program: design of the ENGAGED randomized controlled trial. BMC Public Health 2012; 12(1041).
38. Dolan P, Hallsworth M, Halpern., King D, Vlaev I. Influencing behaviour through public policy. Institute for Government and the Cabinet Office, UK 2010.
39. McKnight-Eily LR, Eaton DK, Lowry R, Croft JB, Presley-Cantrell L, Perry GS. Relationships between hours of sleep and health-risk behaviors in US adolescent students. Preventive Medicine 2011; 53(4–5):271–273.
40. Cappuccio FP, Taggart FM, Kandala NB, Currie A, Peile E, Stranges S, et al. Meta-analysis of short sleep duration and obesity in children and adults. Sleep 2008; 31(5):619–626.
41. Dahl R, Lewin D. Sleep disturbance in children and adolescents with disorders of development: Its significance and management. London, United Kingdom: Mac Keith Press, 2001.
42. O'Driscoll DM, Turton AR, Copland JM, Strauss BJ, Hamilton GS. Energy expenditure in obstructive sleep apnea: validation of a multiple physiological sensor for determination of sleep and wake. Sleep Breath 2013; 17(1):139–146.
43. Krauchi K. The human sleep-wake cycle reconsidered from a thermoregulatory point of view. Physiology and Behavior 2007; 90(2–3):236–245.
44. Krejcar O, Jirka J, Janckulik D. Use of mobile phones as intelligent sensors for sound input analysis and sleep state detection. Sensors 2011; 11(6):6037–6055.

45. Chouvarda I, Koutkias V, Malousi A, Maglaveras N. Grid-enabled biosensor networks for pervasive healthcare. Studies in Health Technology and Informatics 2005; 112:90–99.
46. Rubel P, Fayn J, Simon-Chautemps L, Atoui H, Ohlsson M, Telisson D, et al. New paradigms in telemedicine: ambient intelligence, wearable, pervasive and personalized. Studies in Health Technology and Informatics 2004; 108:123–132.
47. Wang L, Johannessen EA, Hammond PA, Cui L, Reid SW, Cooper JM, et al. A programmable microsystem using system-on-chip for real-time biotelemetry. IEEE Transactions on Biomedical Engineering 2005; 52(7):1251–1260.
48. Jovanov E, Milenkovic A, Otto C, de Groen PC. A wireless body area network of intelligent motion sensors for computer assisted physical rehabilitation. Journal of NeuroEngineering and Rehabilitation 2005; 2(6).
49. Shantaram A, Beyenal H, Raajan R, Veluchamy A, Lewandowski Z. Wireless sensors powered by microbial fuel cells. Environmental Science and Technology 2005; 39(13):5037–5042.
50. Arshak A, Arshak K, Waldron D, Morris D, Korostynska O, Jafer E, et al. Review of the potential of a wireless MEMS and TFT microsystems for the measurement of pressure in the GI tract. Medical Engineering and Physics 2005; 27(5):347–356.
51. Paradiso R, Loriga G, Taccini N. Wearable system for vital signs monitoring. Studies in Health Technology and Informatics 2004; 108:253–259.
52. Borkar S. Design challenges of technology scaling. IEEE Micro 1999; 19(4):23–29.
53. Holloway RH. Capsule pH monitoring: is wireless more? GUT 2005; 54(12):1672–1673.
54. Pandolfino JE, Richter JE, Ours T, Guardino JM, Chapman J, Kahrilas PJ. Ambulatory esophageal pH monitoring using a wireless system. The American Journal of Gastroenterology 2003; 98:740–749.
55. Vo-Dinh T. Biosensors, nanosensors and biochips: frontiers in environmental and medical diagnostics. In: Proceedings of the First International Symposium on Micro and Nano Technology, Hawaii, 2004; 1–6.
56. Morrell MJ. Responsive cortical stimulation for the treatment of medically intractable partial epilepsy. Neurology 2011; 77:1295–1304.
57. Culliford DJ, Maskell J, Beard DJ, Murray DW, Price AJ, Arden NK. Temporal trends in hip and knee replacement in the United Kingdom: 1991 to 2006. Journal of Bone and Joint Surgery 2010; 92(1):130–135.
58. Dalury DF, Pomeroy DL, Gorab RS, Adams MJ. Why are total knee arthroplasties being revised? The Journal of Arthroplasty 2013.
59. Garg SK, Schwartz S, Edelman SV. Improved glucose excursions using an implantable real-time continuous glucose sensor in adults with type 1 diabetes. Diabetes Care 2004; 27(3):734–738.
60. Steil GM, Panteleon AE, Rebrin K. Closed-loop insulin delivery-the path to physiological glucose control. Advanced Drug Delivery Reviews 2004; 56(2):125–144.
61. Ward WK, Wood MD, Casey HM, Quinn MJ, Federiuk IF. An implantable subcutaneous glucose sensor array in ketosis-prone rats: closed loop glycemic control. Artificial Organs 2005; 29(2):131–143.
62. Brown JQ, Srivastava R, McShane MJ. Encapsulation of glucose oxidase and an oxygen-quenched fluorophore in polyelectrolyte-coated calcium alginate microspheres as optical glucose sensor systems. Biosensors and Bioelectronics 2005; 21(1):212–216.
63. Najafi N, Ludomirsky A. Initial animal studies of a wireless, batteryless, MEMS implant for cardiovascular applications. Biomedical Microdevices 2004; 6(1):61–65.
64. Grayson RAC, Shawgo RS, Li Y, Cima MJ. Electronic MEMS for triggered delivery. Advanced Drug Delivery Reviews 2004; 56(2):173–184.
65. Bogue R. MEMS sensors: past, present and future. Sensor Review 2007; 27(1):7–13.
66. Xia YN, Whitesides GM. Soft lithography. Annual Review of Materials Science 1998; 28:153–184.
67. Fedder GK, Howe RT, Liu TJK, Quevy EP. Technologies for cofabricating MEMS and electronics. Proceedings of the IEEE 2008; 96(2):306–322.

68. Lavrik NV, Sepaniak MJ, Datskos PG. Cantilever transducers as a platform for chemical and biological sensors. Review of Scientific Instruments 2004; 75(7):2229–2253.
69. Yazdi N, Ayazi F, Najafi K. Micromachined inertial sensors. Proceedings of the IEEE 1998; 86(8):1640–1659
70. Alper SE, Akin T. A single-crystal silicon symmetrical and decoupled MEMS gyroscope on an insulating substrate. Journal of Microelectromechanical Systems 2005; 14(4):707–717.
71. Barlian AA, Park WT, Mallon JR, Rastegar AJ, Beth LP. Review: semiconductor piezoresistance for Microsystems. Proceedings of the IEEE 2009; 97(3):513–552.
72. Lemkin M, Bose BE. A three-axis micromachined accelerometer with a CMOS position-sense interface and digital offset-trim electronics. IEEE Journal of Solid-State Circuits 1999; 34(4):456–468.
73. Lange K, Rapp BE, Rapp M. Surface acoustic wave biosensors: a review. Analytical and Bioanalytical Chemistry 2008; 391(5):1509–1519.
74. Badilescu S, Packirisamy M. BioMEMS: science and engineering perspectives. CRC Press, 2011.
75. Chin CD, Linder V, Sia SK. Lab-on-a-chip devices for global health: Past studies and future opportunities. Lab on a Chip 2007; 7(1):41–57.
76. Hierlemann A, Brand O, Hagleitner C, Baltes H. Microfabrication techniques for chemical/biosensors. Proceedings of IEEE 2003; 91(6):839–863.
77. Foulds NC, Lowe CR. Enzyme entrapment in electrically conducting polymers – immobilization of glucose oxidase in polypyrole and its application in amperometric glucose sensors. Journal of the Chemical Society – Faraday Transactions 1 1986; 82:1259–1264.
78. Fritz J, Baller MK, Lang HP, Rothuizen H, Vettiger P, Meyer E, et al. Translating biomolecular recognition into nanomachines. Science 2000; 288(5464):316–318.
79. Stutzmann M, Garrido JA, Erickhoff M, Brandt MS. Direct biofunctionalization of semiconductors: A survey Physica Status Solidi A-Applications and Materials Science 2006; 203(14):3424–3437.
80. Ariga K, Nakanishi T, Michinobu T. Immobilization of biomaterials to nano-assembled films (self-assembled monolayers, Langmuir-Blodgett films, and layer-by-layer assemblies) and their related functions. Journal of Nanoscience and Nanotechnology 2006; 6(8):2278–2301.
81. Grieshaber D, Mackenzie R, Voros J, Reimhult E. Electrochemical biosensors – Sensor principles and architectures. Sensors 2008; 8(3):1400–1458.
82. Lee CS, Kim SK, Kim MH. Ion-selective field-effect transistor for biological sensing. Sensors 2009; 9(9):7111–7131.
83. Homola J, Yee SS, Gauglitz G. Surface plasmon resonance sensors: review. Sensors and Actuators B – Chemical 1999; 54(1–2):3–15
84. Bashir R. BioMEMS: state-of-the-art in detection, opportunities and prospects. Advanced Drug Delivery Reviews 2004; 56:1565–1586.
85. Trohman RG, Kim MH, Pinski SL. Cardiac pacing: the state of the art. Lancet 2004; 364 (9446):1701–1719.
86. National pacemaker and ICD database. UK and Ireland. Annual Report 2000.
87. Parkes J, Bryant J, Milne R. Implantable cardioverter defibrillators: arrhythmias. A rapid and systematic review. Health Technology Assessment 2000; 87(5):438–442.
88. Turk DC. Clinical effectiveness and cost-effectiveness of treatments for patients with chronic pain. Clinical Journal of Pain 2002; 18(6):355–365.
89. Sheldon R, Kiff ES, Clarke A, Harris ML, Hamdy S. Sacral nerve stimulation reduces corticoanal excitability in patients with faecal incontinence. British Journal of Surgery 2005; 92(11):1423–1431.
90. Benabid AL, Krack PP, Benazzouz A, Limousin P, Koudsie A, Pollak P. Deep brain stimulation of the subthalamic nucleus for Parkinson's disease: methodologic aspects and clinical criteria. Neurology 2000; 55(12 Suppl 6):S40–44.

91. Chabardes S, Kahane P, Minotti L, Koudsie A, Hirsch E, Benabid AL. Deep brain stimulation in epilepsy with particular reference to the subthalamic nucleus. Epileptic Disorders 2002; 4 Suppl 3:S83–93.
92. Rosenow JM, Tarkin H, Zias E, Sorbera C, Mogilner A. Simultaneous use of bilateral subthalamic nucleus stimulators and an implantable cardiac defibrillator. Case report. Journal of Neurosurgery 2003; 99(1):167–169.
93. Black RD. Recent advances in translational work on implantable sensors. IEEE Sensors Journal 2011; 11(12):3171–3182.
94. Weiler A, Hoffmann RFG, Säthelin AC, Helling HJ, Südkamp NP. Biodegradable implants in sports medicine: the biological case. Arthroscopy. The Journal of Arthroscopic and Related Surgery 2000; 16(3):305–321.
95. Yetkin H, Senkoylu A, Cila E, Ozturk AM, Simsek A. Biodegradable implants in orthopaedics and traumatology. Turkish Journal of Medical Sciences 2000; 30:297–301.
96. Hwang S-W, Tao H, Kim D-H, Cheng H, Song J-K, Rill E, et al. A physically transient form of silicon electronics. Science 2012; 337:1640–1644.
97. Hwang S-W, Huang X, Seo J-H, Song J-K, Kim S, Hage-Ali S, et al. Materials for bioresorbable radio frequency electronics. Advanced Materials 2013; 25(26):3526–3531.
98. Alomainy A, Owadally AS, Hao Y, Parini CG, Nechayev YI, Constantinou CC, et al. Body-centric WLANs for future wearable computers. In: Proceedings of the First International Workshop on Wearable and Implantable Body Sensor Networks, London, UK, 2004.
99. Guyomar D, Badel A, Lefeuvre E, Richard C. Toward energy harvesting using active materials and conversion improvement by nonlinear processing. IEEE Transactions on Ultrasonics, Ferroelectrics and Frequency Control 2005; 52(4):584–595.
100. Mitcheson PD, Green TC, Yeatman EM, Holms AS. Architectures for vibration-driven micropower generators. IEEE Journal of Microelectromechanical Systems 2004; 13(3):429–440.
101. Otis B, Chee YH, Rabaey J. A 400uW Rx, 1.6mW Tx super-regenerative transceiver for wireless sensor networks. In: Proceedings of the IEEE International Solid Circuits Conference, CA, USA, 2005.
102. Neirynck D, Williams C, Nix AR, Beach MA. Wideband channel characterisation for body and personal area networks. In: Proceedings of the First International Workshop on Wearable and Implantable Body Sensor Networks, London, UK, 2004.
103. Yakovlev A, Kim S, Poon AS. Implantable biomedical devices: wireless powering and communication. IEEE Communications Magazine 2012; 5(4):152–159.
104. Kim S, Ho JS, Chen LY, Poon AS. Wireless power transfer to a cardiac implant. Applied Physics Letters 2012; 101(7).
105. Karami MA, Inman DJ. Powering pacemakers from heartbeat vibrations using linear and nonlinear energy harvesters. Applied Physics Letters 2012; 100(4).
106. Denisov A, Yeatman EM. Stepwise microactuators powered by ultrasonic transfer. Procedia Engineering 2011; 25:685–688
107. Xu S, Qin Y, Xu C, Wei Y, Yang R, Wang ZL. Self-powered nanowire devices. Nature Nanotechnology 2010; 5:366–373.
108. Au-Yeung KY, Moon GD, Robertson TL, Dicarlo LA, Epstein MS, Weis SE, et al. Early clinical experience with networked system for promoting patient self-management. American Journal of Managed Care 2011; 17(7):e277–287.
109. Heinzel A, Hebling C, Muller M, Zedda M, Muller C. Fuel cells for low power applications. Journal of Power Sources 2002; 105:250–255.
110. Dyer CK. Fuel cells for portable applications. Journal of Power Sources 2002; 106:31–34.
111. McGrath KM, Prakash GKS, Olah GA. Direct methanol fuel cells. Journal of Industrial and Engineering Chemistry 2004; 10:1063–1080.
112. Shao Z, Haile SM, Ahn J, Ronney PD, Zhan Z, Barnett SA. A thermally self-sustained micro solid-oxide fuel-cell stack with high power density. Nature 2005; 435(7043):795–798.

113. Sasaki S, Karube I. The development of microfabricated biocatalytic fuel cells. Trends in Biotechnology 1999; 17(2):50–52.
114. Soukharev V, Mano N, Heller A. A four-electron O (2)-electroreduction biocatalyst superior to platinum and a biofuel cell operating at 0.88 V. Journal of the American Chemistry Society 2004; 126(27):8368–8369.
115. Arra S, Heinisuo S, Vanhala J. Acoustic power transmission into an implantable device. In: Proceedings of the Second International Workshop on Wearable and Implantable Body Sensor Networks, London, UK, 2005; 60–64.
116. Lee SL, Ali K, Brizzi A, Keegan J, Hao Y, Yang GZ. A whole body statistical shape model for radio frequency simulation. In: Proceedings of the International Conference of the IEEE Engineering in Medicine and Biology Society, Boston, MA. USA, 2011; 7143–7146.
117. Hamdi M, Boudriga N, Abie H, Denko M. Secure wearable and implantable body sensor networks in hazardous environments. In: Proceedings of the International Conference on Data Communication Networking, Athens, Greece, 2010; 1–8.
118. Daniluk K, NiewiadomskaSzynkiewicz E. Energy-efficient security in implantable medical devices. In: Proceedings of the Federated Conference on Computer Science and Information Systems, Wrocław, Poland, 2012; 773–778.
119. Zhou L, Hass ZJ. Securing ad hoc networks. IEEE Network Magazine 1999; 13(6):24–30.
120. Ganesan D, Govindan R, Shenker S, Estrin D. Highly-resilient, energy-efficient multipath routing in wireless sensor networks. Mobile Computing and Communication Review 2002; 1(2):28–36.
121. Cherukuri S, Venkatasubramanian KK, Gupta SKS. BioSec: a biometric based approach for securing communication in wireless networks of biosensors implanted in the human body. In: Proceedings of the International Conference on Parallel Processing Workshops, Kaohsiung, Taiwan, 2003; 432–439.
122. El-Bendary N, Tan Q, Pivot FC, Lam A. Fall detection and prevention for the elderly: A review of trends and challenges. International Journal on Smart Sensing and Intelligent Systems 2013; 6(3):1230–1266.
123. Poh MZ, Swenson NC, Picard RW. A wearable sensor for unobtrusive, long-term assessment of electrodermal activity. IEEE Transactions on Biomedical Engineering 2010; 57(5):1243–1252
124. Patterson D, Fox D, Kautz H, Philipose M. Expressive, tractable and scalable techniques for modeling activities of daily living. In: Proceedings of the Second International Workshop on Ubiquitous Computing for Pervasive Healthcare Applications, Seattle, Washington, USA, 2003.
125. Philipose M, Fishkin KP, Perkowitz M, Patterson DJ, Fox D, Kautz H, et al. Inferring activities from interactions with objects. IEEE Pervasive Computing 2004; 3(4):50–57.
126. Kautz H, Etziono O, Fox D, Weld D. Foundations of assisted cognition systems. Department of Computer Science and Engineering, University of Washington, CSE-020AC-01, 2003.
127. Ravi N, Dandekar N, Mysore P, Littman ML. Activity recognition from accelerometer data. In: Proceedings of the Seventeenth Annual Conference on Innovative Applications of Artificial 2005.
128. Najafi B, Aminian K, Paraschiv-Ionescu A, Loew F, Bula CJ, Robert P. Ambulatory system for human motion analysis using a kinematic sensor: monitoring of daily physical activity in the elderly. IEEE Transactions on Biomedical Engineering 2003; 50(6):711–723.
129. Sung M, Pentland A. Minimally-invasive physiological sensing for human-aware interfaces. In: Proceedings of Human-Computer Interaction International, Las Vegas, Nevada, USA, 2005.
130. Bourke AK, Lyons GM. A threshold-based fall-detection algorithm using a bi-axial gyroscope sensor. Medical Engineering and Physics 2008; 30:84–90.
131. Thiemjarus S, Marukatat S, Poomchoompol P. A method for shoulder range-of-motion estimation using a single wireless sensor node. In: Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Osaka, Japan, 2013.

132. Tanyawiwat N, Thiemjarus S. Design of an assistive communication glove using combined sensory channels. In: Proceedings of the International Conference on Wearable and Implantable Body Sensor Networks, London, UK, 2012.
133. King R, Atallah L, Lo B, Yang GZ. Development of a wireless sensor glove for surgical skills assessment. IEEE Transactions of Information Technology in Biomedicine 2009; 13(5):673–679.
134. Siuru B, "Applying acoustic monitoring to medical diagnostics applications", Sensor Magazine, http://www.sensorsmag.com/articles/0397/acoustic/index.htm, 1997.
135. Liu J, Johns E, Attallah L, Pettitt C, Lo B, Frost G, et al. An intelligent food-intake monitoring system using wearable sensors. In: Proceedings of the Ninth International Conference on Wearable and Implantable Body Sensor Networks, London, UK, 2012; 154–160
136. Liden CB, Wolowicz M, Stivoric J, Teller A, Kasabach C, Vishnubhatla S, et al., Characterization and implications of the sensors incorporated into the SenseWear armband for energy expenditure and activity detection, www.bodybugg.com/pdf/Sensors.pdf.
137. Asada HH, Shaltis P, Reisner A, Rhee S, Hutchinson RC. Mobile monitoring with wearable photoplethysmographic biosensors. IEEE Engineering in Medicine and Biology Magazine 2003; 22(3):28–40.
138. Heller A. Drug-delivering integrated therapeutic systems. In: Proceedings of the Second International Workshop on Wearable and Implantable Body Sensor Networks, London, UK, 2005; 6–11.
139. Giorgiou J, Toumazou C. A 126 microWatt cochlear chip for a totally implantable system. IEEE Journal of Solid-State Circuits 2005; 40(2):430–443.
140. Abidi SS, Goh A. A personalised healthcare information delivery system: pushing customised healthcare information over the WWW. Studies in Health Technology and Informatics 2000; 77:663–667.
141. Backadar T. Ambulatory monitoring-embeddable, wearable, "its all about fashion" studies in wearable electronics. In: Proceedings of the Second International Workshop on Wearable and Implantable Body Sensor Networks, London, UK, 2005; 79–81.
142. Haes AJ, Chang L, Klein WL, Van Duyne RP. Detection of a biomarker for Alzheimer's disease from synthetic and clinical samples using a nanoscale optical biosensor. Journal of the American Chemistry Society 2005; 127(7):2264–2271.
143. Koh DM, Brown G, Temple L, Raja A, Toomey P, Bett N, et al. Rectal cancer: mesorectal lymph nodes at MR imaging with USPIO versus histopathologic findings–initial observations. Radiology 2004; 231(1):91–99.
144. Medintz IL, Clapp AR, Mattoussi H, Goldman ER, Fisher B, Mauro JM. Self-assembled nanoscale biosensors based on quantum dot FRET donors. Nature Materials 2003; 2(9):630–638.
145. Wang J. Nanomachines: fundamentals and applications. Wiley-VCH, 2013.
146. Campuzano S, Orozco J, Kagan D, Guix M, Gao W, Sattayasamitsathit S, et al. Bacterial isolation by lectin-modified microengines. Nano Letters 2012; 12(1):396–401.
147. Martel S. Journey to the center of a tumor. IEEE Spectrum 2012; 49(10):49–53.

Chapter 2
Biosensors and Sensor Systems

Danny O'Hare

2.1 Introduction

This chapter is concerned with the design and operation of devices for the measurement of chemical concentrations in living systems. Whilst the meaning of chemical sensors has been somewhat broadened in recent years, a useful definition from a recent authoritative review [1] is: "*chemical sensors are miniaturised analytical devices that can deliver real time and online information on the presence of specific compounds or ions in complex samples*". Historically, a useful distinction was made between chemical sensors and biosensors in that a biosensor used a biologically derived element (enzyme, antibody, cell, tissue sample etc.) as part of the transduction process. This useful distinction has now been lost; *biosensor* is now often applied to any sensor measuring a chemical concentration in a biological system. Jiri Janata makes a further distinction between sensors and sensor systems, sensors being capable of continuous monitoring and sensor systems providing measurements in discrete steps. Turner, in a recent tutorial review [2] identifies two broad categories of device described under the heading 'biosensor': sophisticated and high-throughput lab-based instrumentation for rapid and accurate analysis of complex biological interactions and components, and portable easy-to-use devices for non-expert use, out of the lab, in the home or in a field environment, 'point-of-care' devices. This chapter will describe both chemical and biosensor (in the historical sense) approaches to the identification and quantification of chemical species in living systems and describe both sensors and sensor systems as described above. It will also outline recent developments, notably new materials with great potential for sensor applications. The underlying principles of the most widely used sensor types are described to enable their proper operation, design of

D. O'Hare (✉)
Department of Bioengineering, Imperial College London, London SW7 2BY, UK
e-mail: d.ohare@imperial.ac.uk

G.-Z. Yang (ed.), *Body Sensor Networks*, DOI 10.1007/978-1-4471-6374-9_2,

control instrumentation, optimisation of sensor design and correct interpretation of the resulting data. Exemplar applications are given to illustrate these principles.

Whilst development and application of chemical sensor and biosensor technology have been major research activities for decades, demographic and economic pressures in the developed world have provided an increased market pull and new opportunities for research and technology. Advances in medical science have coincided with increased life expectancy, a substantial concomitant increase in the proportion of elderly people in the population and massive improvements in the treatment and management of diseases associated with old age. Regardless of the healthcare economics, it is highly likely that there will simply be an insufficient amount of young people to meet the medical and clinical needs of the elderly. Furthermore, as has been demonstrated in the case of Type 1 diabetes, more frequent monitoring, especially outside the clinical environment, leads to an improved quality of life, fewer demands on acute healthcare and lower incidence of the vascular complications associated with diabetes in the long term. Blood glucose monitoring remains an important market, accounting for 85 % of all biosensor sales. Turner [2] claims that around $100 million is required to bring a device to market, given the investment required in high technology plant. At the moment, only blood glucose and pregnancy testing has the volume to attract this level of investment and just four companies (Abbott, Bayer, Roche and Johnson & Johnson) take more than 90 % of the blood glucose market.

The setting for sensor use dramatically alters both the healthcare economics and the technological challenges: it makes a huge difference whether the device is to be operated by a lay person, possibly the patient herself, or by a healthcare professional. Aside from personal use, the most obvious application is in critical care monitoring. The requirements for sensors in critical care have been usefully reviewed by Moore et al. [3]. The rather optimistic wish list is summarised below:

- An accurate and stable sensor for an essential variable; it must be accurate and stable indefinitely. Physicochemical detection must be reversible to minimise sensor recovery and response times;
- Non- or minimally-invasive;
- Continuous monitoring function with the ability to display trends;
- Easy to use and a display that is easily understood;
- Small size and weight;
- Ruggedness and transportability.

In addition to the rather challenging engineering specifications, Moore also points out that the economic value of continuous monitoring has not yet been demonstrated.

An important additional driver is the increased realisation that individual patient responses to drug therapy can vary greatly. This is a particularly acute problem in cancer chemotherapy where time is obviously pressing, but can be important in chronic management of a host of conditions. For example, the widespread use of statins in the management of atherosclerosis and cardiovascular disease is hindered by idiosyncratic adverse effects including myositis and myopathy and, rarely, rhabdomyolysis. Such variable responses to drug therapy in many cases are believed to be due to underlying genetic differences. Rapid identification of an

individual patient's "ideal drug" is therefore important and provides a further application for sensor technology.

Before beginning an exposition of the basic science and technology underpinning sensor design, some overview of the motivations and unique problems posed by biological systems is required.

2.2 Bioanalysis

Recent decades have seen rapid developments and great technological achievements in bioanalysis including the sequencing of the entire human genome. Though the dream was to produce biology's equivalent of the periodic table of the elements [4, 5], the situation has turned out to be more complex. The somewhat surprising discovery of the relatively small number of genes (around 30,000) and the growing realisation of the complex networks involved have resulted in fewer direct applications to human health than originally anticipated. Nonetheless, there has been real insight, which is certain to inform healthcare and likely to lead to new opportunities for medical devices. Proteomics and metabolomics [6] have similarly improved understanding of physiological and metabolic pathways at the cellular level and have led to a more quantitative engineering approach – systems biology – which has the potential to give a quantitative functional understanding of the repertoire of cellular behaviour. DNA sequencing is becoming faster and cheaper with each passing year and new technology on the horizon – particularly nanopore devices – hold out great promise for better fundamental understanding of cellular function in health and disease and also provide great opportunities for personalised medicine. All of these approaches are likely to provide new approaches to drug discovery [7, 8], lead compound identification [9] and to inform personalised approaches to drug therapy [10].

Bioanalysis presents several important problems to the analytical scientist. These range from the difficulty in correctly identifying the research question to more mundane but nonetheless technically challenging issues around materials, stability and calibration. The first problem is to identify the motivation for measuring chemical concentrations.

This field is essentially multidisciplinary, and technical terminology regrettably differs between the different traditional disciplines that contribute to sensor research. This chapter follows the conventions of analytical science, and key jargon is defined below:

2.2.1 Some Jargon

- ***Analyte:*** the target molecular species. This is the molecule we wish to identify and quantify.
- ***Matrix:*** everything else present in our sample apart from the analyte. Matrix interference, where sensor response is inadvertently elicited by so-called

spectator species present in the sample, is a major problem in bioanalysis due to the complexity of biological systems. Matrix interference can usefully be divided into two categories: (i) signal caused by non-target molecules, e.g., ascorbic acid or paracetamol, both of which are oxidized at potentials used for the measurement of glucose, and (ii) sensor inactivation or 'poisoning' due to adsorption of proteins or other surface-active material. For example, albumin is typically present at around 4 % (w/v) in plasma. Adsorption of albumin can occlude or scatter light transmission from fibre optic sensors or prevent electrocatalysis on amperometric or voltammetric sensors (*vide infra*).

- ***Sensitivity:*** the change in sensor output per unit change in analyte concentration. This could be measured in μA mol^{-1} dm^3 for an amperometric biosensor or absorbance units per mole for a spectrometric device.
- ***Limit of Detection (LOD):*** the level of analyte that leads to a sensor signal which is statistically significantly different from the background signal obtained in the absence of analyte. A frequently used definition of LOD is a concentration that gives a signal greater than three times the standard deviation of a blank sample consisting entirely of matrix. This can often be most easily calculated from a calibration-working curve, as the concentration is equivalent to three times the standard deviation of the ordinal intercept. A signal at LOD cannot be related to a specific concentration and can only infer that the target analyte is present, to a given probability, i.e., the sensor is not merely giving out noise. If a numerical concentration is required, the limit of quantitation must be defined, and is typically five times higher than the LOD. Standard deviation of the background is an essential specification of any sensor as this can be applied, multiplied by a suitable coverage factor, by any user according to their needs.
- ***Selectivity:*** the ability of the analytical method to respond only to the target analyte. Strict determination of selectivity is frequently a problem in bioanalysis where the real instantaneous composition of the sample cannot be determined in practice. The expected error can be quantified for each expected interferent as:

$$Maximum\ \ error = \frac{Effect\ \ of\ \ interferent}{Effect\ \ of\ \ analyte} \times 100\% \tag{2.1}$$

where the interferent concentration is maximum expected in the sample and the analyte concentration is the minimum expected. Where there are more than one expected interferents, these can be summed (assuming no synergistic effects). This approach works well for well-characterised analytes such as human blood samples where the composition is relatively well known from decades of measurements over millions of individuals. It is disappointing that relatively few published papers take the trouble to calculate this key parameter, which is obviously essential in determining fitness for purpose. Where sufficient data is available to enable its calculation, the result is frequently disappointing for the aspirant investigator.

Other key parameters are response times (usually quantified as 10–90 % response time to a step change in concentration, or the time constant of a presumed monoexponential response to a step change in concentration) and, frequently ignored, the recovery time or reversibility of response. Rapid reversibility is usually a contradictory requirement to high selectivity due to the nature of the interactions involved; molecular level selectivity typically requires the analyte to approach within a bond length of the sensor.

2.2.2 *Bioanalysis – What Does Chemical Concentration Mean in Biology?*

Concepts of chemical concentration – the amount of material either by mass, moles or number of particles per unit volume – derive from chemistry which takes place typically in volumes of cm^3 and upwards with at least concentrations that are millimolar. We are, therefore, typically considering at least 10^{16} particles, and statistical fluctuations from one region to another are negligible. Concepts based on the essentially homogeneous nature of such solutions translate poorly to biology where heterogeneity and compartmentalisation exist on every length scale from the whole organism down to the subcellular domain. Cells are only able to sense and respond to their immediate environment. A typical mammalian cell is around 10 μm in diameter and if the target analyte concentration is sub nanomolar, which is frequently the case, then fewer than 1,000 molecules will be present. Under these circumstances, assumptions of homogeneity and the law of mass action simply cannot be valid. We would expect, and indeed do find, considerable variations in concentration when measuring at the level of single cells. Some of this variation must be due to natural biological variation (cells are not mass produced according to six sigma processes!) but, inevitably, much of the variation must be due to stochastic variations in the number of signalling molecules detected by the cells themselves.

A further complication is that many biologically significant molecules are produced and consumed at discrete locations (membrane bound enzymes and transporter proteins), which are small even on the length scale of the single cell. The precise *concentration* recorded will depend heavily on the precise position and the sampling volume. In living creatures it is usually not possible to be sure which part of the plume of concentration the sensor is responding to. Heterogeneity and localised sources and sinks greatly complicate the interpretation of concentration data which, for in vivo measurement requires prior definition of the length scale of interest and engineered devices that can deliver the appropriate resolution.

A key decision that the scientist needs to take is whether to attempt the measurement in situ, on-the-fly in real time or to remove what one hopes is a representative sample and preserve it for more leisurely analysis under more controlled conditions. In situ measurements in the physiological system can be

both influenced by and influence the response to the measurement device or probe. The immune system will normally mount a foreign body response [11], initially consisting of protein (and other biomolecule) deposition, which can impair device function. Fibrous capsule formation may seal off the measurement device over a period of several hours to several days and acute inflammatory responses may mean that the results are dominated by the effects of the measurement itself. Some methods of separation of the biological analytes can therefore be a potentially simplifying procedure in bioanalysis and for point-of-care applications; lab-chip devices are obviously attractive options [12].

Rapid response times are also required in vivo due to the time varying concentrations arising from continuous interactions in the system. Good temporal resolution is frequently inimical to stability and noise rejection and optimisation is always necessary. In contrast, the *ex situ* experiment requires isolation and preservation of the sample. Such destructive analysis necessarily limits the sample size – there are absolute limits of sample size if the patient is to survive the measurement.

Limited sample size is an important motivation for the development of miniaturised systems. Separation, clean-up and the preparation of, usually, a homogeneous solution greatly simplifies the chemical measurement step, though preserving the integrity of the sample and ensuring that analyte recovery is high and consistent are not trivial engineering challenges. However, the problems of the representative sample and the relevance to the underlying physiology remain. The conventional analytical approaches of using certified reference materials or standard solutions to validate the analytical measurement and the individual steps often cannot be applied to the overall process. Standard addition may be possible on isolated tissue, but usually cannot be attempted on a living organism.

In summary, there are typically three motivations for measuring chemical concentrations in living systems: (i) statistical correlation with expensive history-taking or known disease condition as an aid to diagnosis, for example rapid detection and quantification of known biomarkers; (ii) legal requirements, such as the strict limits on blood-alcohol levels for driving in most jurisdictions and (iii) fundamental research in physiology and biochemistry. Whilst the specifications for (i) and (ii) are relatively straightforward, new devices and new understanding of physiology require an essentially collaborative approach between clinicians, biologists and engineers and physical scientists to synthesise a wholly new approach to understanding biological functional behaviour in health and disease.

The transduction of chemical concentration into an electrical signal (usually a voltage), which can be recorded and interpreted, can be based on any of the fundamental properties of the molecules under investigation. One obvious characteristic is the interaction with electromagnetic radiation in the form of (i) infra red, which is characteristic of the bonds, and (ii) visible or ultraviolet light, characteristic of the molecular orbital energies. The advantages include a rigorous physical relationship between the absorption of radiation and the quantum mechanical properties of the molecule under investigation.

2.3 Molecular Recognition

Amperometric and voltammetric devices can select the analyte by applying an appropriate electrode potential. This can be further refined by using more advanced signal processing routines, also described below in Sect. 2.4.4. Similarly, recognition is intrinsic to the development of the membrane potential and the core technology employed in ion selective electrodes. This is typically achieved by ion exchange, neutral carriers or glassy materials. Molecular-level recognition for a dynamic device such as an amperometric sensor is not efficient if simply based on a "lock and key" mechanism. Some larger order change in the receptor molecule must be detected. This need not be in the molecular structure, though this helps, but could, for example, be a change in the free energy relationships as exploited in ion selective electrodes. All intermolecular forces can be important in recognition ranging, in order of increasing magnitude, from London dispersion/Van der Waals forces, dipole interactions, hydrogen bonding, ion-dipole interactions to electrostatic attraction. The range and distance dependence of these forces varies greatly. Ideally, these interactions should be rapid and reversible to reduce the impact of history on the sensor. If the receptor entity changes its structure as a result of binding the target, this can be detected by, for instance, surface plasmon resonance, changes in ion current through a nanopore and changes in capacitance or conductance. Alternatively, the recognition element can catalyse a reaction whose products can be detected by a chemical sensor.

As the number of potential targets increases, selectivity becomes more challenging and engineering the interface between the sensor and the tissue or solution sample becomes essential. There is a range of strategies that can be adopted. Historically the first of these was to use enzymes. Millions of years of evolution have led to exquisite selectivity, for example the first enzyme to be exploited in a biosensor, glucose oxidase from *Aspergillus niger* only catalyses the reaction of oxygen with the β-anomer of the D isomer of glucose. However, most enzymes are more promiscuous and certainly less stable, factors which have led to undue optimism in the early stages of biosensor development. Methods for engineering selectivity at the interface are summarized below:

Gas Selective Membranes – If the target analyte is a neutral molecule and the interferent is ionic, then interposing a gas permeable membrane such as PTFE (Teflon) between the test solution and the sensor will prevent the ionic species reaching the working electrode. The condition for this is that the effective pore size must be below two diameters of a water molecule. Ions can go nowhere without their accompanying water molecules that solvate them. A complication with this strategy is that the counter and reference electrode must also be behind the membrane since ions are also the charge carriers between the counter and working electrodes. This principle was first reduced to practice by Leland Clark for what is now universally known as the Clark O_2 electrode. The Clark electrode has been the method of choice for determining blood oxygenation since the late 1950s.

Selective Binding and Catalysis – Should the target analyte be oxidised or reduced at a similar potential to an interfering species, exploiting some selective chemistry of the target species can sometimes be successful. A typical example of this is the nitric oxide sensor first reported by Malinski [13] who used a Ni(II) (porphyrin) modified electrode surface to reduce the operating potential for oxidation of NO.

Amperometric Enzyme Electrodes – The key idea is to exploit the extraordinary selectivity of enzymes, which evolved over millions of years of natural selection. In these systems, there is no direct oxidation of the target analyte by the electrode. The analyte reacts catalytically with the enzyme to produce a reaction product which is then detected. The so-called "first generation" biosensors operate on this basis, the first reported example of which for the determination of glucose was published by Updike and Hicks in 1967 [14]. This approach was commercialised successfully by Yellow Springs Instruments. The underlying chemistry is shown below:

$$Glucose + O_2 \rightarrow Gluconolactone + H_2O_2 \tag{2.2}$$

The enzyme glucose oxidase (GOD) present, initially with its co-factor flavin adenine dinucleotide (FAD) in its oxidised form, is reduced in the process to GOD/$FADH_2$. Re-oxidation of the enzyme in nature is achieved by oxidation by dissolved O_2 which in turn is reduced to hydrogen peroxide. The hydrogen peroxide (H_2O_2) is detected by oxidation on a platinum electrode held at +0.65 V. The enzyme (or rather its cofactor) is reduced in the process and needs to be re-oxidised. The enzyme is immobilised by cross-linking with glutaraldehyde or by an electropolymerised film [15] or even by simple adsorption. An even simpler strategy can be employed where the working electrode is made of a conducting composite material. With the addition of suitable stabilisers such as polyethlyeneimine or dithiothreitol, enzymes can be incorporated into the bulk of the conducting carbon-epoxy composite to provide a cheap, extrudable or printable biosensor [16]. There are several comprehensive reviews of enzyme immobilisation techniques [17]. An ingenious molecular level assembly has been described by Willner [18] where the flavin redox centre is first immobilised followed by spontaneous self-assembly of the apoenzyme onto its co-factor.

The major problem with the first generation biosensors is that there are several common interferents which are also oxidised at +0.65 V, notably uric acid, ascorbate and acetaminophen. An alternative strategy was adopted for the second generation biosensors where it was recognised that the oxygen in the above reaction is in fact regenerating the enzyme. This is shown schematically in Fig. 2.1.

In this figure, the electrode is on the right hand side of the diagram, the test solution on the left. Substrate diffuses from solution (Step 1) through a membrane (where employed) (Step 2) to be oxidised by the enzyme (Step 3). The enzyme must be reduced in this process and needs to be regenerated by oxidation in Step 4. The mediator is then regenerated, in turn, by oxidation at the electrode surface (Step 5). For a concentration sensor, Step 1 or Step 2 needs to be the rate determining step. This ensures that the slope of calibration is not affected if the enzyme denatures

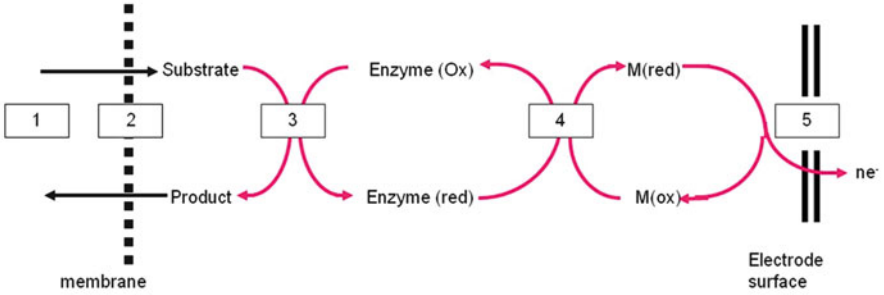


Fig. 2.1 A schematic of an enzyme-mediated electrode reaction showing the coupled mass transport and reaction steps. *1* Mass transport of the target analyte (which is also the enzyme substrate) in the bulk substrate. *2* Permeation through the membrane (if present). *3* Reaction of enzyme in its oxidised form with the substrate. *4* Regeneration of the oxidised enzyme by reaction with oxidised mediator. *5* Re-oxidation of the mediator at the electrode leading to current in the external circuit

slightly or loses activity. The mediator species can be chosen so that it undergoes fast reversible reaction kinetics at a potential where no other redox species are expected to react. Mediators which have been employed for this purpose include benzoquinone, the ferricyanide ion and various derivatives of the iron(II) compound ferrocene. The ethanolamine derivative of ferrocene is the mediator in the enormously commercially successful biosensor for glucose originally developed by Medisense, the ExacTech system. This concept was originally described by Cass et al. [19] using dimethyl ferrocene as a mediator. In this device, the enzyme was chemically immobilised on the surface of a screen printed carbon electrode. It would obviously be a lot simpler if the enzyme could be persuaded to react directly at the electrode surface. This cannot generally be achieved because the electron conduction path between the electrode surface and the redox centre of the enzyme is too great for there to be an appreciable tunnelling current. Third generation biosensors involve no directly added mediator species. There have been two broadly successful approaches – using electrodes made of low dimensional conducting charge transfer salts of tetracyanoquinodimethane (TCNQ) and redox wired enzymes. The former strategy was first described by Kulys and developed by Albery et al. – the most successful compound being the charge transfer salt of TCNQ and tetrathiafulvalene (TTF) (Fig. 2.2).

The mechanism was the subject of heated dispute for some time, it being believed that direct electron transfer was occurring. However, Bartlett was able to show that the TTF insinuates its way into the enzyme structure [20] to enable electronic conduction. A detailed mechanism for electrodes made from these materials has more recently been published by Lennox [21] showing that the mechanism is best understood as a form of heterogeneous mediation, where the mediator species is not soluble in water, but is soluble in the hydrophobic regions of the enzyme. Electrodes based on this technology have been used successfully for long-term studies of glucose metabolism in rats' brains over 10 days [22]. Wired enzymes tackle the problem more directly. Reactive sites in the protein structure are identified (or created by protein

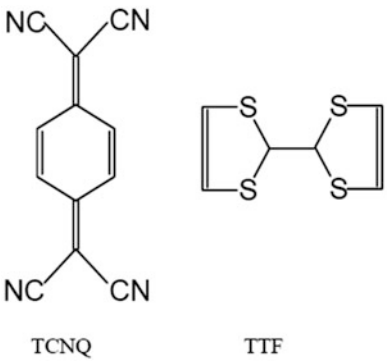


Fig. 2.2 Molecules involved in the classic 'third generation' biosensor

engineering) and reacted with redox active groups such as ferrocene derivatives, an approach now of some commercial significance and originally pioneered by Adam Heller's group [23]. This technology is now being applied with some success to power generation in biofuel cells by Heller [24].

Aptamers are a new class of molecules for recognition and binding based on oligomers of nucleotides, both RNA and DNA [25–27]. Typically 30–50 nucleotides long, they are developed by in vitro selection from libraries of random sequences in a process called SELEX: systematic evolution of ligands by exponential enrichment, first described in 1990 [28]. These binding entities have similarities to antibodies, but also notable advantages including thermal stability, ease of synthesis of appropriate sequences (using the polymerase chain reaction). They have a necessarily lower molecular weight than antibodies and can therefore be loaded onto surfaces (and sensors) at higher molar concentrations which, to some extent, can mitigate a lower binding constant (though extraordinarily high binding affinities can be and have been achieved [29]). Since aptamers are typically linear oligomers with well-defined end-group chemistry, conformation can be easily controlled and effective immobilisation is simpler than antibodies. Unlike antibodies, however, they recognize their targets by undergoing target induced conformational change. This conformational change lends itself naturally to FRET-type detection or fluorescent probe-quencher combinations, but can also be employed in electrochemical aptamer sensors (see Chap. 3) to take advantage of the very strong distance dependence of the tunneling probability in electron transfers. Several recent reviews cover the application of aptamers in biosensing [30–32] including their combination with nanotechnology [33].

Molecularly-imprinted polymers are a wholly synthetic approach. The essential idea is to prepare a polymer film with cavities that selectively accept the target molecule. These polymers are prepared from a monomer in the presence of the target, acting as a template, which is then washed out. The advantages are obvious: non-biological molecules are potentially more stable in storage and use than antibodies or enzymes; the methodology ought to be generalisable, versatile and cheap and the polymers can be prepared under sterile conditions and are likely to be more stable under common sterilisation conditions. Microbeads and nanofibres can

be prepared as well as films, giving rise to a reagent which can be produced in bulk and implemented in a sensor as needed. Applications to sensors in general [34, 35] and to electrochemical sensors in particular [36] have critically reviewed. Electrochemical approaches to polymerisation specifically for sensor applications represent a new approach and this has been reviewed recently [37].

Once molecular recognition has been achieved, the resulting change needs to be transduced into a signal that can be recorded, typically a voltage. Many different technologies have been employed in this role and the subject of recent reviews, for example: optical or photonic devices frequently based on optical fibres [38], thermistors and piezoelectric crystals either in shear mode (detecting mass changes) or using the surface acoustic waves. In addition to these well-established technologies, new sensing modes have arisen in recent years including nanoplasmonics [39]; field-effect organic transistors (ChemFET) [40]; microcantilevers based where target binding induces either a change in resonant frequency or a characteristic deflection [41], and most recently FET sensors based on grapheme [42]. Real world applications are thin on the ground whilst the biocompatibility and reproducible processing of such a new material is quantified. However, point-of care devices have been reported for prostate-specific antigen [43] and a flexible glucose sensor using CVD graphene has been produced [44]. The advantages of graphene are not yet apparent. The principal problem with all biomolecule-based devices remains the poor stability of the biomolecule itself. Therefore, these papers may be describing solutions to the wrong problem, unless the manufacturing process is greatly simplified and the cost of volume production in short order is addressed. The physiological milieu is an astonishingly hostile medium to work in and its heterogeneous structure, abundance of surface active and highly light absorbing structures has limited the range of applicable transduction modes to FETs, fibre-optic devices and electrochemical sensors.

FETs and fibre-optic devices have made substantial inroads in critical care sensing in recent years [45] (though the healthcare economics remains doubtful except for high value, high volume applications. Bedside optics cost upwards of $20,000, the probes themselves, which must be disposable cost >$300 each) but the dominant technology for implantable devices remains electrochemical and this is the subject of this chapter. A recent review of electrochemical biosensor technology can be found in [46].

2.4 Electrochemical Sensors

Electrochemical methods, electrochemical transduction and electroanalysis offer many advantages for biological measurement. The underlying physical description of electrochemical phenomena are beyond the scope of this chapter, but many excellent introductory [47] or more comprehensive texts [48–50] exist.

2.4.1 Potentiometry

Potentiometry is the electroanalytical technique where an electrode potential or membrane potential relative to a suitable reference electrode can be related to the analyte concentration, or strictly its thermodynamic activity. Potentiometric devices have the following characteristics:

- The system is at equilibrium; no current is passed and the analyte is not consumed;
- Instrumentation is simple, all that is required is a reference electrode and a high impedance voltmeter;
- Selectivity is inherent. The selective element is typically an ion selective membrane or a metal oxide coating;
- The response is logarithmic. Typically, there is a 59 mV change in output per decade change in concentration. This leads to a large linear dynamic range but leads to poor sensitivity in electromagnetically noisy environments.
- The practical dynamic range is hard to define since it is determined by deviations from the log-linear response at both ends of the calibration working curve.

The majority of potentiometric devices are based on the exploitation of equilibrium potentials at selective membranes. Such devices that have found to have broad applicability characteristics are, in many cases, commercially available, e.g. the glass pH electrode and microelectrodes have been used in biological media for decades and several tutorial and introductory books and chapters have been published [51–54].

2.4.1.1 Underlying Principles of Operation

Classically, potentiometry considers reduction-oxidation (redox) equilibria of the following kind:

$$M^{n+} + ne^- \rightleftharpoons M \tag{2.3}$$

Such redox reactions, usually coupled with a solubility equilibrium between the metal ion M^{n+} and some anion, e.g., Cl^- usually only find analytical use as reference electrodes (*vide infra*).

However, most analytically useful devices are based on membrane equilibria and are known as ion selective electrodes, ISEs. These have their origin in the glass pH electrode, found in thousands of general chemistry, biology and pathology labs worldwide since its original description by Cremer in 1906 [55].

The most commonly used ion selective membranes are glass and crystalline solids (for example europium (III) – doped LaF_3 for fluoride selective electrodes) that are supported liquid ion exchangers or ionophores (ion-binding molecules) dissolved in polymers with suitable plasticisers.

Regardless of their physical embodiment, all ion selective electrodes are based on thermodynamic equilibrium across an ion selective membrane. Understanding of the underlying phenomenon is essential if real devices are to be operated successfully. The starting point is that the free energy of all species in all phases must be the same throughout the system for the system to be at equilibrium. The key parameter for ions, which bear an electrical charge, is their *electrochemical potential* $\overline{\mu}$, which is the sum of the partial molar free energy, i.e. the chemical potential μ and the electrical potential ϕ appropriately scaled from electrician's units (volts) to chemist's units (joules per mole) using Faraday's constant, F the charge in coulomb on one mole of electrons.

$$\overline{\mu}_i = \mu_i + zF\phi \tag{2.4}$$

where z is the formal charge on ion i. If we have two solutions α and β separated by an ion selective membrane, the equilibrium condition is:

$$\overline{\mu}_{i(\alpha)} = \overline{\mu}_{i(\beta)} \tag{2.5}$$

Expansion of the electrochemical potential terms leads to:

$$\mu^{\circ}{}_{i(\alpha)} + RTlna_{i(\alpha)} + zF\phi_\alpha = \mu^{\circ}{}_{i(\alpha)} + RTlna_{i(\alpha)} + zF\phi_\alpha \tag{2.6}$$

where subscripts α and β refer to the two solution phases α and β. a_i is the activity of ion i which can be approximated for infinitesimal concentrations to the concentration of the ion. Rearrangement gives an expression for the membrane potential, ϕ_m, the electrical potential difference between phases α and β arising from different ionic activities:

$$\phi_m = \phi_\alpha - \phi_\beta = \frac{RT}{zF} ln\left(\frac{a_{i(\alpha)}}{a_{i(\beta)}}\right) \tag{2.7}$$

Provided therefore that we control the ionic activity on one side of the membrane, the membrane potential relative to a suitable reference electrode will depend only on the ionic activity of the ion in the outside solution. At this stage, it is worth pointing out that the membrane potential will only *free ion concentration* not total ion concentration. This is a unique selling point of this technology and ensures that ISEs will continue to be used, as the ratio of free: total ionic concentration continues to be of interest and biological relevance.

Implementation of a practical device exploiting this phenomenon therefore simply requires (i) reference electrodes to sense the solution potential (ii) a selective membrane and (iii) a high impedance voltmeter. This is shown schematically in Fig. 2.3.

The *reference electrodes* are designed such that the galvani potential difference *between* the terminal connection at the top of the electrode and the solution is essentially independent of the composition of the solution. By far the most widely

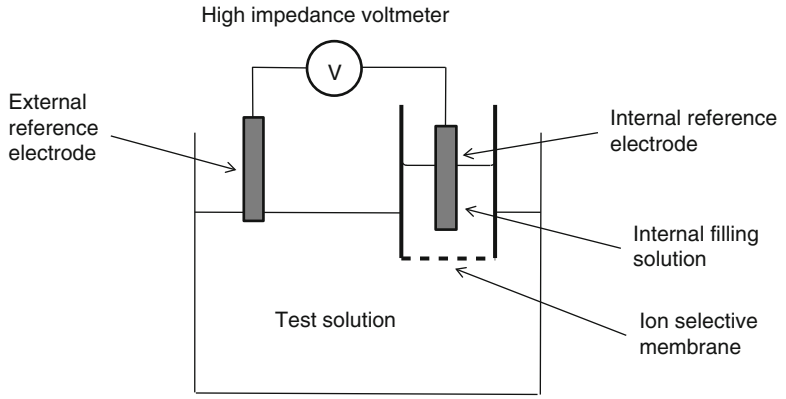


Fig. 2.3 Schematic of an ion selective electrode set-up

used reference electrode in both potentiometry and in amperometry is the silver|silver chloride electrode. This important electrode is based on the following equilibrium:

$$AgCl(s) + e^- \rightleftharpoons Ag(s) + Cl^- \tag{2.8}$$

At equilibrium, the electrochemical potentials of the reactants and products must be equal:

$$\overline{\mu}_{AgCl} + \overline{\mu}_{e^-} = \overline{\mu}_{Ag} + \overline{\mu}_{Cl^-} \tag{2.9}$$

$$\therefore \mu_{AgCl} + (\mu_{e^-} - F\emptyset_M) = \mu_{Ag} + (\mu_{Cl^-} - F\emptyset_s) \tag{2.10}$$

The chemical potential of the chloride ion is given by:

$$\mu_{Cl^-} = \mu^\circ{}_{Cl^-} + RTln([Cl^-]) \tag{2.11}$$

where $[Cl^-]$ is the concentration (strictly activity) of the chloride ion. For pure solids AgCl and Ag, the chemical potentials are equal to their standard chemical potentials. This means that the potential difference between the metal and the solution is given by:

$$\emptyset_M - \emptyset_S = \frac{\Delta\mu^\circ}{F} + \frac{RT}{F} ln\left(\frac{1}{[Cl^-]}\right) \tag{2.12}$$

where $\Delta\mu^\circ$ is a lumped constant of the standard chemical potentials of the silver metal, chloride ion, electron and silver chloride solid. The key consequence of this equation is that, provided that we keep the chloride ion concentration constant, the potential difference between the metal and the solution for this electrode will be constant. This is achieved in practice by placing a chloridised silver wire in a

solution of potassium chloride of a known concentration inside a small glass or plastic tube. Electrical connection to our test solution or internal filling solution is via ionic contact through a fritted glass disc made of either Vycor or a porous polymer that allows ion transport but prevents bulk fluid flow. Therefore, if, as in Fig. 2.3 above, we place a silver-silver chloride electrode either side of the membrane, any change in the measured potential must be due to changes in the membrane potential.

Reference electrodes remain more of a problem for biological application than is generally acknowledged in the literature. Screen-printed Ag | AgCl is cheap enough to be considered disposable and is a scalable technology [56]. Idegami et al. were able to incorporate electrolyte in a sodium alginate paste in the screen-printing process to produce the complete device. Alternative materials have been explored and amongst the most promising are using silver tetramethylbis (benzimidazolium) diiodide [57] and nanoporous platinum solid state layer-by-layer polyelectrolyte junction that have been reported for lab-chip applications and show great promise [58].

The experimental set-up sketched out above will give an idealised logarithmic response to changes in thermodynamic activity that can be represented as:

$$E = E^{\circ} + \frac{RT}{zF} lna_i \tag{2.13}$$

where a_i is the thermodynamic ion activity. This can, under suitable circumstances, be approximated to concentration but it should always be remembered that the underlying response is to activity, not concentration, and that this places constraints on the experimenter.

2.4.1.2 Calibration

The preparation of activity standards is not usually an option for biomedical application of ion selective electrodes. Activity depends strongly on the chemical composition of the sample matrix, especially the ionic strength, I.

$$I = \frac{1}{2}\sum_i c_i z_i^2 \tag{2.14}$$

where c_i is the concentration o ion I and z_i is the absolute charge, including the sign. The relationship between activity and concentration is given by:

$$a_i = \gamma_i c_i \tag{2.15}$$

where γ_i is the ion activity coefficient, essentially a fudge factor that corrects for the non-ideal behaviour of solutes at real concentrations. As the concentration approaches zero, ideal behaviour is better approximated and the ion activity

Table 2.1 Examples of the divergence of activity from concentration for physiologically important ionic species

Concentration/mol dm^{-3}	γ (K^+)	γ (Ca^{2+})	γ (Cl^-)
0.01	0.903	0.675	0.903
0.05	0.820	0.485	0.805
0.1	0.774	0.269	0.741
0.2	0.727	0.224	0.653

coefficient tends to unity. Concentration deviates from activity to a greater extent the greater the ionic strength. Approximate values of $\gamma_{\pm}$, the mean activity coefficient, can be estimated from the ionic strength using the Debye-Hückel theory. For low (sub-millimolar) concentrations, the following relationship can be applied:

$$log_{10}\gamma_{\pm} = -A|z^{+}z^{-}|\sqrt{I} \tag{2.16}$$

where z^+ and z^- are the charges on the cation and anion respectively. Higher concentrations require the use of the Debye-Hückel extended law. However, for real analytical situations, matrix composition cannot usually be determined. Large and potentially important deviations from unity occur at relatively low ionic strengths. Table 2.1 shows the potential extent of this problem for common ions.

The ionic strength of typical biomedical specimens is around 0.2 mol dm^{-3}. Failure to take variations in ionic activity into account can lead to serious errors. This can be true for apparently well-established measurement such as the determination of pH; the ion activity coefficient for H^+ at 37 °C and the physiological ionic strength around 0.83 [59].

So how can selective membrane potentials (at the heart of both ISEs and ion selective field effect transistors) be exploited to determine chemical concentrations? The key is to ensure that calibration solutions have the same ionic strength as the unknown sample. This is relatively easy to achieve for physiological fluids where the ionic and protein compositions are known from decades of measurements over millions of patients. Combining Eqs. 2.13 and 2.15 above yields:

$$E = E^{\circ} + \frac{RT}{zF} lnc_i + \frac{RT}{zF} ln\gamma_i \tag{2.17}$$

so it can be seen that, provided the ionic strength is kept constant, then the measured cell potential can be related to the concentration. This can be achieved in the chemical pathology or clinical biochemistry lab by preparing concentration standards (i.e. solutions of known concentration used to prepare calibration working curves) in *total ionic strength adjustment buffer* (TISAB). This will have a much higher ionic strength than the analyte and usually contain a pH buffer too. Provided that I (TISAB) $>> I$ (analyte) and that the analyte is also diluted with TISAB, the cell potential can be approximated as:

$$E_{cell} = K + \frac{RT}{zF} lnc_i \tag{2.18}$$

where K is a constant that now includes the activity coefficient. For historical reasons, the response of an ISE is usually reported in base 10 logarithm as:

$$E_{cell} = K + \frac{2.303RT}{zF} log_{10} c_i \tag{2.19}$$

At 25 °C this means that the measured cell potential changes by 59 mV per tenfold change in concentration for a singly-charged analyte.

Typically 6–10 calibration standards are prepared, most accurately achieved by dilution with TISAB from a freshly prepared stock solution. The concentration range should bracket the expected concentration range of the unknown samples and be centred on the expected mean value, to minimise confidence limits from the linear least squares fit.

As always with analytical measurement, some quality control is always required. This can involve the "blind" analysis of *certified reference materials* (CRM) where available. Alternatives to the use of CRMs include analysis of one or more samples using a method dependent on a different physical principle or use of standard additions.

The method of standard additions is an important alternative to the use of calibration working curves, especially for ion selective electrodes. In complex or poorly-characterised media, it may not be possible to measure the concentration of, and characterise the response, to all potential interferents. In addition, ISE performance may be impaired by "spectator species" such as plasma proteins. In this method, the analyte sample is added to TISAB and the ISE potential is measured. An aliquot of standard solution, also prepared in TISAB is then added to the prepared sample and the new potential is recorded. Equation 2.18 is then employed. Two potentials and two unknowns enable calculation of the initial concentration in the prepared analyte, enabling calculation of the concentration in the original specimen.

The use of TISAB effectively deals with the thermodynamic response of the ISE membrane. For in vivo or in situ use however there are no easy fixes. Frequently, it may be reasonable to infer a quantitative or semi-quantitative *change* in concentration from an implanted ISE response, but care must be taken to avoid over-interpretation of experimental data.

The logarithmic response of ISEs has important consequences. The dynamic range of ISEs is typically very large and a six order of magnitude range is not unusual. This can be an advantage for environmental analysis where large dynamic ranges are expected. However, for many biomedical analytes, concentrations are highly controlled. The logarithmic response is therefore a big disadvantage since ion selective membranes have impedances that are >10 MΩ. Electrochemical devices are non-linear and have rectifying properties and so the electromagnetically-noisy environment encountered in the healthcare setting can lead to large errors. This can be minimised by careful design, as recently reported for intracellular Ca^{2+} and H^+ microelectrodes using concentric micropipettes to reduce the impedance and give improve response times, around 15 and 5 ms for H^+ and Ca^+ respectively [60] though, as the authors acknowledge, this was inspired by earlier work [61].

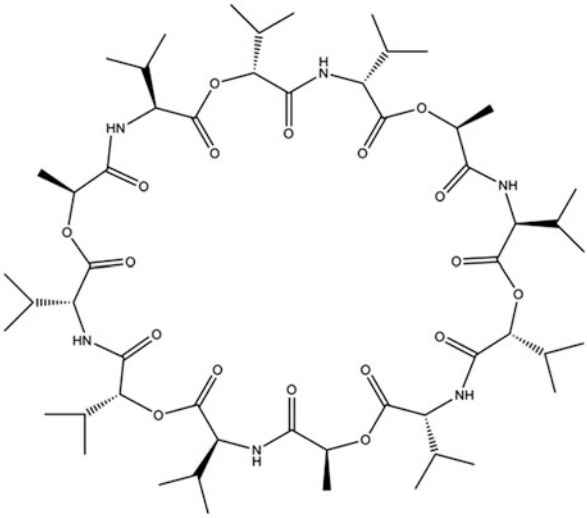


Fig. 2.4 Valinomycin, a bacterially-derived antibiotic and ionophore of choice for K^+ sensors

2.4.1.3 Selectivity

Selectivity of the carrier molecules varies greatly. A quantitative assessment is essential for ascertaining fitness for purpose. Commercially-available ionophores and ISEs usually report selectivity coefficients for single potential interferents. These are defined in a modified version of Eq. 2.19, known as the Nickolsky-Eisenman equation:

$$E = K \pm \frac{2.303RT}{F} log_{10}\left(c_i^{1/z_i} + \sum k_{ij} c_j^{1/z_j}\right) \tag{2.20}$$

where k_{ij} is the *selectivity coefficient* for interfering ion j over the analyte ion i. z_i and z_j are the charges on the analyte and interferent respectively. The key assumption is that there are no synergistic effects between interferents. This is more likely to be true for low concentrations of both analyte and interferent, i.e. there is an adequate concentration of unbound ionophore.

For example, the potassium ion ionophore, valinomycin (Fig. 2.4), when used in a PVC membrane plasticised with dioctyl adipate, has a selectivity coefficient for potassium ions over sodium ions of 6.2×10^{-6}. We can calculate the maximum expected error when this ISE is used to measure potassium ions in plasma. The normal range of potassium ion concentration in plasma is 3.2–5.1 mmol dm^{-3}. The normal range for sodium ions is 135–146 mmol dm^{-3}. The maximum expected error is calculated by substitution into Eq. 2.5:

$$\begin{aligned} \%error &= \frac{k_{K^+Na^+}.c_{Na^+}}{c_{K^+}}.100\% = \frac{6.2 \times 10^{-6} \times 146 \times 10^{-3}}{3.2 \times 10^{-3}} \times 100\% \\ &= 0.028\% \end{aligned} \tag{2.21}$$

The final point to make about selectivity is that a properly functioning ISE will always give a potential whether or not the target analyte is present. This rarely

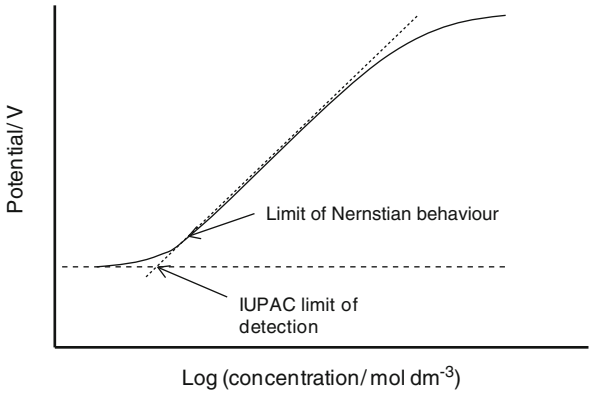


Fig. 2.5 Schematic showing the IUPAC limit of detection for an ISE

presents an issue in biomedical measurement where the concentration of the analyte is unknown but it is known to exist. However, if this is not the case, independent confirmation of the analyte's presence is required using, for example, atomic spectroscopy.

2.4.1.4 Limit of Detection

A further consideration in the use of ISEs is the limit of detection (LOD). For most analytical devices, this is defined as the smallest signal statistically different from the background noise. The logarithmic response of an ISE means that zero concentration is not defined. A different approach is therefore adopted. The log-linear range of the ISE is extrapolated to where it intersects the line describing the zero response. This is the IUPAC agreed definition of LOD of an ISE and is shown schematically in Fig. 2.5.

2.4.1.5 Other Potentiometric Devices

Whilst ISEs (and the related ISFETs) dominate the potentiometric sensor, an important class of potentiometric device for biomedical application is the metal-metal oxide electrode used to measure pH. Glass pH electrodes suffer from obvious disadvantages in the in vivo application: extreme fragility, potential hazard on failure, high electrical impedance and high skill level for manufacturing, all of which militate against commercial, or even research availability. A suitable alternative is the metal-metal oxide electrode which exploits the equilibrium between a hydrated metal oxide and its hydroxide. The advantages of such devices were recognized early and are the subject of a review by Ives [62]. More recent reviews have been published by Głab [63] and O'Hare [64]. The underlying phenomenon exploited in metal-metal oxide electrodes is the measured electrode potential due to the equilibrium between a sparingly soluble salt and its saturated solution, i.e. the

potential depends on the thermodynamic solubility product of the oxide. MMO electrodes are a special case of this kind of electrode since the anion (OH^-) participates in the self-ionisation of the solvent and the equilibrium between unionized water, protons and hydroxide ions. This gives rise to a potential dependence on pH which should therefore be given by:

$$E = E^{\circ}_{M,MO,H^+} - \frac{2.303RT}{F} pH \tag{2.22}$$

where R, T and F have their usual meanings. The standard potential term includes the standard potentials for all the participating species, the solubility product of the metal oxide and the ionisation product of water.

The ideal properties of a MMO electrode have been listed by Ives as:

- The metal must be sufficiently noble as to resist corrosion;
- It must be possible to obtain the metal in a reproducible state;
- The oxide must be stable. (This is incompatible with (1), though in practice, the oxide must only be scarcely soluble);
- It must be possible to obtain the oxide in a reproducible state;
- The oxide must be scarcely soluble yet able to participate in the equilibrium reaction sufficiently rapidly to give an adequate current density.

These properties, though scarcely achievable practically (1 and 5 are strictly contradictory) are useful guides to experimentation.

No MMO system has been found which is well-behaved for all applications, though antimony electrodes have been widely used for many years in medical application, an early example being a 1941 account of continuous recording of the pH of human gastric contents [65]. Relative ease of fabrication using antimony shot melted into a pulled glass capillary has led to application as potentiometric sensing tips in scanning electrochemical microscopy [66]. However, there are several serious drawbacks to using antimony electrodes in vivo: Ives has noted that they must be used in aerated solutions and that Nernstian or even rectilinear responses cannot be relied upon. The solution must not contain oxidising, reducing agents or complexing agents such as citrate, oxalate, tartrate or certain amino acids. There is a response to dissolved oxygen which is caused by localised corrosion for which the cathode reaction is reduction that inevitably leads to sensitivity to stirring.

As a consequence of the shortcomings of the two most widely used pH sensors (glass membrane ISE and the antimony electrode), there has been a substantial recent interest in pH sensors based on hydrated iridium oxide. These devices are of interest due to their reported stability in a wide range of aqueous solutions, low impedance, fast response times and the compatibility of iridium with C-MOS processes allowing the prospect of integrated devices. There has consequently been considerable activity around iridium oxide sensors in recent years. Some aspects of biological applications have been reviewed by O'Hare [64]. This is a genuinely robust piece of technology and, in conjunction with a boron-doped

diamond sensor for histamine (*vide infra*) has been used to elucidate the control pathways for acid production in the stomach of an isolated perfused guinea pig stomach in single sensor measurements [67] and in an array format [68].

There are several methods for preparing these devices: electrochemically generated iridium oxide films (widely known as AIROF – anodic iridium oxide film) [69]; thermally generated iridium oxide (preparative methods are critically compared by Głab [63]); direct electrodeposited iridium oxide from the oxalate complex, developed by Yamanaka [70], and iridium oxide can also be deposited directly using RF sputtering. Thermally deposited and AIROF electrodes appear to be chemically distinct. Iridium oxide deposited using the Yamanaka method has current–voltage characteristics closer to thermally generated iridium oxide. The most convenient method in most applications is that described by Yamanaka, direct electrodeposition by electrolysis of the oxalate complex from alkaline solution:

$$\left[\left[Ir(COO)_2(OH)_4\right]^{2-}\right](aq) + 2OH^- \rightarrow \left[IrO_2(OH)_2.2H_2O\right](s) + 2CO_2 + 2e^- \tag{2.23}$$

where the oxalate ligand is oxidised to CO_2 at the electrode substrate leaving a compact deposit of hydrated iridium oxide on the substrate. Recent work on IrOx devices has coalesced around this method as it is a reliable technique for generating reproducible microsensors, despite its origins, as in electrochromic displays and reliable devices that have been produced in the 3–10 μm range [71]. An improved, faster method based on essentially the same reaction has been reported [72] for the deposition on acid-etched titanium substrate, which show great potential for application in integrated sensing devices, since the usual substrates (gold or platinum) are less compatible with many microfabrication facilities.

Reliable protocols for the three most common methods are given below:

- ***Electrolytic Preparation of AIROF Electrodes*** – Iridium is a dense, brittle and expensive metal, to the extent that it is frequently more convenient to work with small pieces of iridium wire connected to platinum or cheaper material. Iridium wire (0.125 mm diameter, 4–5 mm in length, 99.99+%) was butt-welded to platinum wire in a natural gas/O_2 flame. Spot welding is similarly successful but silver-loaded epoxy shows a high fail rate. The wire needs to be insulated everywhere but the sensing surface. This can be accomplished by dip coating in Epoxylite resin. Additional mechanical strength can be achieved by embedding in a hypodermic needle using low viscosity epoxy resin. Electrode tips were prepared by sawing on the bevel using a low speed saw (Buehler) followed by polishing on emery paper (1,200 grit and 2,500 grit) and aqueous alumina slurry (6 μm, 1 μm and 0.05 μm, Buehler) on polishing cloths with ultrasonic cleaning in water between grades. The oxide film was generated by cycling the potential in sulphuric acid (0.5 mol dm^{-3}) at 2 V s^{-1} for 8,000–12,000 cycles between the potentials of hydrogen and oxygen generation finishing with a 10 mV s^{-1} stopping at the main anodic peak. An iridium rod and Ag | AgCl | 3 M KCl

served as counter and reference electrode respectively. The reference electrode was connected to the cell using a K_2SO_4 (0.3 mol dm^{-3}) salt bridge to minimise chloride ion infiltration. This has been found by present authors and others to be critical in the preparation of stable films. Cyclic voltammograms were recorded at various intervals to assess the extent of oxide film growth. In all cases, the resulting AIROF electrodes were soaked for 48 h in deionised water (>15 MΩ cm) prior to use. For additional stability in biological measurement, we have found that Nafion coating is very successful and barely affects sensitivity or response time. Nafion films were applied and annealed at 120 °C according to the protocol described by Harrison and Moussy [73]. Calibration from pH 3 to 12.1 gave a super-Nernstian response of (69 ± 2) mV per pH unit. Comparison of calibration curves recorded in nitrogen and oxygen-sparged solutions revealed a maximum perturbation of 0.9 mV at pH 7.4. This places an absolute limit on the accuracy of 0.0125 pH units if the oxygen concentration is unknown, though this does of course represent the worst-case scenario.

- ***Thermally Prepared Iridium Oxide Electrodes –*** Iridium wire was annealed in a natural gas flame, straightened and carefully cleaned by sonication in acetone followed by rinsing with deionised water. After drying, one end (approx. 2 mm) was wetted with NaOH solution (1 mol dm^{-3}) and the wire was heated to 800 °C in a muffle furnace for 30 min. This was repeated until a blue-black coating was clearly visible to the naked eye. This typically took five to six applications. The electrode was soaked for 3 days in deionised water before use. All but the electrode tip (approximately 0.5 mm) was insulated using FEP/PTFE dual shrink tubing (Zeuss). Nafion films were applied using the technique described above. Calibration in Britton-Robinson buffer over the physiologically-relevant pH range of 6.5–8 gave a slope of 59.5 mV/pH (r = 0.9999).
- ***Direct Anodic Deposition of Iridium Oxide on Gold*** [67] – Initially, a 75-μm gold wire insulated in Teflon (overall diameter 140 μm, A-M Systems Inc.) was threaded through a 27-gauge hypodermic needle. A copper or silver wire was attached to the gold wire with silver epoxy resin to form an electrical contact. Epoxy resin (Robnor Resins, CY1301 and HY1300) was used to fill the internal volume of the needle and left for 2 days to cure according to the manufacturer's instructions. The lower end of the needle was cut perpendicularly using a diamond saw (Buehler) to expose the 75-μm Au disk microelectrode. Successive polishing with aqueous slurries of 1-, 0.3-, and 0.05-μm alumina in deionized water with rinsing and sonication at each polishing stage was necessary to ensure a flat electrode surface. Cyclic voltammetry (CV) in 0.5 M H_2SO_4 was used to electrochemically clean the gold electrode surface prior to deposition. Using the same technique, the transport-limited currents were recorded in 1–10 mM $Ru(NH_3)_6^{3+}$ in supporting electrolyte to assess the surface of the gold electrodes. When the recorded limiting current had reached the theoretical value for a 75 μm diameter microelectrode in a known concentration of analyte, the electrode was considered ready for use. Anodic electrodeposition of the iridium oxide film onto gold microelectrodes was performed using a deposition solution described by Yamanaka [70]. Briefly, 0.15 g of iridium tetrachloride, 1 mL of 30 % w/w

H_2O_2, and 0.5 g of oxalic acid dihydrate were added gradually in a 100 mL of water at 0.5 h intervals and left to dissolve in a stirred solution. Anhydrous potassium carbonate was then added gradually to the solution until the pH reached ≈10.5 forming a pale yellow solution. The solution was then covered and left at room temperature for 48 h to stabilise until a colour change to pale blue was achieved. This blue solution was stored in the refrigerator and could be used for a few months to successfully produce IrOx films.

The anodic electrodeposition of the IrOx films on gold microelectrodes was achieved amperometrically using a constant potential method. A potential between 0.6 and 0.7 V versus a double junction reference electrode (DJRE) was applied for several minutes to produce a thin, uniform, and defect-free film. The DJRE with a calomel inner junction reference system was used with a sodium sulphate (0.1 M) outer junction solution to prevent chloride ion permeation [64]. Assuming 100 % faradic efficiency of electrodeposition, the total amount of iridium oxide was calculated to be 0.006 μg when the deposition potential, Ed, of 0.65 V was applied for 2 min (current density 0.83 mA cm^{-2}). The coated microelectrodes were washed and placed in deionized water for at least 2 days prior to use to allow redistribution of ions and hydration to occur and a stable open circuit potential to develop. In our lab, once this initial hydration had occurred, the electrodes could be rinsed in de-ionised water and stored dry for months. Re-use only required hydration for a few minutes. Composition and integrity of the film can be tested at regular intervals by examining the cyclic voltammogram in 0.1 mol dm^{-3} H_2SO_4.

Typical performance is shown in Fig. 2.6 (taken from [67]).

2.4.1.6 Recent Developments in Ion Selective Electrodes

For nearly 100 years, research in ion-selective electrodes was largely incremental and focused on improvements in selectivity, novel ion carriers and applications away from the laboratory including point-of-care and environmental applications. This mature technology, which is relevant to biomedical applications, is reviewed in two excellent papers by Pretsch, one of the most significant players in this field [74, 75]. However, there have been significant developments in the last 10 years in both theory and practice which have led to the so called “new wave” ISEs. These significant developments have been the subject of two recent reviews [76, 77].

The conventional theoretical analysis, the classical total equilibrium model essentially followed above, is, of course, a gross simplification of the actual disposition of the ions, carriers and electric fields present in real devices. However, this essential, and deliberate, simplification is entirely adequate to the task of supporting the user of sensor technology. However, more sophisticated analysis is required to account for the variable and time-dependent responses of ISEs in long-term and to provide insight that have since led to improved performance. These theoretical advances are described by Lewenstam [78].

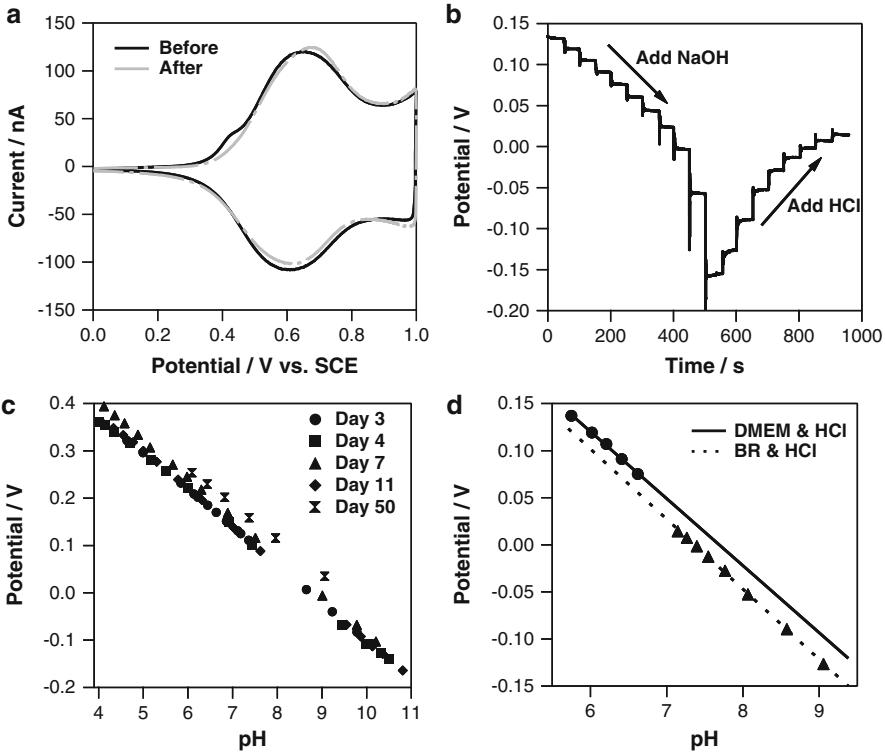


Fig. 2.6 Typical responses of a 75 μm diameter gold microelectrode coated with iridium oxide. (**a**) shows the cyclic voltammetric response in 0.5 M H_2SO_4 before and after exposure to perfused stomach. (**b**) shows the electrode's open-circuit potential response to changes in pH during acid–base titration. (**c**) Typical calibration plots over 50 days showing excellent stability. (**d**) Calibration in Britton-Robinson buffer compared with calibration in tissue culture medium (DMEM) over the physiologically relevant range (Taken from Ref. [67])

The major change in ISEs however is the spectacular improvement in the limits of detection achieved in recent years, such that ISEs are now competitive with atomic spectroscopy. The initial insight was that whilst conventional ISEs had limits of detection around 10^{-6} mol dm^{-3} but picomolar LODs were obtainable with optical sensors using the same ionophores and essentially the same membrane technology (though with a vastly more expensive instrumentation). This led to the hypothesis that primary ions leaching from the membrane determine the LOD. By the simple expedient of incorporating an ion buffer into the internal electrolyte, Pretsch and colleagues were able to extend the limit of detection for a Pb^{2+} ISE from 4×10^{-6} mol dm^{-3} to 5×10^{-12} mol dm^{-3}.

New technology or the application of mature technology from other fields is being brought to bear on ISEs. Solid contact potentiometric sensors remain promising technology and their application to medicine and biology has been touted as early as 2000 [79]. Conventional microfabrication dielectrics have been found to be

suitable for both ISFETs and ISEs [80] and silicon nitride substrate have been used to make miniaturised Na^+ selective electrodes [81]. Solid-state devices have been reported for K^+ and Ca^{2+} based on electropolymerisation of polystyrenesulfonate-doped PEDOT on recessed gold discs insulated with glass, which are then covered with ion selective membranes [82]. Microfabricated devices have found application in cell culture where 2 μm or 6 μm ion selective devices for potassium ions and ammonium were formed in micropipettes microfabricated at the bottom of cell culture wells [83].

Screen printing is also an attractive technology as it is both cheap and widely available and has relatively low start-up and scale up costs. These are important factors which can otherwise restrict the development of new sensors to research-only or high value applications. Getting the reference electrodes to work without the current craft-intensive processes is essential for both ISEs and ISFETs and screen printed miniature solid state devices have been reported [84, 85] including a solid state screen-printed K^+ selective electrode [86].

Kapton-based K^+ and pH flexible microelectrode ISE arrays have been described by Buck [87, 88] and have been used to record on a beating heart during ischaemia. More conventional microfabrication of an ISE chip, complete with built-in reference electrodes has been reported by Uhlig et al. [89]. Usefully, comparative performance data for different membrane polymers and formulations are described and the chip arrays were used in a flow-through format for measurement of K^+ and Ca^{2+} concentrations in urine, human serum and whole blood. The advantages of array sensing are discussed in more detail below. Similarly, Yoon et al. [90] built arrays of ISEs for blood electrolytes (K^+, Na^+. Ca^{2+}, H^+, Cl^-). Whilst technology for blood electrolytes in the clinical setting is mature technology and there is no market pull for improved performance (the major costs are staff salaries and reagents), the novel reference electrode performance using a polyurethane coated reference electrode is reported to be sufficiently stable to reduce the requirement for repeated standardisation between measurements that will simplify operation and improve throughput. Conventional microfabrication gives inherent scalability and potentially reduced costs but the requirement for extensive post-production processing and limited market pull probably explain poor uptake of this technology so far.

2.4.2 Amperometry and Voltammetry

Potentiometric methods are passive and the selectivity is inherent, that is to say, it is built in to the membrane in the case of ISEs. Amperometric and voltammetric methods however, involve applying a non-equilibrium electrical potential and measuring the resulting current or current–voltage relationship to obtain quantitative (in the case of amperometry) or qualitative information. At the core of this technology is the transfer of electrons between the Fermi level of a usually metal electrode and the molecular orbitals of the target analyte. Oxidation involves the loss of electrons from the highest occupied molecular orbital whereas reduction

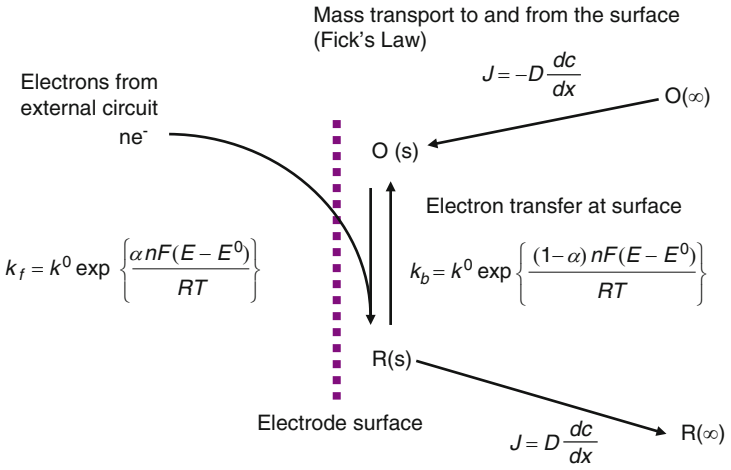


Fig. 2.7 A schematic of the simplest electrode reaction. The *dashed line* represents the electrode surface. The electrolyte solution is to the *right* of the electrode surface showing mass transport to and from the electrode surface (by Brownian motion). To the *left* of the electrode surface is the external circuit

involves electrons being injected into the lowest unoccupied molecular orbital. For some arbitrary pair of compounds where R represents the reduced form and O represents the oxidised form, the reaction can be written as:

$$O + ne^- \rightleftharpoons R \tag{2.24}$$

One or both of R and O can be solution free species or insoluble and bound to the electrode surface, or even be the elemental form of the electrode itself. The oxidation-reduction reaction shown above is the simplest scheme possible where only electron transfer is involved. However, for many real target analytes such as the monoamine neurotransmitters or dissolved gases such as oxygen or nitric oxide, the electron transfer is also associated with adsorption and proton transfers. In addition, there may in some cases also be changes of phase which further complicate an already intricate situation, but the core of an electrode reaction is the transfer of electrons between a molecule in an electrolyte solution and an electrode made of conducting or semiconducting material.

The electron transfer event is a very short range phenomenon; for aqueous systems the electron tunnelling event really only takes place over a few hundred picometres, about the same size as a hydrated metal ion (a hydrated K^+ ion has a radius of 330 pm). This short range has several important theoretical and practical consequences. Firstly, for a molecule to undergo a redox reaction is must be transported to within a bond length of the electrode surface. The simplest scheme describing an electrode reaction is shown in Fig. 2.7.

As can be seen for Fig. 2.7, the rate of mass transport to the surface must be equal to the rate of the electron transfer reaction at the electrode surface. The rate of

electron transfer, suitably scaled from chemists' units to electricians' units using Faraday's constant (*vide supra*) is given by the current through the external circuit. In one dimension (for the sake of clarity) this relationship is given by:

$$\frac{i}{nFA} = J = -D\frac{dc}{dx} \tag{2.25}$$

where F is Faraday's constant (96,485 C mol^{-1}), J is the flux (mol m^{-2} s^{-1}), D is the diffusion coefficient (m^2 s^{-1}) and dc/dx is the concentration gradient. Analytical expressions for the current at an electrode therefore essentially depend on being able to describe the concentration gradient for a specific electrode and boundary conditions. An excellent introduction to analytical and numerical approaches to these sorts of problems has been published by Compton & Banks [91].

The rates of the forward and backward reactions are given by:

$$Forward\ \ rate = k_f(c_O)_{x=0} \tag{2.26}$$

and

$$Backward\ \ rate = k_f(c_R)_{x=0} \tag{2.27}$$

The rate constant for the electron transfer depends exponentially on the applied potential. Whilst there are microscopic quantum mechanical descriptions, the empirical Butler-Volmer relationships with their familiar Boltzmann form are entirely adequate for most purposes:

$$k_b = k_b^{\circ} exp\left(\frac{\propto_A nFE}{RT}\right) \tag{2.28}$$

for the rate constant for the backward electrode reaction, and

$$k_f = k_f^{\circ} exp\left(\frac{-\alpha_c nFE}{RT}\right) \tag{2.29}$$

where n is the number of electrons transferred, R is the gas constant, T is the absolute temperature, E is the applied potential and k° is the standard heterogeneous rate constant. α_A and α_C are the anodic and cathodic transfer coefficients respectively. The can be related to the position of the maximum of the potential energy-reaction coordinate curve and Compton has recently demonstrated that the physical meaning may be associated with bond energies of the transition state. In any case, they sum to unity and typically have a value close to 0.5.

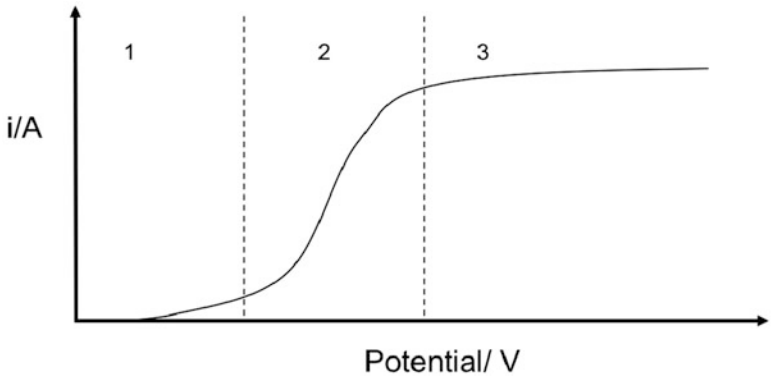


Fig. 2.8 A schematic representation of the current-voltage relationship

At the equilibrium potential, E_e given by the Nernst equation for reversible reactions, whilst the overall current is zero, this is because the backward and forward current densities are equal:

$$j_0 = -j_b = j_f \tag{2.30}$$

where j_0 is the exchange current density, an important measure of the reversibility of the electron transfer reaction. By substitution therefore:

$$j_0 = nFk_f exp\left(\frac{\alpha_A nFE_e}{RT}\right) = -nFk_f exp\left(\frac{-\alpha_C nFE_e}{RT}\right) \tag{2.31}$$

Since the overall current density $j = j_f + j_b$:

$$j = k_C^{\circ} c_O exp\left(\frac{-\alpha_C nFE}{RT}\right) - k_A^{\circ} c_R exp\left(\frac{\alpha_A nFE}{RT}\right) \tag{2.32}$$

Substitution from Eq. 2.31 and defining the overpotential η as the deviation from the equilibrium potential, $\eta = E–E_e$ leads to the Butler-Volmer equation:

$$j = j_0\left\{exp\left(\frac{\alpha_A nF\eta}{RT}\right) - \left(\frac{-\alpha_C nF\eta}{RT}\right)\right\} \tag{2.33}$$

2.4.2.1 Amperometric Methods

Given the equations above, the current-voltage plot of an electrode in the presence of a single electroactive species will initially show an exponential rise as the overpotential is increased. Eventually, diffusion will start to limit the flux to the electrode surface and the current will become independent of the applied voltage. This leads to the sigmoidal form of a typical current-voltage curve illustrated schematically in Fig. 2.8.

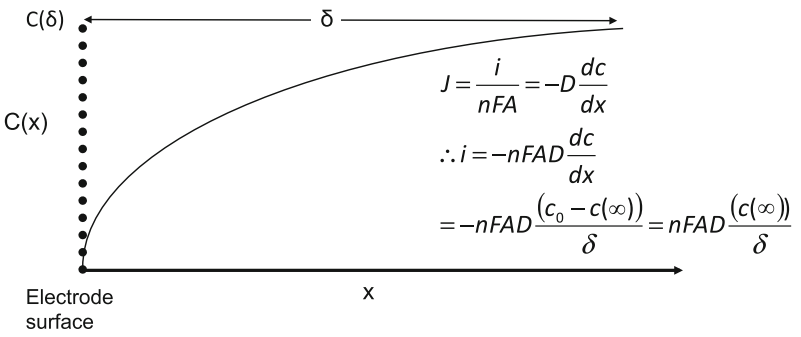


Fig. 2.9 Schematic representation of the concentration profile close in the diffusion-limited regime. Bulk concentrations are denuded close to the electrode surface, in the so-called Nernst layer. In the diffusion limited regime, the surface concentration is maintained at zero due to rapid electron transfer

In Fig. 2.8, region 1 is where the current depends on the rate of electron transfer, and is therefore potential dependent. The form of this part of the curve is given by the Butler-Volmer relationships above. In region III, the current is limited by the ability of diffusion to supply the electrode surface with electroactive material. In region 2, there is mixed control: the rates of mass transport and electron transfer are occurring at broadly similar rates. Region 3, where mass transport is limited, is the region of utility for analytical measurements since the rate of reaction, and therefore the current is proportional to the bulk concentration.

In region II, the surface concentration will be zero, since reactants arriving at the surface will be instantly oxidised or reduced. A conceptual grasp of how amperometric and voltammetric sensors can allow estimation of concentration can be got from consideration of the steady state in the diffusion limit. In this case, the concentration gradient dc/dx can be approximated by:

$$\frac{\Delta c}{\delta} = \frac{c_s - c_\infty}{\delta} \tag{2.34}$$

where c_∞ is the bulk concentration. In the diffusion limit, c_s is zero and δ is the Nernst layer thickness, that region of solution depleted by the electrode reaction. Eq. 2.25 then becomes:

$$i = -nFA\frac{c_\infty}{\delta} \tag{2.35}$$

showing that there is a linear relationship between the measured current in the external circuit and the bulk concentration. The key experimental control then is to ensure that the device is engineered so that δ is kept constant. A graphical representation of this concept is shown in Fig. 2.9.

As clearly seen, bulk movement of the solution would perturb dc/dx and therefore affect the measured current in the external circuit. This would render

the sensor useless (though this phenomenon can be exploited in constant concentration solutions to measure local mass transport rate constants).

There are three modes of mass transport: convection, migration and diffusion. In the biomedical situation, the ionic strength is such that the medium cannot support an electric field (the Debye length is around 0.5 nm) and therefore migration, the movement of a charged particle in an electric field, is not significant. Similarly, most electroanalytical experiments take place in highly conductive ionic media by design, on order to simplify the analysis. Bulk convection can be an important mass transport mechanism – diffusional speeds are of the order of 10 μm s^{-1} so any significant bulk flow will swamp diffusion. However, ultimately, the electrode surface will be in a convective boundary layer, a hypothetical thin film of stagnant solution where at physiological ionic strengths the only mode of mass transport is diffusion, described by Fick's laws. In stagnant solution at room temperature in a typical lab beaker, the boundary layer for natural convection has been estimated by Bockris to be around 0.05 cm [92]. Provided the so-called Nernst layer (that region of the solution where the analyte concentration has been significantly perturbed from the bulk value by the electrode reaction) is much smaller than the convective boundary layer then the problem can be reduced to solutions of a diffusion-reaction equation.

For a useful sensor then, the key engineering target is to ensure that δ is kept constant and ideally known, either from analytical solutions, numerical modelling or by experimental calibration. This is the case whether the applied potential is kept steady, at a potential where the electrode reaction is diffusion limited (amperometry) or systematically varied with time, as in voltammetry. There are three experimental approaches to this and all three have found application in analytical devices.

- **Forced Convection**
  Overwhelm natural convection by using a well-defined forced convection such that the convective boundary layer is significantly smaller than that due to natural convection. Examples include the rotating disc electrode which gives an analytically tractable uniform boundary within which the concentration boundary layer is entirely confined. This is a vitally important technique in sensor development as it allows decoupling of mass transport from the rate of electron transfer. However, it can only be used analytically on extracted fluid samples. Similarly, channel flow sensors can be used in the flow-injection format, again useful for high throughput analysis of multiple patient samples.
- **Membrane-covered Devices**
  This is essentially the complementary approach. A permeable membrane is applied between the electrode surface and the solution (or patient). The membrane does not permit bulk flow. Therefore, provided the diffusion coefficient of the membrane is much smaller than the solution (which will almost invariably be the case), the concentration gradient will be confined to the membrane and the external solution (or tissue) will be unperturbed by the electrode reaction. This

approach can have the additional advantage of preventing surface-active sample components from accessing the electrode surface. Examples of this approach abound in the literature. It is important to recall, however, that the current in the solution is carried by anions and cations, so the membrane must be ion permeable to provide a current path between the working and counter electrodes. This limitation is circumvented in the well-known Clark O_2 electrode, which places the entire electrochemical cell behind a gas permeable membrane, typically cellulose or PTFE. This prevents ionic access to from the test solution to the electrodes, but permits small neutral solutes such as dissolved O_2 though to react at the cathode.

- **Microelectrodes**
  If the electrode is small, the Nernst layer will be correspondingly small, For an inlaid disc microelectrode, 90 % of the diffusion gradient will be contained in a hemisphere of six times the radius of the disc [93]. Recalling Bockris' estimate of 0.05 cm as a typical natural convective boundary layer thickness and allowing a margin of one order of magnitude, a microdisc microelectrode 50 μm in diameter ought to develop a genuine diffusion limited current. This is found to be the case experimentally. Plainly, if there is bulk convective flow, the electrode would need to be smaller still. This important result has other implications. The volume of tissue sampled using a microelectrode must be of a similar dimension, so the effective spatial resolution of a microdisc electrode is also around six times its electroactive radius.

Disk microelectrodes are relatively simple to fabricate in the laboratory by insulating metal wires or carbon fibres. Sectioning with a diamond wafering saw followed by polishing with alumina slurries or diamond lapping compounds reveals the disc. Photolithographic processes can also be used but this usually leads to a recessed disc configuration due to the requirement of co-planar hook-up tracks.

Asymptotic solutions for the microdisc give the diffusion limited current [94] as:

$$i_d = 4nFcDa \tag{2.36}$$

where a is the electrode radius, c is the bulk concentration, n is the number of electrons transferred per mol. of analyte and D is the diffusion coefficient. If bare microdisc electrodes are used in vivo or in tissue samples, the diffusion coefficient is not generally known. It can however be measured in situ (using microelectrode chronoamperometry, see below) but care must be taken since D can be affected by oedema during an inflammatory response or by compression of the tissue due to the insertion of the microelectrode itself. The diffusion limited current at a recessed microdisc [95] (such as an individual element in a microfabricated microelectrode array) is:

$$i_d = \frac{4\pi nFcDa^2}{4L + \pi a} \tag{2.37}$$

it is worth noting here, that the sensitivity of devices based on microelectrodes scales linearly with the electrode radius. This gives microelectrode devices a significant scaling advantage over spectroscopic methods for miniaturisation. Spectroscopic methods essentially count molecules in a given volume so sensitivity will scale with the cube of the linear dimension.

Steady state techniques have the fastest response time of any electroanalytical technique. For electrodes without membranes, the response time is essentially instantaneous since it depends primarily on the diffusion characteristics of the test medium. Electron transfer takes place on the femtosecond time scale. It is this combination of excellent temporal resolution and unparalleled and tuneable spatial resolution that has allowed direct measurement of single vesicles of neurotransmitters to be quantified in real time; an example is detailed below. Steady state techniques offer the advantages of simple instrumentation and simple, often analytical, relationships between the measured current and analyte concentration.

In real applications however, selectivity can present a problem. Microelectrode sensors are typically tested and calibrated in homogeneous pH-buffered solutions containing only the target analyte. There is no guarantee that your biological specimen will be so obliging. In addition to the complexities introduced by compartmentalisation and inhomogeneity (outlined in the introduction above), there may be other (known or unknown) electroactive species. Selectivity in amperometric methods at bare electrode arise entirely from the applied potential. This is characteristic of the electrode material and the physic-chemical properties of the analyte (its molecular orbital energies and its energy of solvation). However, these can all be affected by pH, ionic strength and adsorption on the electrode surface, for example. So the potential identified in the calibration and characterisation may not be correct in vivo. This would not be revealed in a simple amperometric technology unless a current voltage curve was recorded at the start of the experiment, at regular intervals and at the end of the experiments. The principal problem is the presence of known or unknown interferents though. For example, in neurochemical investigations of monoamine neurotransmitters, ascorbic acid is typically present at a concentration that is 100× higher than the target analytes and are oxidisable at similar potentials. Furthermore, the monoamines have redox potentials very close to each other. Applying a steady potential at the diffusion-limited potential of, say, serotonin could result in a current that is augmented due to the unsuspected presence of dopamine, which is oxidised at a lower potential. The only way this can be tested is by running a periodic current-voltage curve or using a non-steady potential programme.

However, for the right biomedical problem, the simple instrumental requirements of amperometry and fast response times are a great advantage. All that is required is a precision voltage source; a current to voltage converter with appropriate sensitivity and some means of recording the signal. This steady applied potential also aids in applying analogue or digital filters to deionise the data.

A representative example now follows from our laboratory on the detection of serotonin release from single vesicles. We have an interest in examining the effects of ageing on release of monoamine neurotransmitters and the gaseous transmitter nitric oxide [96]. Neurotransmitters are chemical messengers which are released

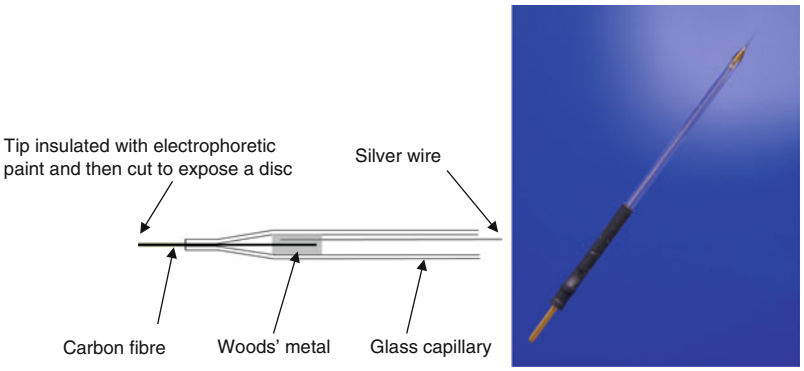


Fig. 2.10 Schematic (*left*) and photograph of a carbon fibre sensor for neurochemical application

from a pre-synaptic neuron and diffuse across the synaptic cleft where another action potential is triggered. There are three types of transmitter: gaseous e.g. nitric oxide and amino acids; peptides such as glutamate or myomodulin and monoamines such as serotonin and noradrenaline. Peptides and monoamines are released through vesicles, subcellular structures originating in the Golgi that fuse with the cell membrane and release their contents. Vesicles typically contain 20,000–50,000 molecules and the release is generally over in 5 ms. This presents a challenging measurement requirement – low concentrations, highly localised release and a requirement for excellent temporal resolution.

An example is presented on how we were able to measure serotonin (5-hydroxy tryptamine, 5-HT) release from an identified neuron in the water snail, *Lymnea stagnalis* using carbon fibre microelectrodes. The microelectrodes were fabricated as follows:

- Clean a 7 μm carbon fibre by sonication in acetone followed by deionised water.
- Place the fibre inside a pulled glass capillary (where the end as been polished to facilitate insertion). This may be aided by using a capillary filled with ethanol.
- Once placed inside the capillary allow approximately 2 mm of the carbon fibre to protrude from the end of the capillary and seal using epoxy resin by capillary action. The resin takes 72 h to set and cure at room temperature.
- Contact using a silver wire via Woods metal.
- The exposed shanks of the protruding tip are then insulated using electrophoretic paint. To coat the carbon fibre a voltage of 2 V was applied for 1 min using a platinum coil as the cathode and the carbon fibre electrode as the anode. Following coating, the electrode was removed by micromanipulator and cured. The anodic paint was then cured after each coating for 20 min at 160 °C. This process was repeated four more times and the voltage was increased to 3, 4, 6 and 8 V for each subsequent coating. The carbon fibre was then cut using a scalpel to expose a carbon fibre disc electrode.

A schematic and photograph of the completed sensor are shown in Fig. 2.10.

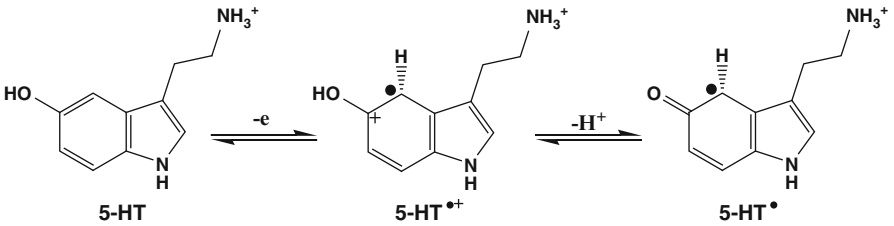


Fig. 2.11 The electrode reaction of serotonin (5-HT). The resulting radical cation is deprotonated to form the neutral radical. This can then react with unreacted serotonin or other radicals to form oligomeric or polymeric deposits on the electrode surface, leading to passivation

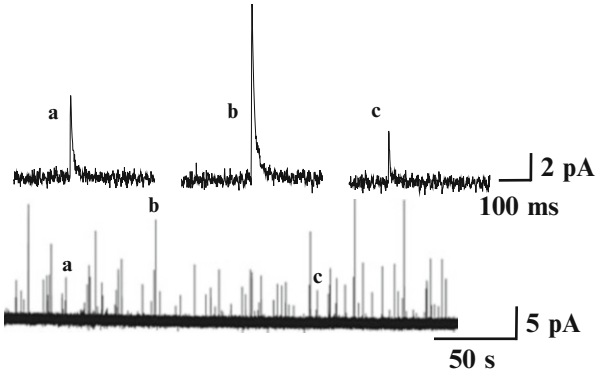


Fig. 2.12 Typical traces showing spontaneous vesicular release of serotonin from the cell body of a neuron in the intact isolated perfused CNS of *Lymnea stagnalis*. The *top trace* shows enlarged single vesicle events

The structure and electrode reaction of serotonin is given above in Fig. 2.11. When a carbon fibre electrode is polarised to +0.7 V versus an Ag/AgCl electrode (see above), it is oxidised at the hydroxyl group to a quinonoid moiety in a two-electron reaction at a rate which is proportional to its concentration.

The sensors were pressed against the surface of chosen cell of the isolated neuronal system to form, what Amatore calls, an artificial synapse [97]. Typical spiky responses are shown above (Fig. 2.12), each spike corresponding to the release of a single vesicle of neurotransmitter under the electrode surface.

The individual vesicular events are analysed for peak height, peak area (which can be related to the total number of molecules detected) and the time constant of decay, which is related to re-uptake by the pre-synaptic cell (shown schematically in Fig. 2.13).

The resolution of these recordings has been clear enough for us to detect changes in neurotransmitter re-uptake kinetics as a function of age. Key findings were that in older animals, serotonin re-uptake was inhibited to increase the peak concentration of transmitter. Similar changes were observed for nitric oxide [98], though in that case, it was the enzymes, rather than re-uptake channels that were responsible. We believe this to be an adaptive change to deal with losses of sensitivity in the post-synaptic cell in ageing. Whilst the sensor results are important on their own, an

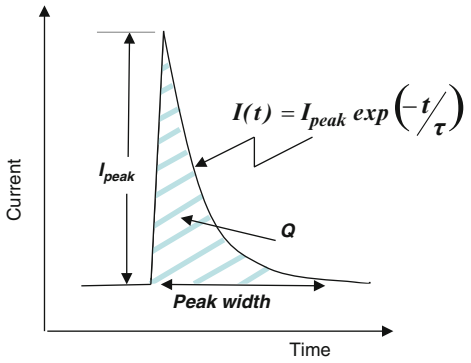


Fig. 2.13 Schematic showing the parameters extracted from each vesicle peak

important part of this work was linking the neurochemical changes to behavioural changes as the animal's age to complete the biological picture [99].

2.4.2.2 Voltammetry and the Use of Non-steady Potential Programmes

Despite the excellent spatial and temporal resolution displayed by steady state voltammetry, there are a number of disadvantages for some applications. Analyte consumption is directly proportional to current. This can be a major disadvantage in oxygen measurement where the biological problems of greatest interest occur is tissues where oxygen concentration is low. Intermittent operation can provide a solution. The limits of detection and sensitivity of the sensors are frequently limited by noise. When the currents are small, as is the case in microelectrode measurements (sub nanoamp currents are typical), the sensitivity may not be adequate. Operating the sensor with a non-steady potential increases the sensitivity by sampling the current when the concentration gradient at the surface is steeper. Operating the sensor in the steady state also raises issues of selectivity. Any molecule which can be electrolysed at or below the applied potential will contribute to the current. This is not necessarily a problem in an anatomically well characterised system, but for many applications, easily oxidisable high concentration components of most biological fluids such as ascorbate (vitamin C) or uric acid present serious problems. Potential programming can be used to confer additional selectivity. Finally, electrode fouling (of which more below in Sect. 2.5.2) can sometimes be overcome by pulsing the electrode potential either to reduce interactions or to oxidise any films formed on the electrode surface. Below, we will consider the most important transient techniques, chronoamperometry, cyclic voltammetry and square wave voltammetry. However, whilst these techniques undoubtedly overcome some problems, they introduce others, most notably capacitive charging.

When a time-varying potential is applied to an electrode, the faradaic current is accompanied by a charging current. This is not simply due to the leads and instrumentation, though these will undoubtedly contribute. The electrical double layer associated with the electrode-electrolyte interface shows capacitor-like

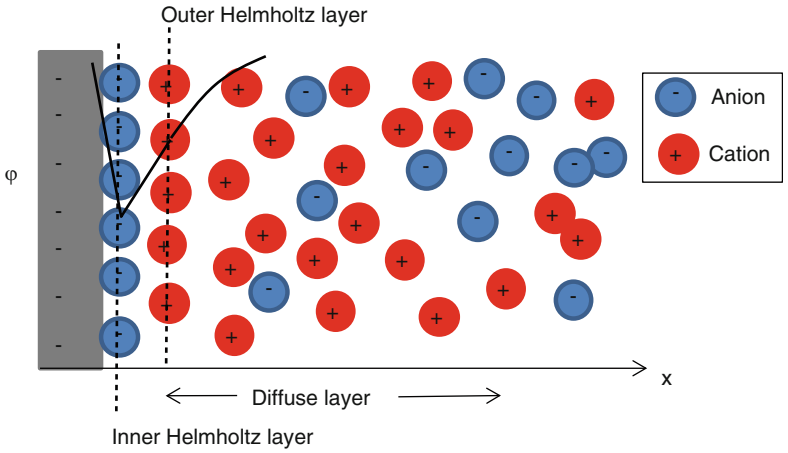


Fig. 2.14 Schematic representation of the electric double layer, the charge separation occurring at an electrified interface between an electrode and an electrolyte. The *solid black line* shows the potential varying linearly through the compact layer of adsorbed ions and exponentially in the diffuse double layer

behaviour. On the solution side of the interface there is an excess of counter ions to balance the charge on the electrode surface, a charge separation that obviously resembles a capacitor, but the charge separation is potential dependent. The dipolar water molecules are also preferentially oriented in the field. A schematic representation is shown above (Fig. 2.14).

The figure shows discrete regions where anions are specifically adsorbed (i.e. without their salvation shell) with their centres on a plane called the *inner Helmholtz layer*. Then there is an excess of solvated cations, balancing the net negative charge on the electrode. On the microscopic scale they cannot be considered as point charges, but can approach the electrode surface no closer than their hydrated radius. A plane through the centres of these ions in the compact double layer is called the *outer Helmholtz layer*. Beyond the outer Helmholtz layer, any remaining charge is balanced by a mobile *diffuse double layer*. The potential $\varphi(x)$ of the diffuse double layer in one dimension is approximated by:

$$\varphi(x) = \varphi^{\circ}\exp(-\kappa x) \tag{2.38}$$

where φ° is the potential at the surface and $1/\kappa$ is the *Debye length*, the screening length scale which depends on the ionic strength:

$$\kappa = \left(\frac{2 \times 10^3 \varepsilon^2 N^2}{\varepsilon\varepsilon_0 kT}\right) \tag{2.39}$$

The Debye length at physiological ionic strength is similar to the length of a chemical bond or a hydrated ionic radius. This matters because molecules will

barely sense the electric field until they are within a bond length and the double layer structure will, to a very good approximation, consist entirely of the inner and outer Helmholtz layer.

When the electrode potential is changed, electrical work must be done to provide the appropriate ion atmosphere and re-orientate the dipoles. This is manifested as a charging current which decays to zero in the steady state. The capacitance of a noble metal electrode is of the order of 20–30 μF cm^{-2}. Since capacitance scales with area, this problem is less severe with smaller electrodes. Some of the newer materials introduced into sensor technology, notably carbon nanotubes and boron-doped diamond, demonstrate much lower capacitance and this can offer additional advantages, though capacitance is always going to be an interference. Many of the more sophisticated and sensitive transient techniques have been designed to minimise the influence double layer charging.

The simplest transient technique is chronoamperometry. The electrode potential is instantaneously changed from one at which no electrolysis occurs, to one that sufficient enough to generate a diffusion limited current. Intermittent operation decreases analyte consumption, the electrode is polarised only when the measurement is required.

The resulting faradaic current rises instantaneously to infinity (or as fast and as high as the instrumentation will allow) as the surface concentration falls to zero. As the concentration gradient relaxes into the solution, the current decays as $t^{-1/2}$. For a large electrode, the current is given by the Cottrell equation which predicts that the current should approach zero at infinite times.

$$I = \frac{nFAc\sqrt{D}}{\sqrt{\pi t}} \tag{2.40}$$

For a disc shaped microelectrode the current asymptotes to a finite non-zero value since, as discussed above, the hemispherical diffusion field is small compared with the natural convection boundary layer. The $t = \infty$ varies with time and two asymptotic solutions exist: $\pi nFcDa$ for short times (where $4Dt/a^2 < 1$) and to $4nFcDa$ for long times [100] as the diffusion to the electrode edge increasingly dominates and the electrode begins to appear as a point sink (Fig. 2.15). The principal advantage of chronoamperometry is that since expressions for slope and intercept on the i vs. $t^{-1/2}$ plot contain both diffusion coefficient and concentration, both of these terms can be obtained from a single experiment. This is a great convenience in biological systems since the diffusion coefficient is generally unknown and likely to be different from a calibration solution. Furthermore, the diffusion coefficient is of intrinsic interest and can reflect tissue hydration. We have used this technique to quantify the effects of tissue hydration in the intervertebral disc (which is affected by mechanical loading) on oxygen transport in the tissue [101]. A further potential benefit is that it may be possible to recondition the electrode surface by applying a cleaning pulse between measurements.

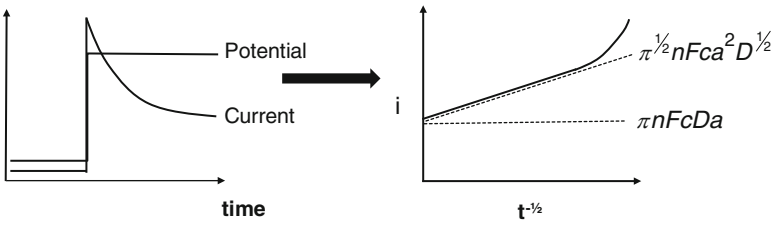


Fig. 2.15 Schematic of microelectrode chronoamperometry. The *left-hand* graph shows the potential step and the resulting current. The *right-hand* graph shows the straight-line response (after the initial capacitive charging current) and the pair of equations describing the slope and intercept

However, the early parts of the current transient are distorted by capacitive charging. The charging current goes as:

$$i(t) = \frac{\Delta E}{R} exp\left(\frac{t}{RC}\right) \tag{2.41}$$

It is essential to establish the RC time constant for charging in a blank solution and only analyse the current for times longer than $3RC$ but less than $4Dt/a^2$.

Microelectrode chronoamperometry can improve sensitivity, decrease analyte consumption (by being switched off between measurements) and give access to the diffusion coefficient independently of the concentration (which can be of critical importance in vivo, for example in the detection of oedema). However, it does not overcome the principal disadvantage of steady state techniques which is that of unknown selectivity in complex samples. Additionally, the sharp edge of the stimulating voltage can provoke action potentials in neurons. These disadvantages are to some extent overcome by other transient techniques and with modern instrumentation and there is no requirement to use only one technique.

More sophisticated transient techniques are not generally suitable for implementation in sensors, though can be useful for characterising both the sensor and the electrode reaction and assessing whether there is any couple solution chemistry occurring. Cyclic voltammetry is a particularly useful "first look" technique but, with the exception of neurotransmitter research [102, 103], has not been widely used in biosensing application due to the relatively high limits of detection of the order of 10^{-5} mol. dm^{-3} for routine applications.

Cyclic voltammetry involves applying a triangular waveform to the working electrode and plotting the resulting current as a function of the instantaneous applied potential. This is dynamic technique in that the diffusion gradient at the electrode is changing continuously with time. This results in a peak-shaped response. The peak arises when the surface concentration falls to zero. The transient current decreases with $t^{-1/2}$ (a useful check that the peak is indeed due to mass transport limitations and not to electrode fouling) as the electrode potential continues to vary. Reversing the potential scan leads to reversal of the electrode

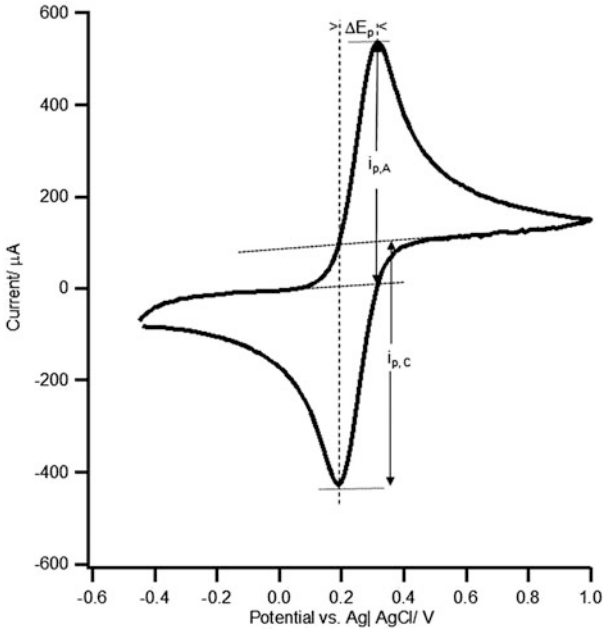


Fig. 2.16 A typical cyclic voltammogram

reaction provided that the reaction is chemically reversible in the potential range examined. A typical cyclic voltammogram (CV) is shown in Fig. 2.16.

The key parameters derived from the voltammogram are the peak currents for the cathodic and anodic reactions, $i_{p,C}$ and $i_{p,A}$; the potential separation between the two peak currents; ΔE_P and the peak width and also the potential difference between the peak current and half the peak current on the rising portion of the *i-V* curve. Diagnosis of thermodynamic reversibility is important since this dictates which form of the Randles-Sevčik equation to use. Exploitation of these equations can be important in assessing biocompatibility (see Sect. 2.5.2).

The principal roles of CV in sensor development and their applications are the characterisation of the electrode reaction and assessing the performance of the electrode. Whilst the capacitance has little analytical value, adsorption of surface-active molecules such as proteins will displace counter ions and decrease the double layer capacitance. This can be calculated directly from the CV from regions where no electrolysis is occurring – the so-called double-layer region. Since $Q = CV$, taking the time derivative gives $i = CdV/dt$. Capacitance does depend on the applied potential but changes in capacitance at any given potential are a useful diagnostic.

In neurochemical applications, the background charging currents can be orders of magnitude larger than the faradaic current. This is usually dealt with using background subtraction. The target neurochemicals are not routinely present in the milieu of the electrode, but released due to electrical stimulation as part of the

normal experimental protocol. The background current-voltage trace is a recorder prior to stimulation. There are several problems with this approach. Background subtraction involves taking the difference between two noisy digitally recorded signals, which is always problematic since the difference may be close to the resolution of the analogue-to-digital conversion. A more serious objection is the assumption that the background current is independent of the faradaic subtraction. It is inconceivable that an electrode reaction which involves adsorption prior to electron transfer and one or more proton transfers does not affect the structure of the double layer. Notwithstanding these objections, which require highly-skilled and critical understanding of the limitations of the technique, the method has led to important insights into neurochemistry which could not have been observed with other techniques.

Differential pulse voltammetry and square wave voltammetry both involve modulation of a ramp or staircase respectively with a train of square pulses. By judiciously selecting the sampling period, the effects of double layer capacitance can be substantially reduced [104]. Both of these techniques offer limits of detection down to the nanomolar but are not continuous and are difficult to implement in the clinical setting. The interested reader is referred to standard electrochemistry texts for further details. Modulating a slow-moving voltage ramp with a sine wave, a technique known as a.c. voltammetry is a promising method that has been widely underused. However, it can be easily implemented using computer controlled instrumentation and is amenable to sophisticated signal processing (see below) and new semi-analytical solutions have been presented.

2.4.3 Instrumentation

Instrumentation for potentiometry is very straightforward and this is one of the appealing aspects of potentiometric devices. A high impedance voltmeter is all that is required along with some means of recording the voltage (usually a computer) and appropriate software for scaling the voltage and relating it to the calibration working curve or prompting user actions in the method of standard additions.

Steady state amperometry, which uses active non-equilibrium potentials, requires a stable voltage source which can respond rapidly to a current load that may vary by many orders of magnitude. Most readout devices (chart recorders, oscilloscopes, analogue-to-digital converters) require the signal to be in the form of a voltage, so some sort of current to voltage conversion is required. In the case of low currents, it may be possible to use a simple two-electrode set-up where the counter electrode also serves as a reference electrode. However, passing any current through the reference electrode can reduce sensor lifetime and if an array of electrodes is to be used, the combined sensor current could cause significant current flow leading to an error in the reference electrode potential (current across an electrified interface can only be sustained by electrolysis which almost inevitably leads to a change in potential, as given by the Nernst equation (*vide supra*)) and may

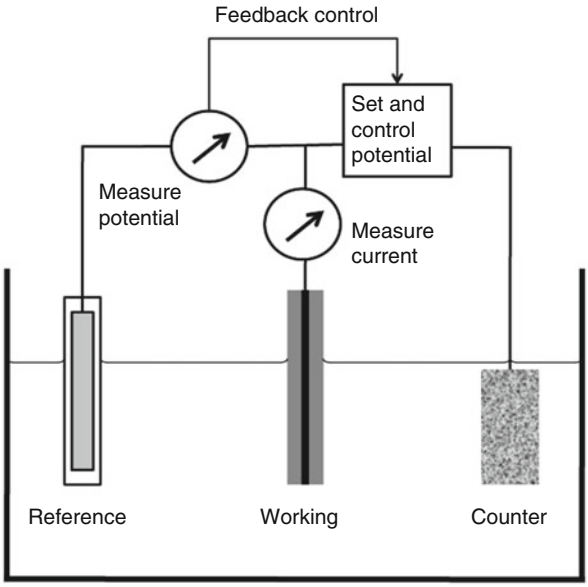


Fig. 2.17 Schematic of the instrumentation required for amperometry

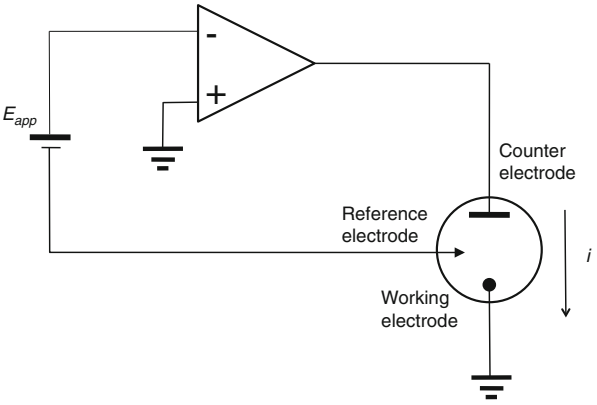


Fig. 2.18 Schematic of a potentiostat circuit based on a voltage follower

introduce hysteresis into the system. In many cases, a three-electrode set up is required. In this case, the electrode potential with respect to the reference electrode is maintained by a control amplifier and a third electrode is introduced to provide a current path. This is shown schematically in Fig. 2.17.

The functions outlined in Fig. 2.17 can be implemented using simple operational amplifier circuits. Usually, the working electrode (sensor) is held at ground or virtual ground and the potential is applied through a control amplifier to which the reference electrode and counter electrode are connected. A simple circuit for achieving this function, based on the voltage follower circuit, is shown in Fig. 2.18.

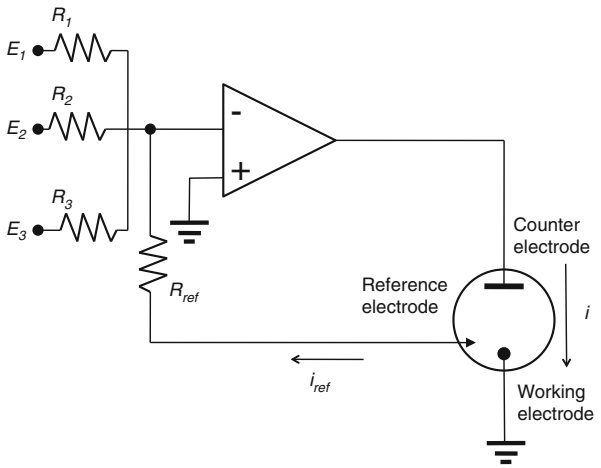


Fig. 2.19 A potentiostat based on a summing amplifier

Whilst this circuit fulfils the essential functions of the potentiostat, in that the reference electrode passes no current, it is not easily adapted for transient techniques where the voltage offset may need to be modulated with a pulse train or an a.c. voltage perturbation. In order to implement this useful function, an op-amp adder circuit can be used (Fig. 2.19):

In this circuit, the voltage applied to the electrochemical cell (or complete sensor) is given by the sum of the inputs to the three resistors (R_1, R_2, R_3), if they are of equal value. A disadvantage is that the reference electrode is now loaded by the resistor, R_{ref}. This can be easily overcome by placing a voltage follower into that limb of the circuit, between the reference electrode and $R_{ref.}$. The complete control amplifier function can now be implemented using a monolithic dual op-amp chip.

Current to voltage conversion is commonly achieved in two ways: (i) passing the current through a high precision measuring resistor and then using standard voltage amplifier circuits to provide adequate gain for interfacing to a chart recorder or analogue-to-digital conversion or (ii) a current follower. This second circuit has the advantage of holding the working electrode at virtual earth. The measuring resistor approach has the advantage of speed and low noise, but the working electrode takes a variable potential above ground. The current follower circuit maintains the working electrode at virtual earth which reduces the capacitance of the working electrode lead (the central conductor and the shield will be at the same potential) and minimises leakage currents, a major consideration when the current can be as low as picoamperes.

It is worth emphasising the advantages of applying the desired potential to the reference electrode and maintaining the working electrode at virtual earth. It is not unusual for electrochemical sensors to only produce nanoamperes or less. The applied voltage is less than 1 V. This implies an effective impedance of $>10^{10}$ Ω. Any paths to earth of less than a teraohm will cause serious errors if the working electrode is not at virtual earth. Furthermore, the non-inverting terminal can be used to drive the shield of the electrode cable further protecting signal integrity.

More recently [105], a new approach to current to voltage conversion has been employed in patch clamp amplifiers for neurophysiology. Developed by Axon, the input stage is a current integrator, thus reducing the effect of random noise. Clearly, the integrator needs to be reset periodically and the complete circuit is considerably more sophisticated and, unlike the circuits outlined above, are beyond the means of most laboratories to implement in home-made devices. Similarly, there is renewed interest in switched capacitor circuits for current to voltage conversion. Again, more sophisticated circuit analysis and design is required than is commonly available to electrochemists. These circuits have major limitations in that they are monopolar devices – the experimenter needs to select either positive or negative currents. This is less of a problem where target analytes have been identified in advance but such circuits would not be suitable for general lab use. Current integration ought to reduce noise right at the beginning of the signal processing chain. The signal is effectively digitised, though, through what is essentially a sample-and-hold circuit and it remains to be seen if these disadvantages are offset by improved signal to noise characteristics.

2.4.4 Signal Processing and Data Analysis

Although there have been tremendous advances in computing power in the last two decades, these have not thus far been translated into significant advances in the processing of data for electrochemical sensors. In fact, computers have largely been used to emulate the traditional signal generator and X-Y chart recorder approaches of half a century ago. Consequently, it is not unusual to record 50,000 pairs of data points in a cyclic voltammetry experiment only to use two or three of these in the analysis e.g. peak potential and current and half-peak potential and current. The analysis of these data then proceeds using the diagnostics developed by Nicholson and Shain in 1964 [106]. Finite difference modelling is more process intensive, but based on the same necessarily simplified models. It is used to test the similarity of the experimental data to predictions based on model reaction schemes. A major barrier is the non-linear nature of electrochemical signals which strictly precludes the use of Fourier transform approaches. Nonetheless, substantial progress has been made in using frequency space interpretations of the entire current–voltage characteristics, principally by Alan Bond and his group [107]. The approach typically involves the uses of a.c. voltammetry: a slow voltage ramp is applied to the electrochemical cell. The ramp allows a selection of the appropriate voltage for the electrochemical system of interest. Conventionally, a small amplitude sine wave is superimposed on the ramp to elicit kinetic and thermodynamic information but large amplitude perturbations provide improved signal to noise, particularly at the higher harmonics and the resulting signals have proven amenable to systematic theoretical analysis [108].

Whilst *ad hoc* modelling has undoubtedly been useful, it seems timely to apply some of the tools developed in other branches of engineering for time series analysis into electrochemistry. We have begun this process by applying the Hilbert transform to the study of immobilised redox species [109] at the surface of electrodes and have extended this work to include freely diffusing species [110, 111]. The aim of this work is to be able to deduce the thermodynamic (E°), kinetic (α, k°) and mass transport (D, concentration) parameters of all species present in solution. The combination of the physicochemical parameters ought to enable unambiguous identification and move electrochemistry away from a correlation-based approach to qualitative analysis and by altering the time scale of the experiment (by, for example, chirping the frequency) resolve redox active species that would otherwise overlap. A major advantage that is already conveyed is that the capacitance can be removed as an offset rather than through background subtraction. This is important since it is widely assumed in cyclic voltammetry that the background current is unaffected by the faradaic reaction. The Hilbert transform technique allows identification of when this cannot be true e.g. when the electrode reaction or spectator species adsorb on the electrode surface.

Furthermore, the technique can be used in reverse, to generate a digital filter that improves the selectivity of even simple devices. We were able to use optimised a.c. voltammetry waveforms to detect physiologically relevant concentrations of dopamine and serotonin (μM concentrations) [112]. These neurotransmitters are challenging to separate in voltammetry and cannot normally be distinguished using cyclic voltammetry because their peaks overlap. Conventional approaches to this problem have involved deposition of selective films on the sensors. Several other factors also complicate what is already a difficult analytical problem: ascorbate (vitamin C) is usually present at concentrations several hundred times higher than the target analytes and is also oxidised at overlapping potentials (serotonin and dopamine have formal potentials of 290 mV and 275 mV respectively); and background subtraction is not indicated since both dopamine and serotonin reaction products foul the electrode surface and displace counter ions leading to time-varying decreases in double layer capacitance. Optimisation of the potential waveform allowed simultaneous detection of serotonin and dopamine in the presence of a hundred-fold excess of ascorbate. These approaches work because although the thermodynamics properties are similar, their kinetics are different from each other and both exhibit such different kinetic characteristics from ascorbate that appropriately optimised waveforms can emphasise one reaction over another. The application of digital signal processing to electroanalysis is currently in it infancy. However, the preliminary results demonstrate their potential in improving analyte identification (through unambiguous determination of the characteristic physicochemical parameters) and improving the selectivity of simple easy-to-make sensors (Fig. 2.20).

New semi-analytic asymptotic solutions for a.c. voltammetry for surface confined species [113] and freely-diffusing species [114] have recently been reported and hold out the prospect of more rational exploration of system parameters.

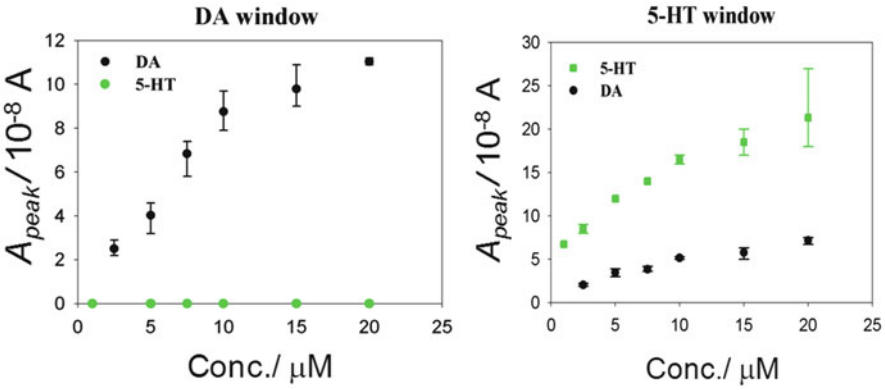


Fig. 2.20 Calibration curves for dopamine and serotonin (5-HT) using the amplitude extracted using the Hilbert transform response. The potential ramp was sine modulated at 150 Hz with an amplitude of 0.4 V. Dopamine and serotonin were distinguished by the choice of ramp potential. Ascorbate interference was removed using the high frequency and high amplitude modulation

2.5 Multiple Sensors and Microsensor Arrays

Microelectrode arrays offer advantages for bioanalysis. These advantages accrue from several sources relating to both the properties of living tissues and the understandable wish to measure more than one analytical variable. Firstly, one of the governing characteristics of living tissues is the large variability from site to site, on whatever length scale. This is discussed in the introduction above. Many disease processes, notably cancer, arise from the aberrant behaviour of one or a few cells. The central value obtained from averaging over a large number of cells will therefore be barely affected, however, the range of values will change. Given the widely differing range of analyte concentrations found in biology, for example the prostate cancer marker PSA, it is likely that the variance is at least as interesting as the central value. Such concepts are likely to be of value in personalised medicine.

Many of the target analytes in bioanalysis are surrogate variables that statistically correlate with the diagnostic target. Measuring more than one analyte or biomarker greatly adds to the confidence and may generate new insights into the underlying molecular mechanisms of disease. For example, simultaneous measurement of pH and histamine concentrations in isolated perfused stomach have elucidated the relative importance of the signalling pathways controlling acid secretion [67] and we have extended this technology to the array format [68].

Microelectrode arrays for recording from brain slices, cultured cells or even from in vivo preparations have been described for many years. Some of these are commercially available (Microchannel Systems GmbH, Ayanda Biosystems, 3Brain etc.). On the face of it, these ought to be easily adapted for electroanalysis – after all, all you really need is an array of inert, individually addressable conducting pads

of suitable dimensions. CMOS processing is mature technology and, in principle, allows for integrated electronics. Consequently, it is tempting to adapt for the apparently simple problem of generating massively parallel microelectrode arrays [115]. However, several problems are immediately evident. The materials used in conventional CMOS processing [116] are rarely suitable for biological applications. Ionic contamination is a major problem and extensive (and expensive) post-lithographic processing using hafnium or platinum, for example, is usually required [117, 118]. Conventional planar photolithography leads to recessed electrode geometry (unless two additional layers are used followed by chemical-mechanical lapping). The electrode hook-up tracks are usually insulated with <1 μm silicon nitride or oxide. Such thin dielectric allows some leakage of the electric field into the test specimen and will contribute to much larger capacitance than observed for conventional microelectrode sensors. This can cause problems with dynamic techniques where the *i-V* characteristics will be dominated by charging currents.

The principal area of failure for all microelectrodes is the metal-insulator seal. Such failures, often arising from unresolved thermal stresses, lead to hairline cracks, often of μm dimensions. Such failures lead to hysteresis and sluggish responses to changes in concentration. These defects can be hard to detect from conventional characterisation, e.g. measurement of the diffusion limited current, because the spatial resolution is of the order of $(2Dt)^{1/2}$ where D is the diffusion coefficient and t is the timescale of the experiment. Since D is around 10^{-10} m^2 s^{-1} and is at a steady state, the *i-V* curve takes around 100 s to record and asperities and defects less than 0.1 mm may not be evident. More rigorous testing procedures are described below. Such tests are not yet standard in the literature and this has led to over-optimistic conclusions in many cases.

Biocompatibility is the other major problem. Many devices reported in the literature are tested with cancer cell lines rather than primary mammalian cells. This is a very poor test for biocompatibility since one of the key characteristics of cancer cells is their ability to grow. Again, this has led to undue optimism about the suitability of many published devices. The other side of biocompatibility is the ability of the sensor to function in the presence of surface active spectator species. There are surprisingly few reports of sensor arrays being used in cell culture medium, the majority being reported in medium free buffer solutions. Most worryingly, some are reported in phosphate-buffered saline which will precipitate Ca^{2+} ions. Genuinely biocompatible sensor arrays are described below, along with rigorous tests for evaluating their performance.

It may be that processes designed for semiconductor microfabrication may be less than suitable for biosensor fabrication. After all, when solid state electronics came along, the manufacturing processes were developed from the ground up and not adapted from vacuum tube technology. A recent paper [119] describes the production of carbon-ring microelectrode arrays prepared from the pyrolysis of acetylene, which, though craft-intensive, could potentially be automated, and the production could thus be scaled up. Compton et al. have published a useful review [120] of fabrication methods, theoretical descriptions and characterisations.

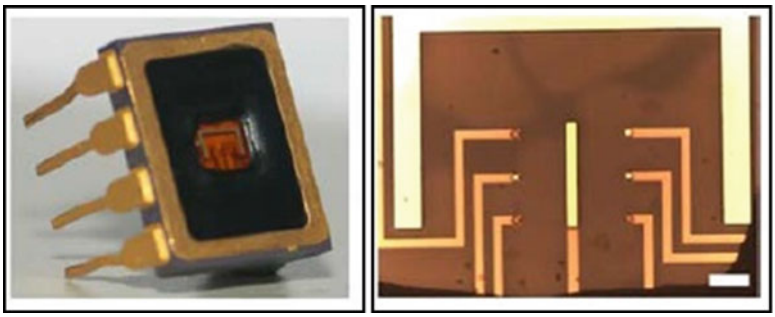


Fig. 2.21 DIL packaged microelectrode array (*left*) and a close up of the 4 mm × 4 mm sensing area

2.5.1 *Microelectrode Arrays for Primary Mammalian Cell Culture*

Over a series of papers, we have described the development of microelectrode arrays to measure signalling molecules. The devices come in two formats: a DIL package [121], and a more recently developed device based on transparent SiO_2 to enable simultaneous observation by light microscopy (see below). The DIL device (Fig. 2.21) was connected to the pins using wire-bonding. The SiO_2 based devices are connected via spring-loaded pins to a Faraday cage (Fig. 2.22) which has a temperature control (using a Peltier device with PID control implemented in LabView) to better than ±0.1 °C.

The devices are development platforms which can be post processed in the laboratory to alter the range of analytes. As bare gold, the arrays can detect nitric oxide (or more generally NO and NO_2^-) using differential pulse voltammetry (DPV) with a peak potential around +0.8 V and dissolved O_2 and hydrogen peroxide with peak potentials around −0.5 and −0.8 V respectively. Modification with iridium oxide (*vide supra*) gives an array of pH sensors and modification with oxidase enzymes and polyphenol allows detection of nutritional markers such as glucose or lactate.

The key to utilising these devices, however, is the biocompatibility from the cellular perspective and extensive physicochemical characterisation in the presence of biological components to test whether the sensors are still working. After a comparison of several different coating materials [122], air-dried fibronectin was shown to have superior properties both for cell culture and for maintaining sensor function over several days in the cell culture. Rehydrated fibronectin has a structure resembling female Velcro (see Fig. 2.23). This is likely to be due to the cysteine groups being well spaced in fibronectin leading to pores which barely affect the accessibility of the sensor surface for small molecules, but effectively repel albumin or exclude albumin or other surface active biopolymers.

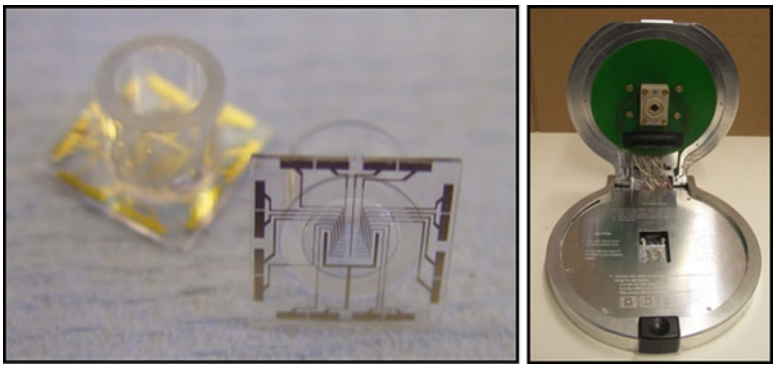


Fig. 2.22 The SiO_2-based array (*left*) and the sprung pin connecting Faraday cage for recording

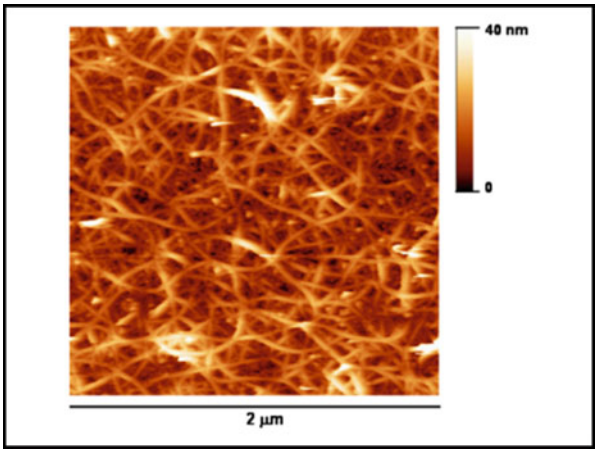


Fig. 2.23 Atomic force micrograph of air-dried fibronectin

2.5.2 Assessing Biocompatibility

An excellent introduction to the seemingly intractable problem of biocompatibility for implantable biosensors and chemical sensors has been published by Vadgama [11] and good reviews for all of the processes involved are covered by Meyerhoff [123].

Morais et al. [124] review the problems specific to implantable glucose sensors: the closed-loop operation of glucose sensors still seems an unsolved engineering challenge, though much progress has been made in recent years. New approaches to long-term biocompatibility of implanted devices have emerged: getting the sensor to integrate into the living tissue after decades of research, where the paradigm has been precisely the opposite, seems like a promising research avenue. These new approaches are in their infancy but include ambitious ideas, even to the extent of encouraging vascularisation by local infusion of vascular endothelial growth factor

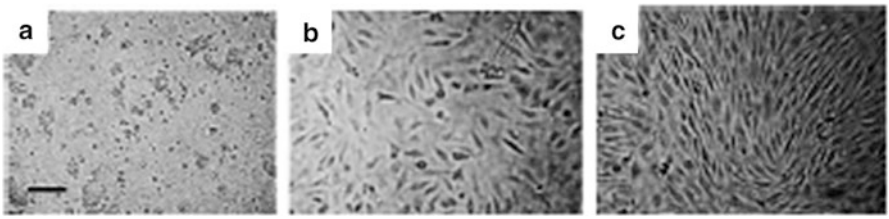


Fig. 2.24 Primary porcine endothelial cells grown on PSS-PLL (**a**), for 24 h (**b**) and 48 h (**c**) on air dried fibronectin coated polystyrene

(VEG-F) [125] or, even more radically, incorporating engineered cells expressing VEG-F into the sensor implantation site [126]. Such radical approaches are unlikely to achieve regulatory approval, but demonstrate the feasibility of incorporating the sensor into the patient. Synthetic constructions able to exploit similar signalling pathways would seem like an important next step.

For our in vitro devices, the problems are less severe and are largely to do with (i) the effects of the sensors on delicate cultured primary mammalian cells and (ii) the effect of culture medium components on sensor performance. We have taken the unusual step of considering both aspects to be important. Furthermore, We have also developed a protocol for rapid testing of biosensors for the likely ability to perform in the presence of biological systems. Firstly, examine the effects of 4 % (w/w) albumin. This is the same concentration of albumin that is found in plasma. Secondly, the rather savage 20 % emulsified chicken liver suspension. Examine the *i-V* characteristics. Key parameters are capacitance, peak width and peak separation. These are discussed in some detail, after we've decided what the sensors do to the cells.

- **What do the devices do to the cells?**
  Primary endothelial cells are adherent and, when healthy, have a characteristic elongated shape. Comparison between fibronectin and the putatively biocompatible synthetic copolymer polystyrene sulphonate-poly-L-lysine (PSS-PLL) is shown above (Fig. 2.24). After 48 h the cells on the fibronectin have grown to confluence and look healthy. In comparison, the cells on the PSS-PLL substrate almost literally curl up and die.
- **Does applied potential affect the cells?**
  We do not expect the electric potential to affect the cells directly because the Debye length, the shielding length scale for the electric field (*vide supra*), in the culture medium is less than 1 nm. However, negative electric potentials could lead to increased pH particularly due to O_2 reduction and, at extreme potentials, hydrolyse the water to form hydrogen. Similarly, at positive potentials, the solution in the vicinity of the electrode will increase if the water is electrolysed to oxygen. Depending on the electrode material, this would be expected to occur for potentials higher than around +1.0 V. We examined these possibilities on a transparent microelectrode array where cultured cells were exposed to different potentials for five minutes. The cells were then exposed to Trypan blue, a dye

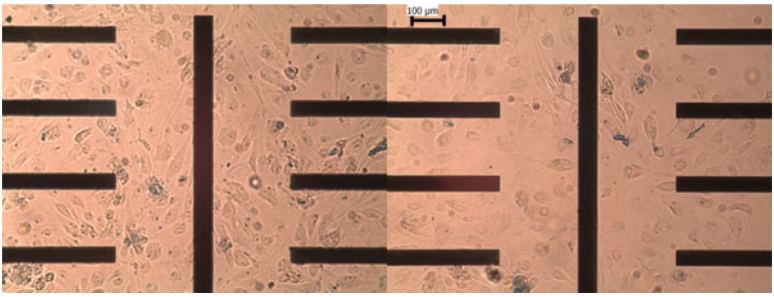


Fig. 2.25 Endothelial cells exposed to +1.5 V (*left*) and −1.0 V (*right*) (The pictures were taken from R. Trouillon, PhD thesis, Imperial College 2010)

which is only taken up by dead cells. Representative results are shown above (Fig. 2.25).

These results show some cell death, presumably due to pH changes close to the electrodes for potentials higher than 1.0 V and lower than −1.0 V.

- **What happens to the electrodes?**
  Adsorption of surface active biomolecules or deliberately added electrode modifiers might be expected to do several things to an electrode: hinder diffusion, inhibit electrocatalysis and affect the partitioning or solubility of the analyte. The effects on diffusion can be quantified using a reversible outer sphere electron transfer couple such as ruthenium (III) hexaammine, $Ru(NH_3)_6^{3+}$. Outer sphere redox is the simplest possible electron transfer process where no bonds are broken or formed and there is no adsorption of reactant or product. Reversibility means that the electron transfer is essentially instantaneous on the time scale of the experiment. Examination of the cyclic voltammograms shows that the peak width is consistent with reversibility (Fig. 2.26a).

  The peak current intensity is however decreased due to a decreased diffusion coefficient in the vicinity of the electrode.

  Adsorption, where the polymer forms an intimate relationship with the electrode surface, would be expected to show several effects: firstly, the capacitance would decrease as the counter ions are displaced by adsorbate. This can be assessed from a.c. impedance measurements by fitting the data to the Randles' model [48] or simply from the so-called double-layer region of the voltammogram:

$$Q = CV \therefore \frac{dQ}{dt} = i = C\frac{dV}{dt} \tag{2.42}$$

assuming the capacitance does not depend on potential (this is definitely not a good assumption, but big changes in capacitance are readily apparent and semi-quantitatively positive for these kind of measurements). Secondly, since adsorption is spontaneous, the free energy must be negative. Inevitably this means that surface sites, which are important for electrocatalytic reactions (where the analyte is

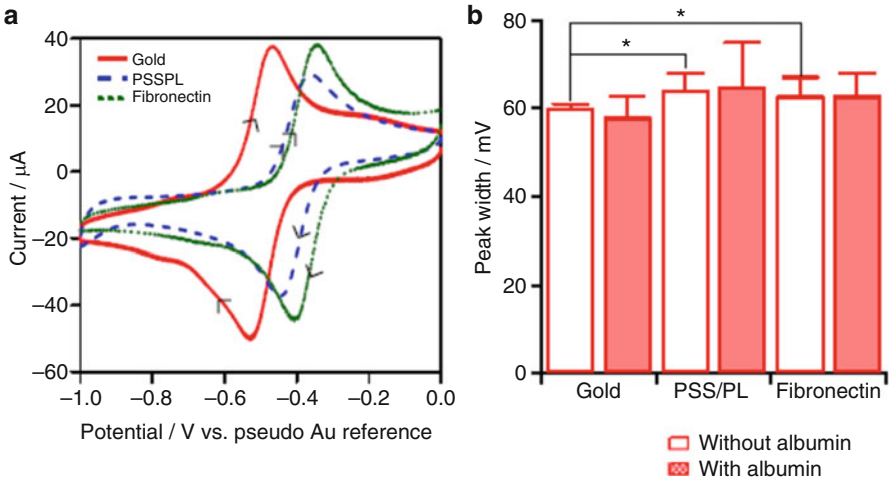


Fig. 2.26 Cyclic voltammograms examining the effect of surface coatings and exposure to albumin

adsorbed before electron transfer, e.g. O_2 reduction), will also be selectively hit by adsorption from proteins. These effects will show up in the peak width in particular or in measurement of the charge transfer resistance from a.c impedance measurements. It was this more rigorous testing that led to the conclusion that fibronectin is the preferred coating material – measurement of dissolved oxygen in fibronectin-coated electrodes though affected by the fibronectin, were not affected by albumin or chicken liver [127].

- **How big are my sensors?**
  This is a non trivial question whose answer depends on the relevant length scale, much in the same way as the measurement of the coastline of Great Britain depends on whether one measures round every single pebble on Brighton beach. Plainly, a sensible answer is required for proper quality control of microfabrication processes and to enable optimisation of the potential waveforms and associated amplification and data processing. Atomic scale measurements of the surface area will, unless the electrode is atomically flat, give larger areas than micrographs or measurements of the diffusion limited current. As mentioned above, the length scale for diffusion-based measurements will be of the order of $(2Dt)^{1/2}$. Measurement will then be no better resolved than a few tens of micrometres in most instances. This becomes a problem when one wishes to assess whether the electrode-insulator seal is intact, whether polishing has worked and if the nanoparticles used in electrode modification have actually deposited successfully in ohmic contact. For measurements based on capacitance, the resolution will be on the scale of the Debye length. This can be a rapid and convenient semi-quantitative measurement. However, accurate determination depends on there being reliable estimates of the specific capacitance of your electrode material in the electrolyte of choice. More reliable techniques for gold

electrodes involve measuring the area of gold oxide reduction peak. This peak area, conveniently generated during the electrochemical cleaning of the electrode in sulfuric acid (0.1 M) and presented in *i-t* form, can be related to near atomic scale area since 390 μC cm^{-2} of charge is passed on during its reduction [128]. Similar approaches can be employed for Pt electrodes, this time using the known unit cell size of β-hydride peaks and Faraday's law.

The meaning of electrode area is discussed in some detail in a report [129] prepared for the International Union of Pure and Applied Chemistry (IUPAC). This report is also an excellent source and critical review of methods commonly used to assess electrode area.

- **What can we use the arrays for?**
  We have used these devices to look at endothelial cell responses to angiogenin [130] and other growth factors [131]. The multiple sensor format has allowed signal averaging and real estimation of the range of cellular responses to stimulation. Because the devices were mass produced (using conventional lift-off processes), we were able to run many experiments in parallel and use well-established drugs to demonstrate the intracellular pathways in nitric oxide release. More recently we have applied similar technology to the study of host-pathogen interactions, specifically the release of nitric oxide from macrophages exposed to the protective antigen from *Bacillus anthracis* [132]. The potential for parallel experimentation and near real time and quantitative data offer significant advantages over fluorescence microscopy and enable interrogation of intracellular pathways.

2.6 New Materials

Several new materials have emerged in recent years that have great potential as sensor materials. In particular, novel forms of carbon including boron-doped diamond (BDD), carbon nanotubes (CNTs) and graphene have all been investigated for the properties as electrode materials.

Boron-doped diamond (BDD) is produced using P-CVD processes and retains many of the attractive properties of intrinsic diamond – high thermal conductivity, chemical inertness, lubricity – whilst doping at high levels ($[B] > 10^{20}$ cm^{-3}) leads to metallic conductivity [133]. A particularly attractive property of BDD is its relative resistance to the adsorption of either electrode reaction products or surface active biomolecules, presumably due to the sp^3 hybridisation of the carbon which leaves no opportunity to accept or donate electrons. Its resistance to adsorption has been compared favourably to glassy carbon [134] and the effects of doping levels on biocompatibility have been documented [135]. The wide potential window of BDD is another unique property. In principle this ought to mean that a wider range of analytes is available since hydrogen generation or oxygen generation at the negative and positive potential limits respectively, are hindered due to the non-electrocatalytic properties. In practice, however, the advantages of this wide

potential window have largely failed to be realised since many target analytes also depend upon an electrocatalytic electrode mechanism. Nonetheless, the low capacitance and long-term stability augur well for major applications in bioanalysis and environmental application. Electroanalytical applications of BDD in biology were reviewed by Swain [136, 137] and there have been important accounts of BDD in hostile environments such as the gut [138].

Much of the early work on BDD was hampered by poor reproducibility of the starting material and variable surface pre-treatments used by different investigators. The role of sp^2 impurities, grain boundaries, heterogeneities and terminating groups on the surface has now been elucidated using highly local measurements in a recent paper by Macpherson [139]. This has greatly clarified the important technical issues, identified the key parameters for electroanalysis and, in collaboration with Element 6, has led to reproducible commercially available electroanalytical grade BDD.

Carbon nanotubes are another promising material for electroanalysis. Their extraordinary mechanical properties, ballistic conductance in a large fraction of fibres and anomalously low capacitance make them attractive for biological applications. The tiny size (an astonishing 1–3 nm in diameter for pristine single-walled nanotubes) also holds out the possibility of highly localised measurements. Again, similarly to the early BDD work, considerable variation in performance is reported in the literature. In this instance, the role of catalyst impurities [140] (usually transition metals) was widely unrecognized in early reports. This was further confounded by the use of oxidizing reagents such as nitric acid which introduce oxygen functionality at the edge and break the nanotubes into smaller fragments. All of these factors affect the electrochemical performance. The role of the thin film behaviour of nanotube-modified electrodes also confounded early interpretations [141]. These early controversies are summarised and resolved in a good review [142]. Nanotube mats or sparse networks have been prepared using photoresist to mask of the catalysts. These pristine single walled nanotube devices allow the exploration of the unique properties of the nanotubes themselves. At a high density, the mats behave like complete metal films, despite the nanotubes occupying only a low percentage of the surface [143]. This gives very low capacitance, very high rates of mass transport and the ability to detect nanomolar concentrations using simple techniques such as cyclic voltammetry as demonstrated for the biologically significant serotonin [144]. In addition to their direct use as sensors, carbon nanotubes can also be used to template other materials [145] and have even been suggested for implantable devices [146].

Graphene is a relative newcomer to electrochemistry, though the subject of intense investigation. Again, purity and condition of the source material are confounding factors in interpreting some of the prematurely published accounts. The electrochemical properties of well-characterised graphene, free from copper contamination do not seem to differ greatly from basal plane graphite, though the high conductivity and optical transparency may offer advantages in some applications. Electrochemical applications have been recently reviewed [147] and applications in field effect transistor sensors have been reported [42].

2.7 Future Perspectives and Research Challenges

The goal of implantable complete devices comprising of sensors, instrumentation, signal processing, power and wireless data transmission remains in the future, but substantial progress continues to be made. Obviously, such complete devices will require new ways of thinking about the other key components, apart from the sensors. Batteries remain the most likely technology and power harvesting or other approaches remain research topics. Battery form is important and the ability to mass manufacture, or scale-up lab-based technology is critical. Recent progress has been reported for the Finnish-developed 'Enfucell softBattery' based on $Zn\text{-}MnO_2\text{-}ZnCl_2$ robust chemistry which is now commercially available in a suitable size (0.7 mm thick, the smallest is 42 × 60 mm), in a 10–90 mAh capacity and manufactured using reel-to-reel processes along with a 1–2 year shelf life. A device has been reported that is fabricated from materials described by the authors [148] as "completely edible" though more fastidious diners may baulk at silver nanowires and poly(glycerol-co-sebacate)-cinnamate. Graphene probably has more potential as a power source component than sensor [149], though the way forward may be metal-free and an all polymer PEDOT device has been reported [150].

Cheap, printable flexible displays are also an active research area. Electrochromic displays on PET substrates using organic transistors have been described consisting with demonstration of an 8 × 8 pixel display manufactured using solution processing based on standard printing and coating [151].

System integration remains the 'holy grail' but is still rarely reported, presumably because of the skill mix required for implementation transcends traditional disciplinary boundaries and the engineering problems remain significant on all aspects, ranging from biocompatibility and sensor stability (probably the most difficult challenge), through to power management, wireless data transmission and presentation of the data in a clinically meaningful form. A complete implantable device [152], the Nano-tera i-IronIC has recently been reported by de Micheli (from EPFL) at the DATE13 conference. At only 14 mm long, it consists of five sensors and a radio transmitter. The device is implanted just below the skin with an external 0.1 W battery patch. Peer-reviewed publication is eagerly awaited to see if the science lives up to the press release.

There continues to be major advances in biosensors for biomedical application. Whilst we now have a better understanding of long-standing problems of poor biocompatibility of typical sensor materials and poor stability of implanted devices, there are no magic bullets and the interface between the sensor and the biology remains the major problem. New materials, in particular, present both new opportunities and new challenges. Array technology offers a broader range of analytes, better stability through sensor redundancy and a better understanding of the role of analyte variance in biology.

References

1. McDonagh C, Burke CS, MacCraith BD, 'Optical chemical sensors' *Chem. Rev.* 108 400–422 (2008).
2. Turner APF, 'Biosensors: sense and sensibility' *Chem. Soc. Rev.* 42 31284–3196 (2013).
3. Moore FA, McKinley BA, 'The next generation in shock resuscitation' *Lancet* 363 1988–1996 (2004).
4. Sanders GHW & Manz A, 'Chip-based microsystems for genomic and proteomic analysis' *Trac Trends in Analytical Chemistry*, 19(6) 364–378 (2000).
5. Lander ES 'The new genomics: global views of biology' *Science* 274 5287 (1996).
6. Nielsen J and Oliver S 'The next wave in metabolome analysis' *Trends in Biotechnology* 23(11) 544–546 (2005).
7. Croston GE 'Functional cell-based uHTS in chemical genomic drug discovery'. *Trends in Biotechnology* 20(3) 110–115(2002).
8. Dittrich PS and Manz A, 'Lab-on-a-chip: microfluidics in drug discovery' *Nature Reviews Drug Discovery* 5(3): p. 210–218 (2006).
9. Wunberg T, et al., 'Improving the hit-to-lead process: data-driven assessment of drug-like and lead-like screening hits' *Drug Discovery Today*,. 11(3–4) 175–180 (2006).
10. Hong J, Edel JB and deMello AJ, 'Micro- and nanofluidic systems for high-throughput biological screening'. *Drug Discovery Today* 14(3–4)134–146 (2009).
11. Vadgama PM, Schoeleber in P.N. Bartlett (ed.) 'Bioelectrochemistry: fundamentals, experimental techniques and applications' Wiley 2008, Chapter 13.
12. Seo J-H, Leow PL, Cho S-H, Lim H-W, Kim J-Y, Patel BA, Park J-G & O'Hare D, 'Development of inlaid electrodes for whole column electrochemical detection in HPLC' *Lab on a Chip* 9 2238–2244 (2009).
13. Malinski T & Taha Z, *Nature* 358 676–678 (1992).
14. Updike JW & Hicks JP, *Nature* 214 986–988 (1967).
15. Bartlett PN & Caruana DJ, *Analyst* 117 1287 (1992)
16. Khurana M.K, Winlove C.P & O'Hare D. 'Detection Mechanism of Metallised Carbon Epoxy Oxidase Enzyme Based Sensors' *Electroanalysis* 15 1023–1030 (2003).
17. Barker SA in Turner APF *et al.* (eds.) *Biosensors: Fundamentals and Applications* Oxford University Press, Oxford 1987. Ch. 6.
18. Willner I, *Science* 229 1877 (2003).
19. Cass AEG, Davis G, Francis GD, Hill HAO, Aston WJ, Higgins LJ, Plotkin EV, Scott LD & Turner APF, *Anal. Chem.* 56 657 (1984).
20. Bartlett PN & Bradford VQ, *J.C.S. Chem Comm.* 16 1135–1136 (1990)
21. Zhao S, Korell U, Cuccia L & Lennox RB, *J. Phys. Chem.* 96 5641–5652 (1992).
22. Boutelle MG, Stanford C, Fillenz M, Albery MJ & Bartlett PN, *Neuroscience. Lett.* 72 283–288 (1986).
23. Ye M, Hammerle AJJ, Olstehoorn W, Schumann W, Schmidt AK, Duine JA & Heller A, *Anal. Chem.* 65 238 (1993).
24. Heller A *et al. Acc. Chem. Res.* 23 128 (1990).
25. Song KM, Lee S, Ban C, 'Aptamers and Their Biological Applications' *Sensors* 12 612–631 (2012).
26. Fodey T, Leonard P, O'Mahony J. *et al.* 'Developments in the production of biological and synthetic binders for immunoassay and sensor-based detection of small molecules' TRAC-Trends in analytical chemistry 30 254–269 (2011)
27. Iliuk AB, Hu L, Tao WA, 'Aptamer in Bioanalytical Applications' *Anal. Chem.* 83 4440–4452 (2011).
28. Ellington AD & Szostak JW '*In vitro* selection of RNA molecules that bind specific ligands' *Nature* 346 818–22 (1990).
29. Liu J, Cao Z, Lu Y, 'Functional nucleic acid sensors' *Chem. Rev.* 109 1948–1998 (2009).

30. Tombelli S, Minunni A & Mascini A, 'Analytical applications of aptamers' *Biosensors & Bioelectronics* 20 2424–2434 (2005).
31. Liu, Y, Matharu Z, Howland MC, Revzin A & Simonian AL 'Affinity and enzyme-based biosensors: recent advances and emerging applications in cell analysis and point-of-care testing' *Anal. Bioanal. Chem.* 404 1181–1196 (2012)
32. Zhou J, Battig M R, Wang Y, 'Aptamer-based molecular recognition for biosensor development' *Anal. Bioanal. Chem.* 398 2471–2480 (2010)
33. Palchetti I & Mascini M 'Electrochemical nanomaterial-based nucleic acid aptasensors' *Anal. Bioanal. Chem.* 402 3103–3114 (2012).
34. Yano K, Karube I *TRAC Trends Anal Chem* 18 199–204 (1999).
35. Shimizu KD, Stephenson CJ *Curr. Opin. Chem. Biol.* 14 743–750 (2010)
36. Suryanarayanan V, Wu C-T, Ho K-C *Electroanalysis* 22 1795–1811 (2010).
37. Malitesta C & Mazzotta E, Picca RA, Poma A, Chianella I & Piletsky SA, 'MIP sensors – the electrochemical approach' *Anal. Bioanal. Chem.* 402 1827–1846 (2012).
38. Wang X-D. & Wolfbeis OS, 'Fiber-Optic Chemical Sensors and Biosensors (2008–2012)' *Anal. Chem.* 85 487–508 (2013)
39. Stewart MER, Anderton LB, Thompson J, Maria SK, Gray JA, Rogers & RG Nuzzo 'Nanostructured plasmonic sensors' *Chem. Rev.* 108 494–521 (2008)
40. Kergoat L, Piro B, Berggren M, Horowitz G & Pham M-C 'Advances in organic transistor-based biosensors: from organic electrochemical transistors to electrolyte-gated organic field-effect transistors' *Anal. Bioanal. Chem.* 402 1813–1826 (2012).
41. Goeders KM, Colton JS, Bottomley LA, 'Microcantilevers: Sensing Chemical Interactions via Mechanical Motion' *Chem. Rev.* 108 522–542 (2008).
42. Stine R, Mulvaney SP, Robinson JT, Tamanaha CR & Sheehan PE, 'Fabrication, Optimization, and Use of Graphene Field Effect Sensors' *Anal. Chem.*, 85 509–521 (2013).
43. Kim DJ, Sohn IY, Jung JH, Yoon OJ, Lee NE, Park JS 'Reduced graphene oxide field-effect transistor for label-free femtomolar protein detection' *Biosens Bioelectron.* 41 621–626 (2013).
44. Kwak YH. Choi DS, Kim YN, Kim H, Yoon DH, Ahn S-S, Yang JW, Yang WS, Seo S 'Flexible glucose sensor using CVD-grown graphene-based field effect transistor' *Biosens. Bioelectron.*, 37, 82–87 (2012).
45. McKinley BA, 'ISFET and fiber optic sensor technologies: in vivo experience and critical care monitoring', *Chem. Rev.* 108 826–44 (2008)
46. Kimmel DW, LeBlanc G, Meschievitz ME, and Cliffel DE 'Electrochemical Sensors and Biosensors' *Anal. Chem.* 84 685–707 (2012).
47. Pletcher D 'A first course in electrode processes' 2nd ed. Royal Society of Chemistry, Cambridge 2009.
48. Bard AJ & Faulkner LR 'Electrochemical methods: fundamentals and applications' 2nd ed. Wiley, Chichester 2001.
49. Kissinger P & Heineman WR, 'Laboratory techniques in electroanalytical chemistry' 2nd ed. Marcel Dekker, New York 1996.
50. Pletcher D, Greff R, Peat R, Peter LM, Robinson J, Woodhead Publishing, Cambridge 2001.
51. Purves RD, 'Microelectrode methods for intracellular recording and ionophoresis' Academic Press, London 1981.
52. Amman D, 'Ion-selective microelectrodes: Principles, Design and Applications' Springer-Verlag: Berlin, 1986.
53. Voipio J. In 'pH and Brain Function' Kaila, K., Ransom, B. R., (Eds.) Wiley-Liss: New York, 1998; p 95.
54. Takahashi A, Camacho P, Lechleiter J, 'Measurement of intracellular calcium' Herman, B. *Physiol. Rev.* 1999, *79*, 1089.
55. Cremer M, 'Über die Ursache der elektromotorischen Eigenschaften der Gewebe, zugleich ein Beitrag zur Lehre von Polyphasischen Elektrolytketten.' *Z. Biol.* 47 56 (1906).

56. Idegami K, Chikae M, Nagatani N, Tamiya E, Takamura Y, 'Fabrication and Characterization of Planar Screen-Printed Ag/AgCl Reference Electrode for Disposable Sensor Strip' *Jpn. J. Appl. Phys.,* 49 097001–097003 (2010).
57. Park WJ, Yi Y, Lee J, Lee BC, Park OK, Lee HJ, Lee H, 'N-heterocyclic carbene–silver complex as a novel reference electrode in electrochemical applications' *Talanta* 81 482–485 (2010).
58. Noh J, Park S, Boo H, Kim HC, Chung TD 'Nanoporous platinum solid-state reference electrode with layer-by-layer polyelectrolyte junction for pH sensing chip' *Lab on a Chip* 11 664–671 (2010).
59. Capone S, A de Robertin, C de Stefano, & Scarcella R, 'Ionic strength dependence of formation constants X. Proton activity coefficients at various temperatures and ionic strength and their use in the study of complex equilibria', *Talanta* 34 593–598 (1987).
60. Fedirko N, Svichar N, & Chesler M, 'Fabrication and use of high speed concentric H+ and Ca^{2+} selective microelectrodes suitable for *in vitro* extracellular recording' *J. Neurophysiol.* 96, 919. (2006),
61. Ujec E, Keller O, Machek J, and Pavlik V. Low impedance coaxial K+ selective microelectrodes. *Pfluegers Arch* 382 189–192, 1979
62. Ives DJG in "Reference Electrodes" D.J.G. Ives & G.J. Janz (eds.) Academic Press, New York, 1961, Chapter 7.
63. Głab S, Hulanicki A, Edwall G & F Ingman, 'Metal-metal oxide and metal oxide electrodes as pH sensors' *Crti. Rev. Anal. Chem.* 21 29–47 (1989).
64. O'Hare D, Parker KH & Winlove CP "Metal-metal oxide pH sensors for physiological application" *Medical Engineering & Physics* 28 982–8 (2006)
65. Haggard HW & Greenberg LA, 'An antimony electrode for the continuous recording of the acidity of human gastric contents' *Science,* (1941) 93 479.
66. Horrocks BR, Mirkin MV, Pierce DT, Bard AJ, Nagy G, Toth K, 'Scanning Electrochemical Microscopy XIX. Ion-Selective Potentiometric Microscopy' *Anal. Chem.* (1993) 65 1213.
67. Patel BA, Bitziou E & O'Hare D, 'Simultaneous detection of pH changes and histamine release from oxyntic glands in isolated stomach' *Anal. Chem.* 80 8733–8740 (2008).
68. Bitziou E, O'Hare D & Patel BA 'Spatial changes in acid secretion from isolated stomach tissue using a pH-histamine sensing microarray' *Analyst* 135 482–487 (2010).
69. Hitchman ML, & Ramanathan S, 'Evaluation of iridium oxide electrodes formed by potential cycling as pH probes' *Analyst* 113 35(1988).
70. Yamanaka K, 'Anodically electrodeposited iridium oxide films (AEIROF) from alkaline solutions for electrochromic display devices' *Jap. J. Appl. Phys.* 28 632–37 (1989).
71. Bezbaruah AN; Zhang TC, 'Fabrication of anodically electrodeposited iridium oxide film pH microelectrodes for microenvironmental studies' *Anal. Chem.* 74 5726–5733 (2002).
72. Marzouk SAM 'Improved electrodeposited iridium oxide pH sensor fabricated on etched titanium substrates' *Anal. Chem.* 75 1258–1266 (2003).
73. Moussy F, & Harrison DJ 'Prevention of the Rapid Degradation of Subcutaneously Implanted Ag/AgCl Reference Electrodes Using Polymer Coatings' *Anal. Chem.* 66 674–679 (1994).
74. Bakker EP, Bühlmann P & Pretsch E, 'Carrier-based ion-selective electrodes and bulk optodes 1. General characteristics' *Chem. Rev.* 97 3083–3132 (1997).
75. Bühlmann P, Pretsch E & Bakker E, 'Carrier-based ion-selective electrodes & bulk optodes 2. Ionophores for potentiometric and optical sensors', *Chem. Rev.* 98 1593–1688 (1998).
76. Bakker E & Pretsch E, 'Potentiometric sensors for trace metal analysis' *TrAC Trends in Analytical Chemistry* 24 199–207 (2005).
77. Pretsch E, 'The new wave of ion-selective electrodes' *TrAC, Trends in Analytical Chemistry* 26 46–51 (2007).
78. Bobcka J, Ivaska A & Lewenstam A, 'Potentiometric ion sensors' 108 329–51 (2008).
79. Lindner E & Buck RP, 'Microfabricated potentiometric electrodes and their *in vivo* applications' *Anal. Chem.*, 72, 336A (2000).

80. Bausells J, Errachid A, Zine N, 'Biosensors based on standard dielectric materials' *Proc. Electrochem. Soc*. 2003-1 'Dielectrics in emerging technologies' 48. 2003-1, 48. (2003).
81. Wygladacz K, Durnas M, Parzuchowski, P, Brzozka, Z & Malinowska E. 'Miniaturized sodium-selective sensors based on silicon back-side contact structure with novel self-plasticizing ion-selective membranes' *Sens. Actuators, B* 95 366–372 (2003).
82. Sundfors F, Bereczki R, Bobacka J, Toth K, Ivaska A & Gyurcsányi, R. E. 'Microcavity Based Solid-Contact Ion-Selective Microelectrodes' *Electroanalysis*, *18*, 1372–8 (2006).
83. Guenat OT, Generelli S, de Rooij NF, Koudelka-Hep M, Berthiaume F & Yarmush MC, 'Development of an Array of Ion-Selective Microelectrodes Aimed for the Monitoring of Extracellular Ionic Activities', *Anal. Chem*. 78 7453–60 (2006)
84. Tymecki Ł, Zwierkowska E, Koncki R. 'Screen-printed reference electrodes for potentiometric measurements' *Anal. Chim. Acta*, 526 3–11 (2004).
85. Liao W-Y, & Chou T-C 'Fabrication of a Planar-Form Screen-Printed Solid Electrolyte Modified Ag/AgCl Reference Electrode for Application in a Potentiometric Biosensor' *Anal. Chem*. 78 4219 (2006).
86. Vazquez M, Danielsson P, Bobacka J, Lewenstam A. & Ivaska A. 'Solution-cast films of poly (3,4-ethylenedioxythiophene) as ion-to-electron transducers in all-solid-state ion-selective electrodes' *Sens. Actuators, B* 97 182 (2004).
87. Cosofret V, Erdösy M, Johnson TA, Buck RP, Ash RB, Neuman MR *Anal. Chem*. 1995, *67*, 1647–1653.
88. Lindner E, Buck R. *Anal. Chem*. 2000, 72, 336A–345A.
89. Uhlig A, Lindner E, Teutloff C, Schnakenberg U & Hintsche R 'Miniaturized ion-selective chip electrodes for sensor application' *Anal. Chem.,* 69, 4032–4038 (1997).
90. Yoon HJ, Shin JH, Lee SD, Nam H, Cha GS, Strong TD Brown RB 'Solid state ion sensors with a liquid junction-free polymer membrane-based reference electrode for blood analysis' *Sens. Actuators, B 64*, 8–14(2000).
91. Compton RG & Banks CE 'Understanding voltammetry' World Scientific Press, New Jersey, 2007.
92. Bockris JO'M, Reddy AKN 'Modern Electrochemistry Volume 2, Part 2', Plenum, New York (2000).
93. Wightman RM & Wipf DO, in Bond A (ed.) 'Electroanalytical Chemistry Volume 15' Marcel Dekker, New York 1989, p267.
94. Saito Y *Rev. Polarograph*. 15 177 (1968).
95. Bond, Luscombe AMD, Oldham KB & Zoski CG, *J. Electroanal. Chem*. 249 1–14.
96. Patel BA, Arundell M, Parker KH, Yeoman MS & O'Hare D, 'Microelectrode investigation of neuronal ageing from a single identified neurone' *Physical Chemistry Chemical Physics* 12(34) 10065–10072 (2010)
97. Amatore C, Oleinick AI, Svir A. 'Diffusion from within a Spherical Body with Partially Blocked Surface: Diffusion through a Constant Surface Area' *Chem. Phys. Chem*. 11 149 (2009); 'Reconstruction of Aperture Functions During Full Fusion in Vesicular Exocytosis of Neurotransmitters' *ib*. 159 (2009).
98. Patel BA, Arundell M, Parker KH, Yeoman MS, O'Hare D, 'Detection of nitric oxide release from single neurons in the pond snail, *Lymnaea stagnalis*' *Anal. Chem*. 78(22) 7643–7648 (2006).
99. Yeoman MS, Patel BA, Arundell M, Parker K, O'Hare D 'Synapse-specific changes in serotonin signalling contribute to age-related changes in the feeding behaviour of the pond snail, *Lymnaea*' *J. Neurochem*. 106(4) 1699–1709 (2008).
100. Phillips CG & Jansons KM, *Proc. Roy. Soc. A* 428 431–449.
101. O'Hare D, Parker KH, Winlove CP, *J. Biomed Eng* 13 304 (1991).
102. Justice JB (ed.) *Voltammetry in the Neurosciences* Humana, Clifton, New Jersey 1987.
103. Cahill PS, Walker QD, Finnegan JM, Mickelson GE, Travis ER & Wightman RMW, *Analytical Chemistry* 68 3180–3186 (1996).

104. Compton RG & Banks CE, 'Understanding Voltammetry' World Scientific, New Jersey 2007.
105. Hochstetler SE, Puopolo M, Gustincich S, Raviola E, Wightman RM, 'Real time amperometric measurements of zeptomole quantities of dopamine released from neurons', *Anal. Chem.* 72 489–496 (2006).
106. Bard AJ & Faulkner LR, *Electrochemical Methods* Wiley, New York 1980.
107. Fleming BD, Barlow NL, Zhang J, Bond AM, Armstrong FA, Application of Power Spectra Patterns in Fourier Transform Square Wave Voltammetry to Evaluate Electrode Kinetics of Surface Confined Proteins, *Anal. Chem.,* 78 2948–2956 (2006)
108. Zhang J, Bond AM, Theoretical Studies of Large Amplitude Alternating Current Voltammetry for a Reversible Surface-Confined Electron Transfer Process Coupled to a Pseudo First-Order Electrocatalytic Process, *J. Electroanal. Chem.* 600 23–34 (2007).
109. Anastassiou C, Parker KH & O'Hare D 'Determination of kinetic and thermodynamic parameters of surface confined species through ac voltammetry and a nonstationary signal processing technique: the Hilbert transform' *Anal. Chem.* 77 3357–3364 (2005).
110. Anastassiou CA, Patel BA, Parker KH & O'Hare D, 'Characterisation of ac voltammetric reaction - diffusion dynamics: from patterns to physical parameters', *Anal. Chem.* 78 4383–4389 (2006).
111. Anastassiou CA , Parker KH, O'Hare D, 'Scaling in nonstationary voltammetry representations', *J Phys. Chem. A*, 111, 13053–13060 (2007).
112. Anastassiou CA, Patel BA, Arundell M, Yeoman MS, Parker KH and O'Hare D, 'Novel subsecond voltammetric separation between dopamine and serotonin in the presence of ascorbate' *Anal. Chem.* 78(19) 6990–6998 (2006).
113. Bell CG, Anastassiou CA, O'Hare D, Parker KH & Siggers JH, Theoretical treatment of high-frequency, large-amplitude ac voltammetry applied to ideal surface-confined redox systems, *Electrochimica Acta* 64 71–80 (2012).
114. Bell CG, Anastassiou CA, O'Hare D, Parker KH , Siggers JH 'Large-amplitude ac voltammetry: Theory for reversible redox reactions in the slow scan limit approximation', *Electrochimica Acta* 56(17) 6131–6141 (2011).
115. Heer F, Franks W, Blau A, Taschini S, Ziegler C, Hierlemann A, Baltes H 'CMOS microelectrode array for the monitoring of electrogenic cells' *Biosensors and Bioelectronics* 20(2) 358–366 (2004).
116. Graham AHD, Robbins J, Bowen CR *et al.,* 'Commercialisation of CMOS Integrated Circuit Technology in Multi-Electrode Arrays for Neuroscience and Cell-Based Biosensors' *Sensors* 11(5) 4943–4971 (2011).
117. Graham AHD, Bowen CR, Surguy SM, 'New prototype assembly methods for biosensor integrated circuits' *Medical Engineering & Physics* 33(8) 973–979 (2011).
118. Graham AHD, Surguy SM, Langlois P, Bowen CR, Taylor J et al. Modification of standard CMOS technology for cell-based biosensors, *Biosensors & Bioelectronics*, 31(1) 458–462 (2012).
119. Lin Y, Trouillon R, Svensson MI, Keighron JD, Cans AS and Ewing AG, Carbon-Ring Microelectrode Arrays for Electrochemical Imaging of Single Cell Exocytosis: Fabrication and Characterization. *Anal. Chem.* 84 2949–2954 (2012).
120. Huang X-J, O'Mahony AM, & Compton RG, 'Microelectrode Arrays for Electrochemistry: Approaches to Fabrication', *Small*, 5 776–788 (2009)
121. Patel BA, Arundell M, Quek RGW, Harvey SLR, Ellis IR, Florence MM, Cass AEG, Schor AM, O'Hare D, 'Individual addressable microelectrode array for monitoring oxygen and nitric oxide release' *Anal. Bioanal. Chem.* 390(5) 1379–1387 (2008).
122. Trouillon R, Combs Z, Patel BA and O'Hare D, 'Comparative study of the effect of various electrode membranes on biofouling and electrochemical measurements' *Electrochem. Comm.* 11 1409–1413 (2009).
123. Frost M & Meyerhoff ME, '*In vivo sensors: tackling biocompatibility*' *Anal. Chem.* 78 7370–7377 (2006).

124. Morais JM, Papadimitrakopoulos IF, and Burgess DJ 'Biomaterials/Tissue Interactions: Possible Solutions to Overcome Foreign Body Response' *American Association of Pharmaceutical Scientists J*. 12 188–196 (2010).
125. Ward WK, Quinn MJ, Wood MD, Tiekotte KLr, Sudha Pidikiti, Jennifer A. Gallagher *Biosensors 7 Bioelectronics* 19 155–163 (2003) 'Vascularizing the tissue surrounding a model biosensor: how localized is the effect of a subcutaneous infusion of vascular endothelial growth factor (VEGF)?'
126. Klueh U, Dorsky DI, Kreutzer DL, 'Enhancement of implantable glucose sensor function in vivo using gene transfer-induced neovascularization' *Biomaterials* 26 1155–63 (2005).
127. Trouillon R, Cheung C, Patel BA & O'Hare D, 'Comparative study of poly(styrenesulphonate)/poly(L-lysine) and fibronectin as biofouling-preventing layers in dissolved oxygen electrochemical measurements' *Analyst* 134 784–93 (2009).
128. Ressinea A, Vaz-Domínguez C, Fernandez VM, De Lacey AL, Laurell T, Ruzgas T, Shleev S 'Bioelectrochemical studies of azurin and laccase confined in three-dimensional chips based on gold-modified nano-/microstructured silicon' *Biosensors and Bioelectronics* 25 1001–1007 (2010).
129. Trasatti S and Petriim OA 'Real surface area measurements in electrochemistry' *Pure & Appl. Chem.,* 63, 71 1–734 (1991)
130. Trouillon R, Kang, D-K, Park H, Chang S-I & O'Hare D, 'Angiogenin induces nitric oxide synthesis in endothelial cells by PI-3 kinases and nuclear translocation pathways' *Biochemistry* 49 3282–3288 (2010).
131. Trouillon R, Cheung C, Patel BA, O'Hare D, 'Electrochemical study of the intracellular transduction of vascular endothelial growth factor induced nitric oxide synthase activity using a multi-channel biocompatible microelectrode array' *Biochim. Biophys. Acta (Gen. Subj.)*, 1800(9) 929–936 (2010).
132. Trouillon R, Williamson ED, Saint RJ, O'Hare D, 'Electrochemical detection of the binding of *Bacillus anthracis* protective antigen (PA) to the membrane receptor on macrophages through release of nitric oxide' *Biosensors and Bioelectronics* 38 138–144 (2012).
133. Patten HV, Meadows KE, Hutton LA, Iacobini JG, Battistel D, McKelvey K, Colburn AW, Newton ME, Macpherson JV, and Unwin PR, 'Electrochemical Mapping Reveals Direct Correlation between Heterogeneous Electron-Transfer Kinetics and Local Density of States in Diamond Electrodes' Angew. Chem. Int. Ed. 51, 7002–7006 (2012)
134. Trouillon R & O'Hare D 'Comparison of glassy carbon and boron doped diamond electrodes: resistance to biofouling' *Electrochimica Acta* 55(22) 6586–6595 (2010).
135. Trouillon R, O'Hare D, Einaga Y, 'Effect of doping level on the biological stability of hydrogenated boron doped diamond electrodes' *Physical Chemistry Chemical Physics* 13 5422–5429 (2011).
136. Park J, Quaiserova-Mocko V, Patel BA, Novotny M, Liu A, Bian X, Galligan JJ, Swain G M 'Diamond microelectrodes for in vitro electroanalytical measurements: current status and remaining challenges' *Analyst*, 133 17–24 (2008).
137. Park J, Show Y, Quaiserova V, Galligan JJ, Fink GD, Swain GM, 'Diamond microelectrodes for use in biological environments' *J. Electroanal. Chem*. 583 56–68 (2005).
138. Patel BA 'Electroanalytical approaches to study signaling mechanisms in the gastrointestinal tract' *Neurogastroenterlogy & Motility* 23 595–605
139. Hutton LA, Iacobini JG, Bitziou E, Channon RB, Newton ME, Macpherson JV, 'Examination of the factors affecting the electrochemical performance of oxygen-terminated polycrystalline boron-doped diamond electrodes' *Anal. Chem.* in press dx.doi.org/10.1021/ac401042t|A.
140. Jones CP, Jurkschat K, Crossley A, Compton RG, Riehl BL, Banks CE, 'Use of high-purity metal-catalyst-free multiwalled carbon nanotubes to avoid potential experimental misinterpretations' *Langmuir*, 23 9501–9504 (2007).
141. Sims MJ, Rees NV, Dickinson EJF, Compton RG, 'Effects of thin-layer diffusion in the electrochemical detection of nicotine on basal plane pyrolytic graphite (BPPG) electrodes modified with layers of multi-walled carbon nanotubes (MWCNT-BPPG)' *Sensors and Actuators B*, 144, (2010), 153–158.

142. Dumitrescu I, Unwin PR, and Macpherson JV, 'Electrochemistry at Carbon Nanotubes: Perspectives and Issues' *Chem. Commun*., 45, 6886–6901 (2009).
143. P. Bertoncello JP, Edgeworth JV, Macphersonand Unwin PR 'Trace level cyclic voltammetry facilitated by single-walled carbon nanotube network electrodes' *J. Am. Chem. Soc*. 129 10982–10983 (2007).
144. Güell AG, Meadows KE, Unwin PR, and Macpherson JV, 'Trace voltammetric detection of serotonin at carbon electrodes: comparison of glassy carbon, boron doped diamond and carbon nanotube network electrodes' *Phys. Chem. Chem. Phys*. 12 10108–10114 (2010).
145. Dudin PV, Snowden ME, Macpherson JV, and Unwin PR 'Electrochemistry at nanoscale electrodes (NSEs): individual single walled carbon nanotubes (SWNTs) and SWNT-templated metal nanowires' *ACS Nano*, 5 10017–10025 (2011)
146. Boero C, Carrara S, de Micheli G 'Long-term Biosensors for Metabolite Monitoring by using Carbon Nanotubes' *Sensors & Transducers Journal,* 125, , 229–237 (2011).
147. Brownson DAC, Kampouris DK, Banks CE, 'Graphene electrochemistry: fundamental concepts through to prominent applications' *Chem. Soc. Rev.,* 41 6944–6976 (2012).
148. Kim YJ, Chun S-E, Whitacreab J & Bettinger CJ, 'Self-deployable current sources fabricated from edible materials' *J. Mater. Chem. B* DOI: 10.1039/c3tb20183j in press.
149. Grande, Chundi L, Wei VT, Bower D, Andrew C, P, Ryhanen, T , 'Graphene for energy harvesting/storage devices and printed electronics' *Particuology* 10 1–8 (2012)
150. Xuan Y, Sandberg M, Berggren M, Crispin X, *Organic Electronics* 13 632–637 (2012).
151. Kawahara J, Andersson Ersman P, Nilsson D, Katoh K, Nakata Y, Sandberg M, Nilsson M, Gustafsson G & Berggren M, 'Flexible active matrix addressed displays manufactured by printing and coating techniques' *J. Polym. Sci. B Polym. Phys*. 51 265–271 (2013).
152. http://www.nano-tera.ch/news/#anchor50

Chapter 3
Biosensor Design with Molecular Engineering and Nanotechnology

Thao T. Le, Christopher J. Johnson, Jakub Trzebinski, and Anthony E.G. Cass

3.1 Introduction

The concept of a biosensor is well established and the idea of integrating a molecular recognition layer with a base sensor, such that analyte binding or reaction at the former results in a measureable change (in current or voltage) in the latter has seen many ingenious embodiments. Despite this and the huge amount of research worldwide in biosensors, as outlined in Chap. 2, many challenges still remain in building reliable, long-lived biosensors, especially in the hostile environment of the human body. The enormous potential for in vivo sensing of pathophysiological molecules over time and space has led to many attempts to achieve this and the rewards, both in terms of clinical benefit and improved understanding, cannot be underestimated. As new tools for producing biosensors become available, they are rapidly recruited. In recent years, developments in two areas of science and engineering have provided new opportunities to look again at how biosensors are built and deployed. These developments were not driven by the needs of those building and using biosensors but by much broader scientific and technological trends, which nonetheless have found ready applicability in this area.

One of the trends is an increasing knowledge of the structural factors that determine function in biological macromolecules and the other is the appreciation that the properties of materials with a characteristic length scale from 1 to 100 nm are not those expected from dividing macroscopic materials into smaller pieces nor those expected from adding atoms or molecules together. The first of these trends is sometimes referred to as *biomolecular engineering* and the second as *nanotechnology*.

There are many drivers for the use of molecular engineering and nanotechnology in the design and application of biosensors and the past decade has seen these tools

T.T. Le • C.J. Johnson • J. Trzebinski • A.E.G. Cass (✉)
Department of Chemistry and Institute of Biomedical Engineering,
Imperial College London, London SW7 2AZ, UK
e-mail: t.cass@imperial.ac.uk

G.-Z. Yang (ed.), *Body Sensor Networks*, DOI 10.1007/978-1-4471-6374-9_3,

being used extensively and demonstrated in a variety of formats particularly to try and surmount the many challenges presented by continuous operation in the human body. Ever since the first publications on chemical sensors and biosensors the great prize has been to obtain reliable measurements of key molecular markers in real time in vivo. Whilst continuous, reliable sensing is not a unique requirement of in vivo sensors (process monitoring is another example), the human body presents an especially challenging environment for sensors both through its response to the presence of a foreign body and in the complexity and variability of the background chemistry that it exhibits. Whilst it is likely that there will not be a "breakthrough" in improving biocompatibility, a steady increase in our understanding of the factors that determine the interplay between tissue and sensors will continue to see improvements in this area.

3.2 Biomolecular Engineering for Biosensors

Biomolecular engineering has been widely employed to create molecules with desirable properties for sensing applications. Those biomolecules of interest include both proteins and nucleic acids, where they perform the role of molecular recognition in sensing platforms but can also be extended to immobilisation and signal transduction functions. In this section we will discuss the rationale of employing biomolecular engineering and describe the techniques used to create engineered biomolecules for sensing applications [1–3].

Three broad approaches to biomolecular engineering have become established, the first is often referred to as '*de novo* design' and aspires to produce a sequence that folds or adopts a defined 3-D structure that has the desired functions. The second approach is 'rational design' which starts from a known 3-D structure and then makes changes that alter function in a predictable fashion. In the final approach, 'evolutionary design', a large population of slightly different molecules undergo successive rounds of selection for members that have the desired function. Each of these strategies has advantages and limitations and the task of the biomolecular engineer is to design and implement the molecular changes that lead ultimately to the desired performance. Unfortunately, the field has not yet gained the same level of predictability that its civil, mechanical or electrical counterparts have. There are many reasons for this including the inherent complexity of biomolecules, however one of the most challenging features of these large molecules is that there are many different structures with very similar energies, most of which are non-functional and easily convert from the functional state.

3.2.1 Engineering Proteins by Rational Design

Proteins have often been referred to as 'molecular machines' [4] with properties acquired through the process of evolution. This means their 'fitness for purpose' is

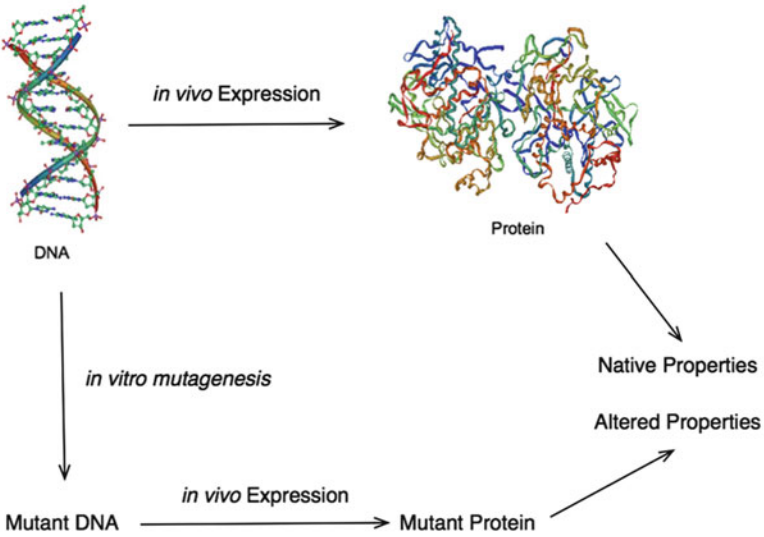


Fig. 3.1 Workflow for the production of an engineered protein via DNA modification

defined by their physiological function and in particular the ability of the organisms in which they are found to reproduce. Unfortunately, the protein's physiologically evolved function often does not match the requirements for its use as a sensing reagent, either in vitro or in vivo. Protein engineering has therefore been used to improve performance or to add new properties to existing molecules as this is often much more straightforward than *de novo* design [5, 6]. The existing protein structure acts as a scaffold onto which new or altered fragments are added. Changes that can be affected in this way include enhancing the binding affinity and specificity or chemical and physical robustness, both of which are important as recognition molecules whilst the addition of new fragments often involves the fusion with other protein domains, peptide sequences or functional groups and are typically used for immobilisation and signal transduction purposes.

Protein engineering by rational design can be achieved through well-established recombinant DNA technology. Based on the so-called "Central Dogma", protein engineering takes advantage of the precise information flow from DNA to protein where a sequence of nucleotide units in the former code for a sequence of amino acids in the latter [7] and it is that amino acid sequence that then folds into a defined 3-D structure, which in turn gives rise to the molecular function. The corollary of this is that a change in DNA sequence results in a possible change in protein function. Low cost DNA synthesis along with enzyme tools for precisely manipulating DNA means that changes in the nucleotide sequence of the DNA can be made with high precision and the very low costs of DNA sequencing mean that the changes can be readily confirmed. Once the DNA sequence of the gene has been changed, the altered protein is then produced in a host organism, most typically the bacterium *E. coli* as a recombinant protein. Figure 3.1 shows the workflow for

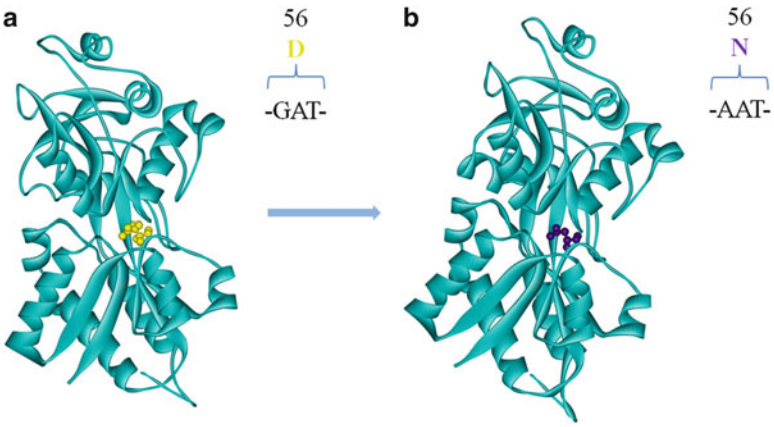


Fig. 3.2 X-ray structure of phosphate binding protein (PBP). (**a**) Wild-type (*wt*) and (**b**) Engineered phosphate binding protein where aspartic acid (**D**) at position 56 was changed to asparagine (**N**) (D56N mutation). The amino acid residue at the position was displayed as "stick and ball" model in *yellow* for the aspartic acid residue in the wild type and in *purple* for the asparagine residue in the engineered form. The images were created from Protein Data Bank by WebLab ViewerLite with (**a**) wt PBP using file 1IXH and (**b**) engineered PBP using file 1IXI

rational protein engineering and Fig. 3.2 gives an example of engineering the phosphate binding protein.

Only alterations at certain positions are effective and therefore decisions on which position(s) of the amino acid sequence should be changed often made through careful analysis of the 3-D structure to search for clues regarding relationships of amino acid residues, structures and functions, often with the help of computer modelling [8–10].

In addition to creating point mutations (i.e. single amino acid changes), protein engineering can be used to fuse protein domains together to create new molecules with a combination of properties of the constituent units (chimaeras). Examples of this approach include fusions of signal reporters or tags for affinity for immobilisation. Figure 3.3 shows maltose binding protein (MBP) fused with green fluorescent protein (GFP).

It is often necessary to introduce short linker sequences between the two proteins being fused, these should be hydrophilic and flexible. A 6–8 amino acid linker comprising of alternating glycine and serine residues is often used for this purpose. It has long been known that enzyme catalysed reactions where the product of one enzyme is the substrate for a second can achieve greater catalytic efficiency when the two (or more) assemble into a complex and so facilitate 'substrate channelling'. A similar effect is frequently seen where the two enzymes are co-immobilised on a surface and genetically, fusing two enzymes together has the same effect.

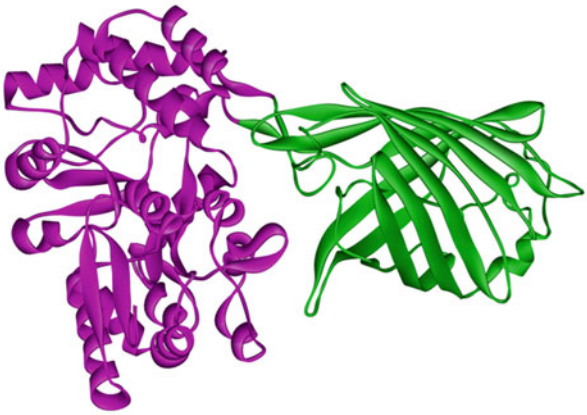


Fig. 3.3 X-ray structures of MBP-GFP fusion, with the MBP shown in *magenta* whilst the GFP is in *green*. The image was created from Protein Data Bank (file 3OSQ) using WebLab ViewerLite

3.2.2 Engineering Proteins by Evolutionary Design

Another approach that does not rely on knowledge of the 3-D structure of an existing protein scaffold or the relationship between sequence and function is often referred to as 'evolutionary design' [11]. This again starts with the gene for the wild type protein but then generates a series of random mutations throughout the gene. Although there are several ways of doing this, the so-called error prone PCR method has proven to be effective and can be controlled to give an error rate of typically 1–6 $\times$ 10^{-3}, i.e., 1–6 nucleotides are changed per thousand. Error prone PCR entails in vitro amplification of the DNA under conditions where a low frequency of error (0.1–0.6 %) occurs in the process and this generates a population of protein variants. The variants will be primarily less or non-functional compared to the wild type, however there will be a small proportion of slightly improved variants and if these can be enriched through screening or selection, then over successive '*generations*' a significantly improved variant will emerge. Evolutionary design is particularly useful where our understanding of structure function relationships may not be detailed enough to rationally design the most advantageous changes and unexpected consequences may arise from changes in protein sequence. Evolutionary designs can also result in mutations distant from the expected site of action such that these would be difficult to predict. Evolutionary design does require good screening methods and the sequence 'space' being screened may not be large enough to encompass all the effective mutants.

3.2.3 Nucleic Acid Aptamers

Aptamers are short single sequences of RNA or DNA molecules. The term 'aptamer', which derives from the Latin word 'aptus' translated as 'fit' speaks for itself. This type of oligonucleotide-based recognition molecule has seen a rapid increase

in sensing applications. In fact, they may even be favoured over the protein-based molecular receptors due to their resilience to chemical and physical conditions (their binding characteristics can easily be restored following denaturation). In addition, labels for signal transduction and linkers for immobilisation can be introduced at specific sites during synthesis.

Although there are aptamers that occur naturally, these native aptamers do not meet the demand for sensing applications, being restricted to a limited number of ligands. Engineering methods with aptamers have used similar approaches to proteins; relying on the natural binding motifs and making defined changes (rational design) whilst the other starts from random sequences and selects and enriches for binders (evolutionary design). It is the latter from which virtually all aptamers have been derived.

The invention of this evolutionary method to create new aptamers is called SELEX (Systematic Evolution of Ligands by Exponential Enrichment) and the first aptamers generated in this way were reported in 1990 and opened a new chapter for biomolecular engineering for sensing applications. Aptamers are selected from oligonucleotide libraries in vitro through a cyclic process of (i) target binding, (ii) isolation of bound sequences and (iii) amplification of bound sequences. A selection of both DNA and RNA aptamers starts with a chemically synthesised DNA library. A library typically consists of a random region of several tens of nucleotides (30–60), flanked by constant regions used for polymerase chain reaction amplification. A DNA library is most easily generated by solid phase synthesis and can be converted to an RNA library through transcription if a RNA polymerase promoter is included in one of the constant regions. A typical library contains from 10^{10} to 10^{15} different sequences. The key to the selection process is the separation of the complexes of target molecules with bound sequences from the non-binding sequences ('partition'). There are many ways to achieve this including membranes filtration, affinity chromatography, magnetic particle separation or electrophoresis. After partition, the bound sequences are amplified to produce an enriched library for the next round of the selection. RNA selections require extra steps (reverse transcription of RNA to DNA and transcription of amplified DNA back to RNA). Multiple selection rounds result in the preferential enrichment of the sequences that bind the target with the highest affinities. A round of DNA aptamer selection is illustrated in Fig. 3.4.

Hundreds of aptamers that bind different types of targets including small molecules, peptides, proteins and even whole cells, with binding affinities ranging from 10^{-6} to 10^{-12} M, have been created since the first aptamers were selected over two decades ago [12–16]. Binding characterisation of aptamers proved they bind their targets not only with high affinity but also with high specificity; being able to distinguish closely related molecules. A famous example is the aptamer that can discriminate between theophylline and caffeine (the two molecules are different in only a methyl group) [15, 17]. The reason for this could be aptamers only adopt their defined structures upon binding to their targets, allowing them to 'fit' their targets' with molecular precision. Figure 3.4 shows the X-ray diffraction structure of a DNA aptamer binding to thrombin, in which the aptamer has a G-quartet structure and experiments have been carried out to confirm that this defined structure only exists in the presence of thrombin [18, 19].

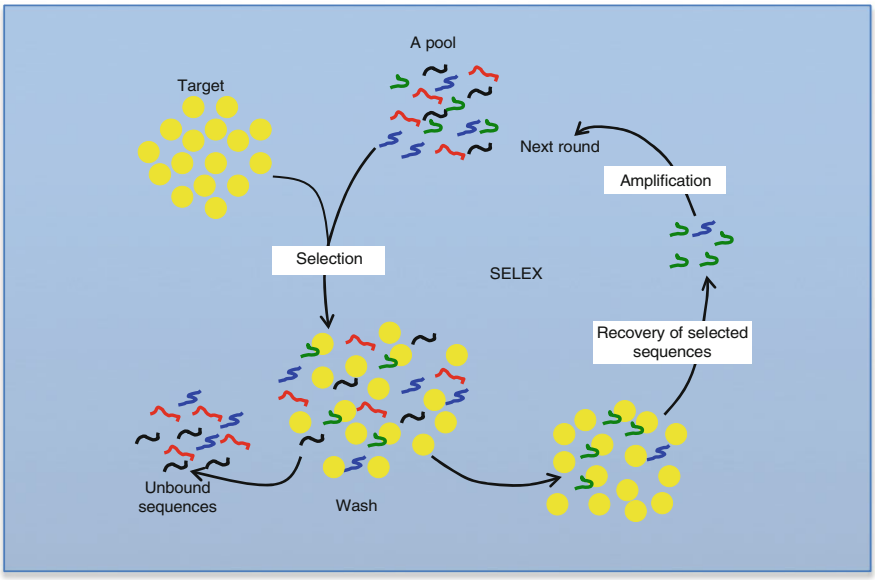


Fig. 3.4 A general scheme for aptamer selection. A random nucleic acid library (beginning with 10^{10}–10^{15} molecules) is incubated with a target (in *yellow*). Unbound molecules are washed away from the binding complexes. The bound molecules are then eluted from the target and amplified to make a library for a next round of the selection

Studies on other aptamers also confirmed this high fidelity in aptamer binding, making them ideal recognition molecules [17, 20, 21]. A huge advantage of aptamers over proteins that should be noted, is that when binding sequences are discovered, they can be readily produced by established solid phase synthesis at relatively low cost. In addition, straightforward chemical incorporation of diverse functional groups such as those for immobilisation and reporter groups for signal transduction during their synthesis provides additional benefits for this type of molecule in the biosensor field (Fig. 3.5).

3.3 Biosensor Applications

One biomolecular engineering approach to improving the molecular components of biosensors treats the molecule as a collection of modules:

- A signal transduction module that converts the physicochemical change associated with molecular recognition into an electrical signal.
- A recognition module that determines the specificity and affinity of the interaction between the biomolecule and the analyte.
- An immobilisation module that mediates the attachment of the recognition molecule to the surface of the sensor.

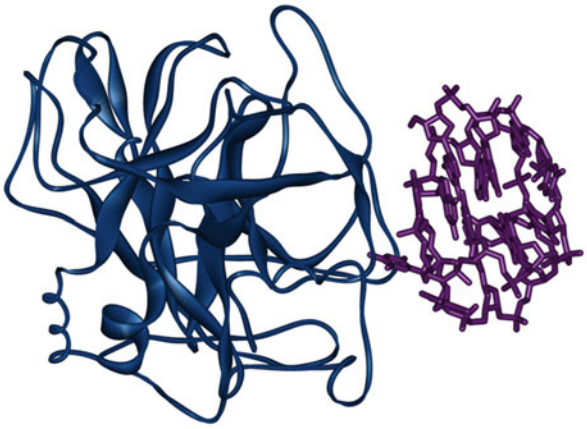


Fig. 3.5 X-ray structure of the anti-thrombin aptamer in complex with its target. The anti-thrombin aptamer shown in *purple* is bound to its target shown in *blue*. The image created from Protein Data Bank (file 4DIH) using WebLab ViewerLite

3.3.1 The Signal Transduction Module

Biosensors are built on the premise that the biomolecular recognition reaction can be transduced into an electrical signal. Whilst this has clearly been demonstrated with the wild type proteins, protein engineering can be used to enhance this transduction. Given the widespread use of electrochemical biosensors, the engineering of redox enzymes to improve electrochemical coupling has been pursued on a number of fronts. In most redox enzymes, the electron transfer to and from substrates occurs deep within the protein matrix as this then avoids redox reactions with other cellular components and biochemical 'short circuits'. Unfortunately such protection also makes heterogeneous electron transfer slow and therefore artificial electron transfer 'mediators' are used to shuttle electrons between the active site and the electrode. Mediators may be either soluble [22], polymeric [23] or protein bound [24] and there have been several reports of engineering redox proteins to improve their electrical communication with mediators or electrodes. The reaction of mediators with redox enzymes appears to proceed through a typical Michaelis (enzyme-substrate) complex consistent with a specific interaction between the two. Sadeghi et al. [25] showed that in the case of cytochrome c peroxidase (cCP), mutations that changed surface residues differentially affected the rate of reaction with ferrocene, depending on both the ferrocene and the mutation. Some combinations showed overall rates of electron transfer greater than that of wild type cCP with its natural substrate (cytochrome c). A site directed covalent attachment of a mediator can be achieved using the same cysteine mutagenesis and chemical modification approach as described in the next section, for fluorescent dyes and enzymes so modified include cytochrome P450cam and trimethylamine dehydrogenase [26]. A rather different approach was adopted by Chen et al. who added an oligo(lysine) sequence to the end of glucose oxidase and then modified this with ferrocene mediators [27]. Although glucose oxidase has native lysine residues, they are unreactive in the folded form of the enzyme [28] and this engineered form

shows an extended linear range and better stability than the wild type enzyme with soluble mediators. Where two (or more) enzymes act sequentially to generate a signal the optimised design of biosensors can be problematic, especially with regard to controlling the loading of the enzymes on the surface, as there may be differential loss of function. Moreover, we know that in vivo such sequential reactions often involve close physical association of the enzymes (a process known as substrate channelling) [29]. In designing a biosensor for maltose determination based on the sequential reactions of glucoamylase and glucose oxidase, Zhou et al. genetically fused the two enzymes together with a resulting improvement in performance [30].

Electron transfer can also be mediated through metallic nanoparticles and whilst this is discussed at greater length in the next section, mention is made here of a protein engineering approach to 'wire' glucose oxidase to an electrode via a site specifically bound gold nanoparticle, again using cysteine mutagenesis [31].

For aptamers, diverse signal reporter groups can be readily engineered in either during solid phase synthesis or post-synthetically where there are 5′ or 3′ modifications such as addition of thiol or amine groups. This easy incorporation of the reporter groups facilitates electrochemical, optical or mass detection. As previously mentioned on a number of occasions in this section, aptamers often undergo significant conformational changes, adopting defined three dimensional structures upon binding to their targets. This conformational coupling property paves the ways for optical biosensors based on changes in fluorescence intensity or polarisation where an organic fluorophore is introduced into the aptamer molecule [32]. In some cases, a pair of fluorescence donor/receptor groups can be incorporated into an aptamer molecule as conformational changes can bring certain sites on the aptamer molecule closer/further from each other modulating a FRET signal [18]. In particular, the design of aptamer beacons facilitates this optical method [33].

Electrochemical aptasensors are particularly attractive as the associated instrumentation can be simple, low cost and low power and readily miniaturised. Chemically, a variety of electroactive groups such as methylene blue or ferrocene can be incorporated into the aptamers during their synthesis or post synthetically to facilitate this detection method [34, 35]. The design of aptamer beacons is also well suited to electrochemical biosensors [36, 37] and is discussed further in the nanotechnology section of this chapter. This is a consequence of the strongly distance dependent nature of surface electron transfer.

3.3.2 The Molecular Recognition Module

In biosensors, molecular recognition lies at the heart of their function. Protein engineering and nucleic acid aptamer selection have been used to tailor such recognition biomolecules to accommodate ligands or substrates for which there are no natural binding or catalytic sites. In proteins, altering the analyte recognition site module affects changes in both specificity and affinity. Modification of affinity is probably the most straightforward to achieve, particularly where this involves

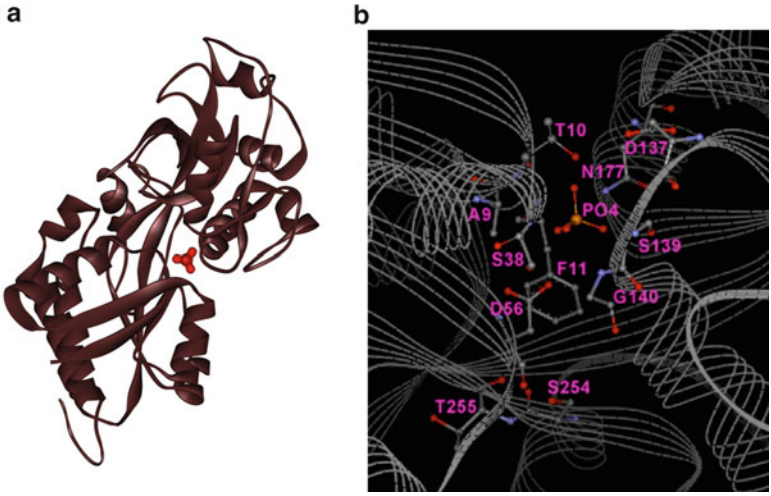


Fig. 3.6 (**a**) X-ray structure of the phosphate binding protein in a complex with its ligand. The protein shown brown whilst its ligand shown *red*. The image was created from Protein Data Bank (file 1IXH) using WebLab ViewerLite. (**b**) Key amino acid residues in the binding pocket that change the binding characteristics of the protein

improving the affinity to lower detection limits. On the other hand, changes in specificity can also be engineered either by relatively small changes in the amino acid residues that interact with the analyte or through wholesale resculpting of the binding site. Changes in specificity can be used to extend or alter the range of molecules that are sensed. As discussed earlier a rational design approach is often used to change properties of an existing protein scaffolds. The insight obtained from 3-D structures, therefore, can be particularly valuable as it reveals key residues involved in analyte recognition sites or binding pockets. We have used this approach with the phosphate binding protein, making mutations in the binding pocket and the hinge region as shown in Fig. 3.6.

In this case the aim was to lower the affinity for phosphate from its native value of 100 nM to a range more typical of physiological concentrations and as could be seen from. Figure 3.7a showing an approximately three orders of magnitude shift in the sensing range for phosphate was achieved whilst Fig. 3.7b showing we have altered the specificity of the phosphate binding protein to increase its affinity for arsenate as compared to the wild type enzyme (Oaew PhD Thesis Imperial College London).

Unlike proteins, where most changes in specificity or affinity come through modifying naturally occurring binding sites, aptamers by nature of the SELEX process having binding sites generated *de novo*. Taking advantage of counter selection to remove sequences that bind closely related ligands it is possible to achieve very high specificity. Figure 3.8 shows the structure of the RNA aptamer that binds theophylline tightly ($K_d = 300$ nM) with a 10,000-fold less tight binding to caffeine, from which it differs by only a methyl group. This aptamer has been used in biosensors to detect theophylline [38, 39].

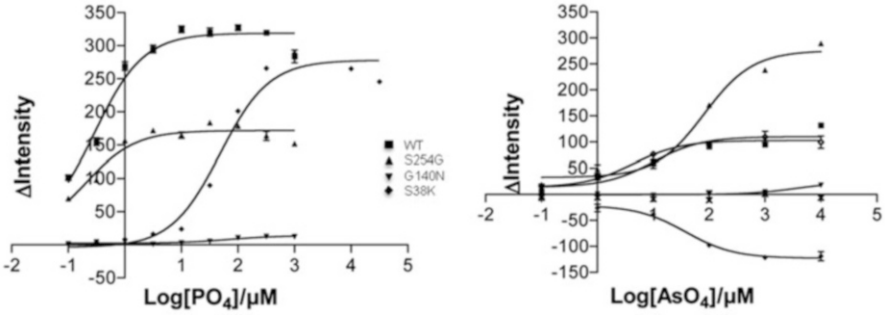


Fig. 3.7 Binding curves of different mutations of the engineered phosphate binding protein to phosphate (PO_4^{3-}) and to arsenate (AsO_4^{3-})

Fig. 3.8 NMR structure of the theophylline binding aptamer in complex with its target. The aptamer, shown in *blue*, bound its target, shown in *red*. The image was created from Protein Data Bank (file 1O15) using WebLab ViewerLite

Although aptamers to many hundreds of targets have been generated, SELEX is quite a labour intensive process and there are on-going efforts to improve throughput by using automated liquid handling systems, better selection protocols and combined techniques such as capillary electrophoresis, microfluidics and next generation sequencing [40–43].

3.3.3 Immobilisation Module

Biosensors, by their very nature, require the biorecognition molecules to be at or near the surface and for implantable sensors this immobilisation has to be virtually irreversible (i.e. a covalent attachment). Early biosensors typically used rather

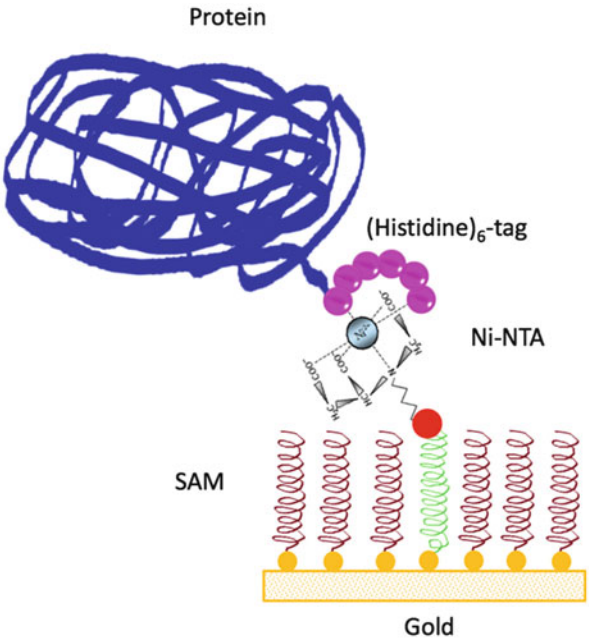


Fig. 3.9 A schematic diagram for engineering a protein with a $(histidine)_6$-tag for immobilisation on Ni-NTA-SAM coated gold surfaces

harsh immobilisation chemistries originating from chemistries used in enzyme bioreactors. These were quite effective where the enzymes had a high catalytic activity, were stable and where the sensor surface was relatively large. As sensors have become smaller and use less robust biomolecules, the maintenance of activity is a key issue in determining performance. Indeed, work is developing rapidly on nanoscale sensors [44, 45] and the issues of factors determining theoretical detection limits have been discussed [46]. Higher retention of activity means that greater control over the interaction between biomolecule and surface is necessary and one approach to this is to use protein engineering to introduce surface specific binding motifs. The related issue of analysing the protein-surface interactions is also of importance and has been reviewed [47]. The control of protein immobilisation can, in principle, be achieved by creating complementary chemistries on the protein and the surface: the former through fusion tags and the latter via chemical modification. Early work used tags already developed for affinity chromatography purification such as the hexa(histidine) sequence that binds to metal chelate surfaces such as nitrilotriacetate (NTA) [48]. An example of proteins engineered with the histidine tag for site-directed and controlled immobilisation on self-assembled monolayer (SAM) modified gold surfaces [49] is illustrated in Fig. 3.9.

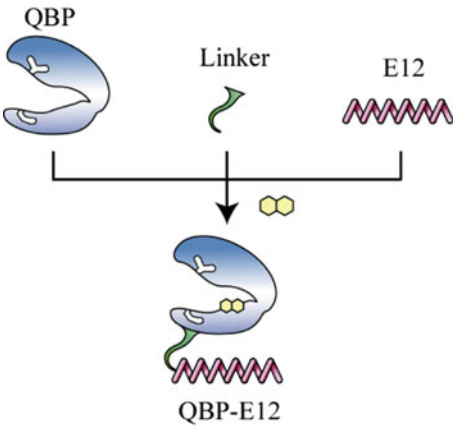


Fig. 3.10 A schematic diagram for engineering glutamine binding protein (QBP) for fluorescent signalling and self-assembly on hydrophobic surfaces (Reproduced from Wada et al. [5] with permission)

Thiol termination (either N- or C-) has also been used to immobilise proteins directly to gold electrodes and the resulting device shows improved performance over simply adsorbed or randomly covalently coupled molecules [50]. Cysteine and histidine tagged proteins also bind covalently to electrophilic conducting polymers such as poly(aniline) [51]. Other engineered tags that have been used for controlled immobilisation include the StrepTag (a streptavidin binding sequence). In the latter case, a flexible glycine-serine spacer was used to separate the protein from the tag, minimising steric clashes between the two molecules [52].

Hydrophobic surfaces can be particularly challenging for protein immobilisation as they have little specific reactivity and immobilisation by adsorption often results in surface induced aggregation [53]. Where such surfaces need to be modified, the use of the 'E12' tag has proven valuable [54] and can be combined with engineering the protein for both immobilisation and signal transduction [5] as shown in Fig. 3.10 for the glutamine binding protein (QBP). The inclusion of a flexible linker between the tag and protein again improved the performance. In addition to a higher protein loading of this hydrophilic protein on polystyrene, the tag immobilised protein showed essentially no loss of binding or signalling activity compared to the untagged protein in solution as shown in Fig. 3.11.

A particular advantage of aptamers is that functional groups such as biotin (for streptavidin-coated surface), thiol (gold) or amino (-COOH surface) as well as other moieties can be readily incorporated during their synthesis, especially with DNA aptamers. With RNA aptamers, incorporation of these functional groups is more expensive and therefore other approaches are often used. One quite generic example is to select a sequence that binds to streptavidin and then fuse this to a target binding sequence giving a molecule that has two binding sites. This is illustrated in Fig. 3.12 for a combined streptavidin/theophylline binder where the fused sequence shows little difference in affinity for the two ligands as compared to the separate sequences [55].

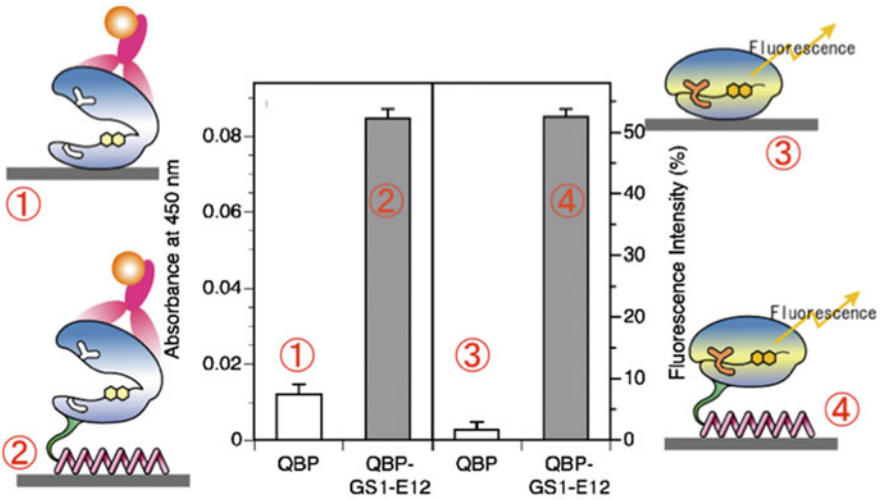


Fig. 3.11 The effect of the E12 linker on surface loading and activity of QBP on hydrophobic surfaces. The *left hand panel* shows the protein loading as assessed from an immunoassay. The engineered protein has about an order of magnitude higher loading. The *right hand panel* shows that only the E12 immobilised protein is functional as judged from the change in fluorescence (Reproduced from Wada et al. [5] with permission)

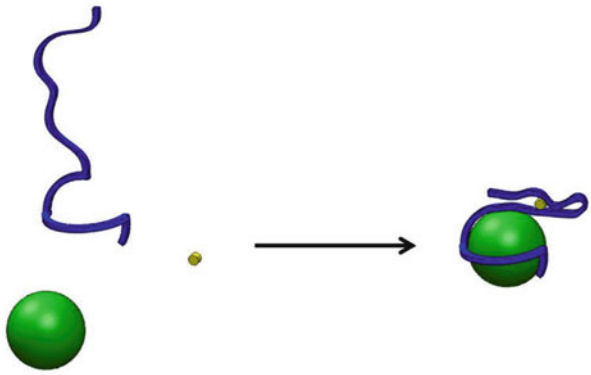


Fig. 3.12 A schematic diagram of a bifunctional aptamer (*blue*) that bind both streptavidin (*green*) and theophylline (*yellow*)

3.4 Nanotechnology

The term nanobiosensors has, in recent years, come to be widely used for devices that typically employ nanomaterials, nanostructured surfaces or nanoscale components to improve performance compared to conventional mesoscale materials [56]. The sensors themselves are usually much larger than the active material for ease of handling, as an example. In this section, we will be largely concerned with the use of nanomaterials as the active sensing surface. The primary focus will be on electrochemical sensors.

3.4.1 Miniaturisation and Scaling Laws: Nanoscale Devices and Performance Enhancements of Biosensors

Capable of being easily embedded and integrated into the measurement electronics, electrochemical sensors provide accurate, fast and inexpensive transducers in biosensor designs. Electrochemical biosensors can be potentiometric, amperometric, impedimetric or coulometric in design. Nanoscale electrodes and electrodes made from nanomaterial composites are attracting considerable attention in biosensors as they typically demonstrate enhanced kinetic and thermodynamic performance when compared to macroelectrodes [57]. The comparable feature size of electrodes to molecular recognition elements, their large surface area, biocompatibility, versatile surface chemistry and in some instances improved electrocatalytic properties have all been employed either collectively or individually to create devices with enhanced performance including improved spatial and temporal resolution; lower limits of detection (LOD) and enhanced dynamic ranges; improved stability and reduced cytotoxicity. Because of the small size of nanoscale devices, sensor arrays can also be produced that enable multiplexed measurements from low sample volumes.

For example, if using a redox enzyme as the molecular recognition element in a sensor, direct electron transfer from the enzyme to common macroscopic electrodes is very rarely observed because the active site of most redox enzymes are located deep within a hydrophobic cavity of the enzyme and as electron transfer rates are strongly distance dependent and decay exponentially. Therefore small molecular synthetic redox mediators are used to 'shuttle' electrons from the enzyme to the electrode. These work remarkably well and are used in modern glucose test strips that are sold in the billions each year. However, two reasons why mediators might want to be avoided, especially in in vivo sensors, are concerns over the toxicity of many mediators and the observation that the redox potential of the mediator is usually higher than that of the enzyme's electron donating group and therefore the operating potential of the sensor could also result in non-specific currents from oxidation of interfering species such as ascorbate, urate or acetaminophen. Operating at a lower working potential, close to that of the redox potential of the enzyme, leads to a better selectivity for the sensor. These third generation biosensors often exploit nanomaterials to enable direct electron transfer.

The application of nanoelectrode nanocomposites in biosensors has been made possible by the discovery of nanoscopic allotropes of carbon and improvements in wet chemical and lithographic methods for reliably producing metallic and semiconductor nanoparticles and nanostructures in large volumes and at acceptable costs. Examples of nanosensors and their sensing applications will be given in the remainder of this section with a focus on the scaling laws.

The overall response of a biosensor is typically determined by either chemical reaction or mass transport effects and depending on the sensor design and working conditions its response will be limited by one or the other of these processes. Therefore, the effects of miniaturisation and scaling laws of both elements must be considered to determine the overall performance of nanoscale sensors.

Decreasing the area of a sensor to the nanoscale decreases both the flux of analyte molecules reaching the sensor surface and the number of molecular recognition molecules available to bind to or react with the analyte. Using analytical calculations and finite element simulations Sheehan and Whitman [46] examined mass transport effects on affinity biosensors of different geometries at the nanoscale. For a system where a biomolecule irreversibly adsorbs onto a sensor surface once it arrives from solution their work showed that for a one femtomolar solution of a molecule with a diffusion coefficient of 150 $\mu m^2 s^{-1}$ (equivalent to a ssDNA molecule 20 bases in length) it would take a single molecule 1 h, 12 h and 2 weeks to arrive at a hemispherical sensor with a diameter of 1 μm, 100 nm and 10 nm, respectively. For a 10 nm diameter electrode they also showed it would take an astonishing month for the tenth molecule to arrive. For biomolecules with smaller diffusion constants such as a 160 kDa antibody the accumulation times are even longer. Interestingly, the accumulation rates are dependent on the sensor geometry as well as critical dimension size. For example for the same concentration and diffusion constant values as above it takes just under 1 day for the tenth molecule to accumulate at a nanowire with a 10 nm diameter and 10 μm length (modelled as a hemicylinder, as nanowire sensors often lie on a substrate surface). The authors conclude that without directed transport (e.g. convection or electrophoresis) of biomolecules, individual nanoscale sensors will be limited to picomolar-order sensitivity for practical time scales.

Assuming irreversible binding between an analyte and a receptor defines an upper theoretical limit in sensor performance but in practice the receptor binding sites may not be saturated at equilibrium and this will depend on the association (k_{ass}) and dissociation (k_{diss}) rates of the reaction between the analyte and receptor. Squires et al. have studied how binding kinetics in dilute analyte concentrations become a limiting effect for small area nanosensors with a limited number of receptors available for binding [58]. Under equilibrium conditions for a sensor that is 'reaction limited' rather than 'mass transport' limited, the fraction of bound receptors, b_{eq}, is given by

$$\frac{b_{\mathrm{eq}}}{b_m} = \frac{c/K_D}{1 + c/K_D} \tag{3.1}$$

where b_m is the concentration of receptors on the sensor surface and K_D is the equilibrium dissociation constant ($K_D = k_{diss}/k_{ass}$) and c is the analyte concentration, where the concentration at the sensor surface is equal to the bulk concentration. Concentrated solutions ($c \gg K_D$) effectively saturate the sensor, whereas dilute solutions ($c \ll K_D$) bind only a small fraction ($b_{\mathrm{eq}} \approx cb_m/K_D \ll b_m$). This has a major impact for low area nanosensors detecting low analyte concentrations as a theoretical situation were less than one target molecule is bound is readily achieved – in reality this implies a target molecule will be bound to the sensor for a certain fraction of the time. For nanosensors with area A, a critical concentration, c^*, at which only one target molecular binds the sensor in equilibrium, can be calculated

$$c^* = \frac{K_D}{b_m A} \tag{3.2}$$

Detecting analytes at concentrations that are not appreciably more concentrated than c^* will therefore inherently involve noisy, stochastic single-molecule binding and dissociation events. For a hemicylindrical nanowire 10 nm $\times$ 2 μm with a well ordered receptor layer of small, 20kD receptors (with a K_D of 1 nanomolar), giving a high active site density of 2×10^{12} cm^{-2}, the nanowire will have roughly one thousand binding sites and c^* will equal 1 picomolar. It should be noted that many sensors do not have such a high density of active binding sites, for example, a disorientated and hence only partially functional layer of receptors (as is often the case), b_m will be reduced by an order of magnitude or more.

In some regards, the above treatment of mass transport and binding is rather simplified and does not accurately predict the enhanced performance of some nanosensor systems seen in practiced examples of nanowire sensors that have faster response times than the above theory predicts will be described in the next section. Nevertheless, when designing new nanosensors it is important to be aware of the physical limitations that can be approached at ultra small sizes and low analyte concentrations.

One reason for discrepancy between theory and practice may involve charged analyte molecules and oppositely charged receptor molecules or sensor surfaces accelerating binding, particularly at low ionic strengths where screening lengths are large so the electrostatic 'reach' in to solution is correspondingly greater. For example, in DNA microarrays, where the capture probe has been functionalised at its terminus with biotin and anchored to a surface coated with streptavidin, at low ionic strengths and at pH 6, where streptavidin carries a positive charge, hybridisation of a target sequence at a 1 nanomolar concentration is at least 80-fold quicker than under neutral pH and higher ionic strength conditions [59].

Chemically grown semiconductor nanowires have often been configured as the drain in field-effect transistors (FETs). Because of the high surface to volume ratio of nanowires, any change in surface potential accompanying an analyte binding event, such as DNA hybridisation or protein binding can act as a field-effect gate upon the nanowire, thereby changing is conductance [60]. Both p-type (boron-doped) silicon nanowires and n-type (phosphorus-doped) silicon nanowires can be used in this regard. Thus, detection of PSA, which has an overall negative charge at pH 7.2 (pI = 6.8) results in a measured increase in conductance of the p-type and reduction in conductance for the n-type silicon nanowires. Modifying the NW surface with monoclonal antibodies for Prostate Specific Antigen (PSA), a LOD of 2 femtomolar and a response time of 1–2 min was demonstrated. Using an array of NWs, the authors were able to perform multiplex detection of three cancer marker proteins, free-PSA, carcinoembryonic antigen (CEA) and mucin-1, from a single sample [61].

While the above nanowire sensors require the chemical synthesis of nanowires followed by integration into a sensor, Reed and co-workers demonstrated the fabrication of silicon nanowire FETs using complementary metal oxide semiconductor

(CMOS) techniques to fabricate pre-integrated silicon nanowires 'on-chip' using a wet etching process involving tetramethylammonium hydroxide, to create nanowire structures. Using this single step fabrication technique they were able to create sensors and demonstrate immunodetection of 100 femtomolar concentrations of either mouse immunoglobulin G (IgG) or mouse immunoglobulin A (IgA) with 20–40 s response times [62].

3.4.2 Graphene

Graphene is the latest allotrope of carbon to be exploited in biosensors. Graphene is a single layer of carbon in which the atoms are arranged in hexagons (sp^2-bonded carbon). Isolated in 2004 and earning Andre Geim and Konstantin Novoselov the 2010 Nobel Prize in Physics, graphene has a range of exceptional mechanical, chemical, optical and electronic properties that make it a good material for improving the performance of electrochemical biosensors [63].

Graphene is a one-atom thick zero-band gap semiconductor material consisting of sp^2-bonded carbon with a honeycomb structure, i.e., one layer of graphite. Graphene is a low cost, low environmental impact nanomaterial that is optically transparent and has a high electrical conductivity. As a nanoscale electrode material graphene has been shown to have several advantageous properties for the creation of nanobiosensors not least its high surface area enabling high loadings of enzymes and other biorecognition molecules, its fast electron transfer kinetics, large operating potential window and size and shape which allow for direct electron transfer from redox centers of enzymes.

The specific type of graphene most commonly used in electrochemical sensors is reduced graphene oxide (rGO), which is commercially available. Graphene oxide is produced by the oxidation of graphite by sulphuric acid and potassium permanganate, followed by sonication to exfoliate the graphite oxide into single sheets of graphene oxide. The resulting graphene oxide is then reduced to graphene either chemically (e.g. by the addition of hydrazine) or electrochemically.

The graphene preparation method has a significant impact the final properties of the nanomaterial; hence these final products are referred to as reduced graphene oxide or chemically modified graphene to distinguish them from graphene created from the direct exfoliation of graphite. The reduction of graphene oxide to rGO is often partial and the preparation method leads to lattice defects, graphitic imperfections and functional groups, which is advantageous for electrochemical applications where heterogeneous electron transfer occurs at the edges of the graphene planes or at defects in the basal plane, rather than at perfect basal planes. A useful property of this route is that rGO is hydrophilic which aids handling and sensor preparation steps. This method also has the advantages of being scalable, rapid and cost effective in addition to the beneficial handling versatility of the liquid suspension.

Lin et al. [64] have demonstrated direct electron transfer between glucose oxidase and a glassy carbon electrode coated with a nanocomposite made from

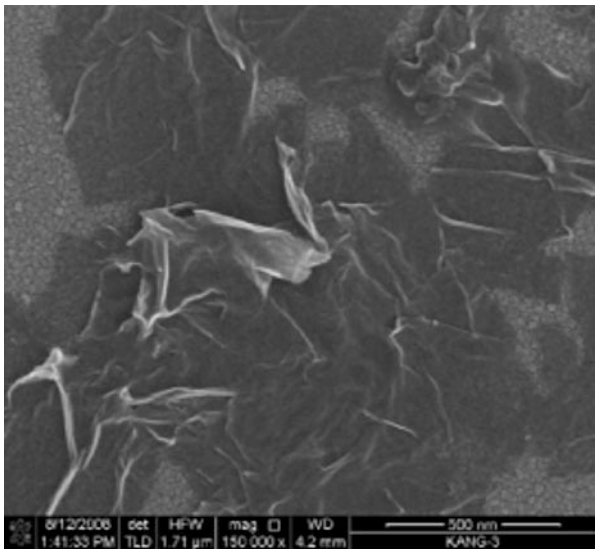


Fig. 3.13 An electron micrograph of graphene sheets in a composite with chitosan (Reproduced from Kang et al. [65] with permission of the copyright holder)

graphene (rGO) and chitosan, a biocompatible polymer derived from crustacean shells (Fig. 3.13). Cyclic voltammograms showed a pair of well-defined redox peaks, which are characteristic of reversible electron transfer of the redox active center (FAD) in glucose oxidase, indicating that a direct electron transfer was occurring.

The formal potential was estimated to be −0.43 V (vs Ag/AgCl), close to the standard electrode potential of FAD/$FADH_2$. A peak-to-peak separation of 69 mV and a linear dependence of peak current on scan rate, indicate that the redox processes of GOx on graphene is quasi-reversible. The electron transfer rate constant was estimated to be 2.8 s^{-1}. Although this is much higher than most values reported on carbon nanotubes it is still an order of magnitude slower than typical outer sphere electron transfer rate constants at solid electrodes. The rGO composite electrodes exhibit a high enzyme loading (1.1×10^{-9} mol/cm^2) presumably due to its high surface area. This is advantageous for increasing the sensitivity of graphene-based biosensors. Finally the system was stable showing response retention of above 95 % after 1 week's storage.

Chemical doping of carbon nanostructures is a means of enhancing their electrical properties by altering their density of energy states (DOES) and the Fermi level. Nitrogen doping of a graphene (rGO)/chitosan electrode provided significantly enhanced oxidation currents for the glucose oxidase detection of glucose (tenfold enhancement compared to undoped electrodes). Electron transfer is sensitive to the surface chemistry and DOES near the Fermi potential. By using nitrogen doping, the Fermi potential was changed and the electron transfer efficiency of N-doped graphene was enhanced.

The detection and measurement of dopamine and serotonin levels is of interest due to their role as neurotransmitters in the central nervous system, hormonal and

cardiovascular systems. Abnormalities in dopamine levels have been linked to disorders such as Parkinson's disease, while deficient levels of serotonin are associated with an irregular appetite, depression and increased pain sensitivity. Both molecules are electrochemically active, however, along with ascorbic acid, which is a coexisting analyte at much higher concentrations, all have overlapping voltammetric responses on macroscopic electrodes and even on carbon nanotube electrodes. This results in poor selectivity and sensitivity for the detection of dopamine and serotonin. Alwarappam et al. [64] have demonstrated that the oxidation peak from dopamine, serotonin and ascorbic acid contained in a single sample can be distinguished in differential pulse voltammetry using a graphene modified glassy carbon electrode.

3.4.3 Nanoelectrochemical Sensors

Shrinking a macroscopic electrode into the nanoscopic size regime has several consequences for its electrochemical behaviour, some of which can be used advantageously to improve the performance of biosensors. The electrochemical behaviour of an amperometric sensor detecting an electrochemically active species in solution relies on the mass transfer of the species to the solution/electrode interface followed by an electron transfer step from the analyte molecule to the electrode. For macroscopic electrodes mass transfer of the analyte from solution to the electrode surface is often the rate-limiting step. In unstirred solutions diffusion is the predominant mass transport mechanism and for electrodes with a diameter larger than the thickness of the Nernst diffusion layer (typically >25 μm) the diffusion to the electrode is described by planar diffusion. In contrast, mass transport to an individual micro- or nanoelectrode is conducted by hemispherical diffusion, which is many times quicker. Biosensors can exploit this increased mass transport behaviour of nanoelectrodes in achieving shorter response times to freely diffusing analytes. A further result of the high mass transport rate due, the electron transfer process at the nanoelectrode is dependent on and is controlled by heterogeneous electrode kinetics. This gives rise to enhanced diffusion controlled faradaic currents and sigmoidal shaped cyclic voltammograms. The increased faradaic currents and decreased charging currents, due to a smaller electrode area, also lead to increased signal to noise ratios and hence improved LODs.

An electrochemical sensor has a resistive (R) value associated with both the sample and the sensor, and also a capacitive (C) value originating from the double layer capacitance of the electrode. These two values partially determine the response time of the sensor and hence its temporal resolution. As the double layer capacitance is proportional to the electrode area, therefore nanoelectrodes exhibit small RC time constants and improved response times compared to macroscopic electrodes. The small physical size of nanoelectrodes can also offer high spatial resolution allowing measurements on a single cell level [60].

The current measured at an electrode is a function of its area. Consequently, the current measured at a single nanoelectrode is very small, often picoamps. Picoamp measurements require high precision low current, low noise potentiostats, with efficient electromagnetic shielding, which is not ideal for many real world applications. Therefore, nanoelectrode arrays are often used that multiply the current from a single nanoelectrode. Though one has to be careful, as if the spacing between the electrodes is too small typically less than the Nernst layer thickness, the array will exhibit planar diffusion under steady state conditions, as the radial diffusion field of each electrode overlaps. This results in the loss of the radial diffusion advantages discussed earlier. A beneficial feature of very small currents flowing through a nanoelectrode is a reduction in the sensor iR drop – even though the resistance of a nanoelectrode is greater than a microelectrode the product of iR is still much smaller for a nanoelectrode. This enables electroanalysis with nanoelectrodes to be performed in high resistance (i.e. low electrolyte and/or nonaqueous) solutions.

Freeman et al. [66] applied advances in semiconductor fabrication technology to produce arrays of metallic nanoband electrodes. These arrays were produced by sandwiching a 50 nm thick platinum layer between an upper silicon nitride and a lower silicon oxide insulating layer. 30 μm square apertures were created through the upper silicon nitride, platinum and into the lower silicon oxide layer, so that there was a 50 nm thick nanoband electrode in the vertical faces of each of the apertures. A 5 × 5 mm footprint contained 1,764 apertures. Dividing the electrode area over thousands of nanobands dramatically alters the electrochemical behaviour of an electrode compared to a single electrode of equivalent area. In chronoamperometric studies the nanoband currents are two orders of magnitude greater than those obtained from a disc electrode of equivalent geometric area, which is due to enhanced mass transport as discussed previously. The high mass transport also results in the nanoband electrode array being insensitive to local hydrodynamic effects (such as stirring and convection), thus significantly reducing the noise levels observed during measurements under these conditions.

Gold nanowire electrodes have also been fabricated and their potential use in biosensors demonstrated. O'Riordan and colleagues [67, 68] have fabricated gold nanowire electrodes (100 nm × 40 μm) using a hybrid electron beam lithography process which are easily contactable with overlaid interconnecting electrodes (Fig. 3.14).

Usually arrays of nanowires are fabricated that can then be addressed individually or collectively depending on the application, for example individually addressable nanowires were used to detect glucose down to 10 μM using a solution based ferrocene carboxylic acid mediated glucose oxidase reaction – these low concentrations may be useful for monitoring glucose levels in saliva [67].

A third method to produce nanoelectrodes is electroless deposition of gold within the pores of track etched polycarbonate membranes. This produces a sensor surface of nanodisk (or nanorods protruding perpendicularly from the surface)

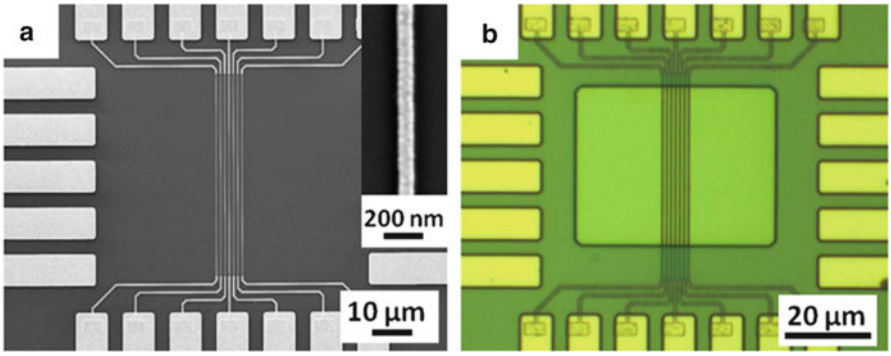


Fig. 3.14 A SEM of a gold nanowire array (**a**) and a single gold nanowire (*inset*) and an optical micrograph of the same device showing connections to connection pads (**b**) (From Dawson and O'Riordan [68] with permission)

electrodes separated by the insulating polymer. These electrodes are interconnected on the rear of the membrane so are referred to as nanoelectrode ensembles or NEEs. Because the distance between each electrode is relatively small compared to the diameter of the electrodes (due to the high pore density), the initial radial diffusion to each electrode very quickly overlaps so that the system exhibits planar diffusion under most measurement conditions. Even working in a planar diffusion regime, NEEs show enhanced electroanalytical detection limits, relative to macroscopic electrodes. This is due to a vastly improved signal to noise ratio (S:N) that can lead to LOD being improved by two-three orders of magnitude. Faradaic currents at the NEE are proportional to the total geometric area of the ensemble (nanodisk electrodes plus insulator area), while the background capacitance current is proportional only to the total area of the nanodisk electrodes. Therefore an array will have a similar faradaic current to a macroscopic electrode of equivalent area, but will have a capacitance current of only a fraction of the value of the capacitance current observed for the same equivalent macroscopic electrode. Using NEEs, Ugo et al. have shown that horse heart cytochrome c can be detected down to 0.5 µM with differential pulse voltammetry (DPV) and calculated a LOD value of 30 nM (calculated with a $3s_b/m$, where s_b is the background standard deviation and m is the sensitivity of the measurement) [69]. This compares to a LOD value of 1 µM for a macroscopic electrode of equivalent area to the NEE, which is a 33× improvement. Interestingly, these results could be obtained either with the absence or presence of the surface modifying molecule 4-4′-bipyridyl, and the authors showed that the direct electrochemistry of cyctochrome c was a result of working at low concentrations rather than an inherent effect of nanoelectrodes. This in turn suggests that the role of the 4-4′-bipyridyl may lie as much in suppressing non-specific adsorption and surface induced denaturation as in any orienting effect on the adsorbed protein.

3.4.4 Graphene Electrochemical Sensors

Graphene is equally transparent to ultraviolet, visible and infrared light. It is 100 times stronger that steel but is also more flexible. Graphene is a highly conductive zero band gap nanomaterial with a large operating potential window and fast electron transfer kinetics. It is also biocompatible, has a high surface area to volume ratio and its surface chemistry is suited to functionalisation with sensor recognition elements.

The properties of graphene, including those that are electrochemical, are strongly dependent on its method of production. Therefore a more precise naming structure is used to characterise graphene depending on its method of production. Reduced graphene oxide (rGO) is most commonly used in electrochemical nanosensors, which is produced by the chemical (e.g., by the addition of hydrazine) or electrochemical reduction of commercially available graphene oxide. The reduction is often only partial and the preparation method leads to many lattice defects, graphitic imperfections and functional groups that are detrimental to many applications of graphene that require pristine graphene sheets. However, fortuitously, these imperfections are a great advantage in electrochemical applications were heterogeneous electron transfer occurs at the edges of the graphene or at defect sites in the basal plane (almost no electron transfer occurs at the basal plane of pure graphene) and the surface functional groups allow for a more diverse range of chemical interactions.

As mentioned at the beginning of this section, nanomaterials enable direct electron transfer from enzyme redox centres to electrode surfaces. Kang et al. [65] have demonstrated direct electron transfer from glucose oxidase to a nanocomposite electrode composed of reduced graphene oxide and chitosan, a biocompatible polymer derived from crustacean shells. Cyclic voltammograms of this sensor showed a pair of well-defined redox peaks with a 69 mV peak-to-peak separation and a linear dependence of peak current on scan rate, which indicates that the direct electron transfer from the redox active centre (FAD) in glucose oxidase to the rGO electrode is reversible. The formal potential was estimated to be −0.43 V (vs Ag/AgCl), which is close to the standard electrode potential of FAD/$FADH_2$. Using the Laviron's model the electron-transfer-rate constant (k_s) was calculated to be 2.8 s^{-1}, which is higher than those reported previously for a MWCNT-chitosan electrodes (1.1 s^{-1}) and comparable to SWCNT-chitosan electrodes (3.0 s^{-1}). The graphene/chitosan electrodes have a large surface area enabling a high loading of glucose oxidase (1.12×10^{-9} mol/cm^2), which helps increase the sensitivity of these sensors leads to longer time performance of the biosensor – the sensor was stable showing a response retention of above 95 % after 1 week storage.

Chemical doping of graphene is a means of enhancing its electrical properties by altering its density of energy states and Fermi level. Nitrogen is considered an excellent element for doping carbon materials as it has a comparable atomic radius and contains five valence electrons available to covalently bond to carbon. Nitrogen doping by plasma treatment of a reduced graphene oxide/chitosan electrode, similar to the sensor described above, resulted in significantly enhanced oxidation currents

when detecting glucose by measuring direct electron transfer from glucose oxidase (ten times greater oxidation currents compared to undoped electrodes [70]). The same nitrogen doped graphene also demonstrated an electrocatalytic activity towards the reduction of hydrogen peroxide (H_2O_2). The reduction current was observed at −0.2 V, compared to a low reduction current at −0.6 V for a glassy carbon electrode. The reduction current was also much higher for N-doped graphene compared to undoped graphene [70].

A second method of achieving catalytic detection of H_2O_2 is by modifying graphene with platinum nanoparticles [71]. Here the graphene acts in a neutral role as a high surface area, highly conductive support for electrocatalytically active platinum nanoparticles. Hydrogen peroxide is oxidised at 0.4 V, compared to ≈0.5 V for a bulk platinum electrode. The detection of hydrogen peroxide was used to create two types of cholesterol biosensors. Firstly, cholesterol oxidase (ChOx) was used to measure free cholesterol and secondly a mixture of cholesterol oxidase and cholesterol esterase (ChEt), which first converts cholesteryl stearate to cholesterol. The sensitivity of the sensor towards cholesterol and cholesteryl stearate was 1.4 and 2.1 μA/μM/cm^2, respectively had a LOD of 200 nM and showed a fast, 4 s response time.

The detection and measurement of dopamine and serotonin levels is of interest due to their role in the central nervous system, hormonal and cardiovascular systems. Abnormalities in dopamine levels lead to disorders such as Parkinson's disease while deficient levels of serotonin leads to irregular appetite, depression, and pain sensations. Both molecules are electrochemically active, however along with ascorbic acid, which is a coexisting analyte at much higher concentrations, all have overlapping voltammetric responses on macroscopic electrodes and even on carbon nanotube electrodes. This results in poor selectivity and sensitivity for the detection of dopamine and serotonin. Li et al. have demonstrated that the oxidation peak from dopamine, serotonin and ascorbic acid contained in a single sample can be differentiated in differential pulse voltammetry using a graphene modified glassy carbon electrode [64].

As a final example of graphene's utility in biosensors, we turn to neuroprosthetic devices. Graphene's chemical stability, biocompatibility (crucial not only for integration with biological systems but also for the operation of field-effect devices without a protective dielectric layer) and the facile integration of graphene electronics with flexible substrates an important requirement for the design of biomedical implants with reduced tissue damage and scarring. The extremely high charge carrier mobility in graphene leads to field-effect transistor (FET) performance that is superior to most known semiconductors. As an important first step in creating such devices, Hess et al. [72] produced an array of 16 graphene-based solution-gated field-effect transistors (G-SGFET) from graphene films grown by chemical vapour deposition (CVD). They then demonstrated that cardiomyocyte-like HL-1 cells could be cultured on top of the array and showed good viability. The G-SGFET array was used to measure the action potential of the cardiomyocyte-like cells and the beat frequency of the action potentials was increased upon the addition of the fight-or-flight hormone, adrenaline.

3.5 Biocompatibility and Implantable Biosensors

The development of nanodevices such as implantable biosensors, which enable the continuous monitoring of biological processes in vivo, is one of the main goals within the emerging interdisciplinary field of nanomedicine. Real time, in vivo measurements are seen as clinically desirable diagnostic tools as they can be extremely helpful in the revealing and understanding of complex disease mechanisms, therefore aiding an improvement in the health-related quality of life [73]. Long lasting biosensor implants are an attractive alternative to conventional diagnostic methods in terms of providing continuous monitoring of various physiological parameters necessary for appropriate diagnosis or treatment [74].

In cardiology, for instance, nanodevices inserted in the body can provide early warnings of heart failure, the signs of which can be detected in slight changes in the relative proportions of different proteins. Similar applications have been recognised in neurology, where chemical monitoring of brain metabolism (using for example glucose and lactate biosensors) can help in the detection of ischaemic symptoms and to direct the course of therapeutic intervention [75]. Continuous measurements of urea would lead to better monitoring of kidney function and disorders associated with it. Determinations of creatinine play a crucial role in the detection of renal and muscular dysfunction. Monitoring of uric acid, the major product of purine breakdown in humans, offers the opportunity to detect disorders associated with altered purine metabolism (gout, hyperuricaemia or Lesch-Nyhan syndrome). Elevated levels of uric acid are observed in a wide range of conditions such as leukaemia, pneumonia, kidney injury, hypertension and ischemia, whereas abnormal concentrations of cholesterol are related to hypertension, hyperthyroidism, anaemia and coronary artery disease [76].

A DNA biosensor, that uses nucleic acid single stranded oligonucleotides (mainly RNA and DNA) as the biological recognition element, has enormous applications as a diagnostic tool for inherited diseases and the rapid detection of pathogenic infections [77]. Their main characteristic is high specificity and selectivity against the target molecule. Moreover constantly decreasing costs of producing synthetic nucleic acid sequences makes such sensors highly attractive to develop and implement [78, 79].

The other potential novel use of nanosensors, besides in vivo monitoring, is in the actual treatment of various diseases. The long-term aim is not only to monitor a wide range of chronic disorders but also to automatically and autonomously interact with pharmaceuticals using small drug delivery devices that can be implanted into the patient in advance of illness. In managing patients with diabetes the continuously measured level of glucose can control insulin delivery from an inserted reservoir. Implantable glucose biosensors have been proposed as key to developing an automatic insulin injection system, that is, an artificial, pancreatic beta cell to maintain a desirable glucose homeostasis [80, 81].

The literature is full of papers on biosensors but only a limited number of them fulfil the criteria of in vivo application. It should be emphasised that performance criteria for in vivo biosensors are not only dependent on the specific analyte, but

also on the intended application. The main property required for clinically usable sensors is accuracy. Accuracy implies a corresponding precision, linearity, sensitivity and specificity with appropriate spatial and temporal resolution. It is necessary to achieve an optimum balance among the parameters of interest for a specific application. It is important to note that the development of in vivo measurement systems is not straightforward. There are numerous problems, difficult to overcome, which prevent the widespread application of implantable biosensors in clinical practice. These include sterilisation, calibration, long-term stability and biocompatibility of the sensor [75].

Sterilisation is a major prerequisite in optimising the in vivo functionality of implantable biosensors for practical use. A useful method of biosensor sterilisation should not only provide microbial verification but also guarantee the functional stability of the sensor. Common approaches for sterilisation include:

- UV or gamma irradiation
- Treatment with antiseptic reagents like alcohol or glutaraldehyde
- Autoclaving
- Oxygen plasma treatment
- Gaseous sterilisation using ethylene oxide

These methods, however, have strong limitations due to their possible influence on bioactive sensor compounds (e.g., enzymes, active coatings or aptamers). Some biosensors cannot withstand thermal sterilisation, as high (>60 °C) temperature causes protein degeneration, for example. Likewise, gaseous sterilisation using ethylene oxide cannot be recommended due to the toxicity as well as the danger of the adsorption of residues of the active agent. Pure oxygen or hydrogen-oxygen plasma treatment can affect protein structure [82]. Liquid sterilisation by antiseptics or sterilisation by gamma irradiation is usually employed but is not always effective as it additionally causes changes of in vitro functionality and polymer structure of the biosensor [83]. Thus, the need for sterilisation should be considered right from the beginning when designing a biosensor for in vivo use.

Consequently, methods of antimicrobial treatment have to be specially adapted. As an example, one possible approach is the combined treatment with hydrogen peroxide solution acting over 4 days with 7 kGy gamma irradiation. Effective methods to produce sterile biosensors should be based not only on final product treatment but should also ensure the presence of bioburden reducing measures in every manufacturing step [83, 84].

The assessment of sensor performance is critically dependent on a reliable calibration procedure. For in vivo applications the output of implanted biosensor has to be related to the actual analyte concentration at the implantation site. Calibration of the device should ideally be done once before use ensuring excellent calibration stability following implantation. However it is often not possible to rely on in vitro calibrations as the basis for in vivo performance owing to drift in the analytical response of the biosensor and/or changes in calibration due to the immunological response of the host tissue [75, 85].

For this reason, one- or two-point in situ calibration methods have been developed. In the one-point calibration procedure the output of the implanted device is related to the blood analyte level measured by the conventional in vitro test method resulting in an in vivo sensitivity coefficient. In the two-point calibration method the plasma analyte level and sensor output reach a new plateau following analyte infusion. In such a case, an in vivo sensitivity coefficient is obtained from sensor readings during the two steady states. The calculated in vivo sensitivity coefficient is then used to determine an apparent analyte concentration from the sensor output and to estimate its variation during changes of concentrations in blood. However the two-point calibration method is time-consuming and requires linear dependence of the sensor signal on the analyte concentration as well as the induction of blood analyte alteration, so its utility for daily clinical practice is problematic [86]. Use of a one- point calibration technique has been shown to provide more accurate estimates of analyte concentration than the two-point calibration technique. Daily in situ one- point recalibrations are suggested to obtain reliable in vivo results. However, if the number of in situ calibrations is excessive there is less value gained from having the sensor implanted in the first place. Furthermore, since background signals in the absence of an analyte can change noticeably with time (e.g. due to scar tissue formation around an implanted biosensor, or simply due to decomposition of the biological sensing element) even frequently used one-point calibrations will not ensure measurements accuracy. One solution to obtaining clinically correct readings is to use a second electrode with no sensitive biological element to determine the exact background. In case of implantable sensors this however, will only succeed if the tissue reaction is comparable in both sensors [85, 87].

Another crucial point in the development of biosensors for in vivo use is their long-term stability [88]. Various factors contribute to the failure of implantable biosensor. There are two possible scenarios for implantable biosensor failure: component failures (such as lead detachment or electrical short-circuit) and failure due to biocompatibility issues (biofouling, hermeticity or encapsulation, electrode passivation, limited life-time of the immobilised enzymes) [89]. Immobilisation of enzymes in gels, membranes or on inert dispersed carriers (usually carbon materials) significantly increase stability and linear response, however at a cost of enzyme lifetime [90].

The major obstacle to the use of implantable biosensors is probably associated with the unavoidable, progressive changes in their function with time caused by the surrounding biological medium. For all in vivo measurements, the implanted device perturbs the environment initiating an inflammatory response in the host [89]. Significant efforts have been made to minimise this in biosensors for intravascular and subcutaneous application [86, 91]. Three different processes can give incorrect analytical results for sensors implanted within the vasculature using a catheter. Firstly, adsorption of proteins on the surface of the biosensor leads to the adhesion and activation of platelets (highly metabolic cells). This event results in an initiation of thrombus on the surface of the implanted device. The presence of adhered platelets on the sensor surface generates a local surface concentration of analyte

species, which is different from the bulk. Similar analytical errors can be generated by the so called 'wall effect' which is caused by placing the implanted device in a region where it can end up touching the blood vessel wall. Positioning of the sensor near metabolically active cells creates localised concentrations of analyte yielding an error pattern identical to that observed by the adhesion and activation of platelets. The third process, which can interfere with sensor readings, is a dramatic fall in blood flow at the implant site due to vasoconstriction around the catheter [85].

Immunological response in the case of subcutaneous implantation consists of three phases: acute response, chronic response and fibrotic encapsulation. During the acute response fouling of the sensor occurs, plasma proteins, particularly fibrinogen, are adsorbed and then phagocytic cells (neutrophils, monocytes, macrophages) surround the surface of the sensor and attempt to encapsulate it as with any foreign body [92]. This results in a higher consumption of oxygen and accelerates glucose metabolism (the 'respiratory burst') this leads to the generation of reactive oxygen species (ROS) such as H_2O_2, NO, OH^-. The subsequent phase of acute response is the release of hydrolytic enzymes from lysosomes present in phagocytes, which aim to further degrade the sensor. The acute inflammatory response phase lasts from 24 to 48 h, after which chronic inflammatory response begins. This reaction continues for 1–2 weeks ultimately resulting in fibrous encapsulation of the implanted material (final stage of wound healing in response to implanted, non-degradable foreign material). The extent and progression of the inflammatory process is dictated by the nature of the implant, specifically size, shape, physical (such as surface structure and morphology) and chemical (presence of charge, hydrophobic or hydrophilic nature, sterilisation) properties [92, 93].

The consequence of foreign body encapsulation is insufficient vascularisation around the biosensor. Due to the absence of dependable flow of blood delivering sample, the implantable device resides in a relatively stagnant environment, therefore diffusion conditions are unsettled and the estimation of true analyte level cannot be correlated to the one in close proximity to the sensor. This critical problem, which seriously limits not only accuracy of determination but also sensor lifetime, has been known for many years and leads to the biocompatibility issue. According to biocompatibility research on materials and sensors, biocompatibility does not mean that the sensor is inert but that it causes minimal perturbation of surrounding living tissue and likewise the in vivo environment does not adversely and significantly influence sensor [75, 88, 89].

The need to diagnose and manage the worldwide health problem of diabetes mellitus has resulted in the most widely studied and arguably most successful implantable sensors to date being amperometric glucose sensors. Generally speaking the two main analytical transduction methods have found application in the design of implantable chemical sensors: electrochemical and optical techniques, however because of their predominance in the literature, amperometric glucose sensors are discussed in this section [86].

The reduction of biofouling has predominantly been accomplished by fabrication of outer membranes that serve as a biocompatible interface, which protects the

underlying enzyme and electrode from the immune system [84]. Designing an appropriate biocompatible coating for biosensors is a particularly complicated problem. The chosen material must retard protein adsorption but simultaneously it must be permeable to the analyte and to reaction products. The material must also not cause excessive enzyme deactivation, therefore organic solvents, radicals of decomposing initiators (e.g. acetophenones), heat and UV light should be avoided. In addition, the biocompatible material should be free from cytotoxic, irritant, sensitising and carcinogenic effects. Polyurethane, Nafion, cellulose acetate, various hydrogels, surfactants, polytetrafluoroethylene, polyvinyl chloride and other materials have been used with varying degrees of effectiveness [89, 91]. Kerner et al. [94] proposed polyurethane as an outer protective membrane for glucose biosensor and showed the ability to monitor glucose for up to 7 h. Poor performance was associated with loss of sensitivity of the sensor caused by low molecular weight substances from the sample diffusing across the polyurethane biocompatible coating. In a similar study Moussy et al. [95] demonstrated a needle-type, electrochemical glucose sensor with Nafion as a protective biocompatible coating and found that protein adsorption caused loss of sensitivity and histological analysis showed limited tissue encapsulation after 14 days subcutaneous implantation in dogs. Ammon et al. [96] utilised a cellulosic material derived from bacterial source and found that the material showed low adsorption of bovine serum albumin and low complement activation. However the bioprotective layer was secured with on an O-ring, a method that cannot be used with miniaturised sensors. Rigby et al. [97] examined the use of a slow moving stream of phosphate-buffered saline solution over the tip of the sensor as the biointerface in the subcutaneous tissue of rats and found a significant reduction of device fouling by protein adsorption. Kros et al. [98] compared various sol-gel derived hybrid materials for use as biocompatible coatings (heparin, polyethylene glycol, dextran, Nafion, and polystyrene). They discovered that fibroblast cell proliferation was dramatically diminished on sol-gel coatings that contained dextran or polystyrene.

Alternatively, a great reduction of inflammatory response in vivo is achieved by surface modification to form hydrogels using hydrophilic polymers. The antifouling character in this case is believed to be due to the ability of these polymers to render the surface extremely hydrophilic, so that proteins have difficulty penetrating the gel because of a tightly bound layer of water. Schmidtke and Heller [99] used this phenomenon in their study and demonstrated that subcutaneous glucose electrodes obtained by 'wiring' glucose oxidase to crosslinked poly(4-vinylpyridine) polymer complexed with osmium (II/III)-bis (2,2′-bipyridine) were less encapsulated.

Instead of non-specific binding reduction, another approach can be used when specific functions inducing intended biological responses are introduced. The most obvious issues are the suppression of a defence reaction near the inserted device and the enhancement of neovascularisation around the biosensor.

There are two strategies for providing a better integration of the sensor within the tissue [89]. The first is the application of the controlled release of drugs that prevent inflammation and inhibit fibrosis in favour of the growth of vascularised tissue that would not severely impair delivery of blood analytes to the sensor. The local

drug release provides benefits by reducing systemic side effects and improving therapeutic response. There are some potential strategies for delivering molecules (e.g. immunosuppressant or anti-inflammatory drugs) at the near proximity to (or directly from) the sensor. The most common methods is to incorporate such drugs in biodegradable membranes [100, 101] or particles which are slowly released from the membrane [88].

There are two keys to the potential success of this approach. Firstly, the presence of the released immobilised agent within or on the outer layer of the sensor should not perturb the analytical response of the device. Secondly, the loading of the agent within or on the outer surface of the sensor must be adequate to ensure its activity for long-term in vivo application. It is commonly known that nitric oxide is a naturally occurring anti-platelet agent, therefore its release chemistry, which mimics a natural physiological process, could provide important conditions for improving the analytical performance of intravascularly inserted sensors [85, 102]. Schoenfish et al. demonstrated that the adhesion of proteins is limited when the electrochemical oxygen sensor is covered with a polymer that slowly releases physiological levels of NO [103]. Subsequently the prospect of using NO release sol-gel outer coatings for glucose sensors has been reported, with biosensors fabricated incorporating NO release chemistry [85].

Another interesting approach to minimise the foreign body response to the implantable sensor was employed by Croce et al. [91]. It is speculated that the by-product of the catalytic reaction between the enzyme and the glucose – hydrogen peroxide – is released from the implantable sensors at a dosage that enhances the foreign body response. Thus, to minimise this effect, catalase – an enzyme that decomposes H_2O_2 – was entrapped between the glucose oxidase and the host tissue. This allowed decomposition of potentially harmful compound before its escape into the surrounding tissues. It should be noted however that the sensor, through its oxidation to oxygen, also consumes hydrogen peroxide and therefore the latter's effect will depend on the relative diffusion rates towards and away from the electrode surface.

Another approach is to promote the growth of vascularised tissue by modifying the sensor surface. These modifications can either involve adding certain functional groups to alter the surface chemistry, or by controlling the topography surface through processing, in order to favour the ingrowth of vascularised tissue. It has been suggested that insufficient vascularisation surrounding the biosensor decreases the appropriate analyte concentration at the implant side. However this effect is alleviated after a few days when the foreign body capsule has matured enough to provide ingrowth of tissue bearing a rich supply of capillaries (angiogenesis process) directly to the surface of the biosensor. Improvement of neovascularisation can be achieved by incorporating an angiogenesis factor such as a vascular growth factor or adding a specially structured poly(tetrafluoroethylene) membrane to the sensor surface. The most important requirement of this approach is the maintenance of proper analyte transport through the multilayer coatings; therefore any additional membrane applied within the biosensors must be extremely thin or sufficiently porous [89, 104, 105].

The concept of initiation and modulation of angiogenesis was used by Updike et al. in their study. They developed an electrochemical glucose sensor consisting of angiogenic, bioprotective and enzyme layers. The bioprotective membrane reduces the defence mechanism caused by macrophages, whereas the outermost angiogenic layer promotes the development of new blood vessels on the sensor surface. It should be pointed out that at the same time these additional coatings of the biosensor do not affect diffusion of the analyte and hence sensitivity or response time [106].

In conclusion, biocompatibility is a key challenge for implantable biosensors and, although many attempts have been made to overcome this issue, it still remains challenging. Designing a sensor capable of withstanding the harsh environment inside the body and being able to accurately perform the measurements remains a holy grail for researchers in this field.

An alternative to partially overcome the biocompatibility issues is to design a sensor that is able to measure the analyte concentration in minimally invasive fashion in e.g. interstitial fluid (ISF) saliva or tears. Such sensor needs to be able to access the fluid in question without causing damage (or causing minimal damage) to surrounding tissue for relatively short time and ideally is removed before triggering the immune response. This can be achieved when using microprobes, which can be attached as a 'digital plaster'. Such structures are long enough to penetrate the top skin layer (the stratum corneum) and access the underlying ISF, but not long enough to reach the nerve endings or capillary bed, hence they are often described as minimally invasive. Various shapes and sizes have been described in the literature, however they can be broadly divided into two categories: microspikes [107, 108] and microneedles [109–111], the main difference being the latter has the inner channel or lumen that allows extraction of the fluid. As a result, using microneedles allows one to perform the actual diagnostic test in situ whereas in case of microspikes the measurement is non-extractive and practically taken in vivo. Extracting ISF minimises the risks associated with infection by introducing the biosensing element (which can sometimes be toxic) outside the body. On the other hand, since the concentration of analytes in ISF usually lags behind those in the blood by about 6–15 min, extracting ISF introduces additional delay to the measurement [86, 112]. In case of microspikes, their tips can be functionalised with biorecognition elements and so measure the analyte concentration at a side of insertion, consequently minimising lag. Moreover, this approach also reduces the risk of clogging the lumen with tissue/cell debris. However, it also introduces the risk of sensor substrate (e.g. biorecognition element or electrode ions) leaking into the surrounding tissue, which can be toxic and induce an immune or inflammatory response [107, 113].

3.6 Conclusions

Despite the many advances in nanomaterials and devices that we have seen over the past decade and the increasing power of molecular engineering in producing new sensing molecules the application of these technologies to implantable sensors has

not yet lead to significant advances in this area. The body is an inherently hostile environment for implanted molecular sensors where, unlike devices that do not need a molecular exchange for their operation (e.g. a pacemaker or replacement knee joint) sealing off the device in a biocompatible coating eliminates the very molecular communication needed for function. Moreover there are regulatory issues in introducing both nanomaterials and engineered biomolecules into patients that still need to be addressed, although there is no inherent reason why they cannot be, especially when they lead to much improved performance of the sensor.

In the end, however, achieving the vision of stably implanted, functional and reliable sensors will be as much about understanding and working with the body's response to foreign objects as it will be about new materials and molecules.

Acknowledgements We thank Dr Sanjiv Sharma for many informative and illuminating discussions.

References

1. Hellinga HW, Marvin JS. Protein engineering and the development of generic biosensors. Trends in Biotechnology 1998; 16(4):183–189.
2. Kohn JE, Plaxco KW. Engineering a signal transduction mechanism for protein-based biosensors. Proceedings of the National Academy of Sciences of the United States of America 2005; 102(31):10841–10845.
3. Song S, Wang L, Li J, Fan C, Zhao J. Aptamer-based biosensors. TrAC Trends in Analytical Chemistry 2008; 27(2):108–117.
4. Goodsell DS. Bionanotechnology today. In: Bionanotechnology. John Wiley and Sons, Inc., 2004;227–294.
5. Wada A, Mie M, Aizawa M, Lahoud P, Cass AEG, Kobatake E. Design and construction of glutamine binding proteins with a self-adhering capability to unmodified hydrophobic surfaces as reagentless fluorescence sensing devices. Journal of the American Chemical Society 2003; 125(52):16228–16234.
6. Wu Q, Storrier GD, Pariente F, Wang Y, Shapleigh JP, Abruña HD. A nitrite biosensor based on a maltose binding protein nitrite reductase fusion immobilized on an electropolymerized film of a pyrrole-derived bipyridinium. Analytical Chemistry 1997; 69(23):4856–4863.
7. Kuryian J, Konforti B, Wemmer D. The molecules of life. New York: Garland Science, 2012.
8. Dwyer MA, Hellinga HW. Periplasmic binding proteins: a versatile superfamily for protein engineering. Current Opinion in Structural Biology 2004; 14(4):495–504.
9. Hackel BJ, Sathirachinda A, Gambhir SS. Designed hydrophilic and charge mutations of the fibronectin domain: towards tailored protein biodistribution. Protein Engineering Design and Selection 2012; 25(10):639–648.
10. Saven JG. Computational protein design: engineering molecular diversity, nonnatural enzymes, nonbiological cofactor complexes, and membrane proteins. Current Opinion in Chemical Biology 2011; 15(3):452–457.
11. Brustad EM, Arnold FH. Optimizing non-natural protein function with directed evolution. Current Opinion in Chemical Biology 2011; 15(2):201–210.
12. McKeague M, DeRosa MC. Challenges and opportunities for small molecule aptamer development. Journal of Nucleic Acids 2012; 2012:20.
13. Strehlitz B, Nikolaus N, Stoltenburg R. Protein detection with aptamer biosensors. Sensors 2008; 8(7):4296–4307.

14. Xu W, Ellington AD. Anti-peptide aptamers recognize amino acid sequence and bind a protein epitope. Proceedings of the National Academy of Sciences 1996; 93(15):7475–7480.
15. Zimmermann GR, Jenison RD, Wick CL, Simorre J-P, Pardi A. Interlocking structural motifs mediate molecular discrimination by a theophylline-binding RNA. Nature Structural and Molecular Biology 1997; 4(8):644–649.
16. Fang X, Sen A, Vicens M, Tan W. Synthetic DNA aptamers to detect protein molecular variants in a high-throughput fluorescence quenching assay. ChemBioChem 2003; 4(9):829–834.
17. Zimmermann GR, Wick CL, Shields TP, Jenison RD, Pardi A. Molecular interactions and metal binding in the theophylline-binding core of an RNA aptamer. RNA 2000; 6(5):659–667.
18. Li JJ, Fang X, Tan W. Molecular aptamer beacons for real-time protein recognition. Biochemical and Biophysical Research Communications 2002; 292(1):31–40.
19. Macaya RF, Schultze P, Smith FW, Roe JA, Feigon J. Thrombin-binding DNA aptamer forms a unimolecular quadruplex structure in solution. Proceedings of the National Academy of Sciences 1993; 90(8):3745–3749.
20. Cho EJ, Lee J-W, Ellington AD. Applications of aptamers as sensors. Annual Review of Analytical Chemistry 2009; 2(1):241–264.
21. Sussman D, Nix JC, Wilson C. The structural basis for molecular recognition by the vitamin B 12 RNA aptamer. Nature Structural and Molecular Biology 2000; 7(1):53–57.
22. Forrow NJ, Sanghera GS, Walters SJ. The influence of structure in the reaction of electrochemically generated ferrocenium derivatives with reduced glucose oxidase. Journal of the Chemical Society, Dalton Transactions 2002; (16):3187–3194.
23. Kase Y, Muguruma H. Amperometric glucose biosensor based on mediated electron transfer between immobilized glucose oxidase and plasmapolymerized thin film of dimethylaminomethylferrocene on sputtered gold electrode. Analytical Sciences 2004; 20(8):1143–1146.
24. Battaglini F, Bartlett PN, Wang JH. Covalent attachment of osmium complexes to glucose oxidase and the application of the resulting modified enzyme in an enzyme switch responsive to glucose. Analytical Chemistry 2000; 72(3):502–509.
25. Sadeghi SJ, Gilardi G, Cass AEG. Mediated electrochemistry of peroxidases—effects of variations in protein and mediator structures. Biosensors and Bioelectronics 1997; 12(12):1191–1198.
26. Loechel C, Basran A, Basran J, Scrutton NS, Hall EAH. Using trimethylamine dehydrogenase in an enzyme linked amperometric electrode Part 1. Wild-type enzyme redox mediation. Analyst 2003; 128(2):166–172.
27. Chen L-Q, Zhang X-E, Xie W-H, Zhou Y-F, Zhang Z-P, Cass AE. Genetic modification of glucose oxidase for improving performance of an amperometric glucose biosensor. Biosensors and Bioelectronics 2002; 17(10):851–857.
28. Degani Y, Heller A. Direct electrical communication between chemically modified enzymes and metal electrodes. I. Electron transfer from glucose oxidase to metal electrodes via electron relays, bound covalently to the enzyme. The Journal of Physical Chemistry 1987; 91(6):1285–1289.
29. Easterby JS. The analysis of metabolite channelling in multienzyme complexes and multifunctional proteins. Biochemical Journal 1989; 264(2):605–607.
30. Zhou Y-F, Zhang X-E, Liu H, Zhang Z-P, Zhang C-G, Cass AEG. Construction of a fusion enzyme system by gene splicing as a new molecular recognition element for a sequence biosensor. Bioconjugate Chemistry 2001; 12(6):924–931.
31. Holland JT, Lau C, Brozik S, Atanassov P, Banta S. Engineering of glucose oxidase for direct electron transfer via site-specific gold nanoparticle conjugation. Journal of the American Chemical Society 2011; 133(48):19262–19265.
32. Wang Y, Bao L, Liu Z, Pang D-W. Aptamer biosensor based on fluorescence resonance energy transfer from upconverting phosphors to carbon nanoparticles for thrombin detection in human plasma. Analytical Chemistry 2011; 83(21):8130–8137.

33. Hamaguchi N, Ellington A, Stanton M. Aptamer beacons for the direct detection of proteins. Analytical Biochemistry 2001; 294(2):126–131.
34. Bang GS, Cho S, Kim B-G. A novel electrochemical detection method for aptamer biosensors. Biosensors and Bioelectronics 2005; 21(6):863–870.
35. Lai RY, Plaxco KW, Heeger AJ. Aptamer-based electrochemical detection of picomolar platelet-derived growth factor directly in blood serum. Analytical Chemistry 2006; 79(1):229–233.
36. Lee JA, Hwang S, Kwak J, Park SI, Lee SS, Lee K-C. An electrochemical impedance biosensor with aptamer-modified pyrolyzed carbon electrode for label-free protein detection. Sensors and Actuators B: Chemical 2008; 129(1):372–379.
37. Radi A-E, Acero Sánchez JL, Baldrich E, O'Sullivan CK. Reagentless, reusable, ultrasensitive electrochemical molecular beacon aptasensor. Journal of the American Chemical Society 2005; 128(1):117–124.
38. Dong Z-M, Zhao G-C. A theophylline quartz crystal microbalance biosensor based on recognition of RNA aptamer and amplification of signal. Analyst 2013; 138(8):2456–2462.
39. Ferapontova EE, Gothelf KV. Effect of serum on an RNA aptamer-based electrochemical sensor for theophylline. Langmuir 2009; 25(8):4279–4283.
40. Cox JC, Ellington AD. Automated selection of anti-protein aptamers. Bioorganic and Medicinal Chemistry 2001; 9(10):2525–2531.
41. Mosing RK, Bowser MThwnnngp. Isolating aptamers using capillary electrophoresis–SELEX (CE–SELEX). In: Nucleic Acid and Peptide Aptamers : Methods and Protocols. Humana Press, 2009;33–43.
42. Murphy MB, Fuller ST, Richardson PM, Doyle SA. An improved method for the in vitro evolution of aptamers and applications in protein detection and purification. Nucleic Acids Research 2003; 31(18):e110.
43. Schütze T, Wilhelm B, Greiner N, Braun H, Peter F, Mörl M, et al. Probing the SELEX process with next-generation sequencing. PLoS ONE 2011; 6(12):e29604.
44. Jain KK. Nanotechnology in clinical laboratory diagnostics. Clinica Chimica Acta 2005; 358(1–2):37–54.
45. Vaseashta A, Dimova-Malinovska D. Nanostructured and nanoscale devices, sensors and detectors. Science and Technology of Advanced Materials 2005; 6(3–4):312–318.
46. Sheehan PE, Whitman LJ. Detection limits for nanoscale biosensors. Nano letters 2005; 5(4):803–807.
47. Halliwell CM. Nanoanalytical measurement of protein orientation on conductive sensor surfaces. Analyst 2004; 129(12):1166–1170.
48. Zhang JK, Cass AEG. A study of His-tagged alkaline phosphatase immobilisation on a nanoporous nickel– titanium dioxide film. Analytical Biochemistry 2001; 292(2):307–310.
49. Le TT, Wilde CP, Grossman N, Cass AEG. A simple method for controlled immobilization of proteins on modified SAMs. Physical Chemistry Chemical Physics 2011; 13(12):5271–5278.
50. Shi J-X, Zhang X-E, Xie W-H, Zhou Y-F, Zhang Z-P, Deng J-Y, et al. Improvement of homogeneity of analytical biodevices by gene manipulation. Analytical Chemistry 2003; 76(3):632–638.
51. Halliwell CM, Simon E, Toh C-S, Bartlett PN, Cass AEG. A method for the determination of enzyme mass loading on an electrode surface through radioisotope labelling. Biosensors and Bioelectronics 2002; 17(11–12):965–972.
52. Shao W-H, Zhang X-E, Liu H, Zhang Z-P, Cass AEG. Anchor-chain molecular system for orientation control in enzyme immobilization. Bioconjugate Chemistry 2000; 11(6):822–826.
53. Nygren H, Stenberg M. Surface-induced aggregation of ferritin concentration dependence of adsorption onto a hydrophobic surface. Biophysical Chemistry 1990; 38(1–2):77–85.
54. Sugihara T, Seong GH, Kobatake E, Aizawa M. Genetically synthesized antibody-binding protein self-assembled on hydrophobic matrix. Bioconjugate Chemistry 2000; 11(6):789–794.

55. Le TT, Scott S, Cass AEG. Streptavidin binding bifunctional aptamers and their interaction with low molecular weight ligands. Analytica Chimica Acta 2013; 761:143–148.
56. Khanna VK. Nanosensors. Boca Raton: CRC Press, 2012.
57. Cox JT, Zhang B. Nanoelectrodes: recent advances and new directions. Annual Review of Analytical Chemistry 2012; 5:253–272.
58. Squires TM, Messinger RJ, Manalis SR. Making it stick: convection, reaction and diffusion in surface-based biosensors. Nature Biotechnology 2008; 26(4):417–426.
59. Belosludtsev Y, Belosludtsev I, Iverson B, Lemeshko S, Wiese R, Hogan M, et al. Nearly instantaneous, cation-independent, high selectivity nucleic acid hybridization to DNA microarrays. Biochemical and Biophysical Research Communications 2001; 282(5): 1263–1267.
60. Tian B, Lieber CM. Synthetic nanoelectronic probes for biological cells and tissues. Annual Review of Analytical Chemistry 2013; 6:31–51.
61. Zheng G, Patolsky F, Cui Y, Wang WU, Lieber CM. Multiplexed electrical detection of cancer markers with nanowire sensor arrays. Nature Biotechnology 2005; 23(10):1294–1301.
62. Stern E, Klemic JF, Routenberg DA, Wyrembak PN, Turner-Evans DB, Hamilton AD, et al. Label-free immunodetection with CMOS-compatible semiconducting nanowires. Nature 2007; 445(7127):519–522.
63. Liu Y, Dong X, Chen P. Biological and chemical sensors based on graphene materials. Chemical Society Reviews 2012; 41(6):2283–2307.
64. Alwarappan S, Erdem A, Liu C, Li C-Z. Probing the electrochemical properties of graphene nanosheets for biosensing applications. The Journal of Physical Chemistry C 2009; 113(20): 8853–8857.
65. Kang X, Wang J, Wu H, Aksay IA, Liu J, Lin Y. Glucose oxidase-graphene-chitosan modified electrode for direct electrochemistry and glucose sensing. Biosensors and Bioelectronics 2009; 25(4):901–905.
66. Freeman NJ, Sultana R, Reza N, Woodvine H, Terry JG, Walton AJ, et al. Comparison of the performance of an array of nanoband electrodes with a macro electrode with similar overall area. Physical Chemistry Chemical Physics 2013; 15(21):8112–8118.
67. Dawson K, Baudequin M, O'Riordan A. Single on-chip gold nanowires for electrochemical biosensing of glucose. The Analyst 2011; 136(21):4507–4513.
68. Dawson K, O'Riordan A. Towards nanowire (bio) sensors. Journal of Physics: Conference Series 2011; 307(1):012004.
69. Ugo P, Pepe N, Maria Moretto L, Battagliarin M. Direct voltammetry of cytochrome c at trace concentrations with nanoelectrode ensembles. Journal of Electroanalytical Chemistry 2003; 560(1):51–58.
70. Wang Y, Shao Y, Matson DW, Li J, Lin Y. Nitrogen-doped graphene and its application in electrochemical biosensing. ACS Nano 2010; 4(4):1790–1798.
71. Dey RS, Raj CR. Development of an amperometric cholesterol biosensor based on graphene−Pt nanoparticle hybrid material. The Journal of Physical Chemistry C 2010; 114(49):21427–21433.
72. Hess LH, Jansen M, Maybeck V, Hauf MV, Seifert M, Stutzmann M, et al. Graphene transistor arrays for recording action potentials from electrogenic cells. Advanced Materials 2011; 23(43):5045–5049.
73. Haberzettl CA. Nanomedicine: destination or journey? Nanotechnology 2002; 13(4):R9–R13.
74. Wang J. Analytical electrochemistry, 3rd ed. Hoboken, N.J.: Wiley-VCH, 2006.
75. Wilson GS, Gifford R. Biosensors for real-time in vivo measurements. Biosensors and Bioelectronics 2005; 20(12):2388–2403.
76. Colomer-Farrarons J, Miribel-Catala PL. A CMOS self-powered front-end architecture for subcutaneous event-detector devices: three-electrodes amperometric biosensor approach. Netherlands: Springer 2011.
77. Malhotra BD, Chaubey A. Biosensors for clinical diagnostics industry. Sensors and Actuators B-Chemical 2003; 91(1–3):117–127.

78. Labuda J, Brett AMO, Evtugyn G, Fojta M, Mascini M, Ozsoz M, et al. Electrochemical nucleic acid-based biosensors: concepts, terms, and methodology (IUPAC technical report). Pure and Applied Chemistry 2010; 82(5):1161–1187.
79. Shao YY, Wang J, Wu H, Liu J, Aksay IA, Lin YH. Graphene based electrochemical sensors and biosensors: a review. Electroanalysis 2010; 22(10):1027–1036.
80. Herrero P, Georgiou P, Oliver N, Johnston DG, Toumazou C. A bio-inspired glucose controller based on pancreatic beta-cell physiology. Journal of Diabetes Science and Technology 2012; 6(3):606–616.
81. Renard E. Implantable closed-loop glucose-sensing and insulin delivery: the future for insulin pump therapy. Current Opinion in Pharmacology 2002; 2(6):708–716.
82. Boscariol MR, Moreira AJ, Mansano RD, Kikuchi IS, Pinto TJ. Sterilization by pure oxygen plasma and by oxygen-hydrogen peroxide plasma: an efficacy study. International Journal of Pharmaceutics 2008; 353(1–2):170–175.
83. von Woedtke T, Julich WD, Hartmann V, Stieber M, Abel PU. Sterilization of enzyme glucose sensors: problems and concepts. Biosensors and Bioelectronics 2002; 17(5):373–382.
84. Abel PU, von Woedtke T. Biosensors for in vivo glucose measurement: can we cross the experimental stage. Biosensors and Bioelectronics 2002; 17(11–12):1059–1070.
85. Frost MC, Meyerhoff ME. Implantable chemical sensors for real-time clinical monitoring: progress and challenges. Current Opinion in Chemical Biology 2002; 6(5):633–641.
86. Oliver NS, Toumazou C, Cass AE, Johnston DG. Glucose sensors: a review of current and emerging technology. Diabetic Medicine 2009; 26(3):197–210.
87. Trzebinski J, Moniz AR-B, Sharma S, Burugapalli K, Moussy F, Cass AEG. Hydrogel membrane improves batch-to-batch reproducibility of an enzymatic glucose biosensor. Electroanalysis 2011; 23(12):2789–2795.
88. Yu BZ, Ju YM, West L, Moussy Y, Moussy F. An investigation of long-term performance of minimally invasive glucose biosensors. Diabetes Technology and Therapeutics 2007; 9(3):265–275.
89. Wisniewski N, Moussy F, Reichert WM. Characterization of implantable biosensor membrane biofouling. Fresenius Journal of Analytical Chemistry 2000; 366(6–7):611–621.
90. Long N, Yu B, Moussy Y, Moussy F. Strategies for testing long-term transcutaneous amperometric glucose sensors. Diabetes Technology and Therapeutics 2005; 7(6):927–936.
91. Croce RA, Jr., Vaddiraju S, Kondo J, Wang Y, Zuo L, Zhu K, et al. A miniaturized transcutaneous system for continuous glucose monitoring. Biomedical Microdevices 2013; 15(1):151–160.
92. Onuki Y, Bhardwaj U, Papadimitrakopoulos F, Burgess DJ. A review of the biocompatibility of implantable devices: current challenges to overcome foreign body response. Journal of Diabetes Science and Technology 2008; 2(6):1003–1015.
93. Atanasov P, Yang S, Salehi C, Ghindilis AL, Wilkins E, Schade D. Implantation of a refillable glucose monitoring-telemetry device. Biosensors and Bioelectronics 1997; 12(7): 669–680.
94. Kerner W, Kiwit M, Linke B, Keck FS, Zier H, Pfeiffer EF. The function of a hydrogen peroxide-detecting electroenzymatic glucose electrode is markedly impaired in human sub-cutaneous tissue and plasma. Biosensors and Bioelectronics 1993; 8(9–10):473–482.
95. Moussy F, Harrison DJ, O'Brien DW, Rajotte RV. Performance of subcutaneously implanted needle-type glucose sensors employing a novel trilayer coating. Analytical Chemistry 1993; 65(15):2072–2077.
96. Ammon HPT, Ege W, Oppermann M, Gopel W, Eisele S. Improvement in the long-term stability of an amperometric glucose sensor system by introducing a cellulose membrane of bacterial origin. Analytical Chemistry 1995; 67(2):466–471.
97. Rigby GP, Ahmed S, Horseman G, Vadgama P. In vivo glucose monitoring with open microflow – influences of fluid composition and preliminary evaluation in man. Analytica Chimica Acta 1999; 385(1–3):23–32.

98. Kros A, Gerritsen M, Sprakel VSI, Sommerdijk NAJM, Jansen JA, Nolte RJM. Silica-based hybrid materials as biocompatible coatings for glucose sensors. Sensors and Actuators B-Chemical 2001; 81(1):68–75.
99. Schmidtke DW, Heller A. Accuracy of the one-point in vivo calibration of "wired" glucose oxidase electrodes implanted in jugular veins of rats in periods of rapid rise and decline of the glucose concentration. Analytical Chemistry 1998; 70(10):2149–2155.
100. Koschwanez HE, Yap FY, Klitzman B, Reichert WM. In vitro and in vivo characterization of porous poly-L-lactic acid coatings for subcutaneously implanted glucose sensors. Journal of Biomedical Materials Research Part A 2008; 87(3):792–807.
101. Yu B, Wang C, Ju YM, West L, Harmon J, Moussy Y, et al. Use of hydrogel coating to improve the performance of implanted glucose sensors. Biosensors and Bioelectronics 2008; 23(8):1278–1284.
102. Frost MC, Batchelor MM, Lee YM, Zhang HP, Kang YJ, Oh BK, et al. Preparation and characterization of implantable sensors with nitric oxide release coatings. Microchemical Journal 2003; 74(3):277–288.
103. Schoenfisch MH, Mowery KA, Rader MV, Baliga N, Wahr JA, Meyerhoff ME. Improving the thromboresistivity of chemical sensors via nitric oxide release: fabrication and in vivo evaluation of NO-releasing oxygen-sensing catheters. Analytical Chemistry 2000; 72(6):1119–1126.
104. Klueh U, Dorsky DI, Kreutzer DL. Enhancement of implantable glucose sensor function in vivo using gene transfer-induced neovascularization. Biomaterials 2005; 26(10):1155–1163.
105. Ward WK, Quinn MJ, Wood MD, Tiekotter KL, Pidikiti S, Gallagher JA. Vascularizing the tissue surrounding a model biosensor: how localized is the effect of a subcutaneous infusion of Vascular Endothelial Growth Factor (VEGF)? Biosensors and Bioelectronics 2003; 19(3): 155–163.
106. Updike SJ, Shults MC, Gilligan BJ, Rhodes RK. A subcutaneous glucose sensor with improved longevity, dynamic range, and stability of calibration. Diabetes Care 2000; 23(2):208–214.
107. Trzebinski J, Sharma S, Radomska-Botelho Moniz A, Michelakis K, Zhang Y, Cass AEG. Microfluidic device to investigate factors affecting performance in biosensors designed for transdermal applications. Lab on a Chip 2012; 12(2):348–352.
108. El-Laboudi A, Oliver NS, Cass A, Johnston D. Use of microneedle array devices for continuous glucose monitoring: a review. Diabetes Technology and Therapeutics 2013; 15(1):101–115.
109. Mukerjee E, Collins SD, Isseroff RR, Smith RL. Microneedle array for transdermal biological fluid extraction and in situ analysis. Sensors and Actuators A-Physical 2004; 114(2–3):267–275.
110. Tsuchiya K, Nakanishi N, Uetsuji Y, Nakamachi E. Development of blood extraction system for health monitoring system. Biomedical Microdevices 2005; 7(4):347–353.
111. Windmiller JR, Zhou N, Chuang MC, Valdes-Ramirez G, Santhosh P, Miller PR, et al. Microneedle array-based carbon paste amperometric sensors and biosensors. Analyst 2011; 136(9):1846–1851.
112. Boyne MS, Silver DM, Kaplan J, Saudek CD. Timing of changes in interstitial and venous blood glucose measured with a continuous subcutaneous glucose sensor. Diabetes 2003; 52(11):2790–2794.
113. Moniz AR, Michelakis K, Trzebinski J, Sharma S, Johnston DG, Oliver N, et al. Minimally invasive enzyme microprobes: an alternative approach for continuous glucose monitoring. Journal of Diabetes Science and Technology 2012; 6(2):479–480.

Chapter 4
Wireless Communication

Henry Higgins

4.1 Introduction

Increasingly, sophisticated implanted medical devices are integrating wireless technology to support an ever-expanding range of therapeutic and diagnostic applications. For example, an implanted heart pacemaker or cardiac defibrillator enabled with a wireless link allows a physician to monitor more easily a patient's response to therapy and adjust device performance as required. With a network of in-body and on-body sensors, muscles can be stimulated to help restore lost limb function. Similarly, a radio-controlled valve in the urinary tract, operated on-demand by the patient, will restore bladder control.

Drug manufacturers are also interested in patient monitoring during treatment to regulate dosages and detect side effects. Introducing a new drug is a costly endeavour with considerable risks, as drugs will be pulled from the market even if a small number of patients have an adverse reaction. By monitoring the internal chemistry, patients susceptible to side effects could be identified earlier in the treatment process and alternative therapies could be considered. This would then benefit the patient, and reduce the risk of a drug being withdrawn that may yet assist others.

Whether it is a pacemaker communicating patient health and performance data to a base station, or a BSN integrating a number of devices, these new applications require a reliable, wireless communication link between implanted devices in the patient's skin to a clinician. The wireless link can be used to interrogate the implant at either irregular intervals, on a regularly scheduled basis or provide near constant communication. A one-way wireless link may be used to obtain patient health or device performance data from the implant, while a two-way link allows external reprogramming of an implanted device.

H. Higgins (✉)
Caldicot, Monmouthshire NP26 4JD, UK
e-mail: henry_m_higgins@hotmail.com

G.-Z. Yang (ed.), *Body Sensor Networks*, DOI 10.1007/978-1-4471-6374-9_4,

This chapter will discuss two types of communication links: the inductive loop and *Radio Frequency* (RF) communication. Widely in use today, the inductive loop is useful for transferring small packets of data without requiring an implanted power source (battery). While an RF system requires an implanted battery source, it is capable of transferring larger packets of data within a shorter time period and over greater distances. RF-based communication will be the main topic of this chapter.

While wireless communication through air has been extensively documented, communication from implanted devices through the human body is a new area of study. This chapter will discuss body properties and their effect on radio propagation. The human body is an uninviting and often hostile environment for a wireless signal. One of the most important considerations for implanted devices is physical size, meaning in-body communication system designs are restricted to an extremely small antenna that needs to be characterised to enable it to be effectively coupled to the transceiver. A significant portion of this chapter is devoted to antenna measurement and coupling circuit design, as it is critical to the success of an implanted RF system.

4.2 Inductive Coupling

Before discussing in-body RF communication, it is important to understand inductive coupling. Several applications still use electromagnetic coupling to provide a communication link to implanted devices by using an external coil held very close to the patient that couples to a coil implanted just below the skin surface. The implant is powered by the coupled magnetic field and requires no battery for communication. As well as providing power, this alternating field can also be used to transfer data into the implant. Data is transferred from the implanted device by altering the impedance of the implanted loop that is detected by the external coil and electronics. This type of communication is commonly used to identify animals that have been injected with an electronic tag.

Electromagnetic induction is used when continuous, long-term communication is required, such as for a cochlear implant used to restore hearing. A cochlear implant includes a coil placed below the skin behind the ear and an external energising coil. Using a magnet in one coil and an iron puck in the other, the coils are kept in alignment without requiring any external support. Power for the implant and encoded sound is transmitted across the small gap. The power source is an external battery that can be changed with ease and the implant can remain in place for many years.

Another application using electromagnetic coupling is in the treatment of an *Abdominal Aortic Aneurysm* (AAA). In this situation, a shaped tube is inserted into the patient through a “keyhole” in the groin and placing it over the affected area. To evaluate the patient’s health, a pressure sensor is included that can be interrogated at any time for many years. As a result, major abdominal surgery is replaced

with a relatively unobtrusive procedure and the patient can be easily monitored. Electromagnetic coupling is also used to "fine-tune" pacemakers and measure intracranial pressure.

The base band for electromagnetic communication is typically 13.56 MHz or 28 MHz, with other frequencies also available. Its use is subject to regulation for maximum *Specific Absorption Rate* (SAR); two important standards are ANSI C95.1 and ICNIRP.

Inductive coupling achieves the best power transfer when using large transmit and receive coils, meaning it is impractical when space is an issue or devices are to be implanted deep within the patient. This technique does not support a very high data rate and cannot initiate a communication session from inside of the body [1].

4.3 RF Communication in the Body

Compared with inductive coupling, RF communication dramatically increases bandwidth and enables a two-way data link that allows an implant to initiate a communication session. This requires an implanted battery, electronics and suitable antenna. While some in-body communication systems initially used the *Industrial Scientific and Medical* (ISM) [2] bands, the *Medical Implant Communication System* (MICS) [3] band of 403–405 MHz is gaining worldwide acceptance. This band has a power limit of 25 μW in air and is split into 300 kHz wide channels.

The human body is a medium that poses numerous wireless transmission challenges. Unlike air, the body is composed of varied components that are not predictable and will change as the patient ages, gains or loses weight, or even changes in posture. More details on the impact of the changing nature of the human body on wireless communications are discussed by Johansson [4]. Although there are simple formulas for designing free-air communications, it is very difficult to calculate performance for an in-body communication system, as each individual is different. To compound the design challenge, the location of the implanted device is also variable. A surgeon fits the implant into the best position to perform its primary function, with little consideration for wireless performance. This means an implanted RF communication system must operate in a wide variety of environments and positions that can change with time.

The typical dielectric constant (ε_r), conductivity (σ) and characteristic impedance (Z_0) properties of muscle and fat are illustrated in Table 4.1. The table demonstrates that these two mediums are very different and that properties change with frequency.

The dielectric constant has an effect on the wavelength of a signal. In air, the wavelength can be found from Eq. 4.1 where $\varepsilon_r = 1$. However, in a different medium the wavelength is reduced as Eq. 4.2.

Table 4.1 Body electrical properties

	Muscle			Fat		
Frequency (MHz)	ε_r	$\sigma(Sm^{-1})$	$Z_0(\Omega)$	ε_r	$\sigma(Sm^{-1})$	$Z_0(\Omega)$
100	66.2	0.73	31.6	12.7	0.07	92.4
400	58.0	0.82	43.7	11.6	0.08	108
900	56.0	0.97	48.2	11.3	0.11	111

Source: FCC and William Scanlon, Queens University Belfast

Table 4.2 Penetration depth of tissue, where penetration depth is when field intensity decreases by e^{-1}

Frequency (MHz)	Muscle (mm)	Fat (mm)
100	75.1	339
400	51.5	229
900	41.6	163

Source: William Scanlon, Queens University Belfast

$$\lambda = 300\frac{10^6}{f} \tag{4.1}$$

where λ is the wavelength in air in meters and f is frequency in Hz.

$$\lambda_{medium} = \frac{\lambda}{\sqrt{\varepsilon_r}} \tag{4.2}$$

where λ_{medium} is the wavelength in the medium. At 403 MHz, the wavelength in air is 744 mm, but in muscle with $\varepsilon_r = 50$ the $\lambda_{medium} = 150$ mm. This is of significant help in designing implanted antennas where physical size is an important consideration. The conductivity of muscle is 0.82 Sm^{-1} – this is more than air, which is almost zero. The effect of this is similar to surrounding the implant with seawater that will attenuate the signal as it passes through. This results in reduced penetration depth as shown in Table 4.2.

The characteristic impedance (Z_0) is relevant when it changes, such as at the fat-muscle boundary. This will cause part of the signal to be reflected by a term known as reflection co-coefficient Γ, found from Eq. 4.3.

$$\Gamma = \frac{Z_o - Z_r}{Z_o + Z_r} \tag{4.3}$$

where (Z_0) is the impedance of free space (377 Ω), and (Z_r) is the impedance of medium in Ω.

If the layers are thick enough, considering the frequency of propagation, this results in a signal being reflected of magnitude Γ of incident signal power. So for a muscle-fat boundary, $\Gamma = 80$ % of the signal is reflected as shown in Fig. 4.1. If the signal is incident at the Bragg Angle, all of the incident signal will be reflected. This is shown in Fig. 4.1.

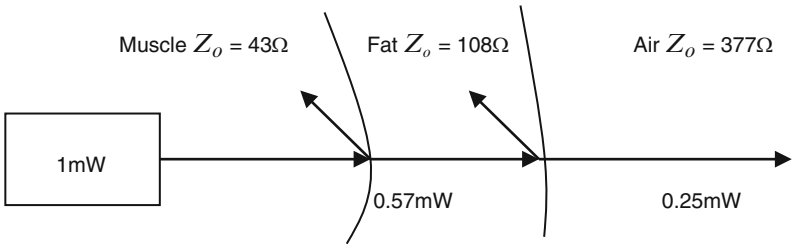
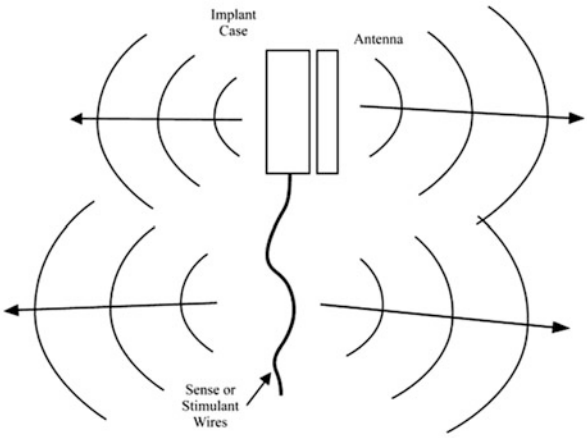


Fig. 4.1 Signal reflection at muscle – fat – air boundaries

Fig. 4.2 Radiation from an implant



As an implant does not have an earth (ground), the case or other wires will also radiate. This means that signals will be radiated from the antenna and other structures associated with the implant, as shown in Fig. 4.2.

To summarise, a signal travelling through the body will suffer attenuation and reflection from various boundary changes.

4.4 Implanted Transceiver

There are several integrated transceivers that will operate over the MICS band [5], many of which have a high current when operating. If left running, the transceiver would deplete the battery very soon and require changing or recharging. Both of which are an inconvenience at the very least and potentially costly with a risk to the patient. Typically, the current during receive is similar to that in transmit so it is not practical to leave the receiver operating constantly while waiting for communication to be initiated from outside of the body. The external transmitted signal must not exceed −16 dBm so the implanted receiver must be sensitive. Clearly, the

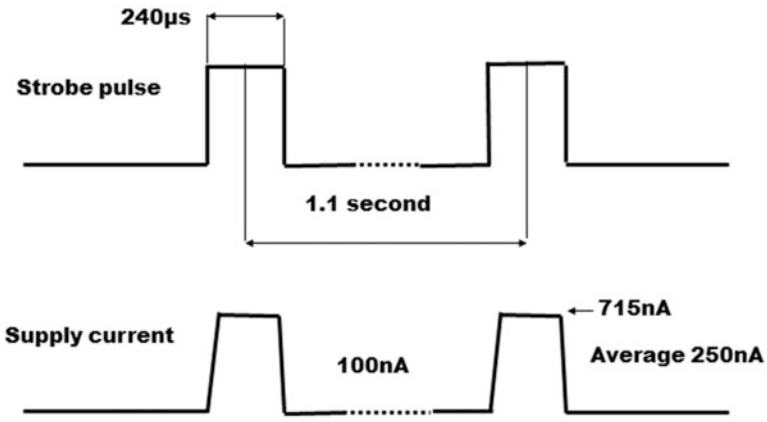


Fig. 4.3 2.45 GHz receiver current – waiting for wake-up (Source: Microsemi (previously Zarlink) Semiconductor)

transceiver must operate virtually on demand to enable therapy to be changed or patient data to be downloaded. The received signal level at 403 MHz within the body will be very low and the antenna far from optimum for that frequency, therefore a receiver will need to be as sensitive as possible. However a higher sensitivity has the price of higher current consumption.

If a transceiver is in "sleep" mode and taking very little current, then it needs to be woken up reliably. This can be done by switching the receiver on for a short period of time of every second (or other interval). During this short switch on period, the receiver must detect a wake up signal and power up the rest of the transceiver. When the data transfer is completed, the transceiver can return to sleep mode.

To further reduce the current consumption, the wake up can be done using an ISM frequency around 2.45 GHz where the transmit power is up to 100 dBm (depending on national regulations) that means the wake up receiver can be less sensitive and draw less current. Figure 4.3 shows an example of the wake up receiver current of the Microsemi (previously Zarlink) ZL70101. This IC has a 2.45 GHz wake up receiver and MICS band transceiver that is switched off during sleep mode. The 2.45 GHz wake up receiver draws 715 nA when switched on but is only active for 240 µs each 1.1 s. This reduces the average sleep current to 250 nA. The wake up function does not require transmission at 2.45 GHz, since it is receive only. The on/off ratio can be varied to meet different circumstances.

To summarise, the wake up receiver is not the same as the MICS transceiver (although it may be on the same silicon), as it has a lower sensitivity. It relies on a higher transmit power to operate and is only switched on periodically for a short time. The transceiver needs to communicate with the application where an SPI interface is often used. This is often found on microcontrollers for data transfer that can be used to set up the transceiver; send data to be transmitted and receive data from outside of the body. If a microcontroller is used to control the wireless functions and/or the

application, then its current must be included in the power budget. A processor operating at a high clock frequency will process instructions faster than one with a lower frequency clock but will require more current from the battery. Some processors also have a sleep function to minimise power, however care must be taken with wake up routines. Another consideration for a processor clock is to be sure that no harmonics appear in the communication band. The receiver may respond to high order harmonics. Using a common clock derived from the communication function can avoid this.

A discrete design is a useful way of proving out a concept and to obtain useful operating data but for an implant to be used in significant numbers, an integrated solution may be needed. This is going to be an expensive undertaking and will take considerable time that need to be included in any project planning.

4.5 Antenna Design

Of the many books written on the subject of antennas, almost all of them describe operation in air. While some of the principles still hold true for in-body use, as described above, the surrounding medium is very different. A good basic reference is provided by Kraus [6].

A half-wave dipole for 403 MHz in air will be 372 mm, but in muscle it is reduced to 52.6 mm. Thus a classical resonant antenna is not feasible even with the reduced dimensions caused by body tissue. The first problem is that each body will have a different ε_r that may change as the patient changes weight or the implant moves. This will cause the antenna to be non-resonant and operate with reduced efficacy. As resonant antennas are also often too large for in-body use, this design is not practical.

An in-body antenna needs to be tuneable with an intelligent transceiver and routine. This will enable the antenna coupling circuit to be optimised and the best signal strength to be obtained. Often, the size constraints dictate the choice of a non-resonant antenna. A non-resonant antenna will have a lower gain and therefore be less sensitive on receive and radiate less of the power generated by the transmitter. Operating in the MICS band of 403 MHz, almost any antenna that will fit comfortably within the body will not be as good as one designed to operate in air with no space constraints. This makes design of the antenna coupling circuit even more important.

Antenna options are also dictated by the location of the implant. For example, a urethra valve (artificial bladder sphincter) needs to be replaced, without surgery, at regular intervals. The available diameter is 4–6 mm and the length is restricted. This rules out a patch antenna, and it would be difficult to keep a monopole or dipole in place even if they would fit. The best option is to integrate a helical antenna into the shape of the valve implant. The design equations are found in a paper by Krall [7] and two lab versions are shown in Fig. 4.4. This type of antenna

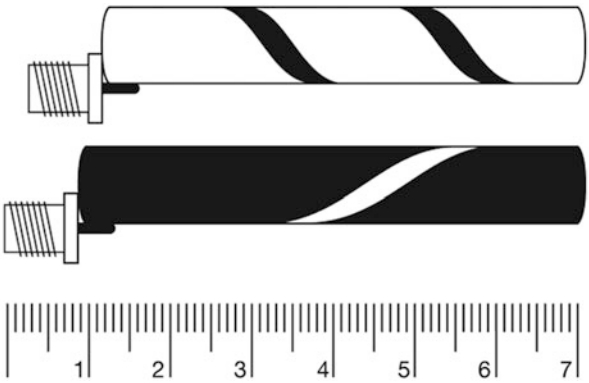


Fig. 4.4 Helix antenna copper conductor on a ceramic substrate

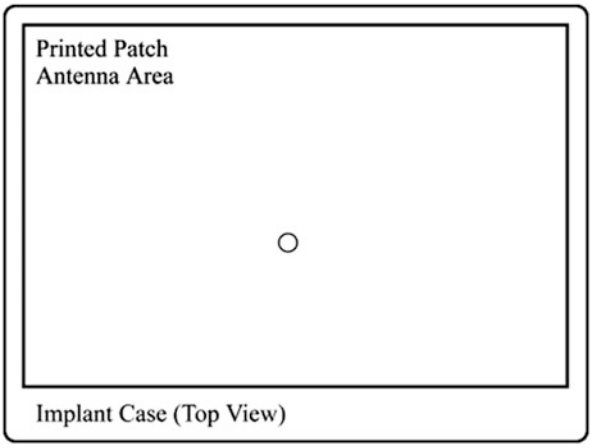


Fig. 4.5 Patch with through-hole connection

may also be of use in an oesophagus probe; the conductor could be printed or evaporated onto the surface of the valve.

A patch antenna can be used when the implant is flat and there is no room to deploy a short wire. Patch antennas comprise of a flat substrate coated on both sides with conductor. The substrate is typically alumina or a similar body-compatible material (biocompatibility will be discussed later), with platinum or platinum/iridium coating both surfaces. The upper surface is the active face and is connected to the transceiver and the back face is typically connected to the implant 0V. The connection to the transceiver needs to pass through the case where the hermetic seal is maintained, requiring a feed-through. The feed-through must have no filter capacitors present; these are common on other devices. The connection to the top (active) surface can be by a hole through the substrate (Fig. 4.5) or by a wire connected to the top (Fig. 4.6). The back face can be connected to the case with conductive epoxy, if it is attached to 0V, or by wire.

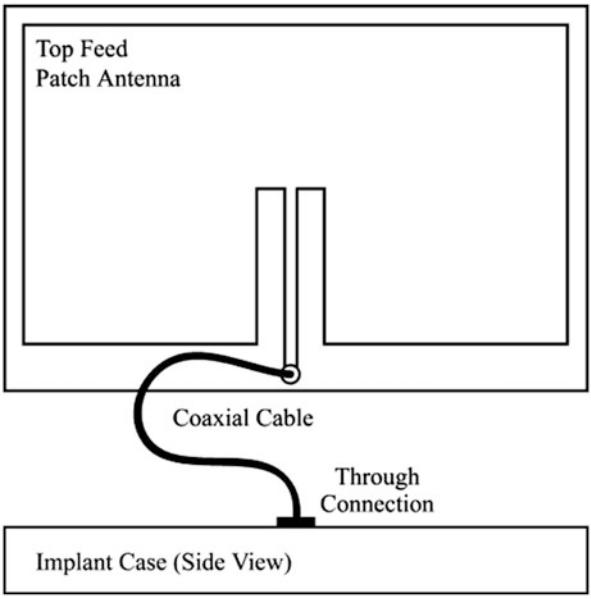


Fig. 4.6 Patch with wired top surface connection

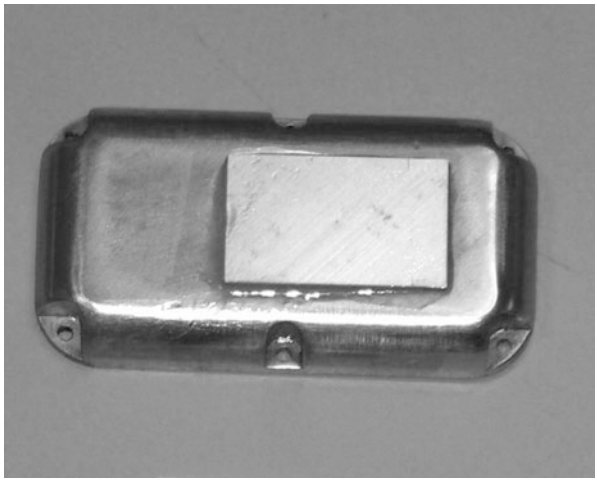


Fig. 4.7 Patch antenna mounted on experimental implant case (Source: Microsemi (previously Zarlink) Semiconductor)

A patch antenna will be electrically larger than its physical size because it will be immersed in a high ε_r medium. It can be made to appear even larger electrically if the substrate is of higher ε_r, such as Titania or Zirconia. Further reading on small antennas can be found in Fujimoto [8] and PCB antennas by Lee and Chen [9]. Other antenna reading includes ARRL [10] and Kraus [6]. An example of a patch mounted on a test implant is shown in Fig. 4.7 where the patch is 19.5 $\times$ 32 $\times$ 1.0 mm using an alumina substrate. This was successfully used in propagation tests for signals transmitted to and from a live patient.

Fig. 4.8 Loop antenna encased in epoxy mounted on experimental implant case (Source: Microsemi (previously Zarlink) Semiconductor)

Another type of design to be considered is the *Planar Inverted F* (PIF) antenna commonly used in small mobile phones. A small phone and an implanted device share many performance challenges. In this case, the antenna is mounted onto the case and is encased in a suitable non-conductive material for mechanical stability. This type of antenna is small but found to be effective in some applications.

The off-resonance antennas have a low radiation resistance, typically in the order of a few Ohms for a patch. Better radiation is achieved with a higher radiation resistance that generally require a larger structure.

A loop antenna is an option that can be deployed away from the implant case or other metal. The loop antenna operates mostly with the magnetic field, whereas the dipole, patch, monopole, etc. operate mostly with the electric field. The loop antenna delivers comparable performance to that of a dipole, but with a considerably smaller size. Furthermore, the magnetic permeability of muscle or fat is very similar to that of air, unlike the dielectric constant that varies significantly as described above. This property enables an antenna to be built and used with much less need for retuning. A loop antenna does need to be mounted away from the case and on a biocompatible structure. Equations 4.4 and 4.5 relate to small and large loops, other equations exist for multi-turn loop designs.

$$R_{rad} = 31200\left(A/\lambda^2\right)^2 \tag{4.4}$$

for $A \leq \lambda^2/100$, where R_{rad} is radiation resistance, A is the loop area and λ is the wavelength in the medium.

$$R_{rad} = 3270\left(A/\lambda^2\right)^2 \tag{4.5}$$

for $A > \lambda^2/100$.

Figure 4.8 shows a loop antenna encased in epoxy, it measured 5 × 15 mm and was a single loop. The electrical properties of the antenna were better than that of a patch and can be integrated into a biocompatible epoxy block.

An important measurement of an antenna performance is the return loss – how much of the signal sent to the antenna is reflected back and how much is radiated. An ideal value for Γ is 0, with 1 representing total reflection. One feature of a small antenna is that it may have a good return loss but radiate poorly, meaning radiation efficiency must also be measured.

4.6 Antenna Testing

Before designing a matching network for the antenna/transceiver interface it is necessary to measure the impedance of the antenna within a representative medium. Testing an implant antenna in-air has limited use and non-living tissue does not have the same properties as the human body, so a body phantom is used. A mixture of water, sodium chloride, sugar and *Hydroxyl Ethyl Cellulose* (HEC) will mimic muscle or brain tissue [11] for the frequency range 100 MHz to GHz (see Table 4.3 below).

4.6.1 Antenna Impedance and Radiation Resistance Measurement

Knowing the impedance of the antenna is critical for the design and operation of the in-body communication system. The following example is for a patch but can be adapted for other antenna types.

Radiation resistance can be measured with an antenna immersed in body phantom liquid in a Perspex cylinder. This can prove difficult as the liquid may leak into the measurement jig and affect the results. As an alternative, a patch, for example, is mounted on a copper plate and pushed against a bag of the phantom (Fig. 4.9). If the real part of the impedance is $>10\ \Omega$ then a network analyser can be used, however, the real part is often lower, requiring a different method. A way of measuring an antenna with low radiation resistance is described in some detail [6], as it will often be the case that an implanted antenna will be difficult to measure directly on a network analyser.

Table 4.3 Body tissue recipes

	% By weight	% By weight
Ingredient	100 MHz–1 GHz	1.5–2.5 GHz
Water	52.4	45.3
Sugar	45.0	54.3
Salt (NaCl)	1.5	0.0
HEC	1.1	0.4

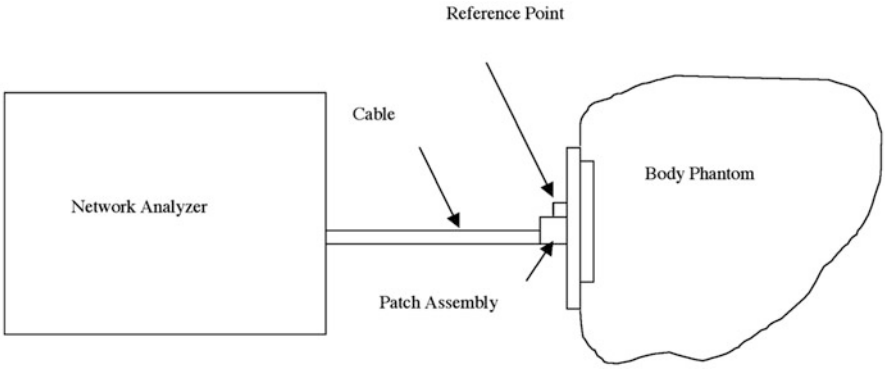


Fig. 4.9 Network analyser test set-up for impedance real part >10 Ω

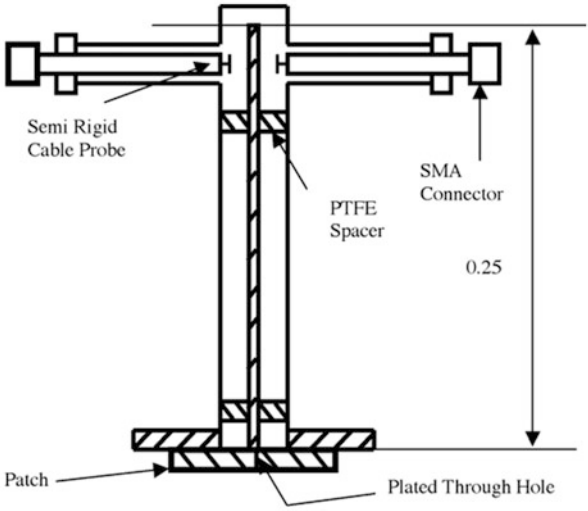


Fig. 4.10 Quarter wavelength 400 MHz line

4.6.2 Quarter Wave Line Impedance Measurement

One alternative to using a network analyser directly is to couple a signal into a quarter wavelength line (or odd integer multiples of quarter wavelength) and measure the signal loss. From this, the change in resonant frequency and the Q (quality factor) can be used to determine the patch impedance.

The resonant line and coupling structure is shown in Fig. 4.10. This is a cross-section through the centre of the assembly showing the internal construction. The length of the centre conductor is 0.25 λ. The impedance is defined by the diameters of the centre conductor and the outer tube.

The feeds comprise of RG402, 50 Ω, semi-rigid cable terminated with SMA connectors and clamped in place. The bottom of the centre bar has a split collet to

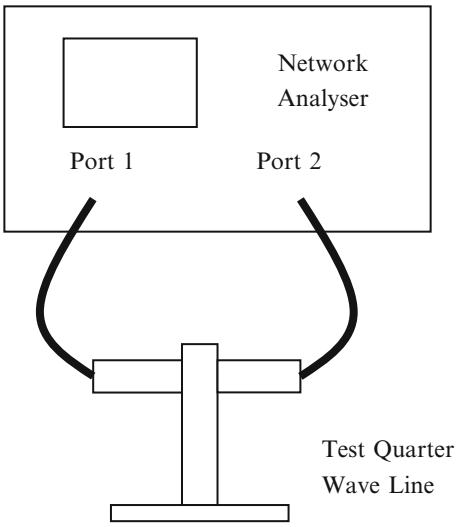


Fig. 4.11 Resonant line measurements

enable the test sample to be attached. The spacers are of PTFE, the rest of the jig is copper.

A reference short circuit plate is also needed to calibrate the line, and this consists of a brass block in contact with the centre conductor and makes good contact to the bottom plate.

The line is attached to a network analyser, as shown in Fig. 4.11, and is set to measure the S21 (transmitted signal from Port 1 to Port 2). The S21 measurement will result in a peak in the transmission at the resonant frequency. Further measurements of the peak are used to determine the Q of the antenna. The resonant frequency and Q of the line are measured. Q is defined as

$$Q = \frac{f_{centre}}{B_{3dB}} \tag{4.6}$$

where f_{centre} is centre frequency and B_{3dB} is 3 dB bandwidth. With the line impedance, the loss can be derived from

$$\frac{R_{loss}}{Z_o} = \frac{N\pi}{4Q} \tag{4.7}$$

where N is the number of quarter wavelengths, and R_{loss} is the resistance of the line. Having found the losses, the reference short is replaced with a test antenna. Resonant frequency and Q are measured again and the radiation resistance is

$$\frac{R_{measure}}{Z_o} = \frac{N\pi}{4Q} \tag{4.8}$$

and

$$R_{rad} = R_{measure} - R_{loss} \tag{4.9}$$

Once the real part of the impedance is known, the imaginary part can be found. This is best done with simulation software using a model of the transmission line created with a load of R_{rad} in parallel, or series, with a capacitor (or inductor as appropriate). The capacitor value is tuned to give the same resonant frequency as measured with the network analyser. It is not necessary to simulate the line feeds, only to measure from the end of the line.

This technique can be adapted for other types of antennas by using a length of semi-rigid coax cut to length or printed transmission line terminated with the reference short and antenna.

Even if an antenna has a known impedance, or is resonant at the required frequency, this does not always mean that it will radiate effectively. It is essential to measure the performance using a body phantom with the transceiver in the case and sense or stimulant wires attached. The Wheeler Cap is also useful in measuring antenna efficiency.

4.7 Matching Network

Once the impedance of the patch is known within a reference liquid, it can be matched to the transceiver. When implanted, a transceiver will be capable of optimisation if it has a built-in variable tuning element. This is typically an array of capacitors that can be switched in or out across an RF terminal. This will enable the implant to be retuned each time it is used to help counter some of the effects of the implant moving or body changes.

A typical transceiver circuit is intended for use in the MICS frequency band of 402–405 MHz. The maximum radiated power allowed in this band is 25 μW (−16 dBm), but antenna gains in implants are usually very low and more power can be generated at the transmitter to compensate for loss through the body. Additionally, link budgets are frequently tight such that the maximum allowed radiated power is required.

The antenna tuning circuits are required to present an optimum load impedance to the transmitter and voltage step-up to the receiver. It should be noted that this does not necessarily lead to a conjugate impedance match.

4.7.1 Transmitter Tuning

A typical transmitter is capable of producing an output signal into the load of 2 V peak-peak maximum, with a maximum peak current of about 10 mA. Maximum DC to RF conversion efficiency is obtained when using the full voltage swing; as

for any particular power output, this will require minimum supply current. The tuning circuit is also required to provide a degree of harmonic rejection for regulatory reasons; exactly how much is dependent upon the antenna gain and impedance at the harmonic frequencies. Maximum efficiency is obtained when the output devices are loaded with a purely resistive load, but because of the effects of stray capacity, as well as the provision of internal variable tuning capacity on the output of the transmitter, the actual load presented to the transmitter output pad is required to be inductive.

The efficiency, η, of the tuning network is determined by the ratio of unloaded $Q(Q_u)$ to loaded $Q(Q_w)$, i.e.,

$$\eta = \left(1 - \frac{Q_w}{Q_u}\right) 100\% \tag{4.10}$$

The use of too low a value of Q_w should be avoided, however, as the harmonic attenuation of the network will be reduced. Usually, a value of Q_w between 10 and 15 is a reasonable compromise. Nevertheless, the harmonic attenuation may well be greater than what would appear at first sight to be available from such a low value of Q. This is because the impedance presented by the network at the harmonic frequencies may be much lower than at the fundamental frequency, resulting in the harmonic current generating a much lower voltage across the input of the network.

Too high a value of working Q should also be avoided. This is because the increased Q leads to an increase in circulating current within the circuit, and losses are proportional to the square of the circulating current thus doubling the Q increases the actual power lost by four times.

The first step in designing the transmitter-tuning network is to determine the required RF output power, P_0. This is derived from the required radiated power, more so than the antenna gain and matching circuit losses. Since the latter are unknown at this stage, the process is iterative.

To determine the resistance presented to the transmitter, first establish the maximum voltage swing by

$$R_L = \frac{\left(0.7071 V_{pp}/2\right)^2}{P_o} \tag{4.11}$$

where R_L is load presented to the transmitter. For $2V_{pp}$ Eq. 4.11 becomes Eq. 4.12.

$$R_L = 0.5/P_o \tag{4.12}$$

For a maximum available current of 10 mA, the lowest value of R_L is 200 Ω. A circuit can now be chosen that provides load impedance to the power amplifier of $0.5/P_0$ Ω in parallel with a tuning capacitance. In many cases, the maximum power will be limited either by the available power supply current or by regulatory limitations on maximum radiated power.

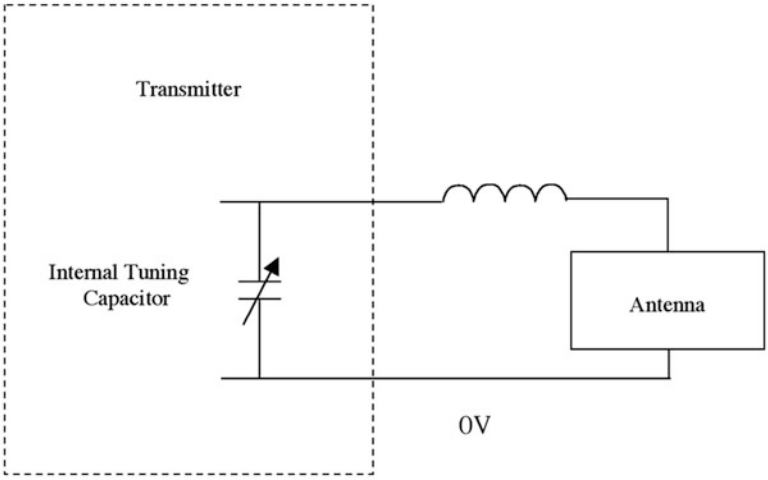


Fig. 4.12 L network matching

4.7.2 The L Network

The simplest network is the L network shown in Fig. 4.12 as it uses only two components. Easing design, the capacitive arm is internal to the device and only an external inductor is required.

There are certain limitations in such an approach. The antenna impedance is constrained in terms of the amount of inductive reactance it can have, and the working Q is given by

$$Q_w = \sqrt{R_L/R_s - 1} \tag{4.13}$$

where R_L is the parallel load resistance presented to the device, and R_s is the sum of the inductor loss resistance and the resistive part of the antenna impedance. R_s must be less than R_L. Note that especially in the case of electrically small antennas, this resistance is not usually the radiation resistance of the antenna as other losses may dominate.

When the antenna impedance is capacitive, the value of inductance can be increased to resonate the antenna circuit. For the L network, the following equations apply:

$$X'_L = R_{ant}\sqrt{R_L/R_{ant} - 1} \tag{4.14}$$

$$X_C = R_L\sqrt{R_L/R_{ant} - 1} \tag{4.15}$$

The total value of inductive reactance is $X_L = X'_L + X_{ant}$, where X_{ant} is negative for capacitive antenna impedance and positive for an inductive antenna impedance.

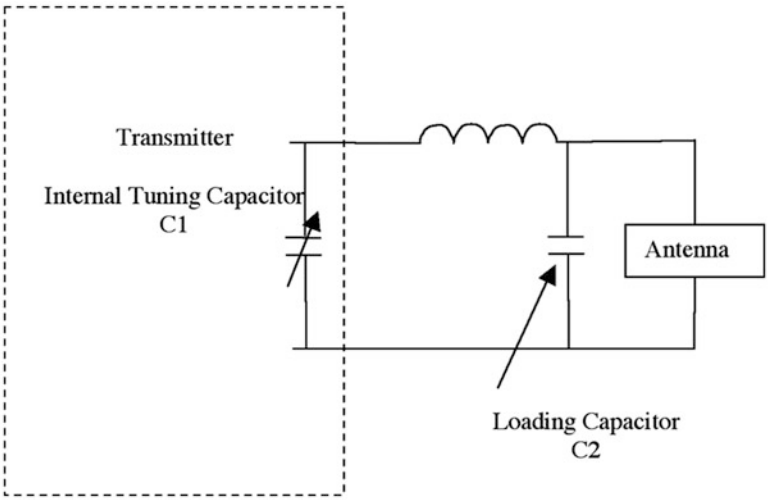


Fig. 4.13 The π network

The required value of X_c will typically lie between 60 and 390 Ω at 403 MHz for tuning with the internal capacitor. Lower values of X_C can be accommodated with an external capacitor. The effects of the inductance of the bond wire connecting the IC output pad to the inductance must also be considered. This inductance will be of the order of 1nH/mm. The resulting effect should not be ignored.

4.7.3 The π Network

The π network shown in Fig. 4.13 has certain advantages in that the Q of the circuit is relatively independent of the impedance transformation required. It is also able to handle antennas with greater inductance than the L network. The Q_w is R_L/X_{C_1}, meaning an external capacitor will be needed, and the tuning range of the internal tuning circuit will be limited, since values of Xc_1 of 20–50 Ω will generally be required.

Equations for the π network are:

$$X_{C_1} = R_L/Q_w \tag{4.16}$$

$$X_{C_2} = \left(\frac{R_L R_{ant}}{\left(Q_w^2 + 1\right) - R_L/R_{ant}}\right)^{\frac{1}{2}} \tag{4.17}$$

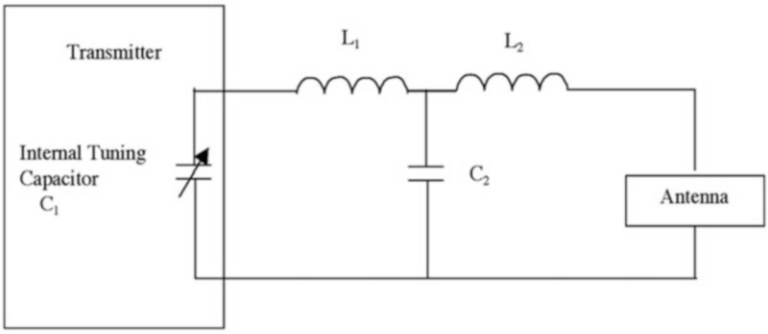


Fig. 4.14 The 'T' or 'π-L' network

$$X_L = \frac{Q_w R_L + R_L R_{ant}/X_{C_2}}{Q_w^2 + 1} \tag{4.18}$$

In practice, the value of X_{C_2} is decreased by the amount of parallel capacity of the antenna (in which case the network appears to be an L network), or is increased by the amount of capacity needed to resonate an inductive antenna. When R_{ant} is small, the value of C_2 may become excessively large, thus limiting the application of this circuit. When R_L is high, the working Q may be higher than desirable because of the unavoidable stray capacities across the input to the network. Again, the parasitic inductance of the bond wire from the output pad to the inductor must be accounted for.

4.7.4 The T and π – L Networks

These networks are identical in circuitry. The T network is the 'dual' of the π network, as illustrated in Fig. 4.14. It claims to show somewhat greater efficiencies under certain conditions. The design technique is to treat the circuit as consisting of back to back L networks, with a centre 'image' impedance at C_2 chosen to provide suitable Q_w values. This image impedance is higher than the value of R_L and thus is more subject to stray capacity problems. The equations are:

$$X_{L_1} = R_L Q_w + X_{C_{1s}} \tag{4.19}$$

$$X_{L_2} = R_L B \tag{4.20}$$

$$X_{C_2} = (A/Q_w + B) \tag{4.21}$$

where

$$A = R_L\left(1 + Q_w^2\right) \tag{4.22}$$

and

$$B = \sqrt{(A/R_L) - 1} \tag{4.23}$$

$X_{C_{1s}}$ is the series equivalent of the circuit of R_L and X_{C_1}. The series equivalent of a parallel circuit is found in Eqs. 4.24 and 4.25. The parallel equivalent of a series circuit is shown in Eqs. 4.26 and 4.27.

$$R_s = \frac{R_p}{1 + \left(R_p/X_p\right)^2} \tag{4.24}$$

$$X_s = R_s R_p / X_p \tag{4.25}$$

$$R_p = R_s\left(1 + \left(X_s/R_s\right)^2\right) \tag{4.26}$$

$$X_p = R_s R_p / X_s \tag{4.27}$$

where R_s = resistance of the series equivalent circuit, R_p = resistance of the parallel equivalent circuit, X_s = reactance of the series equivalent circuit, and X_p = reactance of the parallel equivalent circuit. As with the L network, similar limitations to the amount of antenna inductive reactance apply.

As a π-L, the circuit uses a π section to reduce the impedance to an image impedance, usually of value $\sqrt{R_L R_{ant}}$. The L section then reduces this to the antenna impedance, and tunes out any antenna reactance. The same problems of allowable antenna reactance apply as in the case of the L network, but as the impedance transformation ratio is reduced, the Q_w of the L network is also reduced, allowing for an improved efficiency. The extra section also provides increased harmonic rejection as compared with the simple π network.

4.7.5 Parasitic Effects

The effect of stray capacitance and resistance in inductors also need to be considered, because operating in proximity to the *Self-Resonant Frequency* (SRF) of the inductor leads to large changes in Q and apparent inductance. In general, it is recommended that the inductance SRF should be at least three times the operating frequency. Discrepancy between expected and actual harmonic radiation levels when operating with an antenna may often be traced back to this cause.

The actual difference in parameters caused by an approach to self-resonance can be found from the following. For an inductor:

$$R_{apparent} = \frac{R_{actual}}{\left(1 - \lambda^2\right)^2} \tag{4.28}$$

where $R_{apparent}$ and R_{actual} refer to the parasitic resistance of the inductor and

$$L_{apparent} = \frac{L_{actual}}{\left(1 - \lambda^2\right)^2} \tag{4.29}$$

where λ is the ratio of actual frequency to self-resonant frequency that is available from the manufacturer.

It should be noted that the networks suggested all have the property of using a series inductor from the output of the device, thus allowing absorption of the stray inductance caused by the bond wire.

4.7.6 Network Choice

The choice of which network to use is dependent upon a number of factors. When one antenna is required for operation at the MICS band and 2.4 GHz (wake-up), the 403 MHz network must be chosen so that it presents high impedance at 2.4 GHz. Similarly, the 2.4 GHz tuning system must not have appreciable shunting effect at 403 MHz on the antenna. This suggests that an L or T or π-L network has advantages, insofar as the series inductor at the antenna end of the network will present reasonably high impedance to the 2.4 GHz wake-up signal.

Nevertheless, this advantage may not be realised in practice if the series inductor has too low an SRF, and also in the demonstration implementation where an L network is used the inductor has been split into two parts. Connected to the antenna is a small inductor arranged to be a parallel resonant at 2.4 GHz, and thus presenting high impedance at that frequency. The other inductor in series enables the correct amount of inductance at 403 MHz to be achieved, as shown in Fig. 4.15.

C_1 and L_2 are resonant at about 2.4 GHz. Typical values are about 1.9 nH and 2.2 pF, resonating at 2.46 GHz. L_1 is the remainder of the 403 MHz tuning inductance. Without the splitting of the inductance, the SRF of this inductor (commonly in the order of 10's nH) would have been below 2 GHz, and the circuit would therefore have looked like a capacitor at 2.4 GHz.

Whichever network is used, there must be DC isolation to ground through the antenna and this may necessitate the addition of a series capacitor. Predominantly, it is desirable that such a capacitor be added at a medium impedance point. At a high impedance point, the effects of parasitic capacity to ground may cause problems, while at very low impedance points the series resistance of a large capacitor may increase losses to an unacceptable value.

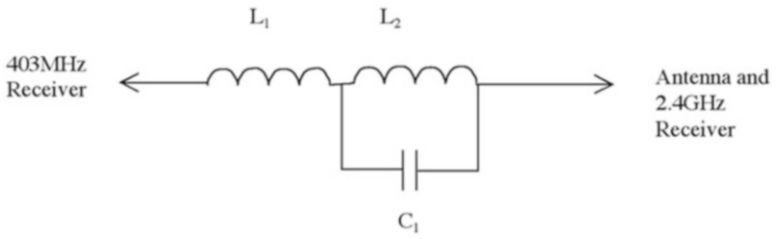


Fig. 4.15 Multi-resonant network

4.7.7 Radio Frequency Losses in Components and Layout Issues

In all cases, after the initial values have been established, substitution of more accurate models of inductors (including the available *Q* and the SRF) are required to establish the circuit efficiency. It may then be necessary to 'fine-tune' the design for a different power output from the transmitter. Overall, losses in the capacitors are negligible in comparison with those in inductors, especially for very small chip inductors or printed inductors mounted very close to a ground plane. Microwave type surfaces mounting chip porcelain or sapphire low loss type capacitors are generally preferred, rather than the lower cost NP0 ceramic types. If a software routine is to be used to tune an antenna in situ, then it must be remembered that a programmable capacitor (such as the Maxim MAX1474) will have a low Q compared to a discrete part. Consequently, it is recommended that the majority of the tuning is done with a discrete capacitor with a variable device in parallel thus obtaining the best Q for the circuit. At present, programmable inductors are an emerging technology.

Equations 4.28 and 4.29 for the apparent values of inductance and resistance of inductors at frequencies removed from the SRF will allow extrapolation of manufacturers' data where necessary.

Here are a few rules that may be useful when laying out a matching network:

- Remember that the track will also influence the performance and should be simulated along with the added components.
- Lay close inductors at 90° to each other. This minimises mutual coupling.
- Take all of the tracking, wiring and feed through into account when designing a matching network.
- As inductor values increase, the self-resonance will be reduced. A high value inductor may behave like a capacitor at high frequency.
- When building a transceiver, be consistent in the choice of component manufacturer and family. A nominal component value may have different parasitics between manufacturers.

4.7.8 Receiver Tuning

The receiver has a very high input impedance, being typically >10 kΩ in parallel with a small capacitance. Conjugate matching of the antenna to the receiver is undesirable because half of the received EMF is lost in the antenna's internal resistance, while the high input impedance of the receiver allows for a degree of voltage step-up. Additionally, conjugate matching does not give the optimum source resistance for lowest noise. The amount of voltage step is proportional to the working Q of the network, and because of the difficulty of maintaining tuning accuracy as the working Q value increases (always assuming that the Q value is not limited by the unloaded Q values of available components), it is not considered desirable to have too high a voltage step-up. Where the step-up is such that the working Q is dominated by the values of Q in the available components, the losses may well reduce the benefits of attempting to obtain a large step-up.

Link budget calculations are based on *Effective Radiated Power* (ERP) and antenna gains. The resulting power available for the receiver should be considered as being an EMF voltage delivered to the input impedance with a source impedance determined by the step-up ratio, after losses in the components have been taken into account. Thus the voltage available at the receiver input will be:

$$E = \left[\left(P_t P_L A_g T_L\right) R_L\right]^{\frac{1}{2}} \tag{4.30}$$

where E is the available RMS EMF, P_t is the transmitted ERP in watts, P_L is the path loss as a ratio, A_σ is the receive antenna gain as a ratio, T_L is the tuning circuit losses as a ratio (which can be derived from Eq. 4.10) and R_L is the resistive part of the load impedance for the transmitter (assuming that the use of a tuning circuit is common to both receiver and transmitter).

Where receive and transmit use separate antennas, the receiver input tuning can be designed for any desired step-up, bearing in mind that the limitations caused by too high a value of Q_w in terms of tuning sensitivity, etc. It must be remembered that the various parasitic components, such as bond wires and stray capacities, must be accounted for when evaluating the performance of any particular circuit.

An implanted antenna working at 403 MHz will always have worse performance than a half wave dipole, so it is essential to optimise and continue to optimise the coupling network. Any loss of signal will result in a reduction in range and/or a reduction in data rate that will ultimately result in a higher current consumption. It should also be remembered that a typical receiver will demand as much, or more current than a low power transmitter, thus it is essential to optimise both tuning regimes even if they are different.

A graphical method can be used to design a matching network as described in Wireless Body Area Networks [12].

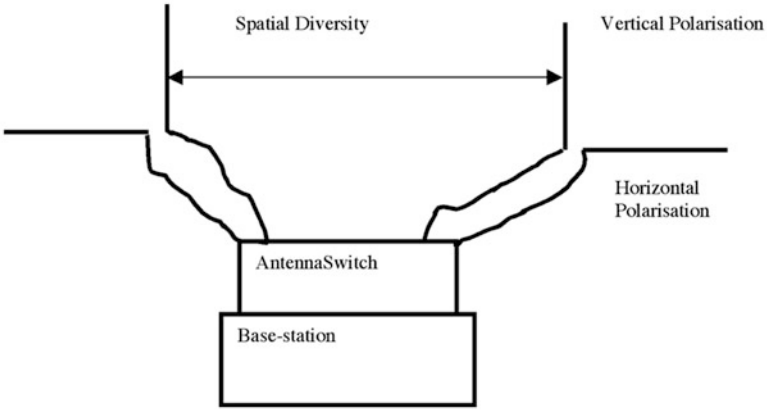


Fig. 4.16 Four antennas for spatial and polarisation diversity

4.7.9 Base Station Antennas

The implanted device and external base station antenna implementations are often very different. In the case of the implant, space considerations may well prevent the use of separate antennas for the frequencies used in communication and 'wake up' (MICS and 2.4 GHz respectively). The base unit will have room for larger antennas and preferably separate antennas for 403 MHz and 2.45 GHz (if used). The 403 MHz transmit and receive paths can be split at the antenna with a switch enabling optimisation of both paths separately. The base unit may use external filters in the receive chain to provide maximum rejection of unwanted signals. The loss of such filters (e.g. a SAW filter) may be offset by using an external RF amplifier. The gain is not too important as long as it is sufficient to overcome the filter losses; such an amplifier will also need a very low noise figure, and a sufficiently large signal handling range.

The base station could also use more than one antenna to overcome the effect of multi-path fading and polarisation, as detailed by Johansson [4], reducing the signal strength. If space permits, an arrangement of four antennas with suitable switching and software optimisation can be employed. This is shown in Fig. 4.16.

4.8 Propagation

The propagation pattern of the antenna, case and any wires for sense or stimulation is required in order to predict the performance of the implant within a body. The Perspex tank filled with a liquid as described above is a useful first representation of a body. This can be done in a lab, or preferably in an anechoic chamber, with care taken to seal the test transceiver from the liquid.

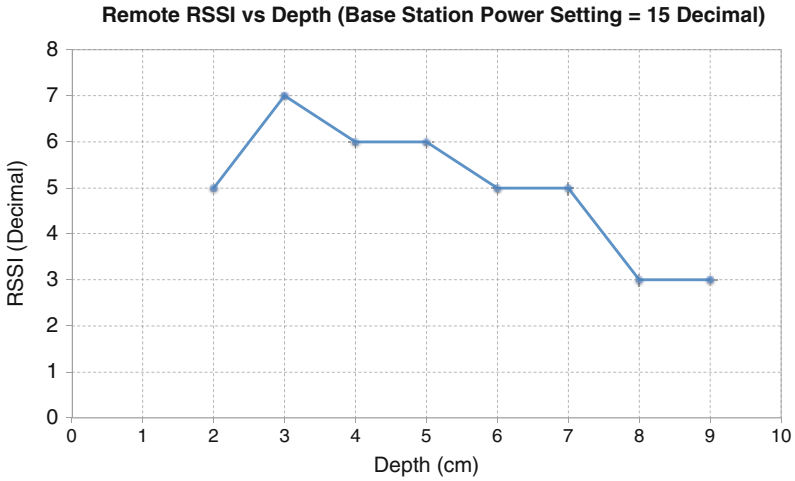


Fig. 4.17 Signal strength received at implanted receiver vs. depth (Source: Microsemi (previously Zarlink) Semiconductor)

Radiation patterns are made with the body phantom using a self-contained transmitter immersed in the liquid. If the antenna were to be attached to a cable then it would contribute to the radiation pattern. This can be minimised, but not eliminated, with the addition of ferrite beads. The patch attached to an implant case within a body does not have an earth (ground) connection, meaning the case will radiate in anti-phase to the patch. This requires the electronics to be self-powered and measured as a whole.

Measurements should be taken with the test device immersed in the liquid and rotated on a horizontal and vertical axis. If vertical rotation is difficult then the 90° points should be measured. Horizontal rotation is straightforward as the test device can be rotated with the tank.

Another important aspect of propagation is polarisation. Human and animal testing has found that the body will cause polarisation of the signal along its long length [4]. This has also been observed with the tank measurements. It has been found that with a vertical tank the polarisation also tends to vertical.

The signal strength will vary with depth within a body. Figure 4.17 shows the relative signal strength, at the implant, for various depths within a body where it can be seen that it increases initially with depth and then declines. This is thought to be due to the body forming a parasitic antenna that improves the performance of, in this case, a small patch antenna. As the transceiver operates from deeper into the body then the effect of being surrounded by a weakly conductive medium will cause the signal to be absorbed. The effect of depth on the error rate of the signal also shows that the link improves then becomes worse with depth.

There are various 3D electromagnetic software packages available that will simulate propagation through the body with commercially available software

from organisations such as Agilent, CST and Ansoft. These have limited use as they only simulate one position, require a body or partial body model and can take several hours to run. Despite this, however, they do produce a detailed image of the progression and attenuation of a signal that may be of use. Any simulation should be backed up with a body phantom or live trial.

There are also several texts on the subject of propagation through the body [13].

4.9 Materials

An implant case is typically titanium or implant-grade stainless steel. In-body wires are either platinum or platinum/iridium that have conductivity in the order of 9.52 MSm^{-1} and 5.2 MSm^{-1} respectively. In comparison, the conductivity of copper, considered one of the best conductors, is 58 MSm^{-1}. At present these are the only two conductors that are used outside of the implant case. Metals such as silver and copper are toxic and blood will corrode gold. This low value of electrical conductivity will impede the performance of the antenna, as some energy will be absorbed by the resistance of the metal. Therefore, it is necessary to maximise the thickness of the conductor to minimise the added resistance and losses.

The substrate needs to be non-toxic, mechanically stable and insoluble in blood or other body liquid. Alumina is a material found to be acceptable. Other substrates that have been considered include titania, zirconia and multi-layer substrates. Care must be taken to ensure the suitability of the materials.

The entire implant is often coated in a passive material such as Parylene. Table 4.4 shows Parylene [14] has good water resistant properties compared to other commercially available materials and is acceptable for in-body use. Typical coatings are in the order of a few microns thick and will have no effect on the RF performance of the antenna. Coating cannot be used to isolate a conductor, such as silver from the body, as blood will dissolve most plastics and organic coating and thus make it porous.

Table 4.4 Water uptake and other parameters of various polymers, noting these are not all biocompatible [16]

Material	ε_r	Loss tangent	Water absorption %
Parylene (C type)	2.9	0.013	0.01
Polyether ketone	3.4	0.005	0.11
Polyether imide	3.2	0.0026	0.25
Polyether ether ketone	3.3	0.0035	0.11

4.10 Environment

The human body may be considered a benign thermal and mechanical environment, with a temperature varying by just ±2 °C and layers of fat and muscle that will partially absorb shocks. However, regulatory approval for implanted medical devices is extremely stringent. Implant grade components need to work over the full military temperature range and be able to withstand shock and vibration. The assembled implant needs to survive the wide storage temperature range that it may be exposed to. An implant is also subject to harsh mechanical testing, including a drop test from 2 m onto concrete for each of six faces (see EN45502). A layer of silicone may be sufficient to absorb the shock of impact.

4.11 External Transceiver (Base Station)

An in-body communication system relies on a base station that will transmit and receive signals from the implant and relay them to a user interface, such as a personal computer as shown in Fig. 4.18. Less rigorous size restrictions on a base station mean larger antennas can be used. The power limit, including any antenna gain, remains at 25 μW. The RF environment within a hospital or doctor's office may be even more challenging than inside the human body, meaning adjacent channel signals need to be filtered out using a SAW filter (or similar).

In the example shown, in Fig. 4.18, the base station has a USB interface that also provides the power to the Base Station thus eliminating the need for an additional power supply. Software needs to be written to operate the base station and an easy-to-operate *Graphical User Interface* (GUI) is also required. The interface for clinician use would typically show the user identification, download and upload data. For system development a more detailed interface is necessary.

When designing a base station for use in a sterile hospital environment, care must be taken to avoid corners or rough surfaces that cannot be thoroughly cleaned. Professionals familiar with infection control should be consulted throughout the design.

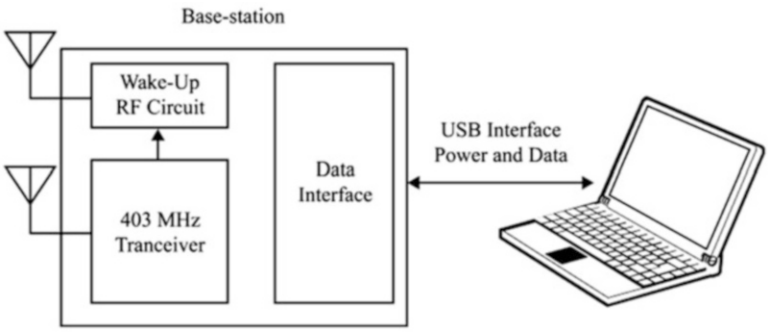


Fig. 4.18 Base station and PC

4.12 Power Considerations

Implants are often designed to consume minimal battery power to extend their useful operating life. For example, a pacemaker may be expected to operate for up to 10 years. Adding a radio link to an implant will cause an additional battery drain that needs to be minimised. As a simple rule of thumb, current demand will increase with frequency, transmitter power, receiver sensitivity and processing power. The receiver must be sensitive enough to detect the incoming signal with an acceptable error rate. A high error rate may be corrected by resending the data and error correction, but this increases power consumption. The transmitter must also produce enough power for the base station to receive with a low raw error rate.

Leaving a receiver on to permanently listen for the base station transmission would require a current in the order of 2–3 mA – an inappropriate power depletion of the battery. The implant needs to detect the base station transmission and start a data exchange session at short notice, on demand, all the while draining minimal battery current. One way to reduce the average current is to switch on a simple receiver for a short time at regular intervals. This is known as the "wake-up" receiver.

The wake-up receiver can use either the MICS band, or the ISM band where more radiated power is permitted. A typical example would be 2.45 GHz wherein excess of 100 mW may be radiated (country dependant). The losses through the body will be greater than for the MICS band, but additional power will compensate and the antenna will be closer to the desired size. It is possible to use the same implant antenna for both MICS and 2.45 GHz with care in the matching networks. The wake-up receiver will be switched on at regular intervals for a short period, known as strobe mode. If a signal is detected, the wake-up receiver will switch on and detect if it is a genuine "wake-up" signal. This is a digital code designed to wake up the implant. If this code is not detected the receiver will revert to strobe mode.

Once the code is verified then the remainder of the implant communication system, which normally includes the crystal oscillator, the *Media Access Controller* (MAC) that controls the operation of the part, and the phase lock loop, will power-up. Once the wake-up is complete an acknowledgment is transmitted. An example of a wake-up sequence and power consumption is shown in Fig. 4.19. Once the acknowledgment is received a data transfer session can begin.

When a data transfer session has finished, the part reverts to sleep mode with the wake receiver strobing.

4.13 Miniaturised Construction

The space within the body is very limited so every effort must be made to miniaturise the electronic assembly. Considerable space savings can be made by using unpackaged silicon die rather than packaged parts. Connection to the

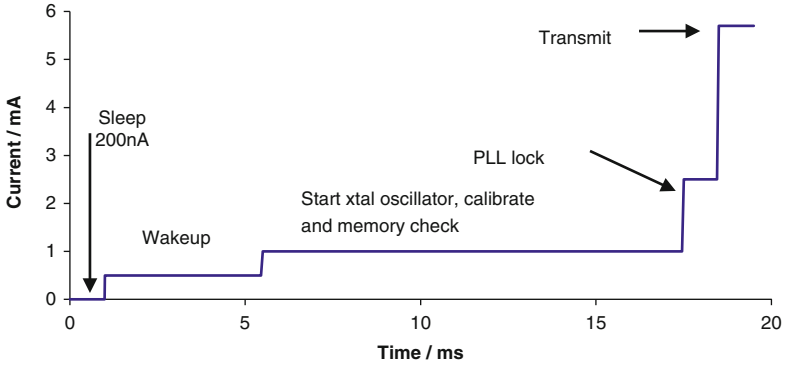


Fig. 4.19 Wake-up sequence and current consumption for a typical implant RF transceiver (Source: Microsemi (previously Zarlink) Semiconductor)

substrate can be made by wire bonding or a flip chip technique. Passive components can take up a significant area on a supporting substrate even if the smallest of packages are used. There are ways to integrate passive components into a silicon substrate, as shown by the PPM2 programme [15] for example. Using a silicon substrate it is possible to integrate passive components, add tracking and have connections from the back to the front of the material (through silicon vias) – this is important when miniaturising the interconnections. There are several commercial organisations that offer this type of service.

4.13.1 Battery Challenges

In many applications, an ideal battery is the one that gives a constant voltage for as long as possible and the user can change the battery when the device stops working. In the case of implanted medical devices, this is obviously not possible.

The lithium-iodine cell, most commonly used in pacemakers, has a very different behaviour. The battery can be modelled as a voltage of about 2.8 V in series with a resistor. The series resistor has a value of about 500 Ω at the beginning of the battery life, and increases slowly to end up at 10–20 kΩ towards the end of the battery life. Assuming a constant average current drain, the resulting battery voltage for the pacemaker electronics starts off at 2.8 V and then gradually decreases with time towards 2.0 V, when the pacemaker battery should be replaced. It is then quite easy to measure the internal resistance of the battery, and the doctor and patient can be alerted 12 months before the battery needs to be replaced.

Figure 4.20 shows a voltage versus time comparison for a typical watch battery and a lithium-iodine battery used in a pacemaker. Though the patient and the doctor benefit from this battery behaviour, it is easy to see the challenges this poses for the designer of an implant system. On top of designing electronics that demand

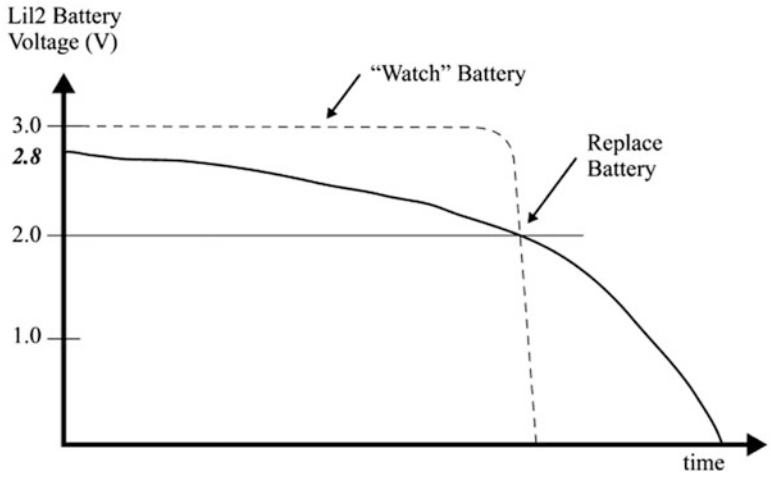


Fig. 4.20 Voltage versus time comparison for a typical watch battery and a Lil2 battery used in a pacemaker

extremely little current, the designer must also cope with a voltage variation over a long operating life of the device.

The other problem for the designer is the presence of a 500 Ω resistor in series with the voltage source. If the transceiver draws 5 mA during transmit the voltage drop across the resistor will be 5 mA × 500 Ω = 2.5 V which is almost all of the battery voltage. The power source during a data transfer session must be primarily from a capacitor.

4.14 Defibrillation Pulse and X-rays

An implant within the chest cavity may need to survive a defibrillation pulse. As well as the external defibrillator, *Implantable Cardioverter-Defibrillators* (ICDs) that deliver a pulse directly to the heart are becoming more common. The pulse can be biphasic with a peak of 800 V applied to the heart, and last several milliseconds. A pacemaker, internal heart monitor, ICD or other chest cavity implant, along with an antenna and delicate transmitter and receiver electronics, will need to survive the pulse.

Care must be taken in the design of the matching network to reduce the energy reaching the electronics to within its capability using electrostatic damage protection diodes. Additional protection diodes should be used with care as they will add capacitance and will be part of the matching network. The protection needs to be designed with the knowledge of the transmitter/receiver electronics capability and the expected defibrillation pulse amplitude and duration.

More details can be found in an online article by the American Heart Association [16].

If a patient has an X-ray examination that will include an implant, then care must be taken to ensure its survival. Radiation can result in temporary loss of function or even long-term damage [17].

4.15 Link Budget

The link budget determines if the link will work by taking into account transmit and receive powers, antenna gains, path losses and receiver sensitivity. The signal-to-noise ratio will determine the un-corrected bit error rate for a given range, i.e.,

$$S/N(dB) = \mathrm{P}_t(dBW) + 204(dBW) - 10\log(B) + G_r(dBi) - P(dB) + G_t(dBi) \tag{4.31}$$

where $P_t(dBW)$ is the transmit power in dB Watts; $204(dBW)$ is thermal noise power for a 1 Hz bandwidth; B is the bandwidth in Hz; $G_r(dBi)$ is the receive antenna gain in dB; P is the path loss in dB, this includes the free space (P_f) and body (P_b) losses, where $P = P_f + P_b$, that can be considerable and G_t is the transmitter antenna gain in dB.

The path loss is comprised of losses through the body, which can be in the order of 20 dB, and the free space loss is:

$$P_f = (\lambda/4\pi d)^2 \tag{4.32}$$

There may also be losses from multi-path propagation causing fading. From the above it is clear that with a low upper limit on transmit power, significant body losses and a low implant antenna gain, low noise design of the receiver is critical. Any energy loss through the matching network will adversely affect P_t so it is essential that this continues to be optimised. The reliability of the link will be improved with the addition of error correction and with re-transmission of data that is in error.

Communication between an implant and a body worn transceiver will not have the free space loss or multipath problems, however, the signal that will emerge from the body will usually be the shortest route, so the positioning of a body worn device is important.

4.16 Electro-stimulation: A Non-MICS Example

This is an example of a one-way wireless link that is used to move a capsule through the digestive system. This does not use the MICS band as there is no requirement to receive data back from inside of the body.

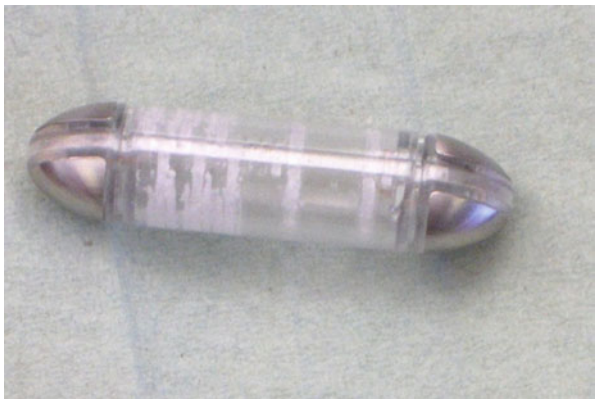


Fig. 4.21 Capsule and electrodes

Electro-stimulation of nerve or muscle is a potential use for an implanted device. Stimulation of a nerve requires a signal of 1 mA, or less, but the stimulation of a muscle may require the order of 10 mA. An area of interest is the digestive tract where endoscopes provide useful information of the upper and lower extremities and the camera pill provides images all the way through. Today's generation of camera capsules contain tiny cameras, a light source and an ultra-low power radio frequency (RF) transmitter which transmits images to an external data recorder as the vitamin-sized capsule passes naturally through the patient's oesophagus and small bowel. The patient then returns the data receiver to the clinician, who uses expert software to detect abnormalities.

The camera capsule is proving to be a very useful tool to diagnose disorders such as obscure gastrointestinal bleeding (OGIB), Crohn's disease, coeliac disease, tumours and a range of other small bowel disorders. Nevertheless, current capsule technology relies on the patient's natural body movements to propel the capsule, meaning there is no opportunity to return to investigate areas of interest. Work carried out by Imperial College and Zarlink (now Microsemi) as part of the European Union funded NEMO project investigated ways to stimulate the oesophagus and small bowel to contract and move a capsule under wireless control.

The oesophagus and small bowel will force a packet of food (or other object) through its length with peristaltic waves that work only in one direction. The objective of the NEMO programme is to use the body's own muscle to steer a capsule, without requiring external moving parts. By electrically stimulating the wall of the oesophagus and small bowel, contractions can force the smooth capsule to move or remain stationary.

A capsule was fabricated with a body of smooth, non-toxic plastic with a pair of stainless steel electrodes at each end, Fig. 4.21. The resistance of the oesophagus or small bowel between a pair of electrodes is in the order of 2–10 kΩ. It was found that a current of 5–7 mA was required to cause contraction and thereby propel the capsule. If the current was too high the oesophagus would contract around the whole capsule preventing movement. The current was applied in pulses of the

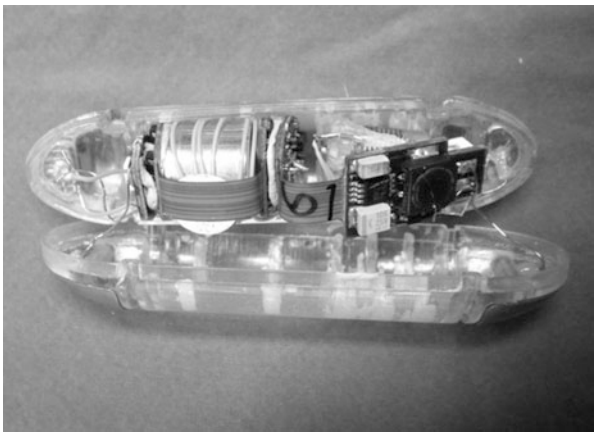


Fig. 4.22 Capsule antenna and electronics

same polarity with no reversal. This current required an open circuit voltage of 20 V to be generated. Circuitry was designed to maintain a constant current (not constant voltage) pulse train.

The capsule electronics contained a battery, DC-DC converter, a constant current pulse generator, an RF receiver, signal decoder and antenna. All of these had to fit into a capsule that would then fit into the oesophagus. The DC-DC converter and battery presented challenges as the excess drain from the battery would cause the receiver to stop working. The battery chosen was a Duracell DL1/3 N with care taken to limit the peak drain.

The antenna for this consisted of a coil of wire wrapped round the internal PCB; there was also a ferrite core inductor as part of the receiver circuit that also acted to detect the RF signal. This enabled a capsule to be steered, by electro stimulation and under RF control, within the oesophagus and small bowel of a live patient. Figure 4.22 shows an example of the capsule. The antenna was much smaller than optimum but operated with a short distance to the transmitter.

The wireless link operated at 28 MHz as components were readily available and there was no requirement to transmit data from the capsule at that stage. The antenna was a loop of wire and a ferrite core inductor. The circuit was built on flex-rigid PCB using discrete components and is shown in Fig. 4.22. The range from the external transmitter to the capsule was typically less than 2 m.

The capsule was successfully controlled within an oesophagus and small bowel producing both forward and reverse movement.

It was observed that the anaesthetic regime affects movement. If atropine is used then movement will improve 1 h after administration. The depth of anaesthetic using isoflurane gas made no detectable difference. Movement was unaffected by the depth of anaesthetic using propofol.

Electro-stimulation can be applied to many body functions for therapeutic and diagnostic purposes and may need to feedback data. To be of use electro-stimulation circuitry needs to be integrated and optimised for a specific purpose. The generation of a sufficient DC voltage is a far from trivial task.

This programme showed not only the potential for wireless control of electro-stimulation but also that it can be archived with what would normally be considered a very small antenna.

4.17 Conclusions

The implant antenna is critical to the operation of the data link, and must be designed as part of the implant to make the best use of the available area. There are several antenna options depending on the given application. Testing to determine antenna characteristics is important to ensure the matching network can be effectively designed. Care should be taken when measuring the impedance of electrically small antennas.

In comparison to strict size limitations on implanted devices, designers should take advantage of the additional space afforded by external base stations for antennas, electronics and filtering. Multiple antennas can be used if there is a polarisation or multi-path fading problem. A very low noise receiver is needed in both the base station and the implant. Error correction can enhance data transfer, but at the cost of longer power-up time and battery drain. The battery will typically not be able to source the peak current used during data transfer, so a large value capacitor will be needed.

Along with wireless performance and power issues, designers must also address a multitude of biocompatibility concerns and regulations governing the design of implanted devices and in-body communication systems. The integration of a high data rate transceiver will enhance the operation and capabilities of existing implanted medical devices, and open the door for new techniques that will lead to improved treatments and better quality of life for patients.

References

1. Finkenzeller K. RFID handbook, 2nd ed. John Wiley and Sons, 2003.
2. ISM band. Federal Communications Commission, Title 47 Volume 1; Telecommunications Section, Part 18; Industrial Scientific and Medical Equipment, Sub Part C, Section 18.305
3. ICS band. Australian Communications Agency Paper, SP 6/03, 2003.
4. Johansson AJ. Wave propagation from medical implants – influence of body shape on propagation pattern. Department of Electroscience, Lund University, Sweden, http://www.tde.lth.se/home/ajn/publications/Houston.pdf
5. ZL70101 http://www.zarlink.com/zarlink/hs/82_ZL70101.htm
6. Kraus JD. Antennas, 2nd ed. McGraw Hill, 1988
7. Krall AD, McCorkle JM, Scarzello JF, Syeles AM. The omni microstrip antenna: a new small antenna. IEEE Transactions on Antennas and Propagation 1979; AP27: 850–853
8. Fujimoto K, Henderson A, Hirasawa K, James JR. Small antennas. Baldok, Hertfordshire: Research Studies Press, 1987
9. Lee FL, Wei C. Microstrip and printed antennas, 1st ed. John Wiley and Sons, 1997

10. Straw RD, ARRL antenna book, Newington, Connecticut: American Radio Relay League.
11. Wojclk J. Tissue recipe calibration requirements. Spectrum Sciences Institute RF Dosimetry Research Board, Nepean, Ontario, Canada, SSI/DRB-TP-D01-003.
12. Yuce M and Khan J Wireless Body Area Networks, 2012; 415–428
13. Yuce M and Khan J Wireless Body Area Networks, Pan Stanford publishing, 2012, pp 323–344
14. Rostami S. Polymer performance – in a tough world. Bulletin – The Technical Journal for Industrial Members of the Welding Institute 2004; 14–16
15. http://www.ppm2project.co.uk/outline.html
16. Manoharan G, Evans N, Allen D, Anderson J, Adgey J. Comparing the efficacy and safety of a novel monophasic waveform delivered by the passive implantable atrial defibrillator with biphasic waveforms in cardioversion of atrial fibrillation. Circulation 2004; 109: 1686–1692
17. Prutchi D, Prince J and Stotts L. X- and Gamma-Ray Hardness of Floating Gate EEPROM Technology As Applied to Medical Implanted Devices. IEE Trans on Components and packaging Technology, Vol 22, No 3, September 1999.

Chapter 5
Network Topologies, Communication Protocols, and Standards

Javier Espina, Thomas Falck, Athanasia Panousopoulou, Lars Schmitt, Oliver Mülhens, and Guang-Zhong Yang

5.1 Network Topologies

Every network has a topology that determines the way in which different devices of the network are arranged and how they communicate with each other. Here we need to distinguish between physical and logical topologies. The former refers to the physical layout of the network, i.e., the way that devices are physically connected to the network, either through actual cables or direct wireless communication links. By contrast, the logical topology of a network refers to the manner that data flows through the network from one node to the other without worrying about the physical interconnection of the devices for transporting a packet from a source to a destination device. The two lower layers of the *Open Systems Interconnection* (OSI) reference model [1], the physical and data link layer, define the physical topology of a network, while the network layer is responsible for the logical topology.

Table 5.1 provides an overview of the most common topologies applicable to wireless sensor networks. Each topology has its advantages and disadvantages regarding network characteristics such as latency, robustness, capacity and the complexity of data routing and processing as shown in Table 5.2. The star-mesh hybrid topology combines the advantages of the star topology with those of the mesh topology, providing the highest degree of mobility for star clusters.

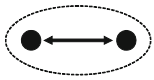
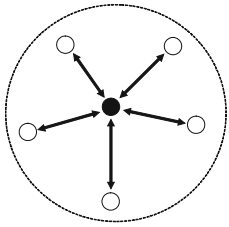
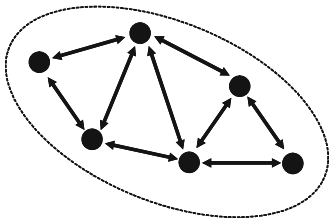
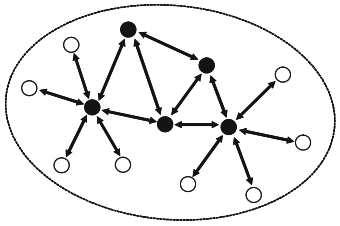
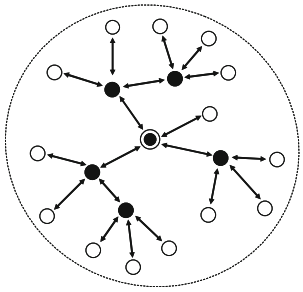
J. Espina • T. Falck (✉) • L. Schmitt • O. Mülhens
Philips Group Innovation, 5656 AE Eindhoven, The Netherlands
e-mail: thomas.falck@philips.com

A. Panousopoulou
The Hamlyn Centre, Imperial College London, London SW7 2AZ, UK

G.-Z. Yang, PhD
The Hamlyn Centre, Imperial College London, London SW7 2BY, UK
e-mail: g.z.yang@imperial.ac.uk

G.-Z. Yang (ed.), *Body Sensor Networks*, DOI 10.1007/978-1-4471-6374-9_5,

Table 5.1 Common network topologies used in sensor networks

Point-to-point network The simplest topology consists of only two devices directly connected with each other	
Star network All devices are connected to a single central controller often referred to as the coordinator or master. The peripheral nodes are called slaves. Slaves can only communicate with the master. Communication between slaves requires passing all data through the master	
Mesh network Any device can communicate with other devices as long as they are in range of one another ("peer-to-peer network"). Multi-hop networking protocols enable routing of packets from one device to the other on the network	
Star-mesh hybrid network This allows connecting a mesh network with one or more star networks or several star networks with each other. A mixed star and mesh network combines the simplicity of the single-hop star topology with the extendibility and flexibility of the multi-hop mesh topology	
Cluster tree network The cluster tree topology is a special case of a multi-hop mesh network where there is always only a single path between two devices. The first device starting the network becomes the root of the tree. Another device can join the network as "child" of the root node and in turn allow other devices to join the network through that device. Devices are aware of their "parent" node and any "child" nodes. This hierarchical topology reduces routing complexity	

In practice, the topology is a choice of application design. The application developers have to balance sensor node costs, battery drain, complexity of routing, robustness, scalability, latency, mobility, and spatial coverage to meet the characteristics and performance requirements of their applications. In the following sections, we will present some typical application scenarios of BSNs to illustrate the usage of the different network topologies.

Table 5.2 Advantages and disadvantages of different topologies used in sensor networks

Topology	Advantages	Disadvantages
Star	Simplicity Simple and cheap slave nodes Low power consumption of slave nodes Low latency and high bandwidth Centralised systems	Dedicated central node Limited spatial coverage Single point of failure Poor scalability, small number of nodes Asymmetric power consumption (master consumes much more energy than slaves) Inefficient slave-to-slave communication Distributed processing
Mesh	Distributed processing Peer-to-peer communication Very fault tolerant Scalable, many nodes possible Large spatial coverage Low/medium complexity Energy consumption can be balanced among nodes	Nodes used must have same basic functionality, including routing capabilities (may be an overkill in some applications as it increases cost) Complexity of routing High latency and low bandwidth
Star-mesh hybrid	Low/medium complexity (if nodes can be classified as slaves or masters before deployment) Large spatial coverage Low latency and high bandwidth between master and its slaves Good for local actuation or data aggregation High reliability possible Scalable, many nodes possible Power consumption can be balanced among masters and it is asymmetrical between master and slaves Nodes acting as slaves can be relatively inexpensive	High complexity (if all nodes can act as masters) High latency and low bandwidth for multi-hop communication Power consumption is asymmetrical between master and slaves
Cluster tree	Low power consumption of leaf nodes Large spatial coverage area Many nodes possible Large spatial coverage Medium complexity (rerouting is required when a node in the tree dies)	Medium scalability (root of the tree is a bottleneck) Low reliability (node failure affects routing) High latency and low bandwidth Asymmetric power consumption (nodes in the tree backbone consume more power) Nodes used must have same basic functionality, including routing capabilities (may be an overkill in some applications)

5.2 Body Sensor Network Application Scenarios

As mentioned in Chap. 1, wireless sensor networks represent an enabling technology for physiological monitoring and real-time intervention. This has direct applications to continuous, cable-free monitoring for medical purposes (Fig. 5.1). Typical examples include the monitoring of vital signs in intensive care units [3]; remote monitoring of chronically ill patients [4–8]; monitoring of patients in mass casualty situations [9]; monitoring people in their everyday lives to provide early detection and intervention for various types of disease [10]; computer-assisted physical rehabilitation in ambulatory settings [11]; assisted living for the elderly at home [12, 13]. In addition, low-power BSN can also empower unobtrusive human-machine interfaces, resulting smart fabrics and natural interfacing devices for video gaming [14]. In all examined scenarios, sensors range from wearable and implantable to ambient sensors such as positioning devices. Additionally, depending on the application scenarios, BSNs are employed either in a stand-alone context or in combination with pre-existing infrastructure [15].

5.2.1 Stand-Alone Body Sensor Networks

A stand-alone body sensor network consists of small wireless nodes positioned on or inside the patient's body, conjointly providing the functionality for sensing and processing required by the application. In the simplest scenario, a central node gathers and records the readings of the biosensors such as ECG, EMG, EEG, SpO_2, blood flow, and blood pressure over a period of time for subsequent interpretation and trend analysis. The data can be enriched with contextual information by attaching further sensors to the body, such as accelerometers. By providing capabilities for local processing of the measurements and user I/O, the patient is alerted in a timely manner when his/her state of health changes for the worse. Both star- and mesh-based topologies are applicable to this class of application. The star topology implies a centralised architecture where the intelligence of the system is concentrated on a central node. This central node is superior to the peripheral sensors, in terms of resources such as processing, memory, and power. A star network is a common choice when simplicity needs to be combined with high bandwidth. For instance, the UbiMon project [7] takes this approach and utilises a *Personal Digital Assistant* (PDA) as the local processing unit for collecting, displaying and analysing the sensor signals. It is advantageous in situations where a PDA is an inherent part of the system and direct communication between sensors is not required.

Utilising a star topology does not necessarily imply that the central node responsible for the data collection and analysis is a physical part of the stand-alone BSN. The project Smart AtTIRE [16], for example, explores the idea of an unattended BSN, by providing activity monitoring services in an unobtrusive manner.

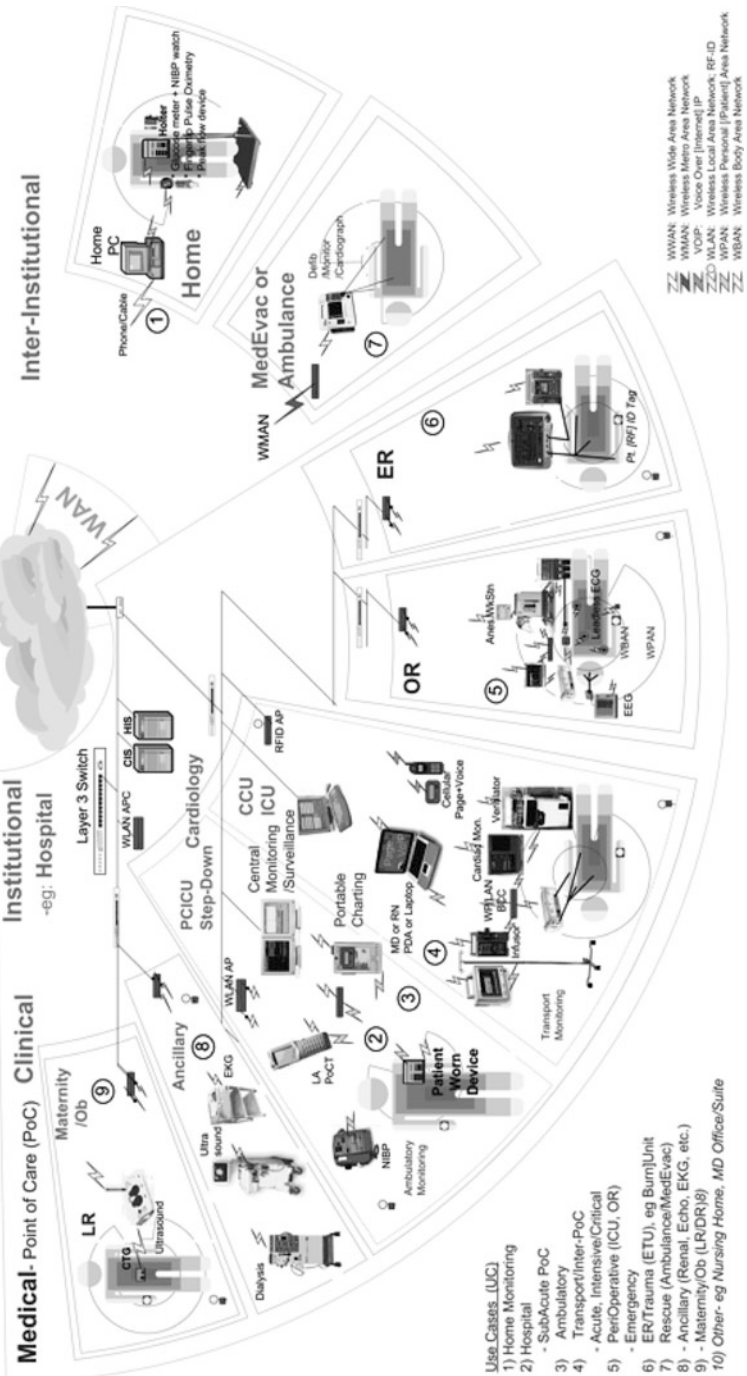


Fig. 5.1 The range of BSN-related applications (Adapted from Ref. [2])

On-body nodes are embedded in garments and locally process the sampled data for detecting activity events. These events are compressed and locally stored. When the BSN is within the range of the central processing unit, a star topology between the on-body nodes and the central unit is formed and the compressed data is automatically uploaded to the latter for further analysis.

While the above approaches respond sufficiently in terms of short term monitoring, a star topology can also benefit from the intelligence of the master node/coordinator of the network for organising an energy-efficient data collection from the slave nodes. This idea is used, for example, in the MERCURY software platform [17] for responding to the needs of long-term physiological monitoring. Specifically, while individual nodes perform sampling and local processing on the raw signals, the base station performs data collection and tuning of the BSN parameters based on the availability of energy resources and radio link quality.

In contrast, a peer-to-peer network is capable of shifting the intelligence towards the sensors, allowing self-organisation and fault-tolerance across the network. The resulting self-contained topologies are useful when the on-the-fly combination of the physiological signals is needed. A characteristic example is the BASUMA project [4]. In particular, the adopted mesh topology serves as a means of realising a non-invasive continuous cuff-less blood pressure sensor. This is achieved by combining the signals from at least one distal pulse wave sensor and a single lead ECG, for indicating the linear relationship between the pulse wave velocity and the blood pressure [18].

The combination of star and mesh topology into hybrid networked schemes allows scalability and access to the same source of information from multiple decision points. This can be advantageous in pre-hospital and in-hospital emergency care, as explored, for example, in the Codeblue platform [19]. More specifically, by adopting a mesh topology, stand-alone BSNs are deployed during emergency and publish vital signs, location and activity. The medical personnel can subscribe to the data of interest. The operational on-body devices cooperate with dynamically routed data from publishers to subscribers, thereby allowing a mesh network to connect to multiple star networks [20].

The discussion so far highlights the fact that the topology of a stand-alone BSN should respond primarily to the needs and requirements of each application. Despite the plethora of related research and applied platforms, there remains an interesting challenge to be considered with regards to wireless body sensor networks. If there are no physical wires connecting the sensors into the network, how is a BSN to be set up by the user? One proposed solution [21] is to equip all sensors with an IR-receiver and to use a setup pen, which emits a unique identifier via IR to limit the scope to a single patient. All sensors receiving the same identifier form a network. The *Active Digital Aura* (ADA) [22] technology uses a similar approach, whereby a tag is worn on the body that capacitively couples a low-frequency RF signal into the body. By modulating this signal with a unique identifier, only sensors attached to the body can pick up the specific tag and form a network.

While standalone BSNs provide immediate access to physiological signals, they are often integrated into smart environments, thereby improving the contextualisation

of the acquired information. In the following sections, representative application scenarios are described.

5.2.2 *Pervasive Sensor Networks*

The potential of BSN alights with the context behind the continuous monitoring of the physiological signals. The combination of wearable and ambient sensing into truly pervasive networks defines a roadmap for the efficient contextualisation of the raw signals. A representative example is the MoteTrack platform [23], which realises decentralised RF-based localisation services based on an ambient radio beacon infrastructure. The location of a mobile wireless node is computed using a received radio signal strength signature from numerous beacon nodes to a database of reference signatures.

MoteTrack has been used for on-demand and instantaneous deployment, with the objective to facilitate first emergency response. Nevertheless, the integration of ambient sensing and physiological signals monitoring is typically met in assistive living environments. Ambient health sensors, responsible for measuring medical-relevant parameters in post-hospital care scenarios, have already been deployed [24–26]. The main characteristic of ambient health sensors is that they can be seamlessly integrated with the living environment, e.g., a chair, a bed, or a personal health area. As such, they can continuously monitor medical conditions and wellbeing, allowing, for instance, the determination of the heart rate of a person sitting in a chair without the need for additional body-worn devices.

The main characteristic of pervasive sensor networks is heterogeneity in terms of topologies, technology and services provided. For instance, in contrast to bio-sensors, ambient sensors are traditionally part of a stationary network (*Ambient Sensor Network*, ASN). Considering the case where direct links can be established between members of the BSN and members of the ASN, the BSN can use the latter as a gateway towards sophisticated decision making systems (as shown in Fig. 5.2). Typical related deployments are based on service discovery and association of a single node to a pre-existing network. When the combined BSN – ASN topology changes, due to the user's mobility, the former has to repeat this discovery and association procedure. This allows an always-on connectivity of the BSN to the ambient network.

It is obvious that forcing all members of a BSN to discover and associate with an ASN would be very inefficient. An alternative is the construction of hybrid star-mesh topologies between the ANS and the BSN. In this case, the members of the BSN form a star topology, by allowing a single sensor to act as a bridge to the ambient sensor network. Furthermore, the bridge sensor alone is required to perform these association procedures when the BSN is in motion. The BSNs deployed on the left part of Fig. 5.2 illustrate this approach for a BSN with a tree or mesh topology. It should be noted that in a mesh-based BSN, one of the

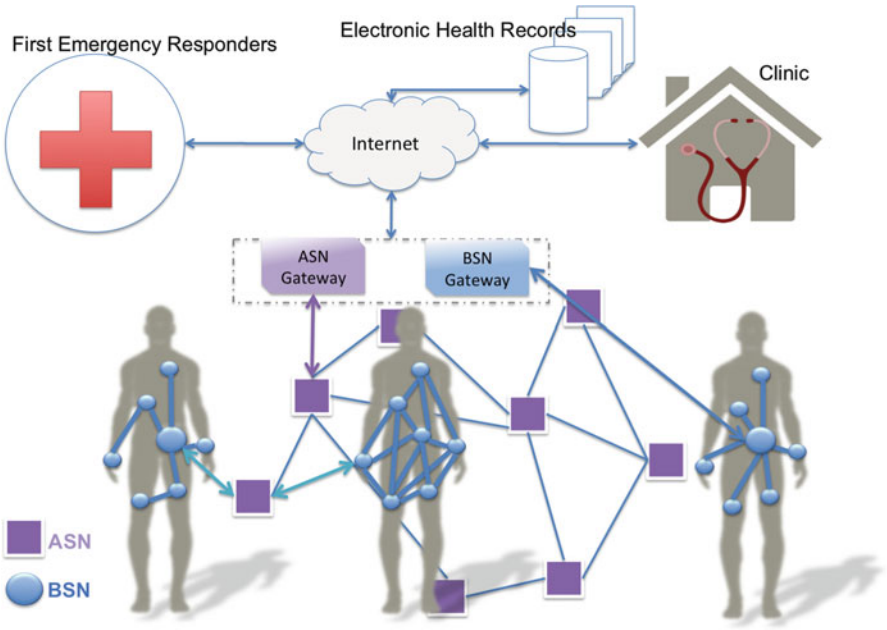


Fig. 5.2 The integration of BSN (in tree, mesh and star topologies) with ASN for healthcare applications

sensors has to be selected to take the role of acting as the bridge to the ambient sensor network [27].

In the general case where ASN and BSN are not compatible, a service-oriented approach should be employed, as it would allow interoperation of different computing devices. In this setup, BSN and ASN have different gateways, while they both need to respond to the same backend infrastructure for the data fusion (right part of Fig. 5.2). As such, a service oriented model needs to be adopted, providing an interoperable mechanism for accessing information from both types of networks. Therefore, a service-oriented architecture for heterogeneous ASN and BSN should incorporate the following resource-aware components [28]:

- *Service Identifier* – which contains information necessary to locate and address a specific service as well as the information required to retrieve further information about a service.
- *Service Interface* – which describes the functional aspects of a service, i.e., it specifies the syntax and order of messages in a service, and the involved data types.
- *Service Binding* – which specifies how the messages of a service are mapped to a concrete communication protocol. Service binding helps to deal with heterogeneity. An example is switching between web-services when communicating with a backend infrastructure and lightweight binding for resource-constrained sensor nodes.

- *Service Characteristics* – which define the core properties of a service and facilitate resource optimisations.
- *Service Constraints* – which restrict the service composition process, by specifying under what conditions services can be used. For example, a healthcare service that sends a patient's heart rate data over a network can specify that the data sent must be encrypted using a predefined algorithm.

5.2.3 Global Healthcare Connectivity

Connected healthcare often requires the integration of standalone BSN and smart environments with larger-scale networked platforms. The objective of this integration is twofold: (a) to expand the access to the information beyond the constraints imposed by the physical range; (b) to promote interoperability in the usage of BSN and integrate them with existing clinical information systems.

In a typical deployment, a health-oriented platform can rely on UMTS/GPRS/3G technology for interconnecting standalone or pervasive BSN with specialised care providers. Representative examples of this architecture include MobiHealth [5], its successor HealthService24 [6] and the MobiCare [29] projects, where a mobile phone serves as the mobile base-station for body-worn sensors. This mobile base-station provides three services: (a) the collection and storage of the received data (data repository); (b) the forwarding of data to a doctor or medical centre (streaming service); (c) the analysis of the data received and the sending of an alarm or a reminder signal to a predefined destination using SMS (feedback service). In contrast to standalone and pervasive BSN architectures, the UMTS/GPRS-based architectures do not allow local processing. In particular, instead of passing only relevant data or alerts to a doctor or a medical centre when detecting a critical event, all sensor readings are piped to a remote data centre where the processing takes place.

Utilising UMTS/GPRS technology as the mediator between BSN and hospital networks has been the traditional approach for realising the vision of connected healthcare. Although this structure expands the functionality of BSN beyond hospitals and specialised healthcare centres, it may limit the activities of the individual, or cause changes in his/her behaviour. Furthermore, the centralised access point may generate single point failures, caused for instance by battery depletion of the mobile gateway. With these considerations in mind, the MEDiSN platform [30] concentrates on a dedicated wireless backbone that is deployed in hospitals for supporting mobile physiological monitoring. In this deployment, the relay backbone network is reconfigured in an ad hoc manner. This allows the on-demand construction of tree topologies between the BSN and specialised clinical information systems. To this end, the routing engine of the Collection Tree Protocol is adapted [31] to satisfy the needs of the components of the wireless backbone. In addition, the mobile stand-alone BSN can subscribe to different relay nodes, according to link quality and quality of service criteria. The resulting

topology in a hospital environment allows for the utilising the same backbone for relaying information from multiple BSNs, in a pervasive and unobtrusive manner.

The combination of BSN with the pre-existing, or rapidly configured network backbone can also be applied to emergency response scenarios. This requires an integrated network platform that combines short-range communications with wireless local area networks and mobile Internet services. The applicability of this architecture has been explored within the framework of AID-N project [32]. AID-N project addresses the development of an advanced collaboration-oriented communication system for coordinating the activities of first line public health and emergency care providers during mass casualty situations. Therefore, the hybrid star-mesh BSN architecture by CodeBlue [19] has been employed and magnified by the connectivity to the emergency response information centre, responsible for coordination actions between pre-, in- and post-hospital care and decision making.

The aforementioned projects highlight the potential of body-centric networked systems for improving healthcare. The related enhancements are applicable to both personalised healthcare systems, which allow cost-effective chronic disease monitoring and rehabilitation, and first-emergency response systems.

5.3 The Standardisation of Wireless Personal and Body Area Networking

The advent of smart wireless sensors that are able to form a BSN would not be possible without the realisation of inexpensive low-power transceivers that allow short range communication and data transmission at low to moderate rates. These are capable of real-time data transmission, within a range of a few meters. In order to achieve cost-effective, flexible and interoperable solutions, it is necessary to avoid proprietary technological approaches, which may be superior for specific applications but may not be optimised in terms of inter-operability. By means of standards, the short-range wireless communications market has a far better chance to proliferate quickly and hence, costs will be driven down at the same time as the product and feature ranges increase.

Current standardisation efforts focus on two primary engineering challenges for interoperable wireless communications: (a) the formation of national and international regulations for the allocation of the electromagnetic spectrum; (b) the definition of the functionalities and interfaces of the layers in the OSI protocol stack [1]. For the latter, the standardisation activities emphasise on the *Physical* (PHY) and the *Medium Access Control* (MAC) layer, whilst defining the interfaces for higher layers, such as the network and application layers.

In the following subsections, we present the regulations on the electromagnetic spectrum allocations for BSN devices, followed by an overview of the dominating standards for short-range connectivity and interoperability to existing infrastructure.

5.3.1 The Wireless Regulatory Environment

The electromagnetic spectrum is regulated by authorized national administrations or agencies, such as Ofcom in the UK [33] and the *Federal Communication Commission* (FCC) in the USA [34]. In addition, international organisations coordinate regulations at supranational level, such as the *Electronic Communications Committee* (ECC) of the *European Conference of Postal and Telecommunications Administrations* (CEPT), which brings together the national administrations of 48 European countries. On a global level regulations are defined by the Radio Regulations of the *International Telecommunication Union* (ITU) [35].

International harmonisation of spectrum regulation enables the reusing of technologies across countries and regions, thereby facilitating their commercial viability. Many body sensor networks use internationally available unlicensed bands for application-generic wireless communication. Although the most popular band is centred at the 2.4 GHz (*Industrial-Scientific-Medical*, ISM) more options are available. More specifically:

- *Ultra-High-Frequency* (UHF) is internationally available (e.g., in 315, 433, 868–928 MHz) but often not in the same frequencies in all regions. This is considered a disadvantage for a widespread deployment. Moreover, some regions impose strict duty cycle restrictions, such as in the 868 MHz band in Europe [36]. Nonetheless, low power systems transmitting small amounts of data may benefit from the good propagation characteristics of the band, when compared to those offered by transmission in higher frequency bands.
- The 2.4 GHz ISM band is available worldwide. Systems operating in the band—unlike in UHF bands—enjoy fewer restrictions on duty cycle or channel spacing. Furthermore, the band offers a significant portion of the frequency spectrum: from 2,400 to 2,483.5 MHz in the USA and Europe, and from 2,471 to 2,497 MHz in Japan. These advantages have made the 2.4 GHz ISM band the most popular for BSN-related applications, and low power radio systems in general. A proof thereof, is its utilisation by widespread technologies such as Bluetooth®, WiFi® (based on the IEEE 802.11b/g/n standards [37]), and ZigBee® (based on the IEEE 802.15.4 standard [38]).
- The 5.8 GHz ISM band is also available worldwide and has similar characteristics to the 2.4 GHz ISM band. Its most common users are WiFi® systems based on the IEEE 802.11a standard [39]. However, it is not preferred for BSN-related applications, due to the poor propagation characteristics in proximity of the human body.
- *Ultra-wideband* (UWB) systems operate at very low radiated power density levels over a large channel bandwidth. Many countries and regions allow UWB operation in parts or the totality of the 3.1–10.6 GHz spectrum band. UWB technology guarantees high communication data rates, low power operation, and radar and ranging capabilities. However, the slow progress in the standardisation process, the cost of initial implementations and the performance, which is characterised significantly lower than initially expected, have undermined the

success of UWB radios [40]. Nevertheless, as explained below, the most recent standardisation efforts invest on UWB, thus highlighting the fact that it also has appealing characteristics for BSN-related applications.

In addition to unlicensed, worldwide-available bands for application-generic communication, BSN can also use the medical-specific spectrum. Implantable medical devices using inductive communication operate worldwide on spectrum bands in the lower part of the frequency spectrum, typically ranging from a few kHz to less than 40 MHz (as mentioned in Chap. 4). Most non-inductive medical implants, such as cardiac pacemakers and defibrillators, operate worldwide without individual spectrum license in the 401–406 MHz spectrum range. Such operations are known as *Ultra Low Power Active Medical Implants* (ULP-AMI) in Europe [36, 41], and *Medical Implant Communications Service* (MICS) in the USA [42] and Japan. More specifically, the original MICS regulation in the USA has been superseded by the Medical Device Radiocommunications Service (MedRadio) [43], which includes the original MICS spectrum at 402–405 MHz and adds adjacent spectrum sub-bands. MedRadio also includes 24 MHz of spectrum for *Medical Micro-Power Networks* (MMNs).

Unfortunately, no spectrum bands are available globally for non-implantable body sensors. However, two regional spectrum regulations are relevant enough to be singled out: *Medical Body Area Networks* (MBAN) and *Wireless Medical Telemetry Service* (WMTS). Both spectrum bands enable wireless diagnosis and treatment of patients. They do not require individual licenses for operation, but often require registration of the hospital site that uses them.

- MBAN consists of on-body wireless sensors and/or actuators that communicate with a monitoring device placed on/around the human body. Such monitoring devices display and process vital sign parameters from MBAN devices and may also forward them, to a nurse station for instance. This is achieved by using other wired or wireless technologies, since MBAN can only be used for short-range communication. MBAN regulation is currently only local. However, it is a very recent regulation that is gaining momentum and has the potential to become worldwide available. The FCC published their MBAN regulation in May 2012, allowing the deployment of MBAN in the 2,360–2,400 MHz spectrum band [44]. Operation in the 2,360–2,390 MHz sub-band is restricted to registered healthcare facilities and requires the use of technical means, known as Control Messages, to ensure that operation outside healthcare facilities is suppressed. Operation in the 2,390–2,400 MHz sub-band has no location restrictions and can therefore also be used in patient homes and ambulances. At the time of writing this edition, ECC and ETSI are in the process of regulating and standardising MBAN in Europe [45]. ECC is considering some spectrum bands within the 2,360–2,500 MHz range for future MBAN operation.
- Similarly to MBAN, WMTS permits bidirectional data communication related to medical care. WMTS is restricted inside hospitals and enables longer-range communications (telemetry), at the expense of higher power consumption.

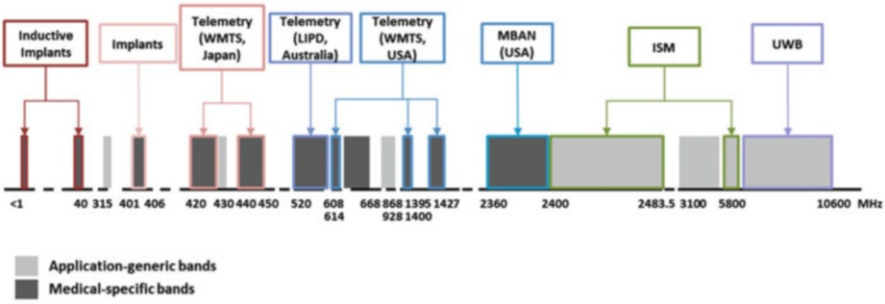


Fig. 5.3 Wireless regulatory environment of BSN

The WMTS market is rather fragmented. WMTS is available in the United States, operating in the spectrum bands 608–614 MHz, 1,395–1,400 MHz and 1,427–1,432 MHz [46]. Similar regulations are in place in Japan, where WMTS operates in the 420–430 MHz and 440–450 MHz bands. In Australia the spectrum available for telemetry transmitters is located in the 520–668 MHz range, under an application-specific subcategory of *Low Interference Potential Devices* (LIPD). No equivalent of WMTS exists in Europe.

As explained above, a number of wireless regulations allow for BSN operation in different spectrum bands. Medical implants operate most often in the 401–406 MHz spectrum band, as small networks based on proprietary radio implementations. For non-implantable BSN, the 2.4 GHz ISM band is most commonly used today since it is adopted by many wireless technology standards. It is possible that in the coming years commercial BSN migrate towards the MBAN band, due to lower spectrum crowdedness derived from its medical-specific nature. As described in Sect. 5.3.2 already two IEEE 802.15-family standards enable wireless communication in the MBAN band.

To conclude, Fig. 5.3 summarises the wireless regulatory environment of BSN.

5.3.2 Wireless Communication Standards

Among the intense activities for standardising wireless communications worldwide, the families of IEEE 802.11 and IEEE 802.15 standards have the largest impact on *Wireless Local Area Networks* (WLAN) and *Personal Area Networks* (WPAN) respectively. In general, the term WLAN refers to systems with a communication range of 10–100 m that have hundreds of mW at their disposal and often interact with a wired infrastructure (LANs). By contrast, the term WPAN refers to systems with a range of less than 10 m for highly mobile devices, such as wireless I/O peripherals with very limited power resources.

IEEE 802.11 [47] defines different wireless LAN technologies, such as the prominent 2.4 GHz, 11 Mbit/s IEEE 802.11b, 2.4 GHz, 54 Mbit/s IEEE 802.11 g, and 5 GHz, 54 Mbit/s IEEE 802.11a.

The standardisation body IEEE 802.15 [48] has issued four major technological releases so far, specifying the functionalities of the PHY and MAC layer of WPAN: (a) IEEE 802.15.1 medium-rate WPAN (derived from the Bluetooth® standard) with up to 3 Mbit/s peak rate; (b) IEEE P802.15.3 high-rate WPAN supporting high data rate personal networking; (c) IEEE 802.15.4 low-rate WPAN, mainly aiming at sensor/actuator networks; (d) IEEE 802.15.6 for short-range wireless communication inside, on or around a human body. In addition to the aforementioned WPAN standards, IEEE has also issued complementary guidelines, addressing both recommended practice for the coexistence of WPAN and WLAN devices (IEEE 802.15.2), as well as for the capability of mesh topologies in WPAN (IEEE 802.15.5).

It should be noted that while IEEE 802.15.1, IEEE 802.15.4 and IEEE 802.15.6 are designed for filling in the technological gap for personal and body-centric networking, the objective of IEEE 802.15.3 is to support multimedia communication. As such, and given the major commercial success of IEEE 802.11 for this set of applications, IEEE 802.15.3 is not used in practice for personal and body-centric networking.

Table 5.3 summarises the members of the IEEE 802.15 family. Derivatives and expansions of the major IEEE 802.15 releases that respond more efficiently to the BSN application requirements are presented in the remaining parts of this section, accompanied by discussions on interference and coexistence issues for body-centric communications.

5.3.3 IEEE 802.15.1: Medium-Rate Wireless Personal Area Networks

Originally developed by the Bluetooth *Special Interest Group* (SIG), this medium-rate standard soon became a synonym for short-range wireless communications worldwide. Upon the finalisation of the Bluetooth V1.1 specification, IEEE adopted and converted it into the IEEE 802.15.1 Standard [49]. This process not only included the conversion of the specification to IEEE format, but also encompassed the addition of an IEEE 802 Logical Link Control interface. In 2005, the original standard was amended [50], supporting faster connections, enhanced error detection, flow control and synchronisation as well as increased robustness against interference. In 2010, the Bluetooth SIG released the Bluetooth Core Configuration 4.0 [51], determining the functionalities for *Basic Rate* (BR) and *Enhanced Data Rate Configuration* (EDR), as well as the *Low Energy Configuration* (LE).

IEEE 802.15.1/Bluetooth BR-EDR emphasises short-range communications with special enhancements for voice data. It dictates operation in the 2.45 GHz ISM frequency band, which is further split into 79 (USA, Europe) or 23 (Japan) RF channels. Each RF channel is 1-MHz wide. The spread of spectrum is relies on the *Frequency Hopping* (FH) technique [59]. Therefore, while employing the entire band, when a frequency hop occurs, the centre of the current transmission frequency switches to the one of another channel.

Table 5.3 Overview of the IEEE 802.15 family

Standard	Description	Initial release revision date	Amendments
IEEE 802.15.1 (Bluetooth) [49–51]	MAC and PHY layer specifications for Wireless Personal Area Networks (WPANs)	2002/ 2005	Bluetooth core configuration v4.0 and bluetooth low energy (2009)
IEEE 802.15.2 [52]	Coexistence of wireless personal area networks with other wireless devices operating in unlicensed frequency bands	2003	In hibernation since 2011
IEEE 802.15.3 [53–55]	MAC and PHY layer specifications for High Rate Wireless Personal Area Networks (HR-WPANs)	2003	802.15.3b (2006): amendment to MAC sub-layer 802.15.3c (2009): millimetre-wave-based alternative physical layer extension
IEEE 802.15.4 [38, 56]	MAC and PHY layer specifications for Low-Rate Wireless Personal Area Networks (LR-WPANs)	2003/ 2006/ 2011	802.15.4.a (2007): PHY layer extension to chirp spectrum techniques and UWB systems 802.15.4c (2009): alternative PHY Extension to support one or more of the Chinese 314–316 MHz, 430–434 MHz, and 779–787 MHz bands 802.15.4d (2009): alternative PHY layer extension to support the Japanese 950 MHz bands 802.15.4e (2012): Amendment 1: MAC Sub-layer 802.15.4f (2012): Active Radio Frequency Identification (RFID) system PHY 802.15.4j (2013): alternative PHY extension to support Medical Body Area Network (MBAN) services operating in the 2,360–2,400 MHz band
IEEE 802.15.5 [57]	Mesh topology capability in wireless personal area networks	2009	–
IEEE 802.15.6 [58]	Wireless body area Networks	2012	–

RFCOMM and Service Discovery Protocol: Representative examples of the upper layers of the Bluetooth stack for higher-level functionalities and interfaces to application profiles.

L2CAP: Utilises HCI to provide data services to the upper protocols. The services provided are related to multiplexing, segmentation and reassembly capabilities.

Host Controller Interface (HCI): Provides a uniform set of interfaces for accessing the upper layers of the Bluetooth stack.

Link Manager Protocol (LMP): Defines the control operations for the baseband and physical layers. LMP is implemented at the LM layer and is carried on the asynchronous connection-less and the broadcast logical transports.

Baseband: Handles packet formats and defines the physical links that can be formed between a master and a slave device. Physical links are only formed a master and a slave. A master can support up to 7 active links with slave devices.

BT Radio: Deals with channel characteristics and determines 4 types of physical channels: (a) inquiry scan, (b) page scan, (c) basic and (d) adapted piconet. Inquiry scan and page scan are used for device discovery and connection, while basic and adapted piconet are used during normal operation.

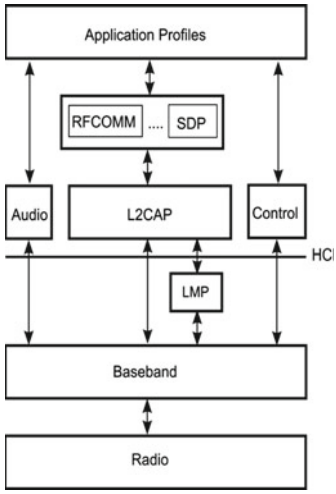


Fig. 5.4 IEEE 802.15.1 (Bluetooth BR/EDR) protocol stack

Bluetooth BR/EDR uses a *Frequency-Hopping/Time-Division-Duplex* (FH/TDD) scheme. A limited number of devices share the same radio channel, by employing the same synchronisation rules and the same frequency hopping sequence. The resulting physical topology is a piconet, which is essentially a star topology. The master of the piconet provides the clock and the frequency-hopping pattern. In particular, the hopping pattern is a pseudo-random ordering of the 79 [23] frequencies in the ISM band, and can be further integrated with the *Adaptive Frequency Hopping* (AFH) mechanism [60]. AFH allows skipping certain frequencies that are used by non-hopping ISM systems.

The hierarchy of layers both for the BR as well as the EDR operational modes is summarised in Fig. 5.4. In particular, Bluetooth Radio, *Baseband* (BB), *Link Manager* (LM) and *Logical Link Control and Adaptation* (L2CAP) are the layers defined by the specification. These layers provide all necessary functionality for devices connection, formation of logical data links and end-to-end services abstraction, via the means of interfaces to Application Profiles.

Bluetooth BR/EDR has a number of valuable features that has made it a widely accepted technology for short-range cable replacement. However, it also has a number of BSN-related limitations. Automatic network formation is not supported, and when the master of an established network moves away, the entire network collapses, which conflicts with the requirements of dynamically changing networks. Starting up a connection is rather slow, i.e., up to the order of 5 s, as it requires channel and page scanning, prior the actual connection setup. Once a Bluetooth inquiry is initiated to look for other Bluetooth devices, it disrupts every on-going communication, such as transmission of an ECG data stream. Furthermore, a Bluetooth inquiry will fail if both devices are simultaneously in Inquiry State.

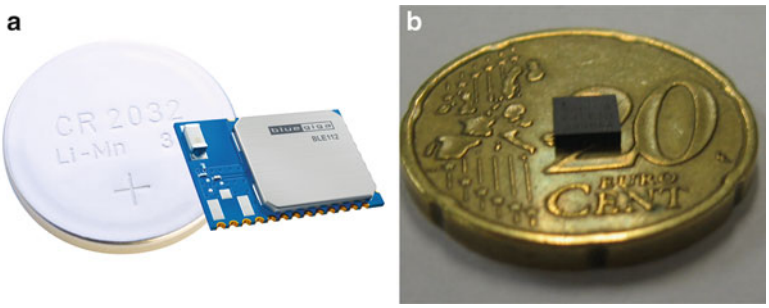


Fig. 5.5 (**a**) Bluetooth LE module by Bluegiga® [61], (**b**) Bluetooth LE module by Nordic® [62]

Moreover, despite the fact that the interconnection of several piconets to form scatternets is defined, these are in their majority met as proprietary solutions. Finally, support of efficient multicast is missing.

5.3.3.1 Bluetooth Low Energy

Bluetooth LE is the complementary form of Bluetooth wireless technology. Similar to the Core configuration defined by Bluetooth BR/EDR, the LE configuration provides device discovery, connection establishment and connection mechanisms. Nevertheless, the LE configuration responds to applications that require energy-efficient and cost-effective solutions.

In addition, depending on the application requirements, devices that implement BR/EDR and LE can communicate with devices implementing both systems as well as devices implementing either system (Fig. 5.5).

Bluetooth LE also operates in the 2.4 GHz ISM band, offering 40 RF channels. Each channel is 2 MHz wide. Bluetooth LE employs the AFH mechanism for spreading the spectrum and offers multiple access both in the frequency (FDMA) and time (TDMA) domains. During FDMA, the 40 physical channels are divided into three advertising channels and 37 data channels. The physical topology between connected devices relies on piconets. The master of the piconet determines the hop interval and the hopping pattern for accessing the 37 physical channels. The TDMA access provides polling-based data transmission between the master and the slaves of the piconet.

The protocol stack for the Bluetooth LE core configuration is presented in Fig. 5.6a. In particular, Bluetooth LE Radio, *Link Layer* (LL) and L2CAP are the layers defined by the specification. In contrast to Bluetooth BR/EDR, LE Radio defines only two physical channels for the formation of a piconet: (a) the LE advertisement broadcast channel, (b) the LE piconet channel. The LE advertisement broadcast channel is comprised of three fixed physical channels. This channel is used during the discovery and interaction of devices prior to their connection. An LE advertisement channel can be shared by any number of devices, also allowing

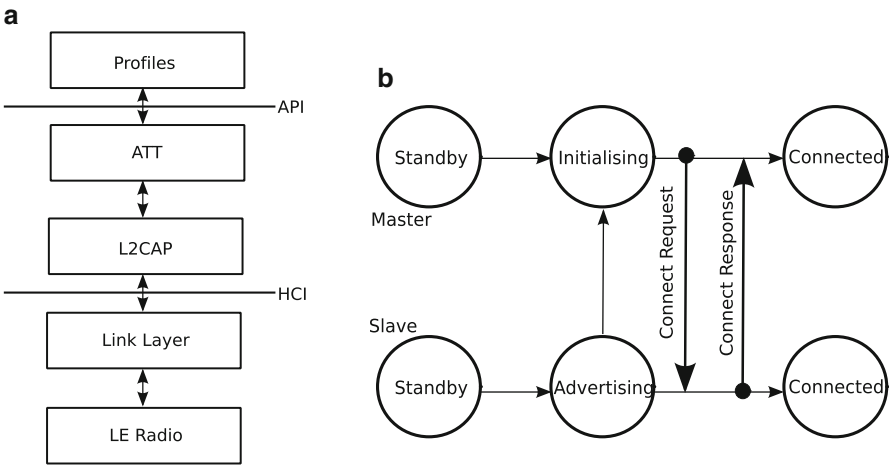


Fig. 5.6 (**a**) Bluetooth LE protocol stack and (**b**) The Bluetooth LE connection setup

connected devices to simultaneously advertise under restrictions. The LE piconet channel is used for communication between connected devices. The access to the LE piconet channel is controlled by the master, which additionally initiates the transmission in connection events and defines the timing of the piconet.

The LE advertisement broadcast and piconet channels support the advertising and active physical links respectively. Advertising links have a short duration, while active physical links yield one-to-one attributes in the master-slave communication. In contrast to the original Bluetooth specification, a slave cannot have more active physical links to more than one master at a time. As such, Bluetooth LE does not support the construction of scatternets. Moreover, there is no theoretical limit on the number of LE active links that can be supported within a piconet. Therefore, the number of active master-slave links per piconet can be significantly larger than 7, which is the limit for the Bluetooth BR/EDR.

On top of the physical links the logical topology of Bluetooth LE is built. In contrast to Bluetooth BR/EDR, which supports five types of logical transports for connection-less/connection-oriented point-to-point services and broadcast services, LE supports only asynchronous traffic both for typical operation and advertisement mode of a piconet. The asynchronous logical connections carry additional control commands generated by the LL protocol for managing the lower layers of the Bluetooth LE core. L2CAP provides the channel-based abstraction between the lower and upper layers of the LE devices. The functionality and capabilities of LE L2CAP are similar to those met in the Bluetooth BR/EDR. Furthermore, L2CAP provides the interface for the *Attribute Protocol* (ATT), which allows the exposition of the attributes of the LE device to its peer devices.

The specification of Bluetooth LE has been driven by the necessity of an energy-efficient alternative to the original standard. As such, when compared to the Bluetooth BR/EDR, the LE configuration manages to overcome some of the BSN-related

limitations. This is achieved by simplifying the protocol stack and by providing support only for reliable best-effort traffic. Therefore, in contrast to Bluetooth BR/EDR, active links are maintained only for as long as data needs to be exchanged between two devices. One of the most significant enhancements made is that slaves connected on the same master do not share the same physical channel. Consequently, the reliability of the data transmission is improved. In addition, the setup of a connection has been significantly simplified (Fig. 5.6b), which allows the completion of this procedure within 6 ms. However, although Bluetooth LE is characterised by 1 Mb/s nominal data rate, it does not support data streaming. As such, it is designated for sending small chunks of data that are capable of exposing the state of the hosting device.

5.3.4 *IEEE P802.15.3: High-Rate Wireless Personal Area Networks*

IEEE 802.15.3 was initially released in 2003, providing the specifications for the PHY and MAC layers of high-rate WPAN [53]. In 2005 the standard was amended, to provide enhancements and corrections related to the MAC sub-layer functionalities [54]. The resulting IEEE P802.15.3 MAC and PHY features are: (a) data rates of 11, 22, 33, 44, and 55 Mbit/s, (b) a MAC protocol that supports asynchronous and *Quality-of-Service* (QoS) isochronous data transfers.

IEEE P802.15.3 PHY layer operates in the 2.4 GHz band, occupying 15 MHz of RF bandwidth per channel. Hence, three or four non-overlapping channels can be accommodated within the available 83 MHz of the 2.4 GHz band. In contrast to IEEE 802.11, IEEE 802.15.3 chooses a single-carrier PHY in an effort to reduce complexity and power drain. Rather than employing spread-spectrum techniques, the original IEEE P802.15.3 PHY harnesses *Trellis-Coded Modulation* (TCM) with multi-bit symbols at 11 MBaud and achieves 11–55 Mbit/s peak data rate over a range of 10–30 m. The original physical layer aims at an RF front-end and baseband processors optimised for short-range transmission, exhibiting a current drain of less than 100 mA (much less than IEEE 802.11) and a small form factor for integration into consumer devices.

The IEEE 802.15.3 MAC employs the rules of CSMA/CA and is partially based on HiperLAN/2. The basic component of the network structure is the Device. Up to 256 Devices can be self-organised in a piconet or scatternets. The operation of each piconet is configured by the Coordinator, which is responsible, among others, for allocating timed access to the channel. In contrast to IEEE 802.15.1 and Bluetooth LE, the members of an IEEE 802.15.3 piconet form mesh topologies for data transmission.

Despite the promised efficacy of IEEE 802.15.3, it soon became apparent that the nominal data rates were not sufficient to fulfil the continuously increasing demands in terms of data rate. After the huge success of IEEE 802.11 WLAN, which by itself became capable of maintaining 54 Mbit/s peak data rate (IEEE 802.11a and 802.11 g, and wired connections like USB 2.0 and IEEE 1394, the market was in quest of a new PHY layer for the existing IEEE 802.15.3 MAC.

This led to the IEEE 802.15.3c alternate PHY [55] which was released in 2009. This amendment to the original standard defined an alternative physical layer, which operates at the 60 GHz band, resulting in higher data rates greater than 5 Gb/s. IEEE 802.15.3c provides three modes for the physical layer: (a) Single Carrier, (b) High Speed Interface, (c) Audio/Video mode. A common signalling mode is additionally defined, in order to enable the communication between devices that operate with different configurations at the PHY layer. Finally, in order to fully accommodate the functionalities of the alternative PHY, IEEE 802.15.3c additionally defines the necessary changes at the MAC layer. These changes allow devices that operate at a specific PHY mode to communicate through a piconet that operates at another PHY mode.

In summary, IEEE P802.15.3 has been released as a standard for low-power high-rate WPANs, featuring a thorough MAC with reservation schemes, powerful sleep modes and even inter-piconet communication support. The initial version of the standard (2003–2005) did not meet the success of the competitive technologies, because of its comparatively low maximum data rate. Considering the operational requirements for a BSN, IEEE P802.15.3c might in many cases be oversized despite its scalable data rates. While the energy-per-bit figures of IEEE P802.15.3c are at first impressive, the overhead for very small BSN packets may introduce concerns over the overall efficiency achievable. It can, however, be very useful when ambient-related information has to be transmitted, e.g. originating from a video camera or if data is transmitted in larger, thus more efficient, bursts.

5.3.5 *IEEE 802.15.4: Low-Rate Wireless Personal Area Networks*

The IEEE 802.15.4 Standard [38] specifies the PHY and MAC layers for low-rate WPANs. IEEE 802.15.4 is the basis for ZigBee®, ISA100.11a, WirelessHART® and MiWi® specifications, which further extend the standard by developing the upper layers. The standard was released in 2003, amended in 2006, and enables communication of data at a maximum rate of 1,000 kbps (Fig. 5.7).

Depending on the national and international spectrum regulations, IEEE 802.15.4 defines the six operational frequency bands: (a) 868 MHz (Europe), (b) 915 MHz (USA), (c) 779 MHz (China), (d) 950 MHz (Japan), (e) 2.4 GHz worldwide, (f) UWB, both in sub-GHz and 3–10 GHz. Depending on the operational frequency the spread spectrum technique is also defined, ranging from *Direct Sequence Spread Spectrum* (DSSS) to *Parallel Sequence Spread Spectrum* (PSSS) and *Chirp Spread Spectrum* (CSS). Table 5.4 summarises the technical characteristics of the resulting PHY specifications.

IEEE 802.15.4 WPANs can form either star or peer-to-peer network topologies. The latter may be used to form arbitrary patterns such as cluster trees. Star topologies have one PAN coordinator, whereas peer-to-peer topologies can have one of more. The role of such a coordinator is to act as primary controller of the

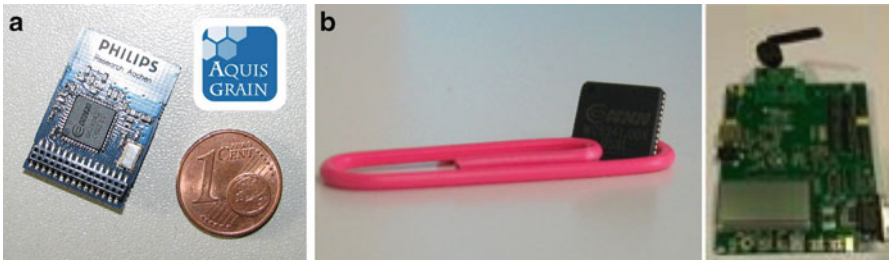


Fig. 5.7 IEEE 802.15.4-compliant modules powered by (**a**) Philips® [22], (**b**) Texas Instruments® [63]

Table 5.4 The PHY specifications supported by IEEE 802.15.4

Band (MHz)	Region	Number of channels	Modulation	Data rate (kbps)	Support
868–868.6	Europe	1	BPSK	20	Mandatory
			ASK	250	Optional
			O-QPSK	100	
779–787	China	8	MPSK	250	Mandatory
			P-QPSK		
902–928	USA	10	BPSK	40	Mandatory
			ASK	250	Optional
			O-QPSK	100	
950–956	Japan	22	BSPK	20	Mandatory
			GFSK	100	
2,400–2,483.5	Worldwide	16	O-QPSK (DSSS)	250	Mandatory
		14	CSS		Optional
			CSS	1,000	
249.6–749.6 (UWB sub-gigahertz)		1	BPM and BPSK	110–27,400 (varying w.r.t. chip rate)	Optional
3,244–4,724 (UWB low band)	Worldwide	4	BPM and BPSK	110–27,400 (varying w.r.t. chip rate)	Optional
5,944–10,234 (UWB high band)		11	BPM and BPSK	110–27,400 (varying w.r.t. chip rate)	Optional

network, by controlling the association of nodes as well as initiating, terminating, or routing communications. To allow for very low-cost low-complexity devices, IEEE 802.15.4 defines *Reduced Function Device* (RFD) and *Full Function Device* (FFD). RFDs implement a subset of the IEEE 802.15.4-defined primitives and cannot act as coordinator, whereas FFDs have a full implementation of IEEE 802.15.4 and can adopt any role in the WPAN.

The MAC specification for the IEEE 802.15.4 provides the basic functionalities that support: (a) the topology construction (PAN) by the means of passive/active

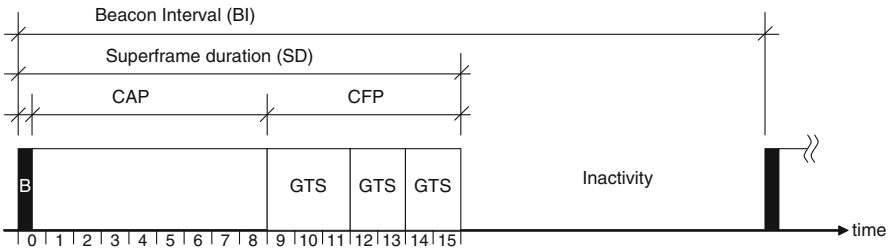


Fig. 5.8 Example of the IEEE 802.15.4 beacon interval

channel scanning and energy detection, (b) the association/de-association of the devices to a specific PAN and (c) the synchronisation of all devices to network beacons, whenever applicable. This set of functionalities cater for low-cost, low-power devices yet flexible enough to enable the delivery of periodic data, intermittent data (such as occasional measurements) and repetitive low-latency data (for instance, real-time ECG streaming). It features a fully handshaked protocol for reliable delivery of data and is able to support extremely low duty cycle (even below 0.1 %) operations efficiently.

Driven by the diversity of the application scenarios, the MAC specification provides two main operational modes: (a) non-beacon enabled, (b) beacon-enabled. The former operational mode employs the CSMA/CA mechanism for accessing the channel, while all devices operating within the same PAN are treated as peers. Operation over the beacon-enabled mode relies on the network synchronisation, which is dictated by the coordinator of the PAN.

Synchronisation is achieved by the means of periodic beacon transmissions, which define the beacon intervals. Each beacon interval is equally segmented into 16 slots. Within a beacon internal the superframe structure allows the nodes to either compete for accessing the channel, by the means of slotted CSMA/CA, or to transmit their data in a contention-free manner. Finally, for low-latency applications and/or applications requiring fixed data rates, the coordinator may dedicate portions of the active superframe to that application. These portions are called *Guaranteed Time Slots* (GTS) and are allocated in the *Contention-Free Period* (CFP). Figure 5.8 presents a representative example of the beacon internal, accompanied by the structure of the superframe, which supports the allocation of three GTS.

It should be noted that all devices operating in the WPAN should have unique 64-bit extended addresses, which can be used for communication within the WPAN. Communication with 16-bit addresses is also possible if the coordinator allocates short addresses for its devices during association to the WPAN. Due to this addressing mechanism, IEEE 802.15.4 offers support for over 65,000 devices in a PAN, in contrast to Bluetooth BR/EDR, which supports only up to seven active devices in a network. The number of WPAN members supported by IEEE 802.15.4 is more than sufficient for any envisaged BSN application.

The general MAC frame format is depicted in Fig. 5.9 along with the PHY frame format. The maximum payload deliverable by a MAC frame containing data is

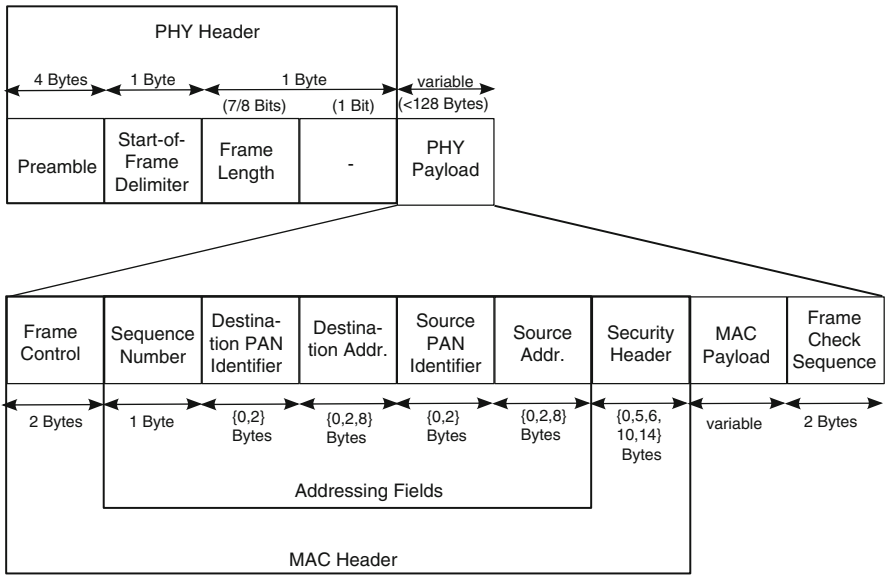


Fig. 5.9 General IEEE 802.15.4 MAC and PHY frame formats

127 bytes, depending on the addressing scheme used and the network topology. In principle, that payload could be directly employed for encapsulating medical data, in case of renouncement to the higher protocol stack layers.

Although a wireless BSN is not always described by as low duty cycle application (i.e. continuous ECG streaming), IEEE 802.15.4 has been extensively used for it. Driven by the appealing combination of low cost and robustness against interference at the 2.4 GHz band, the IEEE 802.15.4 family is expanded to address the specific needs of medical applications. In particular, IEEE 802.15.4j is an amendment of IEEE 802.15.4 that defines the necessary PHY/MAC modifications to allow operation in the FCC MBAN compliant frequency bands (2,360–2,400 MHz, [44]). Two considerations have been taken into account in the 4j amendment: (a) flexibility in the channelisation scheme, thereby accommodating harmonised coexistence with in-band primary/MBAN services, (b) provision of the MAC support that enables MBAN low-power implementations.

In the specification [56], the channelisation scheme defines 15 overlapping channels with a 0.5 MHz guard band at each band edge. In order to dynamically control the MBAN access to the channels in the 2,360–2,390 MHz band, a channel bitmap may be used. By configuring such parameter, a healthcare facility can enforce that all its MBAN devices only have access to the part of the MBAN spectrum authorized by the MBAN coordinator. This enables: compliance with FCC regulations, (b) harmonised coexistence with the primary services, including *Aeronautical Mobile Telemetry* (AMT) and amateur radio services. Moreover, a healthcare facility can configure this channel bitmap to enforce the use of non-overlapping channels to avoid interference caused by using overlapping channels. In particular, for in-hospital deployment, up to seven non-overlapping channels are

provided. For out-of-hospital deployment, three overlapping channels are defined in the 2,390–2,400 MHz band to provide flexibility to avoid co-located primary services (e.g. amateur radio) and other MBAN systems.

With regard to the MAC layer, a new channel switch command is specified to accommodate the need to dynamically vacate a portion of the MBAN spectrum when the MBAN coordinator requests to do so. The PHY Parameter Change *Information Element* (IE) is used by a device to notify another device to change the operating channel and/or band at a certain time indicated in a timestamp. This new feature also allows 802.15.4j users to re-channelise different types of MBAN services with different priorities to avoid collisions and ensure QoS. In addition, the Coordinator Switch feature aims to simplify MBAN device pairing and improve the energy efficacy. Such functionality is, for example, used for orderly and timely reconnecting a patient's sensors—after patient transportation—from the coordinator in a portable monitor to the desired coordinator in a bedside monitor. Finally, it should be noted that IEEE 802.15.4j is backward compatible with IEEE 802.15.4, therefore allowing the reuse of the current IEEE 802.15.4 implementations.

5.3.5.1 The ZigBee Specification

The ZigBee Alliance provides standards-based wireless technology [64]. The Alliance was created in 2002 by the synergy of 25 industrial and research organisations and launched the first specification in 2004. By 2012 ZigBee alliance had 600 members and released application-oriented standards for realising the Internet of Things.

The ZigBee standards provide the necessary means for enabling smart homes, buildings and retails as well as adaptive remote control and energy savings and personalised healthcare systems. These application-driven standards rely on the 2.4 GHz radio frequency and the IEEE 802.15.4 PHY and MAC specifications. As shown in Fig. 5.10, ZigBee builds a *Network* (NWK) layer and an *Application* (APL) layer on the IEEE 802.15.4-defined layers [65]. The network layer provides routing and multi-hop functions needed for creating different logical network topologies. The application layer includes an *Application Support* (APS) sub-layer, the *ZigBee Device Object* (ZDO), and the ZigBee applications defined by the user or designer. Whereas the ZDO is responsible for overall device management, the APS provides servicing to both ZDO and ZigBee applications.

The ZigBee NWK layer supports star, mesh, and cluster tree logical topologies. Its responsibilities include mechanisms used to join and leave the network, to apply security to frames and to route frames to their intended destinations. In addition, the network layer is responsible for the discovery and maintenance of routes between devices as well as for the discovery of one-hop neighbours. ZigBee defines three device types with respect to their networking capabilities:

- *ZigBee coordinator*: The IEEE 802.15.4 PAN coordinator. The ZigBee coordinator is responsible for starting a ZigBee network and assigning network addresses to newly associated devices.

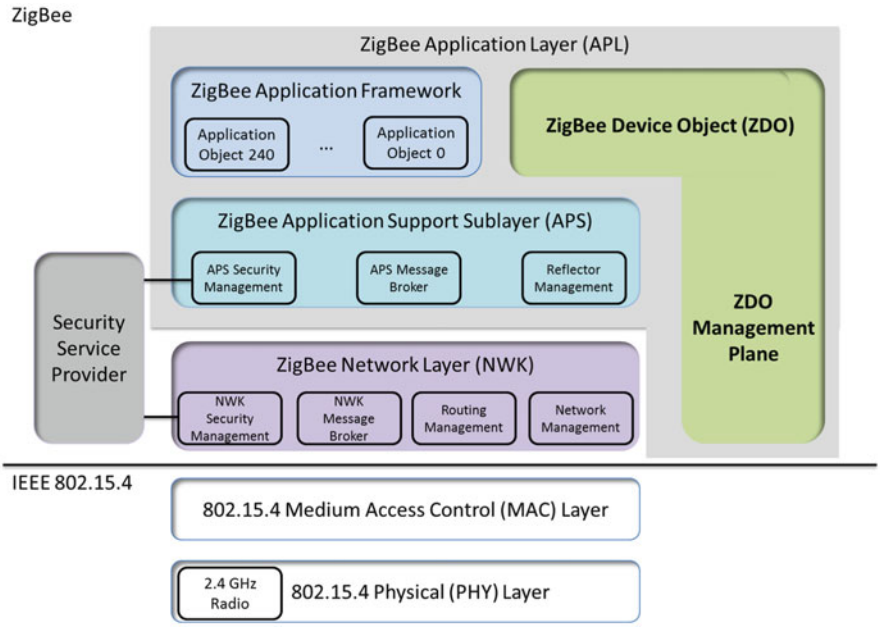


Fig. 5.10 ZigBee stack architecture

- *ZigBee router*: An IEEE 802.15.4 FFD that participates in a ZigBee network and is not the ZigBee coordinator but may act as a coordinator within its personal operating space. A ZigBee router is capable of routing messages between devices and supporting device associations.
- *ZigBee end device*: An IEEE 802.15.4 RFD or FFD that participates in a ZigBee network and is neither the ZigBee coordinator nor a ZigBee router.

The ZigBee APL layer consists of the APS sub-layer, the ZDO, and the manufacturer-defined application objects, which are embedded in the *Application Framework* (AF). The responsibilities of the APS sub-layer include maintaining tables for binding – the logical connection of devices based on their services and needs – and forwarding messages between bound devices. The responsibilities of the ZDO include: defining the role of the device within the network (for instance the ZigBee coordinator or end device); managing the node configuration; initiating and/or responding to binding requests and establishing a secure relationship between network devices. Another responsibility of the ZDO is discovery, which is the ability to determine which other devices are operating in the network. Most importantly, ZDO represents a predefined base class of functionality upon which all applications are written. As such, ZDO creates the necessary level of abstraction that allows the application developer not to worry about the low-level details. The actual applications are therefore implemented on the developer-defined application objects, according to application descriptions specified by ZigBee.

The AF is the environment in which application objects are hosted. An application object sends and receives data over its assigned endpoint, a physical/logical description. Each endpoint enables a single ZigBee device to support up to 240 independent end applications. Endpoints provide ZigBee with a level of sub-addressing additional to network addressing. Applications can be deployed on endpoints 1–240. Endpoint 0 is used by the ZDO for management purposes. Endpoint 255 is used to address all active endpoints (the broadcast endpoint). Endpoints 241–254 are reserved.

The key to interoperability between ZigBee devices of different vendors is the consensus on a profile. ZigBee Application Profiles specify the agreements on messages, message formats and processing actions that enable applications residing on separate ZigBee devices to send commands, request data and process commands/requests to create an interoperable, distributed application. A profile defines the following:

- One or more device description(s): A device descriptor is a description of a specific device within an application segment. For instance, the "Switch Remote Control" and the "Light Sensor Monochromatic" are two device descriptions included in the ZigBee application profile "Home Control" and "Lighting". Each device description is assigned a unique identifier within its profile that is exchanged during the service discovery process carried out by the ZDO.
- Cluster(s): A cluster is a container for one or more attributes, which are, data entities that represent a physical quantity or state. Each cluster is assigned an 8-bit cluster identifier unique within its specific profile. Equally, each of the attributes contained in a cluster is assigned an attribute identifier. An example of a cluster defined within the "Home Control, Lighting" profile is the "Program Light Sensor Monochromatic" cluster, which contains attributes such as "ReportTime" or "MinLevelChange".
- *Service types*: This is the type of AF data service to be used. It can be either a *Key Value Pair* (KVP) or *Generic Message* (MSG) service.

5.3.5.2 The ZigBee Healthcare Profile

In 2009, the ZigBee alliance released the Healthcare Profile, which defines the device descriptions and standard practices implementing healthcare applications on a ZigBee compliant platform [66]. The Healthcare Profile addresses the application requirements related to chronic disease monitoring, personal wellness monitoring and personal fitness monitoring. As such, the ZigBee Healthcare Profile is implemented in networked systems deployed in residents; retirement communities and nursing homes; low-acuity medical care facilities and fitness centres.

Figure 5.11 summarises the device descriptions for a ZigBee Healthcare cluster, accompanied by their ID. The Healthcare Profile addresses the body-centric perspective of chronic disease and personal fitness monitoring in the form of standalone BSN. By contrast, with regard to the personal wellness use case, this profile

Fig. 5.11 The device descriptions defined by ZigBee Healthcare Profile

responds to the ambient sensing that complements the BSN features in a pervasive networks setting.

It should be noted that this profile was architectured with a dedicated consideration on interoperability. Furthermore, the corresponding device descriptions are compliant to the IEEE P11073 standard for point-to-care medical device communication and the design specifications by Continua Alliance. Both IEEE P11073 and Continua Alliance are discussed in Sect. 5.5.

5.3.6 IEEE 802.15.6: Wireless Body Area Networks

In February 2012, IEEE published the IEEE 802.15.6 standard, enabling short-range communications inside, on, or around the human body [58]. In order to address the broad requirements of both medical, as well as non-medical applications, this standard specifies three alternatives for the PHY layer, combined with interoperable operational modes for the MAC layer and robust low-level data encryption and authentication.

Table 5.5 The NB PHY specifications supported by IEEE 802.15.6

Band (MHz)	Region	Number of channels	Modulation	Data rate (kbps)	Support
402–405	Worldwide	10	π/2-DBPSK	75.9	Mandatory
			π/2-DBPSK	151.8	
			π/4-DQPSK	303.6	
			π/8-D8PSK	455.4	Optional
420–450	Japan	12	GMSK	75.9	Mandatory
				151.8	
				187.5	Optional
863–870	Europe	14	π/2-DBPSK	101.2	Mandatory
902–928	US, Korea	60	π/2-DBPSK	202.4	
950–958	Japan	16	π/4-DQPSK	404.8	
			π/8-D8PSK	607.1	Optional
2,360–2,400	US	39	π/2-DBPSK	121.4	Mandatory
			π/2-DBPSK	242.9	
2,400–2,483.5	Worldwide	79	π/2-DBPSK	485.7	
			π/4-DQPSK	971.4	Optional

IEEE 802.15.6 defines three PHY layer specifications, which are: (a) *Narrowband* (NB) PHY, (b) *Ultrawideband* (UWB) PHY, (c) *Human Body Communications* (HBC) PHY. Each PHY layer operates at different frequency bands, which are based on the national and international regulations.

The NB PHY is considered optimal for medical applications, both for wearable and implanted networks. As shown in Table 5.5, the NB PHY determines the operation in several frequency bands, defining the operation of at least ten channels. Depending on the type of modulation, this permits scalable data rates that range from 100 to 1,000 Kpbs, thus enabling the trade-off between rate and range. With regard to nominal operational conditions, a transmitter compliant to NB PHY employs efficient spectrum spreading techniques that are based on repetition and bit interleaving.

The IEEE 802.15.6 UWB PHY enables operation in two different types of technology: (a) *Impulse Radio* UWB (IR-UWB), (b) *Frequency Modulation* UWB (FM-UWB). IR-UWB is specified as mandatory, whereas FM-UWB is defined as optional. The UWB-based specification defines three physical channels in the low band (3–5 GHz) and eight physical channels in the high band (6–10 GHz) with channel bandwidth of 499.2 MHz. For each physical channel, four logical channels are created with unique preamble sequences. In total, eight preamble sequences are defined, half of which should be used for odd number of physical channels and the other half should be used for even number of physical channels. Such arrangement can mitigate co-channel and adjacent channel interference effect.

In the IR-UWB option, time-based hopping is adopted to randomize impulse positions for coexistence. The time hopping sequence is specifically designed to provide guard interval between two consecutive hopping intervals only when

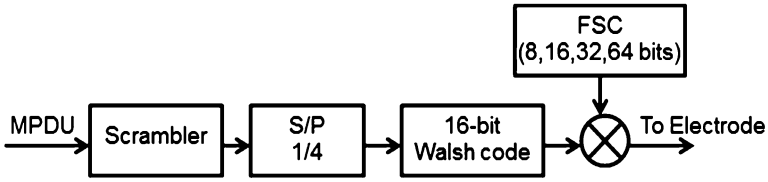


Fig. 5.12 The IEEE 802.15.6 HBC *PHY Service Data Unit* (PSDU) construction block diagram

necessary. Compared to allocating a fixed guard interval in each hopping interval, this method has much less overhead. The modulation schemes supported in the IR-UWB option include *On-Off Keying* (OOK) with either single pulse option or burst pulse option and DPSK with spreading. The FM-UWB option exploits high modulation index of FM to generate an UWB signal. The bits of the *PHY Protocol Data Unit* (PPDU) are first modulated with *Continuous Phase Binary Frequency Shift Keying* (CP-2FSK) and then the resulting narrow band CP-2FSK signal is modulated with wideband frequency modulation to generate a constant-amplitude UWB signal. In summary, the UWB mode is designed to provide a large scope of implementation opportunities for high reliability, low complexity, and ultra-low power operations.

HBC is a technique that uses the human body as the transmission medium for communications via electric fields instead of electromagnetic waves. Devices can thereby communicate without a wire or wireless technology. The user simply touches the device, and the devices are connected to each other. A HBC transceiver is comprised of an electrode and a controller. By the means of capacitive coupling, the electrode can transmit or receive an electrical signal through the human body. The controller generates the data for transmission.

IEEE 802.15.6-HBC PHY operates in the 21 MHz frequency band, characterised by 5.25 MHz bandwidth. The supported data rate ranges from 164 kbps to 1.3125 Mbps. The transmitter block diagram is shown in Fig. 5.12. In particular, data is scrambled and then grouped into 4-bit vectors. Then each 4-bit vector is mapped into a 16-bit Walsh codeword [67]. Finally, the output bits are spread with a *Frequency Shift Code* (FSC), which is generated by repeating [0 1].

IEEE 802.15.6 support only star topologies, which are defined either within the 1-hop or the 2-hop neighbourhood. While the 1-hop star topology is applicable to all operational bands, an IEEE 802.15.6 network operating in the MICS band do not support 2-hop extended star.

The hub of the star is the device responsible for creating/maintaining the BAN and coordinating medium access. With regard to medium access, three policies are defined (Table 5.6): (a) scheduled access; (b) improvised and unscheduled access; (c) random access. Based on these access types, the hub operates in one of the three following modes: (a) Beacon mode with beacon periods (superframes); (b) Non-beacon mode with superframes; (c) Non-beacon mode without superframes. When the BSN operates in either of the first two modes, the hub provides a time reference and corresponding time division to the operational nodes. Conversely, when superframes are not supported, the time division is not necessary.

Table 5.6 The medium access policies provided by IEEE 802.15.6

Medium access policy	Type	Duration of resources allocation	Beacon with superframes	Non-beacon with superframes	Non-beacon without superframes
Scheduled access	Contention – free (pre-negotiated dedicated time slots)	1-periodic/ m-periodic	√	√	–
Improvised/ unscheduled access	On demand slot allocation and data transaction based on polling/posting and round-robin	1-periodic/ m-periodic	√	√	√ (Round-robin)
Random access	Slotted CSMA/SA (NB PHY)/ slotted aloha (UWB PHY)		√	–	√

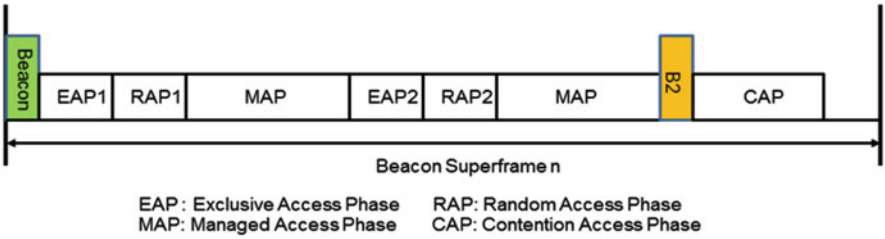


Fig. 5.13 Access phases in an active superframe for beacon mode with superframes

During beacon-with-superframes mode, the hub determines the length of beacon internal by periodically sending beacons. Each beacon internal is divided into active and inactive superframe periods, dictating the normal and sleep operational modes of the nodes respectively. During the active superframe period, the access to the common channel becomes hybrid, as it combines *Exclusively Access Phase* (EAP), *Random Access Phase* (RAP), *Managed Access Phase* (MAP) and *Contention Access Phase* (CAP). The ordering of the different phases is illustrated in Fig. 5.13.

EAP, RAP and CAP rely on random access, and thereby employ either slotted CSMA/CA (NB PHY) or slotted Aloha (UWB PHY). EAP are only used for the hub or a node to send data type frames of the highest user priority (i.e. medical emergency event reports), while RAP and CAP are usually for regular data traffic. By contrast, access during MAP, relies on scheduled or improvised access that allow event- or poll-driven data transmission from or to the hub. Thus, the hub uses MAP in order to provide contention-free time allocation intervals.

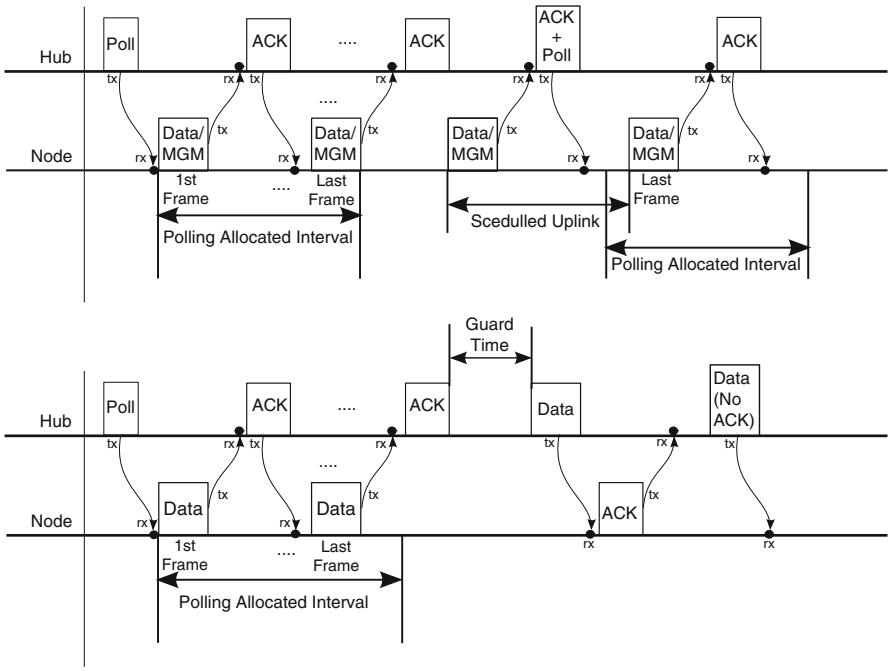


Fig. 5.14 Example of transactions between hub and node, during improvised (*top*) and unscheduled bi-link (*bottom*) intervals

Specifically, the hub may arrange scheduled uplink/downlink/bi-link allocation intervals, unscheduled bi-link allocation intervals, and improvised polled and posted intervals in the MAP phase (Fig. 5.14).

During operation over the non-beacon with superframes mode, the entire period of the active superframe is dedicated to the MAP, thereby allowing the hub to arrange scheduled, unscheduled and improvised allocation intervals. Finally, non-beacon without superframes the nodes, expect for the hub, compete for granting the common medium by employing the rules of either slotted CSMA/CA or slotted Aloha.

It should also be noted that IEEE 802.15.6 also defines the rules for both centralised and distributed guard provisioning, which allows each node to protect their allocated time slots from adjacent transmissions. These can optionally be combined with several MAC features that facilitate BSN coexistence and interference mitigation. These features include: (a) adaptive channel hopping, (b) beacon shifting, (c) active superframe interleaving. More specifically, adaptive hopping is applicable to the NB PHY, excluding the MICS band, and the FM-UWB PHY. This feature allows the adaptive selection of the operating channel set, and thus guarantees the minimum channel separation to achieve frequency diversity. Beacon shifting allows a hub to transmit its beacon in a random time hopping manner. This results in alleviation from potential repeated beacon collisions and scheduled

allocation conflicts between overlapped and adjacent BANs operating on the same channel. Finally, active superframe interleaving provides multiple and co-existing BSN to coordinate the schedules of their active superframes.

5.3.7 Comparison of Technologies

Considering, as a common factor, the need to combine cost effectiveness and energy conservations, Table 5.7 provides an overview of the candidate wireless standards that are considered the most appropriate for BSN-related applications. The results therein summarised highlight the fact that the selection of the technology depends heavily on the needs of the specific BSN application.

Bluetooth LE offers a cost effective and easy-to-implement solution. The data rates proved sufficient for most of the BSN-related applications. The most significant advantage that this technology offers is interoperability. A Bluetooth LE device is compatible with Bluetooth BR/EDR devices, which are typically met in mobile phones, laptops and tablets. Nevertheless, the fact that it does not support data streaming makes this technology suitable for only periodic sampling that does not need to respond to the requirements of waveform data.

Conversely, IEEE 802.15.4 is a well-known technology, with extensive support both in research and industrial communities. One of the strongest advantages of this standard is that it is extremely versatile, in terms of PHY layer support. The CSS complementary technique, along with the UWB characteristics for the PHY layer specification, yields the necessary data resilience for BSN-related applications. Moreover, the most recent amendment (IEEE 802.15.4j) specifically addresses the needs of medical communications, both in controlled and free-living environments, by reserving space in more quiet areas of the spectrum.

IEEE 802.15.6 has a more detailed focus on implant networks or networks operating in close vicinity to the human body. It has a specific focus on implant networks and can also support high QoS demands both for medical, as well as for non-medical applications (IR-UWB). As such, it preserves the versatility offered by its ancestor, and expands it by complementing the MAC layer with well-defined priority levels. Furthermore, each medium access policy addresses a different set of applications. More specifically, the scheduled access provides guaranteed periodic access, allowing nodes to wake up at specific time intervals. Therefore, it is suitable for medical data streaming applications (e.g., ECG waveform data streaming). Improvised access facilitates transfer during medical alarms, variable-rate medical, configuration and network management data. Random access is usually used for episodic medical applications, such as home and fitness monitoring and medical event/alert reporting. A characteristic that may make it less appealing than IEEE 802.15.4 is the fact that it only supports star topologies. Thus, IEEE 802.15.6 may be more prone to single point failures than IEEE 802.15.4. Nevertheless, as it is characterised by extremely low consumption, it is considered a strong candidate for transiting from wearable to truly pervasive implant network structure.

Table 5.7 Summary of the candidate standards for body sensor networking

		IEEE 802.15.4				
	Bluetooth LE				ZigBee	IEEE 802.15.6
PHY layer supported	2.4 GHz (FHSS/ AFH)	2.4 GHz DSSS/ 2.4 GHz CSS	UWB: sub GHZ/ 3–10 GHz	2.36–2.4 GHz (IEEE 802.15.4j)	(supported by 2.4 GHz IEEE 802.15.4)	NB/ UWB/ HBC
Data rate	1 Mbps	250 kbps/ 1 Mbps	110–27,400 kbps (Varying w.r.t to chirp rate)	250 kbps		75.7–971.4 kbps/ 202.5–15,600 kbps/ 164–1,312.5 kbps
Range	10–30 m	~10–30 m	few meters to 30 m (depending on PHY and data rate)	~10–30 m	10–100 m (multi-hop mesh)	<10 m (depending on the PHY)
Network topology	Star	Star, peer-to-peer			+Multi-hop mesh	Star/ extended star
Network size	Unlimited	65,535				65,535
Network join time	~6 ms	<<1 s				<<1 s
Real-time support	No	Guaranteed time slots			Profile–defined priorities	Exclusive access period, 8 priority levels
Protocol complexity	Simple	Simple			Low	Simple
Security	Authentication, encryption					
Peak current consumption	15–20 mA	15–25 mA				<10 mA (expected) (NB PHY)
FEC support	No	Yes				Yes

5.4 Interference and Coexistence

While the standards described above provide the design guidelines for realising technological products for personal and body-area networking, their resilience against interference should also be considered during the network design. Interference is caused by the coexistence of multiple types of networks on the same frequency band and can cause severe degradation on the performance of the network, which is often expressed in terms of *Packet Error Rate* (PER).

The interference level amongst wireless communication systems mainly depends on the transmitted power, channel bandwidth, spectrum spreading mechanism and medium access scheme of each system. Whilst the transmit power of external interferers has an obvious impact on the level of interference induced in a system, the system's own channel bandwidth and spectrum spreading contribute to the robustness against external interference.

Medium access schemes that sense the medium before transmitting (such as CSMA/CA), produce less interference than inflexible time-based medium access schemes like TDMA. Table 5.8 summarises these characteristics for the technologies that are defined at the ISM band.

In particular, considering the coexistence of IEEE 802.15.4/2.4 GHz DSSS PHY with IEEE 802.11, IEEE 802.15.3 and Bluetooth devices, IEEE 802.11 is proved to be more harmful [38, 68–71]. This is due to the following reasons:

(a) The impact of IEEE 802.15.3 on the performance of IEEE 802.15.4 performance is not as severe as the one caused by IEEE 802.11-based networks, due to the fact that IEEE 802.15.3 is not as commercially established as IEEE 802.11.
(b) Bluetooth and Bluetooth LE use FHSS as a spreading spectrum technique. This implies that a narrow band signal is transmitted. By contrast an IEEE 802.15.4

Table 5.8 Coexistence-relevant characteristics of wireless communication standards operating at 2.4 GHz

Standards	Typical transmit power	Channel bandwidth	Spectrum spreading	Medium access
IEEE 802.15.6/ 2.4 GHz PHY	0 dBm	1 MHz	DSSS	CSMA/CA/ ALOHA/ TDMA
IEEE 802.15.4/ 2.4 GHz PHY	0 dBm	2 MHz	DSSS/ CSS	CSMA/CA (TDMA optional)
IEEE 802.11b (WLAN)	14–16 dBm	22 MHz	DSSS	CSMA/CD (polling-based TDMA optional)
IEEE 802.15.1 (Bluetooth)	0 dBm	1 MHz	FHSS (79 channels)	TDMA
Bluetooth LE	0 dBm	2 MHz	FHSS (40 channels)	TDMA
IEEE 802.15.3	8 dBm	15 MHz	none	CSMA/CA (TDMA optional)

DSSS-based transmission relies on the spreading the signal across a larger bandwidth. In addition, while Bluetooth enabled transmitters employ the AFH for avoiding crowded channels, IEEE 802.15.4 employs channel assessment and retransmissions. These differences in the operational principles allow an IEEE 802.15.4 receiver to either recover an interfered signal, or to re-iterate the transmission process.

(c) Considering the case of co-channel operation, the level of transmission power is the most dominant factor. As such, when an IEEE 802.15.4 DSSS-device is moved towards an IEEE 802.11 transmitter, it will be the first to be affected.

In contrast to the IEEE 802.15.4/DSSS case, when the CSS technique is employed, the interference from co-existing channels is less severe [72]. This is due to the fact that the CSS yields higher data rates than DSSS. Consequently, the duty cycle of IEEE CSS devices can be expected to be significantly below the duty cycle of other IEEE 802.15.4 devices. Since the 2.4 GHz ISM band has become an extremely busy medium, a low duty cycle achieved by high data rates can significantly compensate from interference effects.

IEEE 802.15.4 devices that operate in the UWB spectrum allocation have been examined for their coexistence with IEEE 802.16 and IEEE 802.22, which define the specifications for Metropolitan and Regional Area Networks respectively. The results indicate a highly asymmetrical behaviour on the interference [72]. In particular, IEEE 802.16/IEEE 802.22 devices impact UWB PHY devices at a longer range than vice versa. This is due to the high deviation on the operational transmission power. For example, while an UWB PHY device operates at -15 dBm, an IEEE 802.16 device operates at +17 dBm.

IEEE 802.15.6/2.4 GHz devices encounter similar interference effects from IEEE 802.11 and IEEE 802.15.1 devices on their operation as the one experienced by IEEE 802.15.4/2.4 GHz devices. It is also interesting to note the asymmetry between IEEE 802.15.4 and IEEE 802.15.6 devices at the 2.4 GHz band. The results presented in [73] indicate that when the IEEE 802.15.6 is the interferer, in order to achieve PER less than 10^{-2} at the IEEE 802.15.4 device, the physical separation should be 12 m. Conversely, the required separation between an IEEE 802.15.6 and an IEEE 802.15.4 device should be 40 m, when the latter is the interferer.

With regard to the IEEE 802.15.6/UWB PHY, source of interference can either be IEEE 802.15.4 compliant devices or IEEE 802.15.6 UWB co-existing networks. In particular, the preliminary results presented in [74] examine how the receiver sensitivity in IR-UWB and FM-UWB is affected by coexistence with IEEE 802.15.4 UWB systems. IR-UWB devices are not affected by FM-UWB interferences. By contrast, an IEEE 802.15.6 UWB will not be affected by IEEE 802.15.4 UWB as long as either the receiver sensitivity of the former is higher than -28 dBm, or the performance degradation of the receiver sensitivity is lower than 15 dBm.

IEEE 802.15.6 devices that operate at either the MICS band or at the band defined by the HBC PHY are not affected by interference, since their operational bands are explicitly selected to minimise any interference effects. Similar

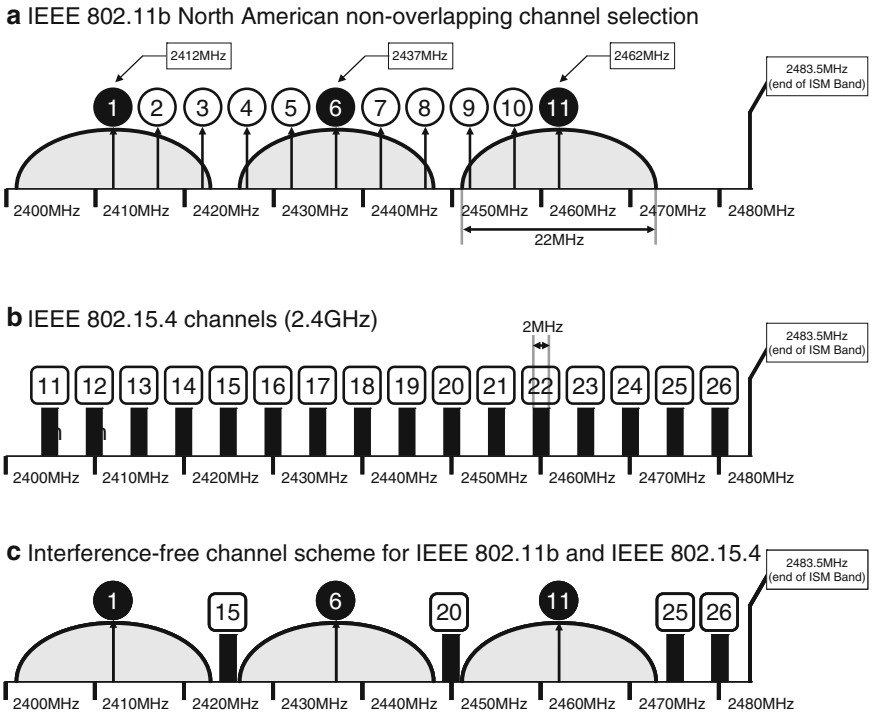


Fig. 5.15 IEEE 802.11b (North America) and IEEE 802.15.4 (2.4 GHz) channel scheme

observations can also be made for the operation of IEEE 802.15.4j-compliant devices, which exploit the quietness of the AMT/Amateur Radio Spectrum.

While all standards in the IEEE 802.15 family are compliant with the rules specified by the IEEE 802.15.2 standard for coexistence with other devices that operate in the same spectrum, IEEE 802.15.4 and IEEE 802.15.6 are additionally protected by additional mechanisms, which characterise their operation. In particular, the clear channel assessment prior transmission; the energy detection of the channels and the low duty cycles contribute to the minimisation of interference effects. Another mechanism that is inherited in the network formation is the channel selection from the coordinator/hub of the network. Practical tests with AquisGrain A1.0 [22] nodes, for example, and laptop computers equipped with standard WLAN (IEEE 802.11b) highlight the efficacy of this policy. More specifically, the results indicate that the interference caused by IEEE 802.11b on an IEEE 802.15.4 BSN is much lower for non-overlapping channels (in Fig. 5.15 IEEE 802.11b channel 1 and IEEE 802.15.4 channel 15) than for overlapping channels (in Fig. 5.15 IEEE 802.11b channel 1 and IEEE 802.15.4 channel 12). Even in case of channel overlap, the measured packet error rate drops noticeably as the offset between the central frequencies of IEEE 802.15.4 and IEEE 802.11b increases.

Table 5.9 depicts the main results of the IEEE 802.11b-induced interference experiments. In view of these results, it is recommended to start a BSN preferably on the channels that fall in the frequency gaps between IEEE 802.11b channels.

Table 5.9 Relative Packet Error Rate (PER) measured at the IEEE 802.15.4 receiver depending on the offset between the central frequencies of IEEE 802.15.4 and IEEE 802.11b. The PER is normalised to the highest PER measured in every test set-up

	12 MHz central	7 MHz central	2 MHz central
	Frequency offset (no overlap)	Frequency offset (overlap)	Frequency offset (full overlap)
No overlap vs. overlap	0.0036	1	n.a.
Overlap vs. full overlap	n.a.	0.098	1

Independently of the scheme used to deploy the interfering technology, it is always recommended to exploit the energy scan functionality and to use a dynamic channel selection, thereby allowing the BSN to continuously scan and switch to less crowded channels. The procedure used to select the BSN channel – either before starting the network or during operation – is not specified by the IEEE 802.15 standards. It is entirely up to the BSN designer how to manage channels to cater for coexistence with other technologies. In addition, quality-of-service mechanisms have been developed and tested for IEEE 802.15.4 to improve the performance of wireless body sensor networks for medical applications in terms of reliability and timeliness [75].

5.5 Healthcare System Integration

Standards are the key for ensuring interoperability between wireless medical sensors and for the integration of body sensor networks into larger health information systems inside or outside the hospital, as, for example, personal telehealth systems. Within the IEEE, interoperability is defined as "the ability of two or more systems or components to exchange information and to use the information that has been exchanged".

Therefore, interoperability involves all layers of the communication stack. At the lower layers, this requires the use of standardized transport technologies, enabling basic connectivity. In the upper layers, standardised application profiles are needed, defining what capabilities of the transport technology have to be used to best support the needs of the connected health domain. Finally, at the application layer, common data models, formats and nomenclatures need to be defined to enable end-to-end data flow and consistent interpretation of the data. As shown in the previous sections, a variety of mature standards for the lower layers (transports) and upper layers (profiles) are available today. In recent years, the definition of standardised application layers for connected health devices is recently being increasingly addressed by international standardisation bodies.

Several standards for application-layer medical information exchange exist today. HL7 (*Health Level Seven*) [76] is a healthcare interoperability standard for the electronic interchange of clinical, financial and administrative information

among independent health care oriented computer systems; e.g., hospital information systems, clinical laboratory systems, enterprise systems and pharmacy systems. "Level Seven" refers to the highest level of the *International Standards Organization*'s (ISO) communications model for the application layer of OSI reference model. DICOM (*Digital Imaging and Communications in Medicine*) [77] is a standard for transmitting medical imaging data, including also the handling, storing and printing thereof.

Standards that facilitate the efficient exchange of vital signs and medical device data, acquired by patient-connected medical devices, for all health environments, are defined within the ISO/IEEE 11073 family of standards. Although parts of this family comprise standards addressing the physical (electrical, synchronisation, cable and connector) and transport characteristics of communication, the main focus is on providing plug and play interoperability on the application layer.

The ISO/IEEE 11073 standards are split into two sub-groups. The first (classic) group of standards addresses point-of-care medical device communication (MDC). These standards are mainly designed for acute monitoring and treatment applications in particular diagnostic, bed or treatment areas in the hospital domain. See [78] for an overview of the standards in the ISO/IEEE 11073 MDC family.

The design objectives of the ISO/IEEE 11073 MDC standards only partly align with the requirements for personal telehealth systems or body sensor network applications, where battery powered devices and sensors demand very low computational complexity and low power consumption. For wireless devices, this implies minimizing transmission time by means of minimal protocol overhead. On the other hand, in personal telehealth applications outside the hospital, the network configuration and user association are rather static and less dynamic than those in typical hospital settings, where frequent network reconfiguration is a strong requirement. Also, in application domains outside the hospital, there are fewer requirements for real-time alarms and streaming, as the setting is less acute. To account for these diverging requirements, a second (new) group of standards, the ISO/IEEE 11073 *Personal Health Device Communication* (PHDC) family, has been created, with the first standards having been released in 2008. During the past years a lot of international industry effort has been and is still spent in extending and maturing this standards family.

5.5.1 ISO/IEEE 11073 Personal Health Device Communication

Similar to the ISO/IEEE 11073 MDC family, the ISO/IEEE 11073 PHDC standards are based on an object-oriented system management paradigm, defined in ISO/IEEE 11073-20601 [79], which represents a generic optimised exchange protocol. An object-oriented data model, the *Domain Information Model* (DIM), is used to specify objects, attributes, and data types for describing personal health data,

Table 5.10 The ISO/IEEE 11073 Personal health device communication series of standards (Status: P = published, D = in draft)

Part	Title	Status
00103	Health informatics — personal health device communication — overview	P
10404	Health informatics — personal health device communication — device specialisation — pulse oximeter	P
10406	... — Basic electrocardiograph (ECG) (1–3-lead ECG)	P
10407	... — Blood pressure monitor	P
10408	... — Thermometer	P
10413	... — Respiration rate monitor	D
10415	... — Weighing scale	P
10417	... — Glucose meter	P
10418	... — International normalised ratio (INR) monitor	P
10419	... — Insulin pump	D
10420	... — Body composition analyser	P
10421	... — Peak expiratory flow monitor	P
10422	... — Urine analyser	D
10423	... — Sleep monitor	D
10424	... — Sleep apnea breathing therapy equipment	D
10425	... — Continuous glucose monitor	D
10441	... — Cardiovascular fitness and activity monitor	P
10442	... — Strength fitness equipment	P
10471	... — Independent living activity hub	P
10472	... — Medication monitor	P
20601	Health informatics — personal health device communication — application profile — optimised exchange protocol	P

such as measurements, device settings, units of measurement, etc. The standardised nomenclature, defined in ISO/IEEE 11073-10101 [80], comprises of a set of numeric codes that identify every item that is communicated between systems. A service model defines access to the data objects of the devices. A communication model manages the connection state machine for message exchange.

The base protocol ISO/IEEE 11073-20601 provides a generic framework and feature set, whereas interoperability for a particular device type is enabled by properly constraining this generic feature set to the needs of a particular device type and by adding device specific nomenclature terms. This is specified in so-called device specialization documents for various personal health devices as part of the ISO/IEEE 11073-104zz series (Table 5.10). The list of devices covered by the –104zz series has been grown over the past years and is expected to further expand in the future.

According to the requirements of the personal telehealth domain, the encoding and parsing of protocol data units using ISO/IEEE 11073 PHDC is efficient due to the concept of precoded messages (message templates can be filled in memory in which only the actual updated values must be copied). A guiding principle for the development of ISO/IEEE 11073 PHDC has also been to place the greater computational burden on aggregators (*managers* in the ISO/IEEE 11073 domain), which

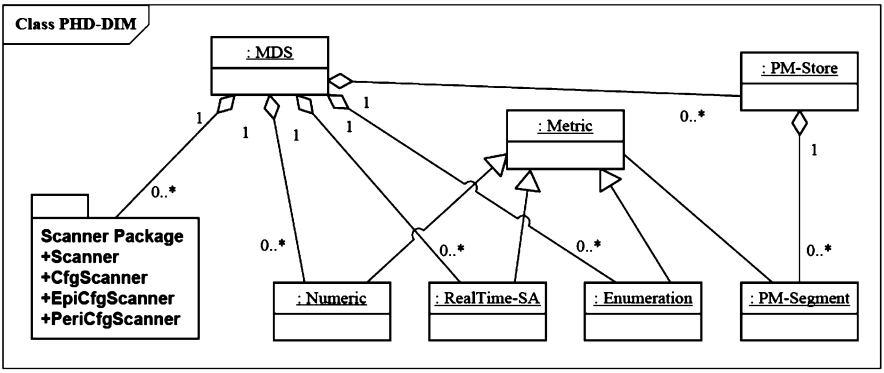


Fig. 5.16 ISO/IEEE 11073 PHDC – Domain information model

have typically richer capabilities in terms of memory and CPU power, compared to the sensor/device side (*agents* in the ISO/IEEE 11073 domain). Managers often use larger batteries or are even mains powered.

Another distinct feature is a mechanism to reduce message overhead for very simple personal health devices having a static configuration.

By negotiating virtually all static context information once in a so-called configuration phase, only the dynamic information is transmitted in a device's measurement report. The term 'static' here refers to information that does not change from measurement to measurement and which is usually exploited for parsing (e.g. attributes denoting the unit of a measurement value, or attributes describing the type of measurement). This mechanism reduces message overhead significantly and hence, results in less time spent for transmission and thus a reduction of transmit power consumption, improving battery life.

In addition to the standardised service and communication model for exchanging data and a standardized nomenclature, the DIM is the key element to model specific device types. It comprises of several different classes for modelling an agent.

An agent device is described as a set of objects that represent the data sources, with each object having one or more attributes. The attributes describe measurement data and device status information that is communicated to a manager as well as elements that control behaviour. Figure 5.16 shows a class diagram of the DIM of an agent along with class relationships, using Unified Modelling Language.

In particular, device types are modelled by defining proper objects that are instances of the classes in Fig. 5.16.

For example, the domain information model (object diagram) of a pulse oximeter according to the device specialization ISO/IEEE 11073-10404 [81] is shown in Fig. 5.17.

The model contains three different types of numeric objects representing the measurements for saturation of peripheral oxygen (SpO_2), pulse rate, and quality of the pulsatile wave. A *Real-time Sample-Array* (RT-SA) object represents the plethysmogram waveform. Three different types of enumeration objects represent

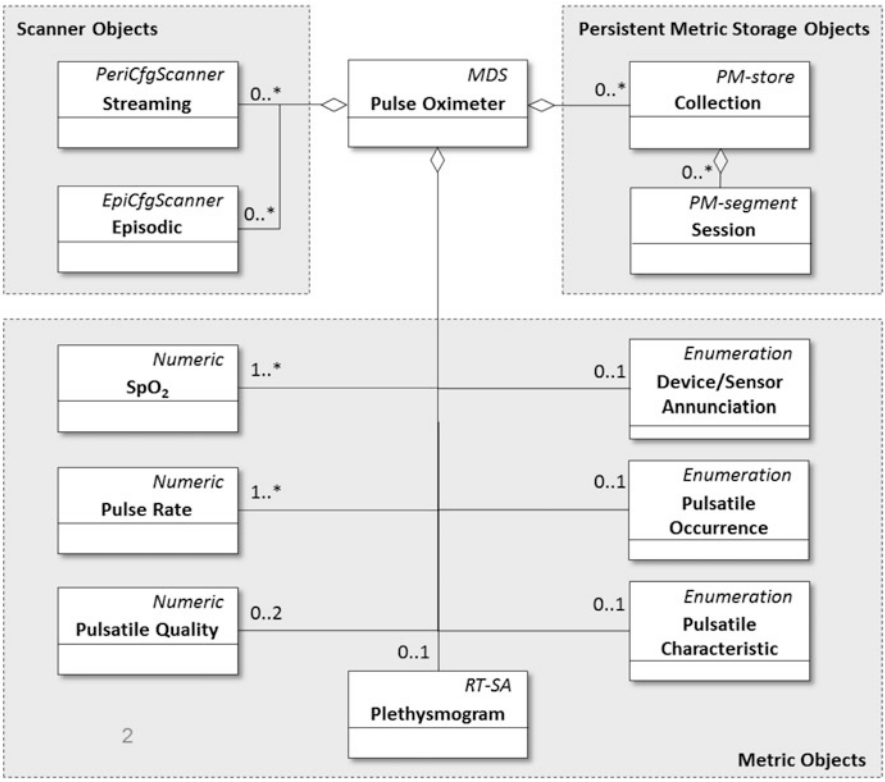


Fig. 5.17 ISO/IEEE 11073-10404 Pulse oximeter – Domain information model

device and sensor annunciation status information (including, for example, sensor displacement, faulty sensor detection, signal irregularities and low-perfusion determination); event information regarding the detection of pulsatile occurrences and additional information about the characteristics of the pulsatile wave. Persistent metric storage objects are used to facilitate long-term data acquisition of several minutes or hours of oximetry data, e.g. during overnight measurements, which can be retrieved from the agent device after the acquisition is complete. The scanner objects are used for enabling efficient grouping of several measurements into a single message payload.

Finally, an example of an encoded data message is shown in Fig. 5.18. The message in binary format represents an unconfirmed event report message of length 58 bytes, sent by a pulse oximeter agent. The last 24 bytes contain the actual SpO_2 and pulse rate measurements, as well as time stamp information. Due to the fact that during operating state the configuration of the agent is known to the manager, static context information such as unit of measurement, type of measurement, and the numerical format of the measurement value does not need to be included in the event report messages. Note, that the highlighted bytes represent the actual varying

```
0xE7 0x00                 APDU CHOICE Type (PrstApdu)
0x00 0x36                 CHOICE.length = 54
0x00 0x34                 OCTET STRING.length = 52
0x00 0x04                 invoke-id = 4
0x01 0x00                 CHOICE(Remote Operation Invoke | Unconfirmed Event Report)
0x00 0x2E                 CHOICE.length = 46
0x00 0x00                 obj-handle = 0 (MDS object)
0x00 0x00 0x00 0x00       event-time = 0
0x0D 0x1D                 event-type = MDC_NOTI_SCAN_REPORT_FIXED
0x00 0x24                 event-info.length = 36
0xF0 0x00                 ScanReportInfoFixed.data-req-id = 0xF000
0x00 0x01                 ScanReportInfoFixed.scan-report-no = 1
0x00 0x02                 ScanReportInfoFixed.obs-scan-fixed.count = 2
0x00 0x1C                 ScanReportInfoFixed.obs-scan-fixed.length = 28
0x00 0x01                 ScanReportInfoFixed.obs-scan-fixed.value[0].obj-handle = 1
0x00 0x0A                 ScanReportInfoFixed.obs-scan-fixed.value[0]. obs-val-data.length = 10
0x00 0x62                 Basic-Nu-Observed-Value = 98 (%)
0x20 0x13 0x02 0x08       Absolute-Time-Stamp = 2013-02-08T12:10:0000
0x12 0x10 0x00 0x00
0x00 0x0A                 ScanReportInfoFixed.obs-scan-fixed.value[1].obj-handle = 10
0x00 0x0A                 ScanReportInfoFixed.obs-scan-fixed.value[1]. obs-val-data.length = 10
0x00 0x48                 Simple-Nu-Observed-Value = 72 (beats/min)
0x20 0x13 0x02 0x08       Absolute-Time-Stamp = 2013-02-08T12:10:0000
0x12 0x10 0x00 0x00
```

Fig. 5.18 An example of encoded message

parts of the message. This is in accordance to the principle of canned messages, which allows the use of predefined message templates. Only the fixed location, varying parts need to be modified before sending. This reduces the parsing efforts at the agent side significantly.

The ISO/IEEE 11073 PHDC family of standards has been adopted by the Continua Health Alliance and is leveraged by its design guidelines to enable a consistent, interoperable data layer concept across the different Continua interfaces.

5.5.2 Continua Health Alliance

The Continua Health Alliance is the leading industry consortium [82] for the advancement of personal telehealth. Its mission: to establish an ecosystem of interoperable personal health systems to better manage health and wellness has attracted to date more than 200 member companies that include payers, health care providers, medical device manufacturers, information technology companies, software and middleware vendors, and silicon systems manufacturers.

The Continua Health Alliance aims to enable end-to-end system interoperability, from device to the care provider and the *Electronic Health Records* (EHR). Their approach is to leverage and integrate existing standards for all layers of the communication stack and for all parts of the overall system between the patient-end and the service- and provider-end. On the device side, the major focus of Continua is on personal healthcare devices like weighing scales, blood pressure monitors, pulse-oximeters, heart-rate monitors, fitness devices like motion and

activity monitors, as well as devices and sensors supporting independent living of the elderly like smoke, gas, and fall detection sensors.

A lack of interoperability is a significant technical barrier that limits the growth of the personal telehealth market. The approach taken by Continua is the development of design guidelines [83] to support interoperable sensors, platforms and services, and a logo and certification program to signify the promise of interoperability to the customer. These high-level guidelines lay the groundwork for the development and deployment of personal health care and wellness devices for home, group care environments and ultimately professional clinical facilities. Continua is not a *Standards Development Organization* (SDO); instead, it works with SDOs like the IEEE, HL7, ISO, *American Telemedicine Association* (ATA), *European Telecommunications Standards Institute* (ETSI) and others, to optimise existing standards to support Continua's vision of interoperable personal health care and wellness devices and systems. Three market segments have been defined: Chronic Disease Management, Fitness and Aging Independently. There are strong commonalities between these three spaces in personal health device type, data acuity and perceived market value for adoption.

In addition to technical barriers, the Continua Health Alliance addresses other barriers as well in order to encourage the growth of the personal telehealth market. On the regulatory side, current safety regulations are not adapted yet to facilitate the creation of multi-vendor solutions. As a consequence, Continua Alliance is working with regulatory agencies to safely and effectively manage diverse vendor solutions. On the financial side, the economic value of personal telehealth has still proven difficult to demonstrate in a scientifically sound manner, which restrains the adjustment of reimbursement models that support personal telehealth. To overcome this, Continua is working with leaders in the healthcare industry to develop new ways of addressing the costs of providing personal telehealth systems, such as new reimbursement models and payment models.

Striving for interoperability requires making choices. From the connectivity standards discussed above Continua has selected those that best match the requirements of the different system interfaces and use cases addressed. As a result Bluetooth and USB have been selected by Continua for its PAN interface, ZigBee for its LAN interface, and *Near Field Communications* (NFC) [84] for its *Touch Area Network* (TAN) interface. To ensure semantic interoperability on the applications layer, the ISO/IEEE 11073 PHDC standards have been selected for all these interfaces.

5.6 Conclusions

Wireless sensor networks are a key technology for pervasive health monitoring. Vital signs such as ECG or SpO_2 are measured by means of body-worn medical sensors or in the future by unobtrusive health sensors that are integrated in the furniture like chairs or beds. The medical data will be enriched with the readings

from ambient sensors providing additional context information, for example location and activity. The mobile, star- or mesh-based body sensor network connects to the ambient multi-hop mesh network resulting in a pervasive wireless hybrid network. Adding a mobile phone or a home gateway to the system provides global connectivity to distant sites such as hospitals or medical service centres.

In the last few years, much effort has been spent by the industry on ensuring interoperability of wireless medical BSNs and their integration into existing healthcare systems. The family of ISO/IEEE 11073 Personal Health Device Communication standards and their adoption by the Continua Health Alliance is a visible sign of the impressive progress that has been accomplished.

In addition, regulatory authorities such as the FCC have recently acknowledged the high value of BSNs for the society by allowing the deployment of *Medical Body Area Networks* (MBAN) in the 2,360–2,400 MHz spectrum, thereby providing a more protected alternative to the crowded 2.4 GHz ISM band.

As of today, there is a range of suitable short-range communication technologies available to enable pervasive health applications based on wireless BSNs. The most promising candidates are Bluetooth Low Energy, IEEE 802.15.4 and IEEE 802.15.6. Bluetooth Low Energy is the first choice for applications requiring connectivity of BSN to mobile phones acting as a gateway. IEEE 802.15.4j is the currently the only medical-specific wireless communication standard allowing BSNs to operate in the new dedicated MBAN spectrum, greatly reducing the risk of interference. Finally, IEEE 802.15.6 enables small form factors and long operating times thanks to its expected low power consumption.

In summary, the advance in wireless communication technologies supported by frequency spectrum regulation and healthcare interoperability standards opens the way for making pervasive healthcare a reality.

References

1. ISO/IEC 7498-1, ISO/IEC international standard, Information technology - open systems interconnection - basic reference model: the basic model, 2nd edition, http://standards.iso.org/ittf/PubliclyAvailableStandards/s020269_ISO_IEC_7498-1_1994(E).zip, 1994.
2. IEEE 11073-00101, Health Informatics—Point-of-Care Medical Device Communication—Technical Report—Guidelines for the Use of RF Wireless Technologies, Draft, October 2005.
3. Philips Medical Systems. Striving for cableless monitoring. In: Ragil C (eds), Philips Medical Perspective Magazine 2005, 8:24–25.
4. Falck T, Espina J, Ebert JP, Dietterle D. BASUMA-the sixth sense for chronically ill patients. In: Proceedings of the Third International Workshop on Wearable and Implantable Body Sensor Networks, Cambridge, USA, 2006.
5. The MobiHealth project, http://www.mobihealth.org/
6. The HealthService24 project, http://www.healthservice24.com/
7. Lo BPL, Yang GZ. Key technical challenges and current implementations of body sensor networks. In: Proceedings of the Second International Workshop on Wearable and Implantable Body Sensor Networks, 2005, 1–5.

8. Anliker U, Ward JA, Lukowicz P, Tröster G, Dolveck F, Baer M, *et al*. AMON: a wearable multiparameter medical monitoring and alert system, IEEE Transactions on Information Technology in Biomedicine 2004, 8(4): 415–427.
9. Malan D, Fulford-Jones T, Welsh M, Moulton S. Codeblue: an ad hoc sensor network infrastructure for emergency medical care. In: Proceedings of the First International Workshop on Wearable and Implantable Body Sensor Networks, 2004, 55–58.
10. MyHeart project, http://www.hitech-projects.com/euprojects/myheart/
11. Jovanov E, Milenkovic A, Otto C, de Groen PC. A wireless body area network of intelligent motion sensors for computer assisted physical rehabilitation. Journal of Neuroengineering and Rehabilitation 2005, 2(1): 6.
12. Eklund JM, Hansen TR, Sprinkle J, Sastry S. Information technology for assisted living at home: building a wireless infrastructure for assisted living. In: Proceedings of the Twenty-Seventh Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 2005.
13. Stanford V. Using pervasive computing to deliver elder care. IEEE Pervasive Computing 2002, 1(1): 10–13.
14. The Wear-a-BAN project, www.wearaban.eu, 2012.
15. Kumar S. Nilsen W. Pavel M., Srivastava M. Mobile Health: Revolutionizing Healthcare Through Transdisciplinary Research. IEEE Computer Magazine 2013, 46(1): 28–35.
16. Ganti R. K., Jayachandran P., Abdelzaher T. F., Stankovic J. A. SATIRE: a software architecture for smart AtTIRE. In: Proceedings of the 4th International Conference on Mobile Systems, Applications and Services, ACM, 2006, 110–123.
17. Lorincz K., Chen, B.-R., Challen G. W., Chowdhury A. R.; Patel S., Bonato P., Welsh M.. Mercury: a wearable sensor network platform for high-fidelity motion analysis. In: Proceedings of the 7th ACM Conference on Embedded Networked Sensor Systems, ACM, 2009, 183–196.
18. Espina J, Falck T, Mühlsteff J, Jin Y, Adán MA, Aubert X. Wearable body sensor network towards continuous cuff-less blood pressure monitoring. In: Proceedings of the 5th International Summer School and Symposium on Medical Devices and Biosensors, 2008.
19. Lorincz K., Malan D.J., Fulford-Jones T.R.F., Nawoj A., Clavel A., Shnayder V., Mainland G., Welsh M., Moulton S., Sensor networks for emergency response: challenges and opportunities. IEEE Pervasive Computing 2004, 3(4): 16–23.
20. Chen B. -R., Muniswamy-Reddy K. -K., Welsh M. Ad hoc multicast routing on resource-limited sensor nodes. In: Proceedings of the 2nd international workshop on Multi-hop Ad Hoc Networks: from Theory to Reality, ACM, 2006, 87–94.
21. Baldus H, Klabunde K, Müsch G. Reliable set-up of medical body-sensor networks. In: Proceedings of the First European Workshop on Wireless Sensor Networks, Berlin, Germany, 2004; Springer LNCS 2920:353–363.
22. Falck T, Baldus H, Espina J, Klabunde K. Plug'n play simplicity for wireless medical body sensors. Mobile Networks and Applications 12(2–3), 143–153.
23. Lorincz K, Welsh M. MoteTrack: A robust, decentralized approach to RF-based location tracking. In: Proceedings of the First International Workshop on Location- and Context-Awareness, Oberpfaffenhofen, Germany, 2005; Springer LNCS 3479:63–82.
24. Philips Research. Philips demonstrates new personal healthcare techniques, Press release, http://www.research.philips.com/newscenter/archive/2005/050623-aachen-healthcare.html, June 2005.
25. The SAPHE project, http://ubimon.doc.ic.ac.uk/saphe/m338.html, 2008.
26. Wood A., Stankovic J., Virone G., Selavo L., Zhimin H., Qiuhua C., Thao D., Yafeng W., Lei F., Stoleru, R. Context-aware wireless sensor networks for assisted living and residential monitoring. IEEE Network 2008, 22(4): 26–33.
27. Tia G., Pesto C., Selavo L., Yin C., Jeong G. K.; Jong H. L., Terzis A., Watt A., Jeng J. Bor-rong C., Lorincz K., Welsh M. Wireless Medical Sensor Networks in Emergency Response: Implementation and Pilot Results. In: Proceedings of the 2008 I.E. Conference on Technologies for Homeland Security 2008,187–192.

28. Atallah L., Lo B., Yang G-Z., Siegemund F. Wirelessly accessible sensor populations (WASP) for elderly care monitoring. In: Proceedings of the Second International Conference on Pervasive Computing Technologies for Healthcare, 2008 (Pervasive Health 2008), 2–7.
29. Chakravorty R. A programmable service architecture for mobile medical care. In: Proceedings of the Fourth Annual IEEE International Conference on Pervasive Computing and Communications Workshops, 2006. (PerCom2006), pp.5.–536.
30. Ko J, Lim J. H., Chen Y., Musvaloiu-E R., Terzis A., Masson G. M., Gao T., Destler W., Selavo L., Dutton, R. P. MEDiSN: Medical emergency detection in sensor networks. ACM Transactions on Embedded. Computing Systems 2010, 10(1): 11:1–11:29.
31. Gnawali O., Fonseca R., Jamieson K., Moss D., Levis P. Collection tree protocol. In: Proceedings of the 7th ACM Conference on Embedded Networked Sensor Systems, 2009, 1–14.
32. The AID-N Project, http://www.jhuapl.edu/aid-n/ (Final Report), 2007.
33. The Independent Regulator and Competition Authority for the UK Communications Industries (Ofcom), www.ofcom.org.uk.
34. Federal Communication Commission, http://www.fcc.gov/
35. International Telecommunication Union. 2012. Radio Regulations. http://www.itu.int/pub/R-REG-RR
36. ERC/REC 70-03, European Radiocommunications Office. ERC recommendation related to the use of short range devices (SRD):http://www.erodocdb.dk/docs/doc98/official/pdf/rec7003e.pdf, May 2013.
37. IEEE 802.11. 2012. IEEE Standard for Information technology - Telecommunications and information exchange between systems Local and metropolitan area networks - Specific requirements Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications. IEEE, March 2012.
38. IEEE 802.15™. IEEE Standard for Local and metropolitan area networks–Part 15.4: Low-Rate Wireless Personal Area Networks (LR-WPANs). IEEE Std 802.15.4-2011 (Revision of IEEE Std 802.15.4-2006), 1–314, 2011.
39. IEEE 802.11a. IEEE Standard for Telecommunications and Information Exchange Between Systems - LAN/MAN Specific Requirements - Part 11: Wireless Medium Access Control (MAC) and physical layer (PHY) specifications: High Speed Physical Layer in the 5 GHz band. IEEE, January 1999.
40. Takahashi D. Tzero Technologies shuts down; That's the End of Ultra-Wide Band. 2009 http://venturebeat.com/2009/02/12/tzero-technologies-shuts-down-thats-the-end-of-ultrawideband/.
41. ETSI EN 301 839. 2009. Electromagnetic compatibility and Radio spectrum Matters (ERM); Short Range Devices (SRD); Ultra Low Power Active Medical Implants (ULP-AMI) and Peripherals (ULP-AMI-P) operating in the frequency range 402 MHz to 405 MHz. ETSI. http://www.etsi.org/deliver/etsi_en/301800_301899/30183901/01.03.01_60/en_30183901v010301p.pdf.
42. Federal Communication Commission (FCC). 1999. 47 CFR 95.601–95.673 Subpart E.: http://wireless.fcc.gov/services/index.htm?job=operations&id=medical_implant.
43. Federal Communication Commission (FCC). 2012. Medical Device Radiocommunications Service (MedRadio). http://www.fcc.gov/encyclopedia/medical-device-radiocommunications-service-medradio.
44. Federal Communication Commission (FCC). 2012. FCC 12-54. Amendment of the Commission's Rules to Provide Spectrum for the Operation of Medical Body Area Networks. First Report and Order and Further Notice of Proposed Rulemaking: http://hraunfoss.fcc.gov/edocs_public/attachmatch/FCC-12-54A1.pdf
45. ETSI TR 101 557. 2012. System Reference document (SRdoc); Medical Body Area Network Systems (MBANSs). http://www.etsi.org/deliver/etsi_tr/101500_101599/101557/01.01.01_60/tr_101557v010101p.pdf.
46. Federal Communication Commission (FCC). 2012. Wireless Medical Telemetry Service (WMTS). http://www.fcc.gov/encyclopedia/wireless-medical-telemetry-service-wmts.

47. The IEEE 802.11™ Wireless Local Area Networks. The working group for WLAN Standards. http://www.ieee802.org/11/.
48. The IEEE 802.15™ Wireless Local Area Networks. The working group for WPAN Standards. http://www.ieee802.org/15/.
49. IEEE 802.15™. IEEE standard for information technology – Telecommunications and information exchange between systems – Local and metropolitan area networks – Specific requirements – Part 15.1: wireless medium access control and physical layer specifications for wireless personal area networks. 2002.
50. IEEE 802.15 ™. IEEE Standard for Information Technology - Telecommunications and Information Exchange Between Systems - Local and Metropolitan Area Networks - Specific Requirements. - Part 15.1: Wireless Medium Access Control (MAC) and Physical Layer (PHY) Specifications for Wireless Personal Area Networks (WPANs). IEEE Std 802.15.1-2005 (Revision of IEEE Std 802.15.1-2002), 1–580, 2005.
51. Bluetooth SIG™. Bluetooth Core Version 4.0. 2010. https://www.bluetooth.org/Technical/Specifications/adopted.htm.
52. IEEE 802. 15 ™. IEEE Recommended Practice for Information Technology - Telecommunications and Information Exchange Between Systems - Local and Metropolitan Area Networks - Specific Requirements Part 15.2: Coexistence of Wireless Personal Area Networks With Other Wireless Devices Operating in Unlicensed Frequency Bands. IEEE Std 802.15.2-2003, 1–115, 2003.
53. IEEE 802.15 ™. IEEE Standard for Information Technology - Telecommunications and Information Exchange Between Systems – Local and Metropolitan Area Networks - Specific Requirements Part 15.3: Wireless Medium Access Control (MAC) and Physical Layer (PHY) Specifications for High Rate Wireless Personal Area Networks (WPANs). IEEE Std 802.15.3-2003, 1–315, 2003.
54. IEEE 802.15™. IEEE Standard for IEEE Amendment to Part 15.3: Wireless Medium Access Control (MAC) and Physical Layer (PHY) Specifications for High Rate Wireless Personal Area Networks (WPAN): Amendment to MAC sublayer. IEEE Std 802.15.3b-2005 (Amendment to IEEE Std 802.15.3-2003, 1–169, 2006.
55. IEEE 802.15™. IEEE Standard for Information technology - Telecommunications and information exchange between systems - Local and metropolitan area networks - Specific requirements. Part 15.3: Wireless Medium Access Control (MAC) and Physical Layer (PHY) Specifications for High Rate Wireless Personal Area Networks (WPANs) Amendment 2: Millimeter-wave-based Alternative Physical Layer Extension. IEEE Std 802.15.3c-2009 (Amendment to IEEE Std 802.15.3-2003), 1–187, 2009.
56. IEEE 802.15™. IEEE Standard for Local and metropolitan area networks—Part 15.4: Low-Rate Wireless Personal Area Networks (LR-WPANs) Amendment 4: Alternative Physical Layer Extension to Support Medical Body Area Network (MBAN) Services Operating in the 2360 MHz – 2400 MHz Band, 2013.
57. IEEE 802.15™. IEEE Recommended Practice for Information technology-Telecommunications and information exchange between systems-Local and metropolitan area networks-Specific requirements Part 15.5: Mesh Topology Capability in Wireless Personal Area Networks (WPANs). IEEE Std 802.15.5-2009, 1–166, 2009.
58. IEEE 802.15™. IEEE Standard for Local and metropolitan area networks - Part 15.6: Wireless Body Area Networks. IEEE Std 802.15.6-2012, 1–271, 2012.
59. Dwola F. Handbook of RF and Wireless Technologies. Newens, 2003. ISBN-13 978-0-7506-7695-3.
60. Popovski P., Yomo H.; Prasad R. Strategies for adaptive frequency hopping in the unlicensed bands. IEEE Wireless Communications, 13(6): 60–67.
61. Bluegiga, www.bluegiga.com.
62. Nordic Semiconductors. Bluetooth Low Energy. http://www.nordicsemi.com.
63. Texas Instruments, Single-Chip 2.4 GHz IEEE 802.15.4 Compliant and ZigBee™ Ready RF Transceiver, http://www.ti.com/product/cc2420.

64. The ZigBee Alliance, http://www.zigbee.org/.
65. The ZigBee 2012 Specification. http://www.zigbee.org/Specifications/ZigBee/Overview.aspx.
66. The ZigBee Healthcare Profile. http://www.zigbee.org/Standards/ZigBeeHealthCare/Overview.aspx.
67. Arora S., Barak B. Computational Complexity: A Modern Approach. Chap. 19. Cambridge University Press, 2009. ISBN-13 978-0521424264.
68. Golmie N, Cypher D, Rebala O. Performance analysis of low rate wireless technologies for medical applications. Computer Communications 2005; 28(10):1 255–1275.
69. Jennic Wireless Microcontrollers. Co-existence of IEEE 802.15.4 at 2.4 GHz. Application Note, 2009.
70. Pollin S., Tan I., Hodge B., Chun C., Bahai A. Harmful Coexistence Between 802.15.4 and 802.11: A Measurement-based Study. In: Proceedings of the 3rd International Conference on Cognitive Radio Oriented Wireless Networks and Communications (CrownCom 2008) 2008, 1–6.
71. LaSorte N.J., Rajab S.A., Refai H.H. Experimental assessment of wireless coexistence for 802.15.4 in the presence of 802.11g/n. In: Proceedings of the 2012 I.E. International Symposium on Electromagnetic Compatibility (EMC2012) 2012, 473–479.
72. IEEE 802.15™. Coexistence analysis of IEEE Std 802.15.4 with other IEEE standards and proposed standards. Technical Report, 2010. http://grouper.ieee.org/groups/802/19/pub/CA/15-10-0808-00-0000-802-15-4-2011-coexistence-analysis.pdf
73. IEEE 802.15™. TG6 Coexistence Assurance Document. Technical Report, 2011. https://mentor.ieee.org/802.15/documents.
74. Hernandez M., Miura R., Coexistence of IEEE Std 802.15.6TM-2012 UWB-PHY with other UWB systems. In: Proceedings of 2012 I.E. International Conference on Ultra-Wideband (ICUWB2012) 2012, 46–50.
75. Garcia J.J., Falck T. Quality of Service for IEEE 802.15.4-based Wireless Body Sensor Networks. In: Proceedings of the Third International Conference on Pervasive Computing Technologies for Healthcare, 2009.
76. HL7 – Health Level 7, http://www.hl7.org/.
77. DICOM – Digital Imaging and Communications in Medicine, http://medical.nema.org/
78. Schmitt L, Falck T, Wartena F, Simons D. Novel ISO/IEEE 11073 Standards for Personal Telehealth Systems Interoperability. In: Proceedings of the 2007 Joint Workshop on High Confidence Medical Devices, Software, and Systems and Medical Device Plug-and-Play Interoperability. 2007; pp. 146–148.
79. ISO/IEEE 11073-20601. 2010. Health informatics—Personal health device communication—Application profile—Optimised Exchange Protocol.
80. ISO/IEEE 11073-10101. 2004. Health informatics – Point-of-care medical device communication – Part 10101: Nomenclature.
81. ISO/IEEE 11073-10404. 2008. Health informatics – Personal health device communication—Part 10404: Device specialization— Pulse oximeter.
82. Continua Health Alliance, http://www.continuaalliance.org/index.html.
83. Continua Health Alliance. Continua Design Guidelines, version 2012.
84. NFC Forum, http://nfc-forum.org/.

Chapter 6
Energy Harvesting and Power Delivery

Eric Yeatman and Paul Mitcheson

6.1 Introduction

As we have seen in previous chapters, the increasing miniaturisation and cost reduction of sensors, circuits and wireless communication components is creating new possibilities for networks of wireless sensors, in wearable and other applications. However, in order for sensors to be wireless, or untethered, this requires not only wireless communication to and from the nodes, but also wireless powering. Batteries, of course, provide this capability in the great majority of portable electronic devices, and thus are the obvious solution also for the wireless sensor node application. However, their need for replacement or recharging introduces a cost and convenience penalty which is already undesirable in larger devices, and is becoming increasingly unacceptable for sensor nodes as the number of these (their ubiquity) grows. As an alternative, therefore, sources which harvest energy from the environment are very attractive. At the same time, the power demands of many electronic functions, wireless communication being a particularly important example in this context, are continuously falling. Although this lightens the demands for batteries, it also makes alternatives based on energy harvesting look more and more realistic.

Where batteries can power a sensor node for its whole expected lifetime without maintenance, and without dominating the node cost or weight, this is likely to remain the favoured solution in most cases, although even there, energy harvesting methods have advantages to offer. The materials required in batteries are often toxic or environmentally unfriendly, adding to the burden of both bio-compatibility for implanted use and to the end-of-life disposal. Where a lifetime beyond what a battery can provide is needed, the "eternal" nature of the harvesting supply clearly becomes particularly favourable.

E. Yeatman (✉) • P. Mitcheson
Department of Electrical and Electronic Engineering, Imperial College London, London SW7 2BY, UK
e-mail: e.yeatman@imperial.ac.uk

G.-Z. Yang (ed.), *Body Sensor Networks*, DOI 10.1007/978-1-4471-6374-9_6,

Table 6.1 Biosensor sample rates

Signal	Sampling rate (Hz)	Resolution (bits)	Data rate (bits/s)
ECG (per lead)	120–250	12	1,440–3,000
Temperature	0.2–2	12	2.4–24
Oximetry	60	12	720
Blood pressure	120	12	1,440
Respiratory rate	20	12	240
Heart rate	10	12	120

Adapted from Ref. [4]

Energy harvesting for supply of wireless electronics is a relatively young research field. In this chapter we review the state-of-art and technology trends, after first discussing the likely energy requirements of sensor nodes and briefly reviewing the capabilities of batteries. Detailed discussion of implementation will focus on inertial energy scavenging, but within this context more generic issues will be considered. This chapter is not intended to be a comprehensive survey; for such a purpose, the works of Starner and Paradiso [1] Kim et al. [2], and Mitcheson et al. [3] are recommended.

6.1.1 Sensor Node Power Requirements

BSN nodes will require power for three main functions: the sensor itself, any signal conditioning or data processing circuitry and the wireless data link. For all of these functions, the power requirements depend strongly on the nature of the measurement. For wearable applications, sensor nodes will usually be monitoring environmental conditions or physiological functions. The data requirements of many such sensors will be modest, since both the resolution and the required update rate are low.

Some examples of typical data rates for measurements suitable for BSN are given in Table 6.1; the highest required rates are a few kbit/s, with most considerably less. Inertial sensors (accelerometers, gyros) are another important node type in BSN – in this case sampling rates are typically 10–100 samples/s per axis, giving total data rates of a few kbit/s or less.

These modest rates are well within the capability of commercially available radio technologies. Popular protocols for BSN include Zigbee, low power Bluetooth and ANT, all having data rates in the range 200–1,000 kbit/s. Their power requirements when active are high (typically mW) compared to likely harvester outputs and will tend to dominate the node power. However, the excess data rate capabilities allow them to be used at low duty cycles, which can reduce the power considerably. In [5], power requirements for the three protocols above are compared; mean currents could be reduced to 10–28 μA (33–92 μW at the 3.3 V supply level), although this was for transmission of only a single byte every 120 s.

The ultimate limits of low power radio indicate that very substantial improvements upon these levels are feasible. Ultra-low power wireless communications is a

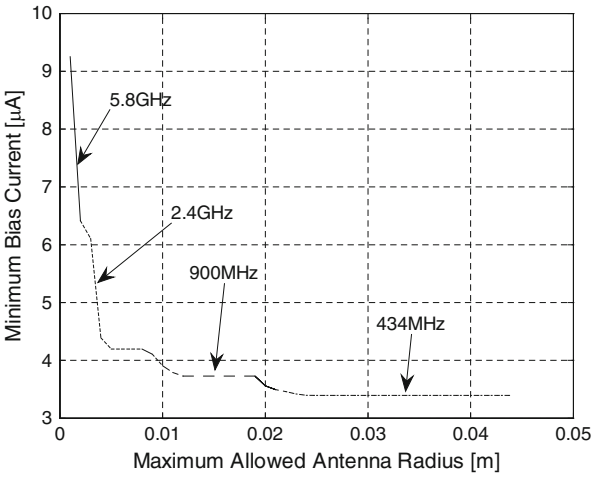


Fig. 6.1 Minimum bias current vs. antenna size for 100 kbit/s transmission over 1 m (From Ref. [6])

major field in its own right, and has been discussed in detail in Chap. 4. The ultimate power limits depend strongly on maximum antenna size, and this has been modelled for a 1 m, 100 kbit/s link, based on a Colpitts oscillator transmitter, in which the antenna coil serves as the inductor in the L-C tank circuit [6]. The results are shown in Fig. 6.1; optimum carrier frequencies are indicated, and for antenna radii of a few mm, required bias currents are in the 4–6 μA range. Since the minimum rail voltage for such a circuit is about 1 V, this corresponds to 4–6 μW. Sub-microwatt transmitter powers are then clearly feasible if a reduced duty cycle is used, given the required data rates as discussed above.

Next we can examine power usage for the sensors themselves. Both temperature and pressure sensing can be done by measuring the voltage drop of a current through a resistor. In principle, the power needs only to be sufficient to overcome thermal noise ($\approx 10^{-20}$ W/Hz at room temperature), and so can be negligible for the low bandwidths of BSN applications (typically 1–100 Hz). Although, to reach these theoretical limits is impractical, even if they can be approached by a few orders of magnitude, the sensor elements in many cases need not make a significant contribution to power consumption compared to the other circuitry.

Many sensor types include active circuitry and so require additional power, but the available devices are reaching ever lower power levels. For example, the ADXL362 3-axis MEMS accelerometer from Analog Devices consumes less than 2 μA at a 100 Hz output data rate, and 270 nA when in motion triggered wake-up mode [7]. Imaging devices typically require substantial power levels, but in [8] a CMOS active pixel image sensor is reported, having 54 × 50 pixels, which consumes only 14.25 μW at 7.4 frames per second. Furthermore, the device incorporates photovoltaic energy harvesting which supplies up to 3.35 μW depending on light levels.

Finally, the interface or signal conditioning electronics will cause some power loading. The most straightforward requirement will be for A-D conversion, and

very low power ADCs have been reported for many years. In 2003, Sauerbrey et al. [9] described a device with power consumption of 0.85 μW for 8-bit sampling at 4 kS/s. More recently, similar power levels have been achieved at 10 kS/s [10]. More functional processing can also be done at low power; in [11] a system-on-chip for biosignal processing is reported with average power consumption of 20 μW per bio-sensing node, and this includes analog amplification, digitisation and some digital signal processing.

Considering BSN nodes as a whole, then, average power consumption of tens of microwatts is achievable with existing technology, and sub-microwatt levels are certainly within the bounds of possibility.

6.1.2 Batteries and Fuel Cells for Sensor Nodes

As stated above, primary or rechargeable batteries are currently used for powering most wireless devices, with primary batteries having somewhat higher energy densities, lower leakage rates and a lower (initial) cost. For body sensor node applications, battery lifetimes of at least a year will be desirable in many instances. With respect to, from our previous analysis, a few μW or tens of μW as an average power requirement, a lifetime of 1 year corresponds to 32 J per μW of average power. Lithium ion batteries have the highest energy density of the main types used for electronics, at about 700–1,400 J/cc for rechargeables [12], with primary cells even higher. Thus, in principle, a power-lifetime of several tens of microwatt-years is achievable for a battery below 1 cc. Consequently, although the finite lifetime remains a disadvantage, and other issues such as operating temperature range, safety and toxicity, may reduce their practicality for BSN applications, batteries remain a very attractive source for sensor nodes. Ultracapacitors are attracting increasing attention for powering small scale electronics, but while their energy densities are much higher than those of conventional capacitors, they remain well below those of batteries [13]. They do, however, offer high lifetime in the sense of the number of charge-discharge cycles.

Exhaustible sources using fuel are also under investigation for small portable electronics, although mainly for higher power levels. The motivation is the very high specific energy of hydrocarbon fuels, e.g. 31 kJ/cc for iso-octane or 16 kJ/cc for pure methanol [14]. As a result, a variety of small heat engines have been developed for portable power applications. In [15], a cm-scale external combustion micro-machined engine is reported which generated up to 200 μW. However, to avoid the moving parts and very high temperatures of a heat engine, fuel cells have been extensively researched for small power supplies. A popular variant for miniaturisation is the direct methanol fuel cell [16]. In these devices, the methanol reacts electrochemically with water at the anode, producing free electrons and protons, the latter being oxidised to water at the cathode after passing through a polymer membrane. Power levels reported were as high as 195 mW/cm^2.

Fuel cells may also provide an attractive type of inexhaustible energy harvesting source for implantable applications. This can be achieved by using bodily fluids as the fuel source, such as glucose and dissolved oxygen in blood. In [17] such a device is reported which is 0.5 mm thick and generates up to 4.4 $\mu W/cm^2$. A key remaining challenge for such devices is operational lifetime.

6.1.3 Ambient Energy Sources

We can move away from finite energy sources by harvesting from the energy types present in the node's environment. There are a variety of such potential sources, and all have been investigated to some degree for energy harvesting applications. The main categories are: motion and vibration, air flow, temperature differences, ambient electromagnetic fields, light and infra-red radiation. In the latter case, solar cells (photovoltaic cells) provide an excellent solution. This is a relatively mature technology, inexpensive and is highly compatible with electronics. The available power levels are typically in the range of 0.1–1 mW per cm^2 in artificial light and 10–100 mW/cm^2 in direct sunlight. Conversion efficiencies of 25 % can be obtained with single crystal silicon devices, or up to 34 % with multi-junction devices [18]. However, the drawback is that the sensor must be in a well-lit location, correctly oriented and free from obstructions. This creates severe limitations for a BSN application.

Gathering radio frequency radiation suffers much less from these geometric limitations. In the VHF and UHF bands [19], for which miniature antennas can operate with reasonable efficiency, field strengths are from about 10^{-2} to 10^3 V/m. We can approximate the power density crudely as E^2/Z_0, where $Z_0 = 377\ \Omega$ is the impedance of free space. For 10 or 1 V/m, for example, this gives 26 or 0.26 $\mu W/cm^2$ respectively. A few V/m thus probably represents the minimum radiation level needed for successful energy harvesting. However, even typical urban environments do not show these levels except in special areas such as in the vicinity of cellular base stations [19]. Recent results suggest that available power levels at, for example, 25–100 m from a GSM base station are below 1 $\mu W/cm^2$ [20]. This suggests that radio frequency harvesting remains an approach with limited applicability for BSN unless a relatively large antenna can be used.

Harvesting thermal energy depends on the presence of temperature differences, e.g. between the surface of the body and the ambient. Applying this temperature difference at the junction between two different materials results in a potential difference, which can drive a current through an external load (the thermo-electric effect). The power available depends on the heat flux through the device and the efficiency; the latter is highly dependent on the temperature difference ΔT achieved, being approximately proportional to ΔT^2. The most popular material for thermo-electric generators (TEGs) is bismuth telluride (Bi_2Te_3), and by using this material, output powers of over 100 $\mu W/cm^2$ at $\Delta T = 10$ K are possible [21]. Practical implementations of thermo-electric generators in BSN applications

have generally reported much lower power levels, although in [22] 1–4 mW was obtained with an array of 16 TEGs sewn into a shirt.

The use of air flow is promising for higher power levels, although with correspondingly higher device size. A micro-engineered axial flow turbine with an integrated electromagnetic generator, having a turbine diameter of 2 cm, has been reported which produced over 1 mW at an air speed of 6 m/s [23]. Further miniaturisation may make a breath powered device feasible.

Harvesting power from vibration or body motion is perhaps the most promising approach, and is being pursued by a large number of research groups. Some advantages of this approach are that devices based on motion harvesting can function both on and in the body, the devices are well suited to relatively efficient transduction techniques, the devices do not require exotic materials and can work in any orientation. We can categorise motion harvesting devices as those that depend on the relative motion between two structures and those that depend only on the absolute motion of the single structure to which the device is attached. The former can offer substantially higher levels of specific power, but will tend to be larger and have a limited number of applicable locations. Heel strike devices, which are installed in the shoe and depend on the force between the landing foot and the ground, are the most investigated of these [24]. The latter we can call inertial devices; these are very flexible as regards size and location, and will be the main focus of the rest of the chapter.

6.2 Inertial Energy Harvesters: Principles and Performance Limits

6.2.1 Energy Extraction Mechanisms for Inertial Generators

The basic operating principle of inertial micro-generators can be described with reference to the generic architecture shown in Fig. 6.2. The inertia of a proof mass m, which is suspended on a spring suspension with spring constant k, causes the mass to move relative to the generator frame with relative displacement $z(t)$ when the frame, with displacement $y(t)$, experiences acceleration. The maximum and minimum values of $z(t)$ are $\pm Z_l$, as imposed by the finite size of the generator. Energy is converted when work is done against the damping force $f(\dot{z})$, which opposes the relative motion of the proof mass and the frame. As discussed above, inertial generators can be used when only one suitable attachment point is available, as they depend on the absolute motion of the frame rather than relative motion between two anchor points. For micro-generators of the size scale of interest for BSN applications, it can normally be assumed that the loading of the "host" structure by the generator is too small to affect the host's motion. This simplifies the analysis, and means that the available power is effectively infinite, in the sense that it is not a limiting factor on the achievable output of the generator.

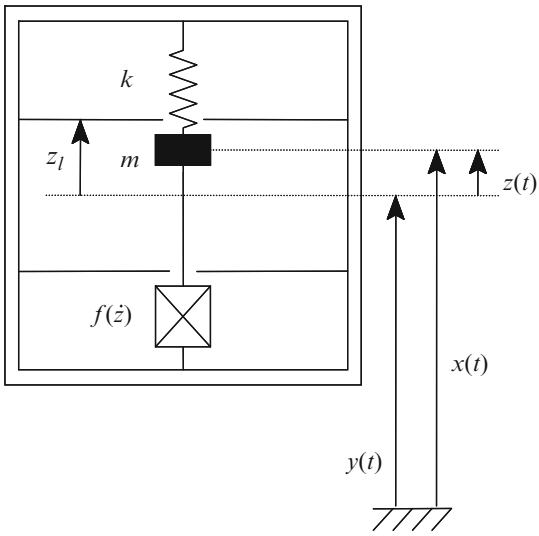


Fig. 6.2 Generic architecture of inertial micro-generators

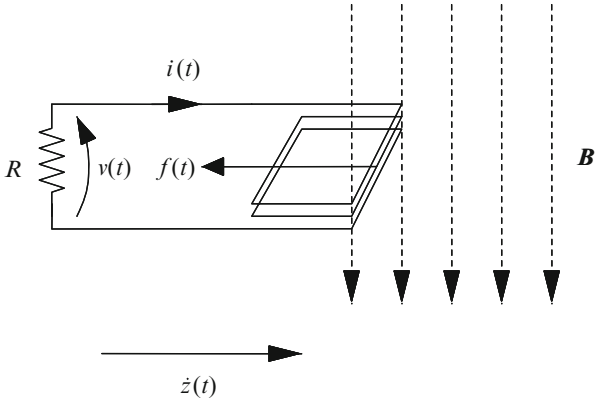


Fig. 6.3 Principle of operation of the electromagnetic transducer

In order to generate useful power, the damper must be an implementation of a suitable mechanical to electrical transduction mechanism. Three such mechanisms have been extensively investigated for this application: electromagnetic, electrostatic and piezoelectric.

Rotating electromagnetic generators have long been in common use, from power levels of a few watts to several hundred megawatts. It is possible to implement the damper of a micro-generator using the same principle, i.e. that described by Faraday's law of induction. This is illustrated in Fig. 6.3. A change of magnetic flux linkage with a coil induces a voltage *v(t)* in the coil, driving a current *i(t)* in the circuit. The combined force *f(t)* on the moving charges in the magnetic field acts to oppose the relative motion which is causing the change in flux linkage, as described by Lenz's law. The mechanical work done against the opposing force is converted

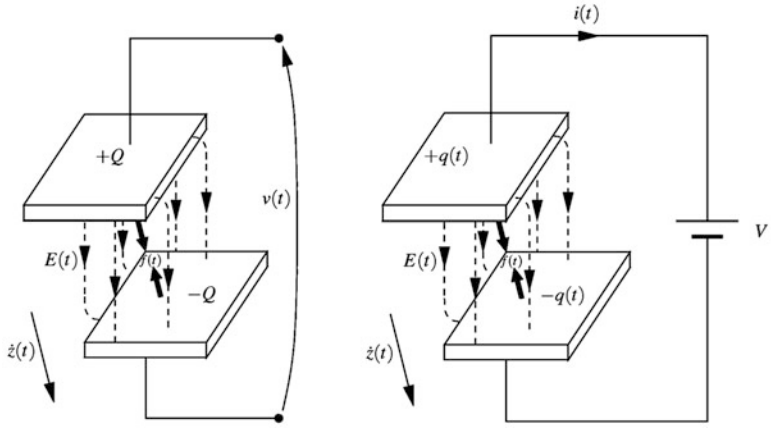


Fig. 6.4 Principle of operation of the electrostatic transducer

to heat in the resistance (external load) of the circuit, and to stored energy in the magnetic field associated with the circuit inductance.

A number of groups have reported micro-generators based on electromagnetic energy harvesting. In [25], for example, a device is described using a coil on a cantilever moving between a pair of permanent magnets. The device volume is only 0.15 cm^3, and 46 μW output power is obtained at about 56 Hz for an excitation of 60 mg (where g = the acceleration of gravity, 9.8 m/s^2). Although the power density is attractive for BSN applications, the frequency is considerably higher than those of body motion, which are typically around 1 Hz. An additional challenge for such devices is caused by the generated voltages being proportional to the rate of change of flux. Since the achievable flux difference over the range of travel of the coil is limited in these small geometries, rapid motion is needed to generate significant voltage. At low frequencies the motion is so slow that even if the generated power levels are acceptable, the voltages are too low for straightforward rectification into useful DC output.

The second transduction method, electrostatic, involves the use of the forces between opposite charges on a pair of electrodes which move relative to each other (i.e. one fixed to the frame, and one to the proof mass). Typically, two possible modes of operation are identified: constant charge and constant voltage. The first involves moving a fixed amount of electric charge through an electric field and thus increasing the electrical potential of that charge, as illustrated in the left diagram of Fig. 6.4. For a parallel plate structure with a variable separation and constant overlap, with a negligible fringing field, the field strength is proportional to the (constant) charge and thus, the energy density of the electric field is independent of plate separation. As the electrode separation increases, additional electrical potential energy is stored in the increased volume of electric field. Alternatively, if the plates are moved relative to each other with a sliding motion at a constant separation, mechanical work is done against the fringing field and

there is an increase in stored electrical energy because the electric field strength increases with the reduction in plate overlap. The energy density of the field (proportional to the square of field strength) increases faster than the volume of the field decreases.

The other extreme of operation is constant voltage, illustrated in the right diagram of Fig. 6.4. Moving the relative positions of the plates (either due to sliding or normal movement) changes the capacitance between the electrodes under a constant voltage. If the plate separation is increased with a fixed overlap, the electric field strength between the plates falls, causing charge to be pushed off the plates into an external circuit as a current flow *i(t)*. If the plates are moved with constant separation and changing overlap, the field strength stays constant, but current is again forced to flow into the source because the volume of the field decreases. In both cases, the mechanical work done is converted into additional electrical potential energy as an increased charge in the voltage source.

Because of practical implementation constraints, such as non-zero conductance (in the constant charge case) and non-ideal voltage sources (in the constant voltage case), real electrostatic transducers work somewhere between these two extremes, albeit in many cases very close to one of them, and both types have been reported in the literature for implementations of micro-generators. For constant charge operation, variable gap motion gives a constant force, while for constant voltage devices, a constant force is provided by variable overlap motion. As will be discussed below, maximising the output power depends on achieving the highest transduction force, throughout the proof mass displacement, that does not actually prevent this displacement. Such power maximisation is not likely to be achieved if the applied force is strongly position dependent. It should also be noted that electrostatic transducers, unlike electromagnetic ones, normally need the application of an initial "priming" voltage to generate the damping force, and this represents an energy cost in the complete cycle. While the priming energy can, in principle, be small compared to the extracted energy, it adds to the device complexity. Electret designs, employing a buried fixed charge to eliminate the need for cyclic pre-charging, have been extensively investigated.

Both variable overlap and variable gap devices have been reported in the literature. The latter, particularly if micro-engineered, are usually not simple parallel plates, but use the so-called "comb-drive" structure of inter-digitated electrodes, as this allows the two electrodes to be defined in the same level on-chip. Such structures are commonly used in MEMS technology for both electrostatic sensing and actuation. A key difficulty in energy harvesting appears to be in achieving the combination of high capacitance and substantial displacement that is needed to maximise power output. High capacitance values are useful as they allow lower voltages to be used for the same level of force.

Probably the most extensively investigated transduction technique for inertial harvesters is the piezoelectric effect [26]. This is a phenomenon whereby a strain in a material produces an electric field across that material, and conversely an applied electric field can produce a mechanical strain. The first of these modes can be used to realise micro-generators. When the material is strained, some of the mechanical

work done on the device is stored as elastic strain energy, and some in the electric field brought about by the space charge. Only a small class of materials exhibits strong piezoelectric effects, and of these the most commonly exploited is the ceramic $PbZr_xTi_{(1-x)}O_3$, or lead zirconate titanate (PZT). This can be machined as a conventional powder-based ceramic into monolithic pieces, or deposited as a thin film for micro-engineered devices.

Piezoelectric devices offer significant advantages, especially since relatively high voltages can be obtained with modest strains and no priming is needed. However, the need for integration of a specialised material, and its associated electrodes, significantly adds to the fabricational challenge for micro-engineered devices, and most reported piezoelectric harvesters are assembled from conventionally engineered parts. Furthermore, the geometric possibilities are limited: the material has low maximum strain, so that a high leverage factor is needed to get large proof mass displacement. Most reported devices have used a layered cantilever structure with the proof mass at the free end. The limited strain can be overcome by using a polymeric piezoelectric such as polyvinylidene fluoride (PVDF), but the electromechanical coupling factors available are much lower than for ceramics.

These transduction methods provide a useful way of classifying devices of inertial energy harvesting. Another useful distinction, particularly for the BSN application regime, is between devices that are mechanically resonant and those that are not. Resonance allows an internal displacement amplitude greater than the source motion amplitude to be achieved, and this can be useful when the source amplitude is slight, as it may well be for high frequency (e.g. machine) vibration. However, human body motion amplitude is likely to be greater in most instances than the dimensions of BSN nodes. Body motion also has a complex and widely varying spectral character, which cannot be effectively exploited by a device with a narrow and fixed operational frequency range.

Many reported miniature inertial energy scavengers have been resonant; this is partially because in MEMS and other small implementations, support of moving parts by flexure suspensions is more practical than other options such as sliding or rolling bearings. Since the proof mass is thus a mass on a spring, it is inherently a resonant structure. However, it is increasingly recognised that broadband operation is essential for many practical applications, including BSN, and researchers are investigating a variety of approaches to achieving this. One approach is to use resonant structures but to make them tuneable. For example in [27], the gap between a magnetic proof mass on a cantilever tip and a tuning magnet is varied; this alters the stiffness of the beam and consequently the resonant frequency. This approach is useful for vibration sources such as rotating machines which have distinct but somewhat varying frequencies, but is unlikely to be suitable for the complex motion of the human body. Another approach is to make the mechanical operation nonlinear, which can result in high oscillation magnitude over an extended range of excitation frequency. In [28], a micromachined electrostatic harvester is implemented with a mechanically nonlinear suspension, and the operational bandwidth is shown to increase by more than an order of magnitude compared to linear operation. Finally, a range of devices have been reported

using frequency up-conversion. Here typically the proof mass moves slowly and non-resonantly, but it transfers its energy to resonant transducer structures such as piezoelectric beams. An example is described in Sect. 6.3 below.

6.2.2 Performance Limits

A comprehensive analytical framework for inertial energy scavengers was reported by the present authors and co-workers [29]. This analysis allows the different architectures to be compared quantitatively, and has derived the achievable power levels and their dependence on both source and device characteristics. Key practical constraints were also analysed. The results of that study are summarised here.

Two resonant and one parametric generator topologies were considered. Of the resonant type, one is damped by a force which is proportional to velocity, the velocity-damped resonant generator (VDRG), and the other is damped by a constant force, the Coulomb-damped resonant generator (CDRG). Of the non-resonant, non-linear generators, only the Coulomb-force parametric-generator (CFPG) is considered here, as the velocity-damped parametric generator was found to be ineffective. Variations of VDRGs and CDRGs have been widely reported in the literature. Broadly speaking, the electromagnetic and piezoelectric devices correspond to VDRGs, and the electrostatic devices correspond to CDRGs. In this analysis, the resonant generators were considered to operate in modes in which the proof mass does not strike the end-stop limits, i.e. $-Z_l < z(t) < Z_l$, and thus the only forces which act on the mass are the inertial, spring and damping forces, and gravity.

The source motion was assumed to be harmonic, with amplitude Y_0 and frequency ω, from which the maximum acceleration a_{max} can easily be derived as $\omega^2 Y_0$. The fundamental parameters determining the generator output are its proof mass m, resonant frequency (if any) ω_n, and the maximum internal displacement Z_l. From very basic considerations we can derive a maximum power for any energy scavenger driven by harmonic source motion. The damping force by which the energy is extracted cannot exceed the inertial force on the proof mass, ma_{max}, otherwise the mass will not move. If energy is extracted in both directions, and taking the internal motion range as being $2Z_l$, we derive a total energy per cycle of $4Z_l m a_{max} = 4Z_l m\omega^2 Y_0$. To convert this to power is simply a matter of dividing by the excitation period $2\pi/\omega$, giving a maximum power:

$$P_{max} = 2Y_0 Z_l \omega^3 m/\pi \tag{6.1}$$

Examples of this limit are shown in Fig. 6.5, for harvesters at sub-cm^3 size, and assuming 1 g of acceleration and a silicon proof mass (2.5 g/cm^3). For body motion frequencies of a few Hz, achievable powers are a few microwatts or less, so this indicates that even for sensors with very low power requirements, motion harvesters will probably need to be above 0.1 cm^3 in volume to be effective, regardless

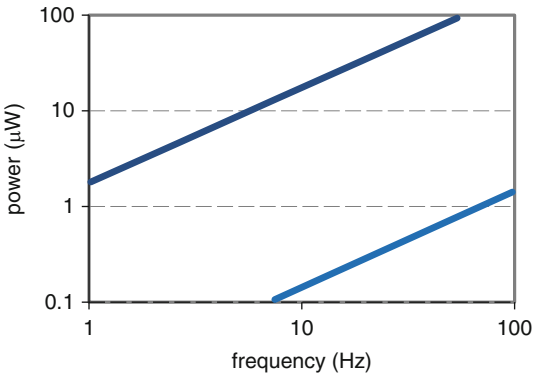


Fig. 6.5 Ultimate power limits for a MEMS scale inertial harvester, for device sizes of 10 × 10 × 2 mm (*upper line*) and 3 × 3 × 0.6 mm (*lower line*)

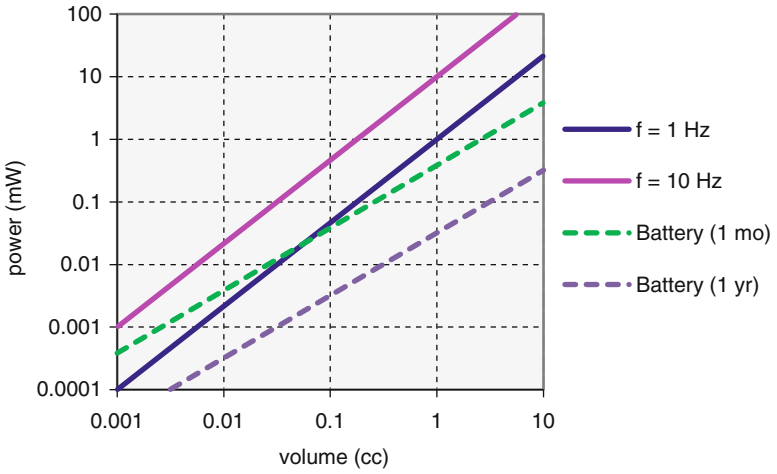


Fig. 6.6 Ultimate power limits for inertial harvesters, for excitation frequencies as indicated, vs. average power of Li ion batteries for lifetimes as indicated

of the device structure or transduction method. This can be seen more clearly in Fig. 6.6, where the ultimate power vs. the size of harvesters operating at 1 or 10 Hz is shown vs. the power of Li ion batteries of that size having 1 month or 1 year operating lifetime. It can be seen that harvesters much below 0.1 cm^3 have difficulty exceeding the power of batteries even for long operating times.

If the proof mass motion is also harmonic, as in a resonant device, the maximum power is in fact somewhat less than that indicated in Eq. 6.1, since the acceleration is not a_{max} for the whole travel and so the transduction force must be reduced accordingly. But Eq. 6.1 does provide, on the basis of fundamental considerations, an upper bound on the power of an inertial energy scavenger of any architecture, construction, transduction mechanism or operating mode. It shows the linear dependence on mass and on travel range, and also the very strong dependence on

frequency, indicating the serious challenge of achieving useful power levels in the low frequency environment of BSN.

In the analysis, it was found that for idealised cases of the three architectures considered, the optimal output power can always be derived as a function of the two dimensionless parameters Z_l/Y_0 and ω/ω_n, and can be normalised to a characteristic power $Y_0{}^2\omega^3 m$.

Analysis of the output power of velocity-damped generators is a matter of integrating the product of the damping force and the incremental displacement and averaging this over a cycle. Then the optimum power can be found by choosing the damping coefficient to maximise this value. However, if resonant motion is assumed without regard to travel limits, a derivation is obtained which implies infinite power at resonance, although a corresponding infinite internal displacement is implied. A realistic assessment requires that the damping force be optimised only up to the limit imposed by the maximum travel range. Thus, the achievable power of an ideal VDRG takes two forms; firstly, if the damping can be optimised without the displacement constraint being breached, and secondly if a higher than optimum damping is required in order to prevent collision with the end-stops. Full solutions to the two cases are given in [29]; for operation at the resonant frequency, both cases reduce to:

$$P_{res} = \frac{1}{2} Y_0 Z_l \omega^3 m \tag{6.2}$$

As anticipated, this is just less than the ultimate limit given by Eq. 6.1; a factor $\pi/4$ less in fact.

The Coulomb damped devices do not form linear systems, because the damping force is discontinuous at the boundaries (where the direction changes), and so analytical solutions are not as straightforward to obtain. Nevertheless, closed form solutions to the equations of motion for the CDRG do exist, from which the optimal damping coefficients, and the achievable power levels, can be derived [29]. Just as for the VDRG, the maximum power depends on whether or not the optimal damping is limited by the internal displacement constraint, but in either case the solution reduces to Eq. 6.2 for operation at resonance.

The analysis of the CFPG is essentially the same as that used to derive Eq. 6.1; the Coulomb (electrostatic) force is constant for the whole travel distance, and so the energy per transit is just the applied force times the travel range. However, a correction is needed to Eq. 6.1 because the force applied, in the case of harmonic source motion, cannot be equal to a_{max} since this acceleration is reached only instantaneously at the extremes of the frame displacement. Thus we reduce the damping force to βa_{max}, where β is a dimensionless coefficient, giving:

$$P_{max} = \frac{2\beta}{\pi} Y_0 Z_0 \omega^3 m \tag{6.3}$$

In this general formulation, the displacement limit of the device, Z_l, has been replaced by the actual internal motion amplitude Z_0. Thus, determination of the

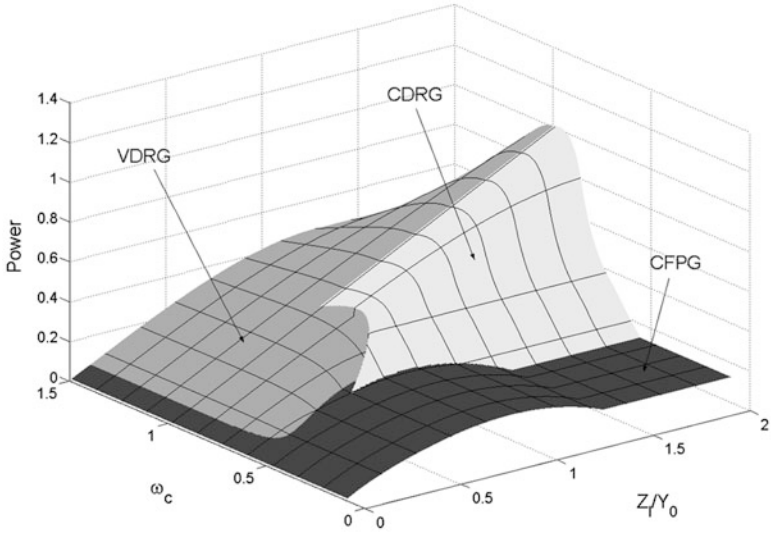


Fig. 6.7 Comparison of inertial energy harvesting architectures, with normalised maximum power output

output power requires not only the optimal value of β, but also the corresponding travel range Z_0, to be determined. For large source displacement amplitudes, however, it can be shown that the optimal β value is that which just allows the full travel range to be traversed, so that $Z_0 = Z_l$. Specifically, this proves to be the case for $Z_l < 0.566Y_0$, i.e. the source motion amplitude is more than about double the internal displacement limit. This will almost certainly be the case for wearable or implanted devices excited by body or limb motion. It may not be the case for implanted devices driven by cardiac motion.

With expressions for the achievable power levels of the three main device architectures, we can compare them and determine which is the most effective for a given operating regime. Figure 6.7 shows the result, indicating the operating regions where each architecture is superior and what the maximum power level is, normalised to $Y_0^2\omega^3 m$.

Several general conclusions can be drawn from Fig. 6.7. For large devices or low source amplitudes ($Z_l/Y_0 > 0.1$), the resonant devices are superior, except where the frequency of operation is more than two times below the achievable resonant frequency, in which case the parametric generator is preferred. The CFPG is superior for all cases where the device size is well below the source motion amplitude. As mentioned above, this is likely to be the case for many BSN applications. Furthermore, the CFPG, being non-resonant, can operate effectively over a wide range of source frequencies (as would be expected with BSN) without the need for dynamic tuning.

All the analysis above has been for harmonic source motion. As stated previously, body motion is of complex and varying spectral form, and for that reason

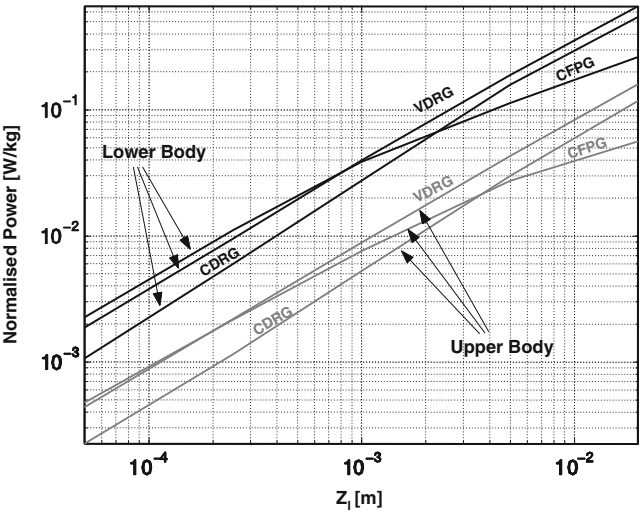


Fig. 6.8 Comparison of architecture performance for generators mounted on the upper and lower body. Output power is normalised by the value of proof mass (From Ref. [30])

analysis of the output of inertial energy scavengers with realistic body motion excitation has been carried out [30]. Motion waveforms were captured for three orientation axes, at each of a number of body locations, using accelerometers, and the power output of the various scavenger architectures, for a range of sizes, were simulated and compared using these waveforms. The results are shown in Fig. 6.8. As anticipated, the CFPG devices are superior for small devices, particularly for lower body locations, where the displacements were greater.

6.3 Inertial Energy Harvesters: Practical Examples

6.3.1 Electrostatic Harvesters

Most of the electrostatic harvesters reported have been micro-engineered, typically in silicon using the techniques of MEMS (micro-electro-mechanical systems). This is because high electric field strengths are needed to get strong transduction forces, and unless these are in narrow gaps, the voltages required are excessive. MEMS is an ideal technology for producing small, well defined gaps between moving mechanical parts. Furthermore, MEMS has been very successful for the production of inertial sensors, particularly accelerometers, and these have very similar structures to inertial harvesters. Both externally primed and electret devices have been reported, and our first example is one of the former [31, 32]. It can be described as a parametric generator according to the classification of Sect. 6.2.2. Having a

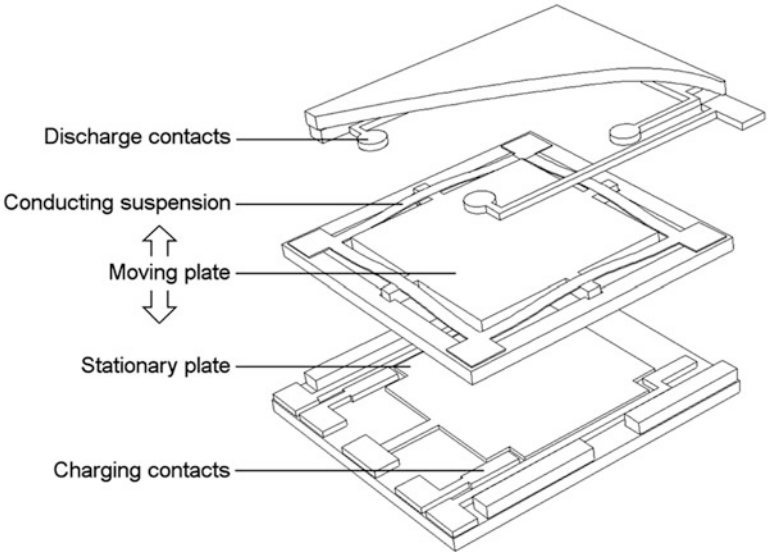


Fig. 6.9 Exploded view of the generator construction (From Ref. [31])

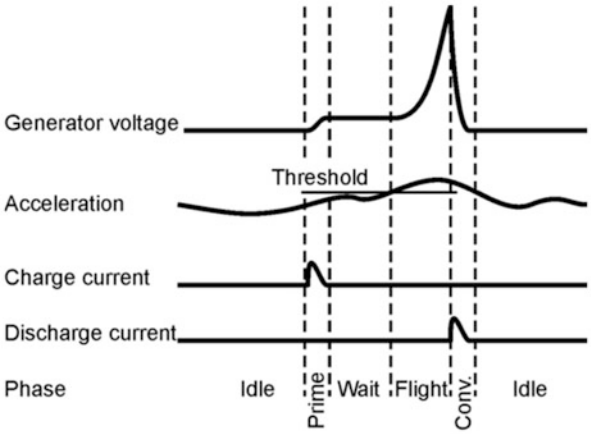


Fig. 6.10 Phases of operation of the electrostatic generator of [31]

nonlinear mode of operation, it is essentially a parallel plate capacitor with varying gap, operating in constant charge mode.

An exploded view of the device structure is shown in Fig. 6.9, with the phases of operation illustrated in Fig. 6.10. The structure consists of a moving plate attached to a frame by a low stiffness suspension, a bottom plate containing a counter-electrode and charging studs, and a top plate with discharge contacts. The operation cycle proceeds as follows:

- At the start of a generation cycle, the capacitor is at its maximum capacitance position, i.e. minimum separation (*idle phase*).
- The capacitor is pre-charged to a relatively low voltage which will give the optimal value of β (i.e. the optimal holding force) for the current operating conditions (pre-charging or *priming phase*).
- The generator frame is accelerated by the input motion. The proof mass moves along with the frame (*wait phase*) until the magnitude of the frame acceleration is sufficient for the inertial force on the proof-mass to overcome the electrostatic force of attraction between the plates, at which point the proof-mass separates from the frame of the generator and starts to move relative to it. At the point of separation the electrical contact between the moving plate and the charging circuit is broken.
- The relative movement proceeds, and increases the volume of the electric field between the capacitor plates as they separate. Because the charge on the plates remains constant, the energy density of the electric field remains constant and so the electrical potential energy stored increases with the volume of field (*flight phase*).
- The moving plate and proof-mass slow down relative to the generator frame as they approach maximum plate separation. Under optimal conditions for electrical energy generation, the relative velocity tends to zero as the maximum displacement is reached.
- Whilst the plates are separating, the voltage across them increases in proportion to the separation (because the electric field strength is constant).
- The variable capacitor, now at its lowest capacitance and highest voltage, is discharged through power conversion circuitry and the energy is available to drive a load (conversion phase).

The device was fabricated using a 3-wafer construction. The central wafer contains a silicon proof mass, forming one plate of the variable capacitor, along with a silicon frame and a polyimide suspension, metallised for electrical contact. The proof mass is about 0.12 g, and measures $\approx 11 \times 11$ mm $\times$ 0.4 mm thick. It is separated from the frame by deep reactive ion etching (DRIE), throughout the whole wafer thickness, after patterning of the suspension. Polyimide was chosen to give the required very low suspension stiffness, as discussed above.

The bottom wafer is glass, to minimise the parasitic capacitance. It includes the fixed electrode of the variable capacitor itself, and also the charging studs and spacers for the moving plate and middle wafer, the studs being deposited by electroplating. These set the minimum gap at about 6 μm, giving a theoretical starting capacitance of $\approx$ 180 pF. The top wafer is also glass, and has studs for discharge. Spacer studs 300 μm thick, fabricated from SU8 polymer on the top and bottom wafers, set the layer separation, and thus the proof mass travel distance. The minimum (discharge position) capacitance was measured at 5.5 pF. Figure 6.11 shows the completed device.

The device was tested on a low frequency shaker platform, for frequencies in the range 10–100 Hz. Motion was monitored using a linear displacement transducer or

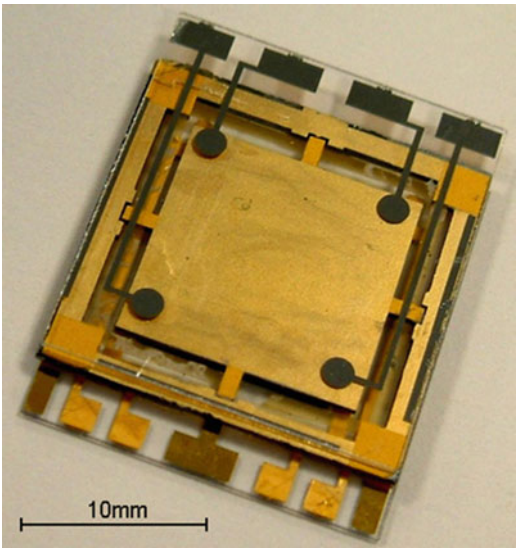


Fig. 6.11 Prototype CFPG fabricated using silicon micromachining

an accelerometer, at lower and higher frequencies respectively. As the pre-charge voltage (and thus the holding force) was increased, the release point occurred later in the cycle, as expected. Depending on operating frequency and amplitude, output voltages of up to 220 V were obtained, corresponding to a net generated power of 120 nJ per cycle.

The power obtained with this device was significantly below theoretically achievable values. This was partly due to motion of the proof mass in unwanted degrees of freedom; in particular, tilting motion. This reduces the capacitance ratio, by decreasing the charging capacitance if the moving plate is not parallel to the fixed plate and does not contact all the charging studs, and by increasing the discharge capacitance. The design also suffers from inherent non-idealities. For example, the proof mass motion occurs in the shortest direction (out-of-plane), and so is very limited. Furthermore, silicon is a relatively low density material and so not ideal for the proof mass.

Subsequently, a variant electrostatic device (Fig. 6.12) has been reported using an external proof mass in the form of a steel pin that rolls across the substrate, which addresses to some extent both these shortcomings [33]. The device was integrated with a temperature sensor as the priming voltage, and a loop antenna acting as the inductor in a resonant tank circuit in which the output energy is discharged, with the output pulse amplitude thus being proportional to the senor output voltage. In this way, a wireless sensor node was successfully demonstrated which requires neither energy storage nor digital circuitry.

A wide range of electret based harvesters have also been reported, and these have typically integrated proof masses moving in the in-plane direction. Recent progress is reviewed in [34], and a typical device is shown in Fig. 6.13. This has

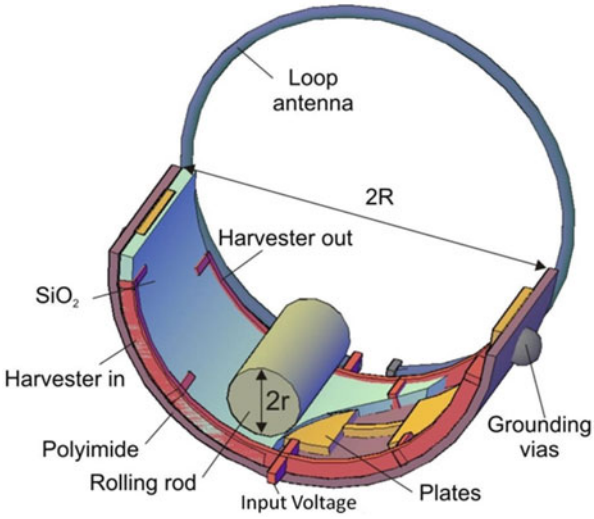


Fig. 6.12 Rolling mass electrostatic harvester with integrated wireless transmission

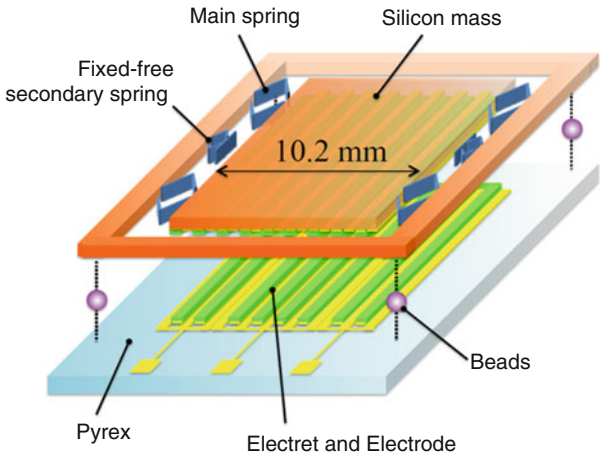


Fig. 6.13 A diagrammatic overview of an electret harvester (From Ref. [34])

dimensions 18.5 mm × 16.5 mm, with a Si proof mass supported by parylene high-aspect-ratio springs. The electrets and electrodes on opposing faces are patterned in stripes as shown, which results in a change in the mirror charge on the electrodes as they move in and out of phase with the electrets. This causes a current to flow in the external circuit. For excitation movement of 2 mm p-p at 63 Hz, an output power of 17 μW was estimated.

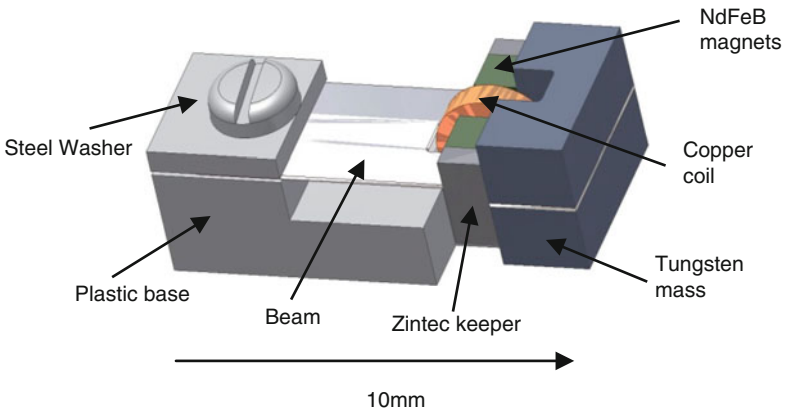


Fig. 6.14 Electromagnetic harvester (From Ref. [25])

6.3.2 *Electromagnetic Harvesters*

The harvester referred to in Sect. 6.2.1 above [25] is a good example of a small electromagnetic harvester, and is illustrated in Fig. 6.14. As a size reference, the cantilever beam shown is 9 mm long. As mentioned, it generates 46 μW output power at about 56 Hz. To maximise the output voltage and in order to make power conversion to DC easier, up to 2,300 turns were used on the coil, resulting in a maximum of 400 mV rms output, which is near the minimum where efficient conversion circuits are possible. However, operation at the lower frequencies available in the human body will lower the output voltage substantially. This device, being resonant, also has a fairly limited bandwidth for BSN.

Electromagnetic transduction is used in virtually all conventional generators, from power stations down to bicycle dynamos. At very small sizes, however, some difficulties arise, such as the increasing ohmic losses in coils as the wire diameter reduces. Smaller electromagnetic harvesters have been reported, such as the planar device illustrated in Fig. 6.15 [35]. This uses in-plane motion of an upper array of planar coils on a printed circuit board acting as the proof mass, moving relative to a multi-pole magnet on the substrate. Output was measured for discrete excitation (by displacing the mass 2 mm and releasing it) rather than oscillation, and 1.1 mJ was generated per excitation, with an estimated efficiency of 9 %. At a true MEMS scale with silicon integration, a number of devices have been reported by various groups, but the power levels have generally been too low for realistic application.

6.3.3 *Piezoelectric Harvesters*

Piezoelectric devices are becoming increasingly popular for small and micro-scale motion energy harvesters. Researchers are increasingly focusing on low frequency excitations which are representative of most realistic applications, and the need for

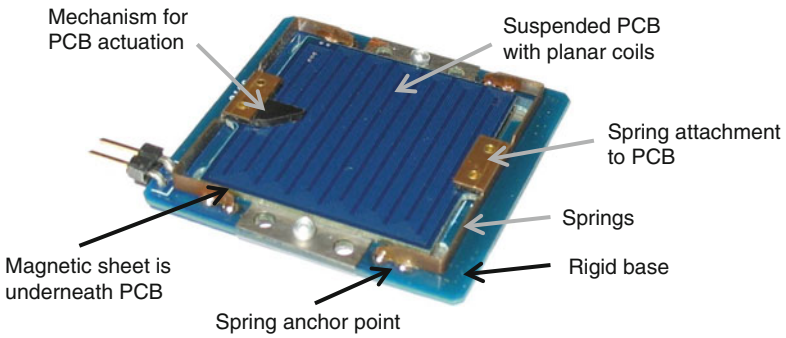


Fig. 6.15 MEMS electromagnetic harvester (From Ref. [35])

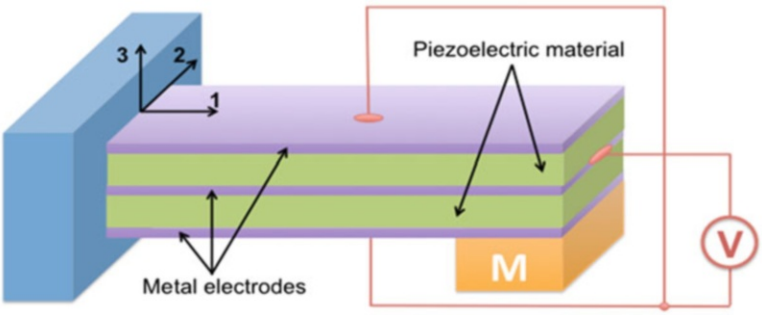


Fig. 6.16 Typical piezoelectric harvester structure (After Ref. [2])

output power to be in a form suitable for efficient conversion to DC to power load electronics or for battery recharging. Thus the relative ease of obtaining significant output voltages from small, low frequency piezoelectric devices is becoming an increasingly key advantage. A typical structure is the vibrating cantilever with a proof mass at the tip, such as that shown in Fig. 6.16 [2].

Such devices have been fabricated in a wide range of sizes, but their resonant frequencies tend to be excessive for BSN applications, particularly for the smaller devices. A table in [2] compares recent devices of this form having active volumes up to about 30 mm^3; of 14 of these, only one had an operating frequency below 100 Hz, and that produced 14 μW at 76 Hz. The narrow bandwidth of such devices is another barrier to their adoption for BSN applications. Bandwidth can be increased in resonant devices by increasing the strength of the electrical damping, but this is generally impractical with piezo devices because of limited electro-mechanical coupling strength.

Coupling strength can be enhanced by the use of active conversion circuits that apply a synchronous pre-bias to the piezo element, and this is discussed further in

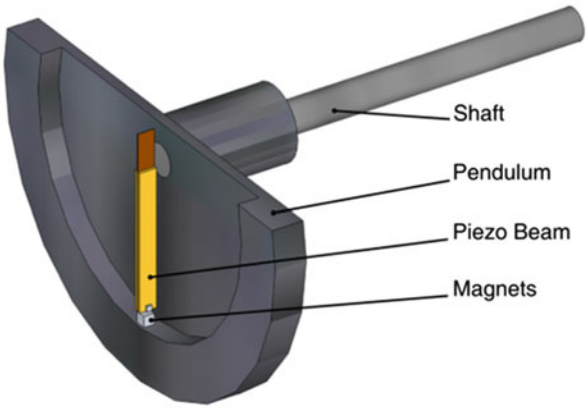


Fig. 6.17 Piezoelectric harvester with rotating proof mass and plucked piezoelectric transduction (From Ref. [36])

Sect. 6.4. Alternatively, both the narrow bandwidth and the high resonant frequency can be overcome using the frequency up-conversion method, as mentioned in 6.2.1. An example is illustrated in Fig. 6.17, from [36]. In this case the proof mass is continuously rotating around a bearing, so it has no resonant frequency or hard travel constraints. This is the same form of proof mass as is used in self-winding wristwatches, including electronic variants such as the Seiko Kinetic, which uses electromagnetic transduction and a high ratio gear train to achieve sufficient voltages from the generator. In the device of [36], the transduction is done using a piezoelectric cantilever beam with a high resonant frequency (compared to the excitation) and a high Q. Both the beam tip and the rotating mass have magnets attached, and as these magnets pass each other the proof mass "plucks" the piezo resonator, which then rings down and passes its stored energy into a conversion circuit. The device has a swept volume of 3.7 cm^3, and produces over 2 μW when excited at 2 Hz. Smaller and more efficient variants are now under development.

6.4 Power Electronics for Energy Harvesters

Power electronic circuits play a critical role in energy harvesting systems. Firstly, together with the transducer, they enable a maximum transfer of power from the source (mechanical or otherwise) by performing an impedance match. Secondly, they allow voltage conversion (up or down) in order to interface the transducer with a storage element (as most harvesting sources are intermittent); and thirdly they can be used to regulate the voltage from a storage element to power the sensing, signal processing and communications aspects of a BSN node. The requirements for the power processor differ significantly between the transduction mechanisms. A brief overview of the requirements and state of the art for each transducer type will now be discussed, while a more detailed review can be found in [37].

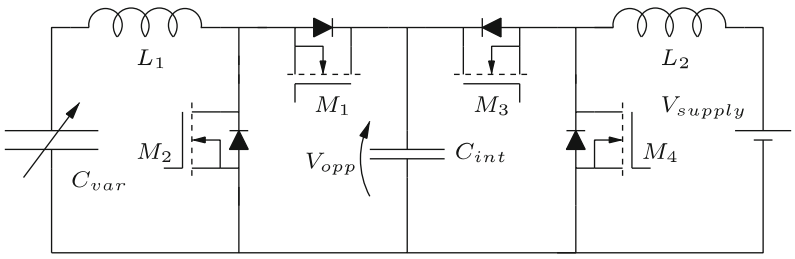


Fig. 6.18 Example power interface circuit for constant voltage electrostatic harvesters (From Ref. [38])

6.4.1 Electrostatic Harvester Interfaces

These devices tend to operate in constant charge or constant voltage mode, typically using MEMS comb-drives with changing overlap for constant voltage mode and either comb-drives or parallel plate capacitors with varying separations in constant charge mode. The most successful electrostatic harvesters now tend to be of the constant voltage type, primed using electrets [34] and so here we will concentrate our discussion on the constant voltage mode of operation.

Figure 6.18 shows a basic circuit topology for interfacing with a constant-voltage mode electrostatic harvester. In such systems, the variable capacitor (C_{var}) is precharged to a voltage which gives the optimal damping force between the electrodes (giving what is effectively a mechanical impedance match to the source), from an intermediate energy storage element, C_{int}, which is kept at the optimal voltage. Mosfets M_1 and M_2 are used to pre-charge the variable capacitor when it is in its maximum capacitance position, and then M_1 is left short circuit whilst the capacitance decreases due to the mechanical motion, pushing charge into the intermediate storage element. At the end of the cycle, the additional energy that has been stored in the intermediate capacitor is discharged into a low voltage element, V_{supply}, by the action of M_3 and M_4.

There are trade-offs in the circuit design. The basic requirements for the mosfets are that they are able to block the maximum voltage on the variable capacitor, which sets the doping and length requirements of the devices; however, the mosfet widths can be optimally chosen: too large and leakage currents decrease the system performance; too small and conduction losses will be significant. When an optimal width is chosen (via simulations across all practical values) the maximum possible effectiveness (a measure of how close the performance of a particular device is to its theoretical limits [3]) of a constant voltage electrostatic harvester can be determined, and is shown in Fig. 6.19.

As can be seen, the maximum effectiveness of electrostatic harvesters (with custom designed mosfets) is acceptable to good across a wide operating envelope of harvester size and acceleration. The compatibility of MEMS fabricated electrostatic harvesters with CMOS, combined with these relatively high effectiveness values when placed in a system with the power processor, makes them a popular harvester solution.

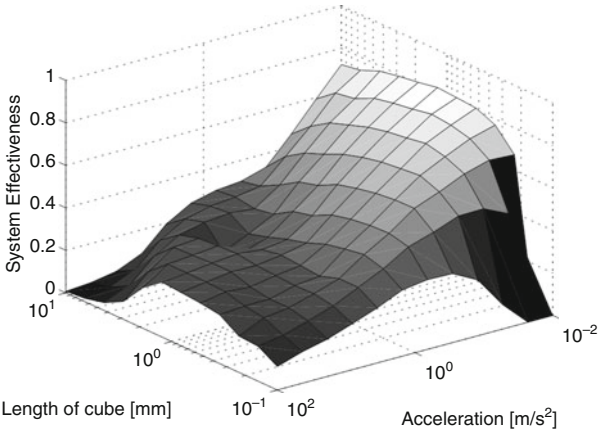


Fig. 6.19 Maximum effectiveness of an electrostatic harvester (1 kHz excitation) (From Ref. [38])

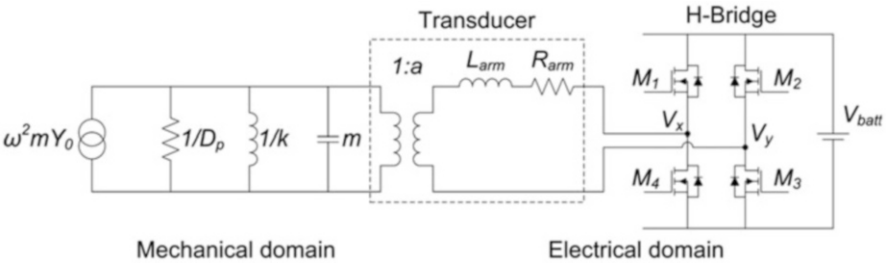


Fig. 6.20 Model of an inertial harvester, electromagnetic transducer and full-bridge for frequency tuning (From Ref. [40])

6.4.2 Electromagnetic Harvester Interfaces

Electromagnetic harvesters possess some desirable characteristics: they tend to be reliable, and with a suitable power electronic interface, their resonant frequency can be tuned [39]. This is very useful in BSN applications where the human activity level, and along with it, the excitation frequency, can change rapidly. The most basic requirement for an electromagnetic harvester is rectification of the generated AC waveform into DC. If this is accomplished using an active rectifier arrangement, such as a full-bridge, both real power, P, and reactive power, Q, can be exchanged between the mechanical and the electrical systems [40]. Control of P and Q (by modifying the input impedance of the H-bridge) allows the damping and resonant frequency to be controlled, respectively.

Figure 6.20 shows an electrical equivalent model of an inertial electromagnetic harvester with a full bridge circuit attached to the electrical terminals of the transducer. The input impedance of the H-bridge can, with suitable operation of the mosfets, be arranged to emulate a reactive impedance (which will have the same

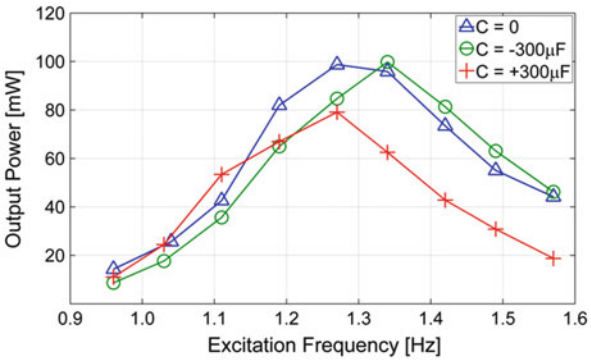


Fig. 6.21 Output power of inertial harvester with emulated reactive loads implemented with a full bridge circuit (From Ref. [40])

effect as modifying the spring constant or the effective proof-mass of the harvester) and thus change the resonant frequency. Example results of this type of frequency tuning system are shown in Fig. 6.21.

As can be seen, at low frequencies, the power is increased by adding a positive capacitive load, corresponding to increasing the effective mass and lowering the resonant frequency, and at higher frequencies, the power is increased by adding a synthesised negative capacitance, effectively reducing the proof mass.

6.4.3 Piezoelectric Energy Harvester Interfaces

In low frequency, high amplitude applications of energy harvesters, as is typically the case for BSN applications, the main obstacle when using the piezoelectric transducer is obtaining the sufficiently high electrical damping forces required to maximise the power generated. When a piezoelectric transducer is connected directly to a simple passive rectifier circuit to charge a battery or capacitor, the total work that can be done by the mechanical force is limited because of the capacitive nature of the piezoelectric element. This can be overcome by using charge modification techniques [41, 42] and a circuit for one of the implementations is shown in Fig. 6.22.

The basic principle for the operation of this scheme is that a charge is placed on the transducer when the proof mass is at either extreme of the motion range, so that the force obtained from this charge is in a direction to oppose the relative motion between the mass and frame. This increases significantly the work that can be done by the transducer and hence the power generated. The value of electrical damping force can be modified by altering the quantity of charge placed on the transducer at each end-point of travel. This is achieved by altering the value of V_{cc} in Fig. 6.22. Techniques such as this have been shown to increase the output power of piezoelectric harvesters by up to ten times [43].

Power electronics is clearly an important part of the energy harvesting system. It is critical in performing the basic functions of rectification and interfacing to energy storage elements, but also, when designed with knowledge of the transduction

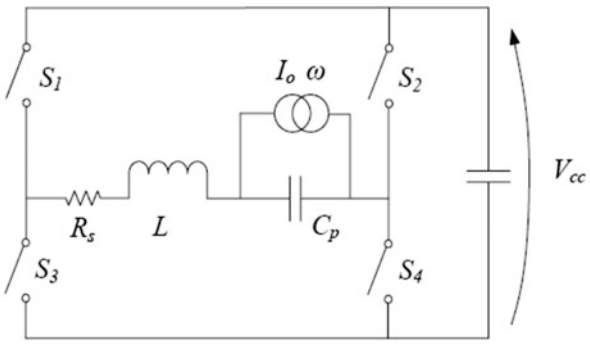


Fig. 6.22 Single supply pre-biasing circuit (From Ref. [42])

mechanism, allows the power density of energy harvesters to be significantly increased and their operational envelope widened.

6.5 Wireless Power Delivery

An alternative method to that of providing a local power supply in the form of an energy harvester or a battery is to deliver power to a sensor node wirelessly. Three methods have been considered for wireless power delivery to sensors and medical implants. In decreasing order of power delivery capability under typical usage, they are:

- Near field inductive power transfer
- Acoustic power transfer
- Far field radiative power transfer

The first of these methods has been used for decades, e.g. for charging electric toothbrushes. These systems rely on relatively strong coupling between the transmitter and receiver, meaning the device is often in physical contact with the charging station, giving little or no mobility when charging. These systems are essentially transformers where primary and secondary couple relatively tightly when the device is placed on the charging station. More recently, there has been significant interest in inductive near-field power transfer using secondary resonance [44]. These systems can achieve high charging efficiency without the coils being in very close proximity, and thus allow some degree of mobility whilst power is being delivered, making the technique highly attractive for powering sensors in the BSN context.

The second method involves transferring power using acoustic energy, and can be used for powering devices through media such as body tissue, as long as the system is impedance matched to the medium in question. The final method, radiative power transfer, involves using a transmitter which radiates RF energy in the usual way, which can then be picked up and rectified by a receiving antenna and rectifier, known as a rectenna [45].

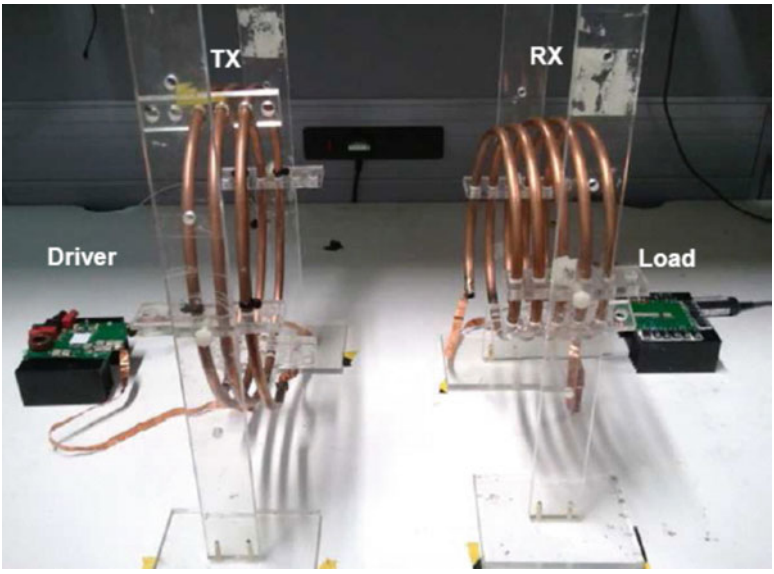


Fig. 6.23 Inductive power transfer setup (From Ref. [45])

6.5.1 Near Field Inductive Power Transfer

Near field inductive power transfer is probably the most common means for wireless power delivery and is already used over short range [46] for the recharging of implanted medical devices. Energy is transferred through a magnetic field produced by the primary which links with the secondary. In essence, this is the same as in a regular transformer, but in this case the iron core has been removed to allow mobility between primary and secondary. This technique is often called wireless power transfer (WPT), or inductive power transfer (IPT). Figure 6.23 shows a wireless power transfer system designed at Imperial College London, which is capable of transferring 200 W over around 30 cm with an efficiency of over 70 %. In this setup, transmit and receive coils are constructed with copper plumbing pipe. The significant challenges involved in the efficient transfer of energy in such a system can be readily understood when the system is examined in the form of a transformer equivalent circuit model as shown in Fig. 6.24.

In a well-constructed iron or ferrite core transformer, the vast majority of the flux produced by the primary coil links with the secondary. This means that the leakage inductance (L_l in Fig. 6.24) can often be negligible and thus all the applied voltage across the primary coil appears across the ideal transformer component. In addition, the magnetising inductance, L_m, is large due to the presence of the highly permeable core. This means that a closely coupled transformer can transfer energy efficiently from primary to secondary, irrespective of frequency (as long as the frequency is

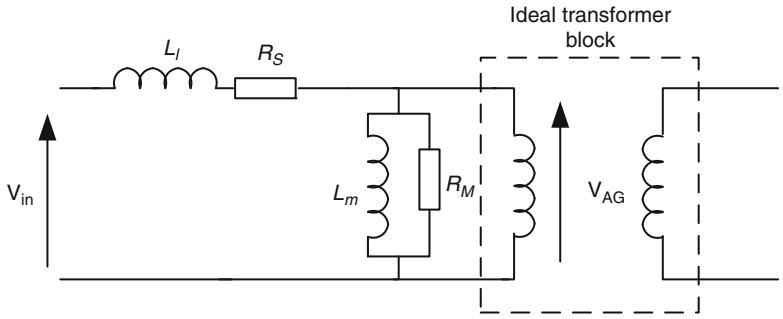


Fig. 6.24 Transformer equivalent circuit with open circuit secondary

low enough so that magnetisation losses in the core remain low). However, when the highly permeable core is removed, two things occur: the coupling between the primary and secondary falls significantly, meaning the leakage inductance becomes non-negligible, and the magnetising inductance falls significantly, meaning that the air-gap voltage, V_{AG}, is significantly less than the driving voltage, V_{in}. This means that the voltage output from the drive circuit (supplying V_{in}) must be much higher than the required air-gap voltage, decreasing the efficiency of the driver. In addition, the low value of magnetising inductance means the drive circuit also has to supply significant current, most of which does not contribute to the transfer of power to the secondary.

In order to make this system efficient to drive, a method must be found to avoid having high currents and high voltages present at the terminals of the drive amplifier. A solution to this is to operate the primary at or close to resonance, where the high magnetising currents can oscillate between the series combination of the magnetising and leakage inductances, and a tank capacitor. This removes the need for the drive circuit to conduct high currents, but the resonant circuit still produces high voltages. In order to shield the drive circuit from these high voltages, a capacitive potential divider can be used, as shown in Fig. 6.25. In addition, the secondary coil should be tuned to resonance in order to maximise power transfer to the secondary (such that the secondary only reflects a real impedance to the primary).

For a system in secondary resonance, with a primary tank Q factor of Q_{TX} and a secondary tank Q factor of Q_{RX}, coupled with a coupling factor of k, the maximum link efficiency (power received at secondary/power input to primary) can be given as:

$$\eta_{link} = \frac{k^2 Q_{TX} Q_{RX}}{\left(1 + \sqrt{1 + k^2 Q_{TX} Q_{RX}}\right)^2} \tag{6.4}$$

In a well coupled transformer with a permeable core, k is almost 1 and the link efficiency tends to 1 for low Q factors. In a typical IPT system for BSN, the

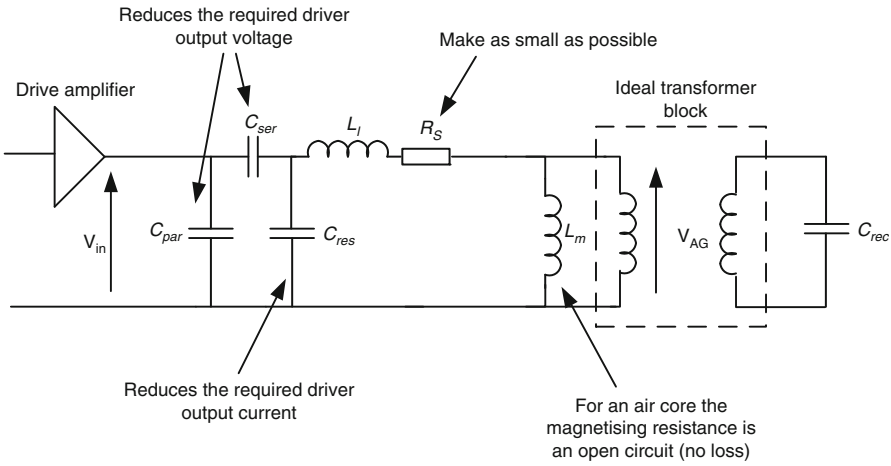


Fig. 6.25 Equivalent circuit of an IPT system using both primary and secondary resonance

secondary coil could be very small, even if the primary coil is large. The coupling factor is defined as:

$$k = \frac{\Phi_{12}}{\Phi_1} \tag{6.5}$$

where Φ_{12} is the flux that links with the secondary from a primary flux of Φ_1. In a typical IPT system for BSN where the secondary coil may be much smaller than the primary, k may be 1 % or lower, meaning that the Q factors must be very high in order to obtain a high link efficiency. The link efficiency as a function of $k^2 Q_{RX} Q_{TX}$ is shown in Fig. 6.26, highlighting the need for high Qs if the coupling factor is low. For a coupling factor of only 1 %, we need each coil Q factor to be 316 to get a 95 % link efficiency. The requirements for high Q factors often drives the frequency up into MHz: the coil Q increases with frequency (as reactance dominates over resistance) until such a point as the combined effect of radiation (the coil behaves like an antenna) and skin effect starts to reduce the Q.

A complete IPT system is shown in Fig. 6.27, where a power amplifier drives a coil (through a suitable passive network as described above). The secondary contains a tuning network and a rectifier, followed by a voltage regulator to power the load.

This emphasises that high Q factors are needed when the coupling factor decreases to very low values, which is especially challenging for small receiver coils. However, such systems have been considered for BSN type applications. For instance, it is possible to send mW level power over around 10 m with 1 m transmit coils and cm size receiver coils whilst keeping the magnetic field strengths within the required limits for human exposure. Such systems are good alternatives for the powering of BSN applications.

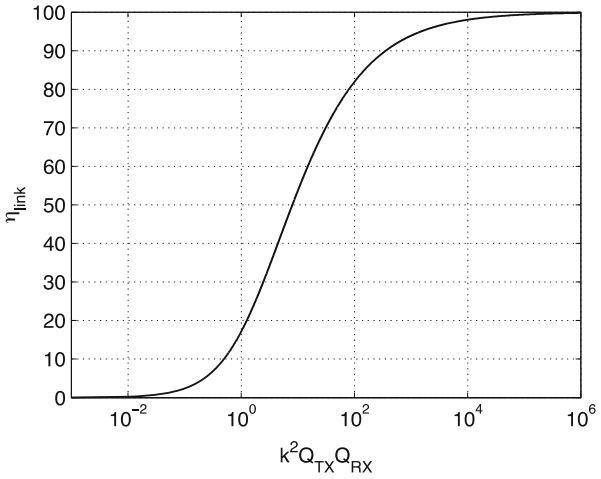


Fig. 6.26 Link efficiency as a function of $k^2Q_{RX}Q_{TX}$ (From Ref. [47])

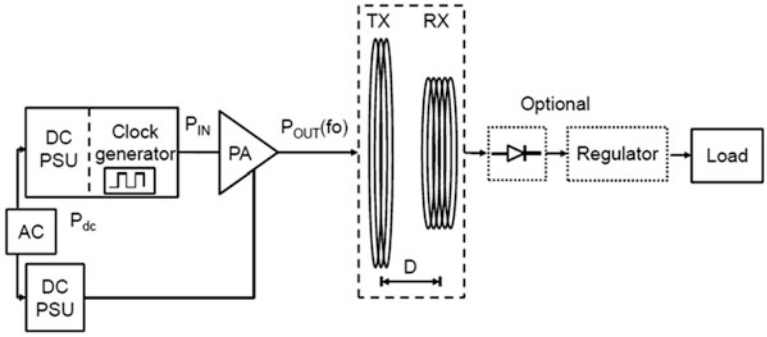


Fig. 6.27 Basic wireless power transfer system. Receiver system can be implanted

6.5.2 *Ultrasonic Power Delivery*

Wireless power delivery is particularly advantageous for implanted devices, where direct access for battery replacement or charging is not desirable. Inductive power transfer, as described above, is used extensively in such applications. However, the coils required are relatively large, and the need for high efficiency requires the receiving coil to be near the skin, necessitating a tethering wire if the implant is deeper in the body. Such tethers add to the medical impact and cost, and reduce safety and reliability. More remote electromagnetic delivery is possible as described above, but there is significant attenuation in body tissue at higher frequencies, and the coupling efficiency drops off rapidly if the receiving antenna is small compared to its separation from the surface.

Ultrasound is an attractive alternative which can overcome these restrictions. Because of the much lower speed of sound compared to radio waves in human

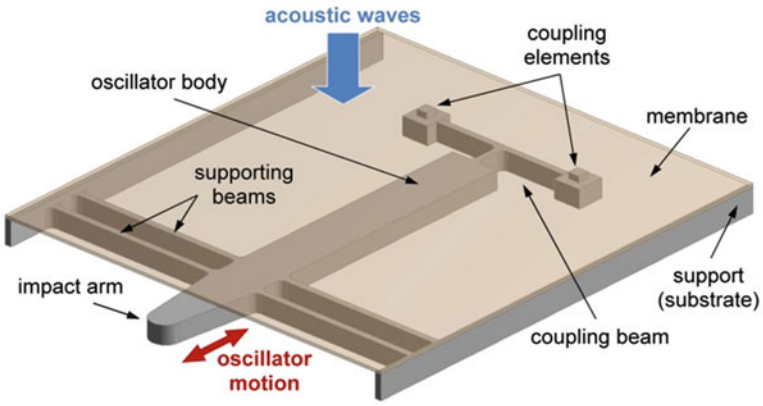


Fig. 6.28 MEMS device for powering actuation by remote ultrasound (From Ref. [50])

tissue (about 1500 m/s in soft tissue), the wavelengths are orders of magnitude less for a given frequency. This gives the possibility of directional transmission and reception even for transducers of modest size. A comparative simulation showed the efficiency of ultrasonic delivery to be an order of magnitude or higher than that of inductive systems, where the distance to the receiving element was more than a few times its diameter [48]. As a result a range of ultrasonic systems have been proposed or investigated for ultrasonic power delivery to medical implants as well as other devices [49].

Ultrasound could also be used to provide power to directly energise mechanical function without intermediate conversion to electricity, for implanted actuators. Such a device has recently been reported [50], in a MEMS implementation in which the received ultrasonic wave vibrates a membrane, which is coupled to a discrete oscillator which impacts a linear actuator (Fig. 6.28). This might be used for applications as drug release, or the fine mechanical adjustment of an implanted prosthetic.

6.5.3 Radiative Power Transfer

This type of power delivery mechanism relies on the use of ambient RF energy from digital TV, mobile phone signals, or WiFi, which can be collected by an antenna and rectified and stored in a battery. As an alternative, RF energy can be beamed from a source, but in that case inductive power delivery may be more favourable. A typical RF energy harvesting system is shown in Fig. 6.29, where an antenna connects to a rectifier through a matching network, which feeds a smoothing capacitor and a power management module which performs maximum power point tracking.

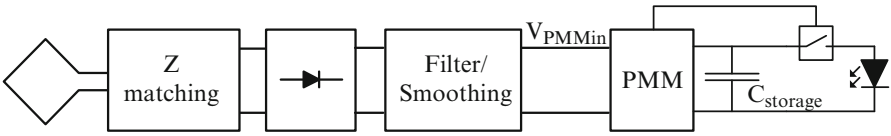


Fig. 6.29 RF energy harvesting system capable of intermittent flashing of an LED (From Ref. [45])

Systems such as these are capable of receiving power levels of around 10 μW/cm^3 [45], which is comparable to mechanical harvesting, but has the potential advantage that these devices contain no moving parts.

6.6 Discussion and Conclusions

6.6.1 What Is Achievable in Body-Sensor Energy Harvesting?

Section 6.2 presented ultimate limits for inertial harvesters, showing that for a given source frequency and amplitude Y_o, the maximum extractable power depends only on the proof mass m and internal displacement amplitude Z_l. In [3] the performance of published devices were compared to this limit and were found to be generally at least an order of magnitude lower, sometimes much lower still. This indicated that at that time, substantial improvements in power output were still achievable. This is still the case, although more recent devices have improved somewhat upon these earlier reports. Another trend that can be seen is that larger devices tend to perform at closer to their ideal levels. This is likely an indication of the technological difficulties encountered at smaller size scales, for example the greater difficulty in achieving high magnetic flux gradients.

It can also be noted that for a given volume, the performance limiting value $Z_l m$ depends on the fraction of volume taken up by the mass, and on the device shape and its direction of internal motion. For a given shape and internal direction, allowing the proof mass to occupy half the volume gives the maximum product $Z_l m$, so that if the device is a cube of dimension a, $Z_l m$ will be given by $\rho a^4/8$, or equivalently $\rho V^{4/3}/8$, with ρ the proof mass density and V the device volume. This scaling of power with $V^{4/3}$ indicates that the achievable power density drops as the devices reduce in size, although not rapidly.

For a given volume, half occupied by the proof mass, the internal displacement limit depends on the device shape or aspect ratio. Ideally, one direction would be elongated, so that a long thin cylinder would be an efficient shape from this point of view. However, the volume, independent of shape, is unlikely to be the key constraint. Also, the fabrication technology may well put limits on shape. For example, in MEMS processing or for other planar techniques, one dimension

(the out-of-plane one) is typically much shorter than the other two. This constrains the device to a shape like that described in Sect. 6.3. In such a case, power would clearly be maximised if the travel is in one of the long directions. Again, however, practical constraints may mitigate against this choice. For example, it is difficult to achieve long travel in a lateral (in-plane) suspension without excessive ease of motion in unwanted axes, both out of plane and rotational.

Assuming that such limitations are overcome and a planar device is constructed with lateral motion, we can use this format to calculate a maximum specific power for inertial devices in BSN applications. We assume an aspect ratio (out-of-plane to in-plane dimension) of 10, and a proof mass specific gravity of 20 (gold); for the source, we assume a frequency 1 Hz and a maximum displacement of 25 cm. Then:

$$P \approx 1.7V^{4/3}\ \mathrm{mW/cm^4} \tag{6.6}$$

For a 0.1 cm^3 device this gives a maximum power of 80 μW; for a cubic millimetre device, only 0.17 μW. This suggests tens of mm^3 are likely to be needed for an inertial harvesting device to be of much value for foreseeable body sensor nodes.

In [51], the likely power densities of inertial and thermoelectric energy harvesters in a body mounted application are compared, for both walking and running subjects. Although the theoretical levels are significantly higher for the inertial devices, when likely effectiveness levels are considered the performance is similar, and somewhat better for thermoelectric devices in the walking case.

6.6.2 *Future Prospects and Trends*

Energy harvesting continues to attract attention from a large number of researchers, and the performance levels achieved continue to rise. Body powered applications, however, remain a great challenge because of the low specific power levels at low frequencies, and so substantial progress will be needed in reducing power requirements, particularly for wireless data transmission, before such solutions become feasible. However, such progress is certainly being made, as reported in other chapters.

The body motion that may power energy harvesting will vary substantially with time, and this variation is unlikely to correspond with the varying power demands of the sensor node. Therefore some energy storage is almost certain to be required. While mechanical energy storage based on MEMS is conceivable, there is little sign of such an approach being developed and so secondary batteries are likely to be used. This is also likely to be the case for other, non-inertial forms of energy harvesting. This suggests that if harvesting methods are successfully exploited, they are likely to be supplementary to, rather than a replacement for, battery technologies. The need for integrated power conditioning circuits with energy harvesting also encourages a trend towards intelligent energy modules, possible incorporating several forms of harvesting as well as storage, power conditioning, and power management electronics.

References

1. Paradiso JA and Starner T, Energy scavenging for mobile and wireless electronics. Pervasive Computing, IEEE, 2005. 4(1): p. 18–27.
2. Kim S-G, Priya S and Kanno I, "Piezoelectric MEMS for energy harvesting", MRS Bulletin 37, 1039–1050, 2012.
3. Mitcheson PD, Yeatman EM, Rao GK, Holmes AS, Green TC, "Energy harvesting from human and machine motion for wireless electronic devices", Proc. IEEE 96 (9), 1457–1486 (2008).
4. Afonso JA, Macedo P, Rocha LA and Correia JH, "Hierarchical Wireless Networks of Body Sensor Networks for Healthcare Applications", in Cruz-Cunha MM, Tavares AJ and Simoes R, Eds., Handbook of Research on Developments in E-Health and Telemedicine: Technological and Social Perspectives, IGI Global, 2010.
5. Dementyev A, Taylor S, Smith J, Hodges S, "Power consumption analysis of Bluetooth Low Energy, Zigbee, and ANT sensor nodes in cyclic sleep scenario", IEEE International Wireless Symposium, Beijing, China, 2013
6. Yates DC, Holmes AS, and Burdett AJ, Optimal transmission frequency for ultralow-power short-range radio links. Circuits and Systems I: Regular Papers, IEEE Transactions on, 2004. 51(7): p. 1405–1413.
7. ADXL362 Data Sheet, Analog Devices, One Technology Way, Norwood, MA 02062-9106, U.S.A., 2012–2013.
8. Ay SU, "A CMOS Energy Harvesting and Imaging (EHI) Active Pixel Sensor (APS) Imager for Retinal Prosthesis", IEEE Trans. Biomedical Circuits and Systems 6, 535–545 (2011).
9. Sauerbrey JD, Schmitt-Landsiedel D, and Thewes R, A 0.5-V 1-μW successive approximation ADC. Solid-State Circuits, IEEE Journal of, 2003. 38(7): p. 1261–1265.
10. Lee S-Y, Cheng C-J, Wang C-P, Lee S-C, "A 1-V 8-bit 0.95mW successive approximation ADC for biosignal acquisition systems", IEEE Int. Symp. on Circuits & Systems, ISCAS 2009, 649–652 (2009).
11. Xie L, Yang G, Mantysalo M, Xu L-L, Jonsson F, Zheng L-R, "Heterogeneous Integration of Bio-Sensing System-on-Chip and Printed Electronics", IEEE J. Emerging and Selected Topics in Circuits and Systems, 2012, 2(4) pp. 672–682.
12. Tie SF, Tan CW, "A review of energy sources and energy management system in electric vehicles", Renewable and Sustainable Energy Reviews 20, 82–102, 2013.
13. Holladay JD, Jones EO, Phelps M, Hu J, "Microfuel processor for use in a miniature power supply", J. Power Sources 108 (1–2), 2002, Pages 21–27.
14. Cho J-H, Lin CS, Richards CD, Richards RF, Ahn J, Ronney PD, "Demonstration of an external combustion micro-heat engine", Proc. Combustion Institute 33(2), 2009, 3099–3105.
15. Wee JH, "A feasibility study on direct methanol fuel cells for laptop computers based on a cost comparison with lithium-ion batteries", J. Power Sources, 173 (2007), pp. 424–43.
16. Kim T, Kwon S, MEMS fuel cell system integrated with a methanol reformer for a portable power source, Sensors and Actuators A: Physical, 154(2), 2009, 204–211.
17. Kerzenmacher S, Kräling U, Metz T, Zengerle R, von Stetten F, A potentially implantable glucose fuel cell with Raney-platinum film electrodes for improved hydrolytic and oxidative stability, J. Power Sources 196 (3), 2011, 1264–1272.
18. Green MA, Emery K, Hishikawa Y, Warta W, Dunlop ED, Solar cell efficiency tables (version 39), Progress in Photovoltaics: Research and Applications 20(1), 12–20, 2012.
19. Mantiply ED, Pohl KR, Poppell SW, and Murphy JA, Summary of measured radiofrequency electric and magnetic fields (10 kHz to 30 GHz) in the general and work environment. Bioelectromagnetics, 1997. 18(8): p. 563–577.
20. Vullers RJM, Schaijk RV, Visser HJ, Pender J, Van Hoof C, Energy Harvesting for Autonomous Wireless Sensor Networks, IEEE Solid-State Circuits Magazine, 2(2), 29–38, 2010.
21. Vullers RJM, van Schaijk R, Doms I, Van Hoof C, Mertens R, Micropower energy harvesting, Solid-State Electronics, 53 (7), 2009, 684–693.

22. Leonov V, Vullers RJM, and Van Hoof C, Thermoelectric generator hidden in a shirt with a fabric radiator, AIP Conf. Proc. 1449, 556–559 (2012).
23. Howey DA, Bansal A and Holmes AS, Design and performance of a centimetre-scale shrouded wind turbine for energy harvesting, Smart Materials and Structures 20 (8), 085021, 2011.
24. Shenck NS and Paradiso JA, Energy scavenging with shoe-mounted piezoelectrics. Micro, IEEE, 2001. 21(3): p. 30–42.
25. Beeby SP, Torah RN, Tudor MJ, Glynne-Jones P, O'Donnell T, Saha CR and Roy S, A micro electromagnetic generator for vibration energy harvesting, J. Micromech. Microeng. 17(7), 1257, 2007.
26. Solymar L and Walsh D, Lectures on the Electrical Properties of Materials. 1993, Oxford, U.K.: Oxford University Press.
27. Ayala-Garcia IN, Mitcheson PD, Yeatman EM, Zhu D, Tudor J, Beeby SP, Magnetic tuning of a kinetic energy harvester using variable reluctance, Sensors and Actuators A: Physical 189, 2013, 266–275.
28. Nguyen DS, Halvorsen E, Jensen GU and Vogl A, Fabrication and characterization of a wideband MEMS energy harvester utilizing nonlinear springs, J. Micromech. Microeng. 20, 125009, 2010
29. Mitcheson PD, Green TC, Yeatman EM, and Holmes AS, Architectures for vibration-driven micropower generators. J. Microelectromechanical Systems, 2004. 13(3): p. 429–440.
30. von Buren T, Mitcheson PD, Green TC, Yeatman EM, Holmes AS, Troster G, Optimization of Inertial Micropower Generators for Human Walking Motion. IEEE Sensors Journal 6 (1), pp. 28–38, 2006.
31. Miao P, Mitcheson PD, Holmes AS, Yeatman E, Green T, and Stark B, MEMS inertial power generators for biomedical applications, Microsyst. Technol., vol. 12, no. 10–11, pp. 1079–1083, Sep. 2006.
32. Mitcheson PD, Miao P, Stark BH, Yeatman EM, Holmes AS, Green TC, MEMS electrostatic micropower generator for low frequency operation. Sensors and Actuators A: Physical, 2004. 115 (2–3): p. 523–529.
33. He C, Kiziroglou ME, Yates DC, Yeatman EM, "A MEMS Self-Powered Sensor and RF Transmission Platform for WSN Nodes", IEEE Sensors Journal, 11 (12), 2011, 3437–3445.
34. Suzuki Y, Recent progress in MEMS electret generator for energy harvesting, IEEJ Trans. Electrical and Electronic Engineering, 6 (2), pages 101–111, 2011.
35. Roundy S, Takahashi E, A planar electromagnetic energy harvesting transducer using a multi-pole magnetic plate, Sensors and Actuators A: Physical, 195, 2013, 98–104.
36. Pillatsch P, Yeatman EM and Holmes AS, Piezoelectric Rotational Energy Harvesting For Body Sensors Using An Oscillating Weight, Proc. Body Sensor Networks 2013, Boston, USA, May 6–9, 2013.
37. Szarka GD, Stark BH, and Burrow SG, Review of Power Conditioning for Kinetic Energy Harvesting Systems. Power Electronics, IEEE Transactions on, 2012. 27(2): p. 803–815.
38. Mitcheson PD and Green TC, Maximum Effectiveness of Electrostatic Energy Harvesters When Coupled to Interface Circuits. Circuits and Systems I: Regular Papers, IEEE Transactions on, 2012. 59(12): p. 3098–3111.
39. Cammarano A, et al., Tuning a resonant energy harvester using a generalized electrical load. Smart Materials and Structures, 2010. 19(5): p. 055003.
40. Mitcheson PD et al., Tuning the Resonant Frequency and Damping of an Electromagnetic Energy Harvester Using Power Electronics. Circuits and Systems II: Express Briefs, IEEE Transactions on, 2011. 58(12): p. 792–796.
41. Badel A, et al., Piezoelectric Energy Harvesting using a Synchronized Switch Technique. Journal of Intelligent Material Systems and Structures, 2006. 17(8–9): p. 831–839.
42. Dicken J, et al., Power-Extraction Circuits for Piezoelectric Energy Harvesters in Miniature and Low-Power Applications. Power Electronics, IEEE Transactions on, 2012. 27(11): p. 4514–4529.

43. Elliott ADT and Mitcheson PD. Power density improvement of a piezoelectric energy harvester through use of a micropower switch-mode interface. in Sensors, 2012 IEEE. 2012.
44. Kurs A et al., Wireless Power Transfer via Strongly Coupled Magnetic Resonances. Science, 2007. 317(5834): p. 83–86.
45. Pinuela M, Mitcheson PD, and Lucyszyn S, Ambient RF Energy Harvesting in Urban and Semi-Urban Environments. Microwave Theory and Techniques, IEEE Transactions on, 2013. 61(7): p. 2715–2726.
46. Koenraad van Schuylenbergh RP, Inductive Powering: Basic Theory and Application to Biomedical Systems. 2009: Springer.
47. Pinuela M, et al., Maximizing DC-to-Load Efficiency for Inductive Power Transfer. Power Electronics, IEEE Transactions on, 2013. 28(5): p. 2437–2447.
48. Denisov A and Yeatman EM, "Ultrasonic vs. inductive power delivery for miniature biomedical implants," in *Body Sensor Networks (BSN), 2010 International Conference on,* 2010, pp. 84–89
49. Roes MGL, Duarte JL, Hendrix MAM and Lomonova EA, "Acoustic Energy Transfer: A Review," *Industrial Electronics, IEEE Transactions on,* vol. 60, pp. 242–248, 2013.
50. Denisov A and Yeatman EM, Battery-less Microdevices for Body Sensor/Actuator Networks, Proc. Body Sensor Networks 2013, Boston, USA, May 6–9, 2013.
51. P.D. Mitcheson, Energy Harvesting for Human Wearable and Implantable Bio-Sensors, *Proc. 32nd IEEE EMBS,* Buenos Aires, Aug 31 – Sept 4, 2010.

Chapter 7
Towards Ultra-low Power Bio-inspired Processing

Leila Shepherd, Timothy G. Constandinou, and Chris Toumazou

7.1 Introduction

The natural world is analogue and yet the modern microelectronic world with which we interact represents real world data using discrete quantities manipulated by logic. In the human space, we are entering a new wave of body-worn biosensor technology for medical diagnostics and therapy. This new trend is beginning to see the processing interface move back to using continuous quantities, which are more or less in line with the biological processes. We label this computational paradigm "bio-inspired" because of the ability of silicon chip technology which enables the use of inherent device physics, allowing us to approach the computational efficiencies of biology. From a conceptual viewpoint, this has led to a number of more specific morphologies including neuromorphic and retinomorphic processing. These have led scientists to model biological systems such as the cochlea and retina and gain not only superior computational resource efficiency (to conventional hearing aid or camera technology), but also an increased understanding of biological and neurological processes.

A similar approach to "chemically-inspired" microelectronics can be used for BSN, which would lead to portable ultra low-power micro-systems capable of faster chemical/biochemical discrimination and interrogation of data, integrated monolithically at the sensor end. In contrast to the digital approach, where each operation is performed through a network of devices operated in a switched fashion, the physics of the elementary device itself, electrical, chemical or electro-chemical, can be exploited to perform the same operation in an analogue way. Therefore, both

L. Shepherd
DNA Electronics Ltd, 56 Wood Lane, London W12 7SB, UK

T.G. Constandinou (✉) • C. Toumazou
Department of Electrical and Electronic Engineering, Imperial College London, London SW7 2BT, UK
e-mail: t.constandinou@imperial.ac.uk

G.-Z. Yang (ed.), *Body Sensor Networks*, DOI 10.1007/978-1-4471-6374-9_7,

the energy per unit computation and silicon real estate are reduced, resulting in significantly increased overall resource efficiency.

This chapter will first look at the motivation for bio-inspired signal processing and discuss the relative merits of analogue and digital signal processing, and the need for hybrid architectures. The concept of applying bio-inspired design methodologies to CMOS-based biosensors will then be introduced. Field-effect transistor (FET-) based sensors will be presented, including a detailed example of the application of analogue processing techniques to these devices. Finally, future directions and applications for biochemically-inspired design will be discussed.

7.2 Bio-inspired Signal Processing

Although modern microelectronic technologies have surpassed our expectations in virtually all areas, there still remains a vast application space of computational problems either too challenging or complex to be solved with conventional means. These applications often require the transformation of data across the boundary between the real (analogue) world and the digital world. The problem arises whenever a system is sampling and acting on real-world data, for example in recognition or identification tasks. Traditional processing techniques find it prohibitively challenging or at best computationally demanding to identify and process complex structures and relationships in vast quantities of ill-conditioned data (for instance data that is low precision, ambiguous and noisy) [1].

Despite significant progress made in hardware processing techniques (e.g. DSP, FPGA), in both computational load and efficiency, the solution to complex recognition tasks still continues to elude us. Furthermore, artificial intelligence, artificial neural networks and fuzzy logic have yet to provide effective and robust solutions for practical sensing applications. However, biological organisms routinely accomplish complex visual tasks such as object recognition and target tracking. For example, a common housefly, with a brain the size of a grain of rice, can outperform our modern multiple gigahertz processors in real-time obstacle avoidance in flight navigation, in addition to countless other perception tasks. Thus fields such as neuromorphic engineering have emerged, aiming to provide a design methodology for tackling such problems using hybrid, distributed processing architectures based on simple primitives. Inspired by biology, modern microelectronic technology is progressing one step closer to finding a workable solution to these problems.

In the context of wearable sensors for BSN, which communicate information over a wireless network, it is vital signs and not necessarily raw data that should be signalled to the user. There is, therefore, scope for bio-inspired analogue processing to take place local to the sensor, before transmission of data, rather than transmit raw data at high accuracy. In employing this approach, the power saving is twofold: both in reducing the communication bandwidth/duty cycle and therefore the transmit power (see Fig. 7.1). Rather than a mere “biosensor” chip, one would have a “biocomputer” chip with the intelligence to extract important features of the data

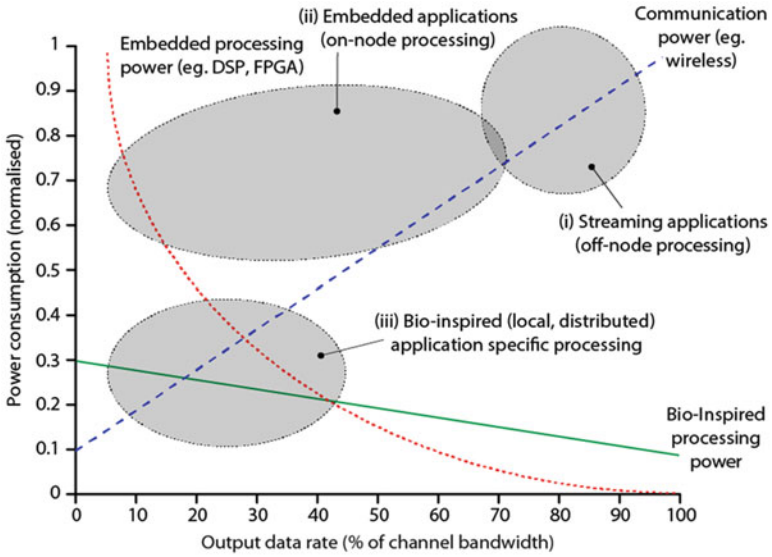


Fig. 7.1 Power budget (processing vs. communication power) – bandwidth trade-off for: (*i*) off-node processing, i.e. streaming raw data, (*ii*) convention-al on-node (embedded) processing, (*iii*) bio-inspired on-node processing

sensed and to discriminate between different scenarios (for example, in brain machine interfaces [2, 3]).

7.3 Analogue Versus Digital Signal Processing

7.3.1 Quantised Data/Time vs. Continuous Data/Time

Information exists in a three-dimensional media with data encoded in time, amplitude and space. Various techniques of data representation exist for processing this information, with spatial information coded universally by the position of the sensor in the three dimensional medium [4]. For example, in electronics, analogue circuits represent data as continuous voltages and currents, varying both in time and intensity. On the other hand, conventional digital electronics use clocks to synchronise activity – data therefore being represented as discrete voltages quantised in time as well as amplitude. Sampled data techniques also exist such as switched-capacitor (SC) [5, 6] and switched-current (SI) [7] that use a clock to sample continuously varying signals and are therefore discrete in time but continuous in amplitude. Such techniques are widely used in signal processing of continuous (analogue) signals, for example in implementing filters for oversampling data converters.

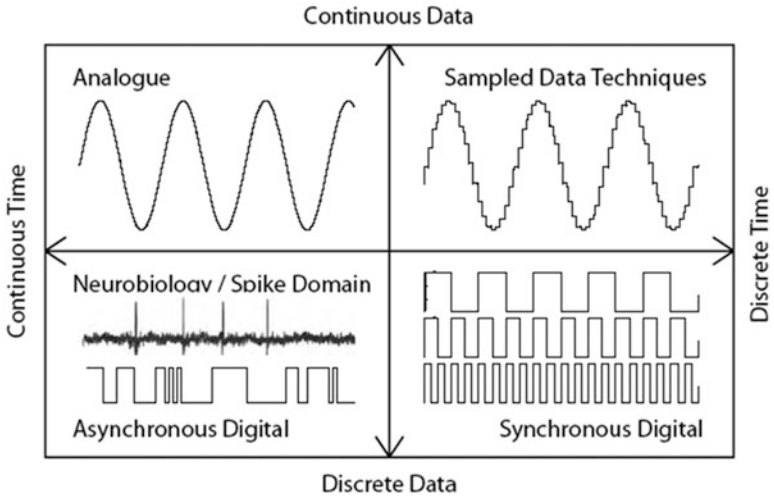


Fig. 7.2 Different data representations in standard microelectronic technologies classified by time (continuous vs. discrete) and quantity (continuous vs. discrete)

Exploring this two-dimensional space, i.e., time and amplitude as shown in Fig. 7.2, for encoding data, a remaining unexploited representation is continuous-time, discrete-data. This is in fact the principle representation of biology- with spiking neurons conveying no data in the shape or amplitude of the action potential, but rather in the timing. This encoding can easily be achieved by using asynchronous digital technology, although it is not widely used in system-level design due to complexity in synthesis.

7.3.2 Analogue/Digital Data Representation

A key debate in the low power electronics community is whether analogue or digital signal processing is more computationally efficient. Much work [1, 8–13] has already gone into this by considering factors such as signal-to-noise ratio (SNR), power consumption, silicon area, channel utilisation and design time.

The general conclusions from the existing research include [11–13]:

1. Analogue processing can be far more computationally efficient than digital signal processing. This is due to the rich mathematical content in the physics of the devices in comparison to the primitive nature of a digital device (a switch). It follows that to achieve similar functionality with digital logic, many more devices need to be used – in fact this can be several orders of magnitude more devices. Moreover, at high activities this results in a significantly higher power consumption. This is because digital logic dissipates, due to both continuous

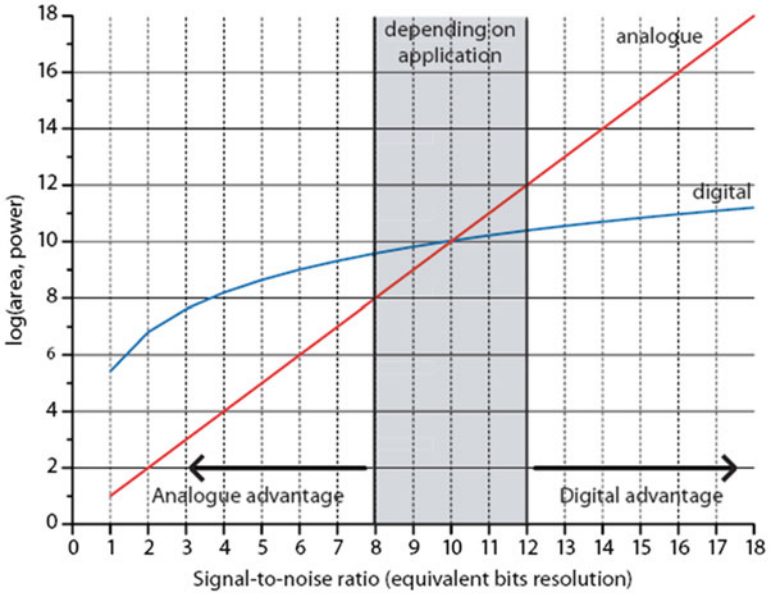


Fig. 7.3 The relative cost of computation using analogue or digital signal processing. The crossover between analogue and digital having the advantage is between 50 and 72 dB SNR (8–12 bits resolution) depending on application and circuit topology

subthreshold "leakage" current (static power) and during switching (dynamic power), whereas analogue devices only have a continuous current supply (static power).

2. Digital processing is more tolerant of noise and cumulative offsets. The continuous nature of analogue signals means they cannot be restored at each stage as discrete signals can. Consequently, any noise or circuit-introduced offset accumulates through cascading and can ultimately deteriorate the signal in complex analogue systems. This reduces the accuracy and dynamic range of such a system for a given power budget. If device geometries are increased and more power is dissipated, analogue systems can be made to perform to higher accuracies, however the computational efficiency of digital systems then tends to be superior.
3. Quantifying these benefits, it can be shown that the cost (silicon area and power consumption) of analogue computation is exponential with respect to SNR, whereas the cost of digital computation is linear. In addition, the starting overhead (at low SNR) of analogue is low, whereas for digital it is high. This sets a trend where the benefits of each method can be divided using SNR alone (see Fig. 7.3 [9, 11–13]). For lower SNRs, analogue techniques can have many orders of magnitude area and power advantage, whereas for higher precision computation digital techniques have the cost advantage.

These conclusions are the result of deriving mathematical expressions to quantify the computational cost based on the fundamental limits of each technique. Although these provide the ultimate theoretical performance of each representation technique, they do not consider implementation issues, with circuit design and wafer processing being far from ideal.

In the subsequent sections, we shall consider qualitative comparisons of various microelectronic representations in performing common computational tasks with implementation issues being considered.

7.3.3 Linear Operations

The most common mathematical computations are in fact linear operations. These include addition, subtraction, multiplication, and division etc. Implementing these computations in different ways can prove beneficial. For example, to add two currents, only a single wire is needed (by Kirchhoff's Current Law), whereas an 8-bit digital implementation would require eight full-adder stages, comprising of a total of at least 228 transistors. Similarly, a multiplication can be achieved using a Gilbert (translinear) multiplier circuit [14] employing only eight transistors biased in the "subthreshold" or "weak inversion" region of operation. Here, the equivalent digital solution would be an 8-bit array multiplier requiring an excess of over 2,000 transistors. In these examples, silicon area can be saved using analogue techniques, however, as always in electronic design, the various trade-offs need to be considered. A qualitative comparison of the most popular techniques used for linear arithmetic computation is illustrated in Table 7.1.

For these comparisons, sampled data techniques have been combined with their respective continuous-time counterparts as these are both based on the same underlying circuit theory. To substantiate this, Furth et al. [10] have shown these continuous-time and sampled-data techniques to follow similar SNR-to-power-consumption relationships.

7.3.4 Non-linear Operations

In most complex processing tasks, the underlying computation tends to be non-linear. This may comprise of an array or bank of linear functions to achieve the overall non-linear behaviour. A qualitative comparison, as previously presented for linear operations, has been formulated for selected common non-linear functions, shown in Table 7.2.

Table 7.1 A qualitative comparison of linear computations implemented using different signal representations

Signal representation	Topology	Silicon area	Power	Accuracy	Noise	Speed	Ref.
Addition, subtraction, summation							
Current-mode analogue[a]	Current addition (KCL)	Best	Best	Good	Excellent	Good	[15]
Voltage-mode analogue[b]	Charge domain (switched-cap)	Good	Good	Good	Excellent	Good	[16]
Digital[c]	Parallel counter, ripple adder	Fair	Good	Excellent	Excellent	Excellent	[17]
Multiplication, division							
Current-mode analogue	Gilbert multiplier	Excellent	Excellent	Excellent	Good	Fair	[14]
Voltage-mode analogue	Flipped voltage followers	Good	Good	Good	Good	Fair	[18]
Digital	Array, tree multiplier	Poor	Fair	Excellent	Excellent	Excellent	[19]
Scaling							
Current-mode analogue	Scaled current mirror	Excellent	Excellent	Good	Fair	Good	–
Voltage-mode analogue	Operational amplifier	Good	Fair	Good	Good	Good	–
Digital	Barrel shift and accumulate	Fair	Good	Excellent	Excellent	Excellent	[20]

[a]Provide maximum resource efficiency (area and power) [10–12]
[b]Provide good all-round performance
[c]Provide highest speed operation and precision [10–12]

7.3.5 *Hybrid System Organisation*

The ultimate goal of using a hybrid approach is to exploit different representation strategies throughout a system – ideally to achieve optimum performance for a given processing task. Most modern applications typically require both analogue and digital techniques to work alongside one another as the bare minimum. Since the real world is analogue, any system requiring a sensor interface requires analogue electronics. On the other hand, as most control systems and communication protocols are digital, any system requiring external interface capability requires digital electronics.

This defines a minimum requirement of one data converter. Therefore, in order to utilise resources effectively, it would be best to use this data conversion to our advantage by using this as the main conversion stage within a system. By using hybrid processing strategies and shifting the data conversion interface, the required conversion accuracy can be relaxed to specifications that lend themselves to micropower techniques. Using previously mentioned signal representation techniques, there exist several architectures that fulfil these criteria as illustrated in Fig. 7.4.

Table 7.2 A qualitative comparison of non-linear computations implemented using different signal representation techniques

Signal representation	Topology	Silicon area	Power	Accuracy	Noise	Speed	Ref.
Comparison, thresholding[a]							
Current-mode analogue	Current comparator	Excellent	Excellent	Good	Fair	Fair	[15]
Voltage-mode analogue	Operational amplifier	Good	Good	Good	Good	Good	–
Digital	Subtractor	Fair	Fair	Good	Excellent	Excellent	[20]
Exponential, logarithm, square, square root[b]							
Current-mode analogue	Translinear circuits	Excellent	Excellent	Good	Good	Fair	[21]
Voltage-mode analogue	Non-linear V to I	Good	Fair	Good	Good	Fair	[22]
Digital	Root/division algorithm	Fair	Fair	Excellent	Excellent	Good	[23]
Filtering, integration, differentiation, fourier transform[c]							
Current-mode analogue	Log domain	Good	Excellent	Good	Excellent	Excellent	[24]
Voltage-mode analogue	Charge domain (switched cap)	Good	Excellent	Fair	Good	Good	[16]
Digital	IIR/FIR filters, FFT	Poor	Fair	Good	Excellent	Good	[25]

[a]The direct comparison of continuous signals makes analogue comparators the most easily implementable, whereas digital comparison techniques are typically implemented using subtraction driven combinational logic

[b]Analogue realisations are based on translinear techniques or exploitation of non-linear component response, whereas digital implementations require either ROM-based lookup tables or synthesis of custom arithmetic-logic-unit (ALU) type hardware

[c]Digital implementation provides better reconfigurability, stability to drift/temperature and low frequency operation

7.4 CMOS-Based Biosensors

Complementary Metal Oxide Semiconductor (CMOS) is the dominant semiconductor technology for fabrication of modern microelectronic components such as microprocessors, memories and application specific integrated circuits (ASICs) on a silicon substrate through a defined sequence of material deposition, doping, lithography and etching [26]. CMOS technology has transformed the electronics industry with a seemingly undiminishing ability to integrate more and more uniform devices of ever-decreasing dimensions onto a single silicon wafer. It is no wonder that a goal of recent miniaturisation trends in electrochemical biosensing has been the fabrication of electrochemical "microsensors" on a CMOS substrate, providing not only low-cost batch fabrication and reproducible sensor characteristics, but also reduced power consumption and rapid sensor response due to reduced device dimensions.

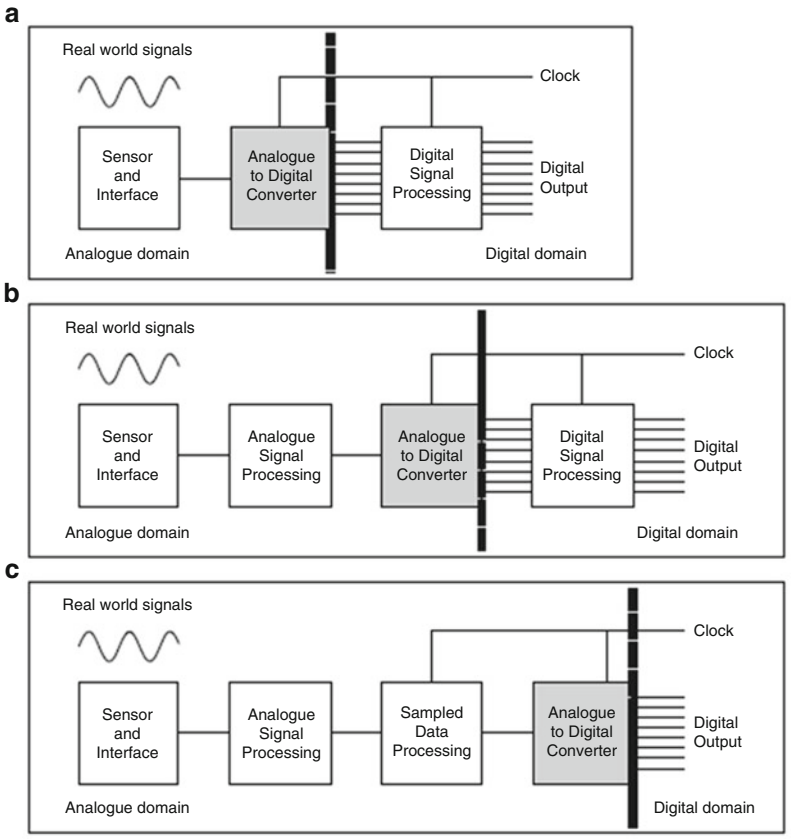


Fig. 7.4 Hybrid processing architectures with a single data conversion stage: (**a**) conventional analogue front-end with digital processor and output, (**b**) hybrid analogue/digital processing platform with digital output, and (**c**) hybrid analogue and sampled data processing platform with digital output

In addition to meeting these key criteria, CMOS-based chemical sensors are amenable to monolithic integration of interface circuitry and thus to the use of the bio-inspired hybrid processing architectures described herein which can potentially extract critical information from noisy signals in an adaptive, intelligent manner.

A key issue to consider is whether or not sensing and processing functions should be combined when this limits one's choice of sensors and materials, which can also introduce expensive sensor-specific process steps to standard commercial CMOS fabrication. Of the integrated sensing and electronic functions published so far, such as [27], none have been shown to offer higher signal to noise ratios than separate sensors and interface electronics, leading to observations by some [28, 29] that integration is not necessarily the most appropriate goal, given that the lifetime of sensors can be short. Keeping sensor and processing circuitry separate allows reusable electronic modules and an optimised fabrication process

for both electronics and sensor. However, the norm for other types of sensors is that integration traditionally results in better SNR. Most high spec sensors, e.g. MEMS microphones, pressure sensors, and even image sensors (APS), have integrated active gain element to significantly improve SNR.

For small systems, which do not require adaptive or intelligent discrimination, disposable off-chip sensors and reusable electronic modules may well be the most cost effective and high performance solution. However, for more complex applications, involving array processing on an array of homo- or heterogeneous sensors, such as bio-inspired learning, adaptivity, feature extraction or redundancy, then large-scale integration of sensors and electronic devices is paramount and CMOS technology is the current state-of-the-art for hybrid architectures.

CMOS-based chemical sensors and biosensors comprise of a physical transducer and a chemical or biological recognition layer. The reversible change in a physicochemical property of the recognition layer (such as mass, volume, optical absorption spectrum, conductivity, temperature) upon interaction with a chemical or biological target is converted by a transducer into an electrical signal such as frequency, current or voltage. Electrochemical sensors are a class of sensors with fast response times based on ionic charge transfer between the recognition layer and the analysed target causing changes in electrical potential or conductivity. One of the most popular types of CMOS-based electrochemical sensor is the family of field-effect transistor (FET-) based devices, whose lowest common denominator, the ion-sensitive field effect transistor (ISFET), will be discussed in more detail in this section.

7.4.1 Ion-Sensitive Field-Effect Transistor (ISFET)

Research into FET-based biochemical sensing began in 1970, when Piet Bergveld proposed a MOSFET without a gate metallisation as an ion-sensitive field effect transistor (ISFET) [30] for measuring the ionic flux around a neuron. Accumulation of charge at the exposed insulator (i.e., oxide) surface, is related to the activity (concentration) of ions in the sample solution, and modulates the threshold voltage of the transistor, causing shifts in the current-voltage (I–V) characteristic of the ISFET (Fig. 7.5). Being a potentiometric sensor, a reference electrode (such as silver/silver chloride) is required so that the phase boundary potential formed at the oxide-electrolyte interface can be measured with respect to a fixed boundary potential in the bulk of the solution. In transistor terms the reference electrode acts as a remote gate, which is capacitively coupled across the electrolyte to the insulator surface.

In the most common ISFET implementation, the silicon dioxide layer is replaced with a double layer insulator, which has better pH-sensitivity and stability to form a pH-ISFET, sensitive primarily to hydrogen ions (Fig. 7.6). Examples of the upper layer of these double insulator structures are silicon nitride, aluminium oxide and tantalum oxide. Though an ISFET by definition is a FET without a gate, a goal of

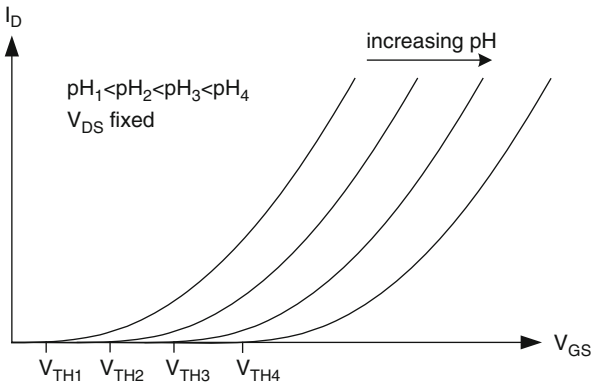


Fig. 7.5 pH-ISFET transconductance characteristics showing that increasing pH has same effect as adjusting threshold voltage of underlying MOSFET

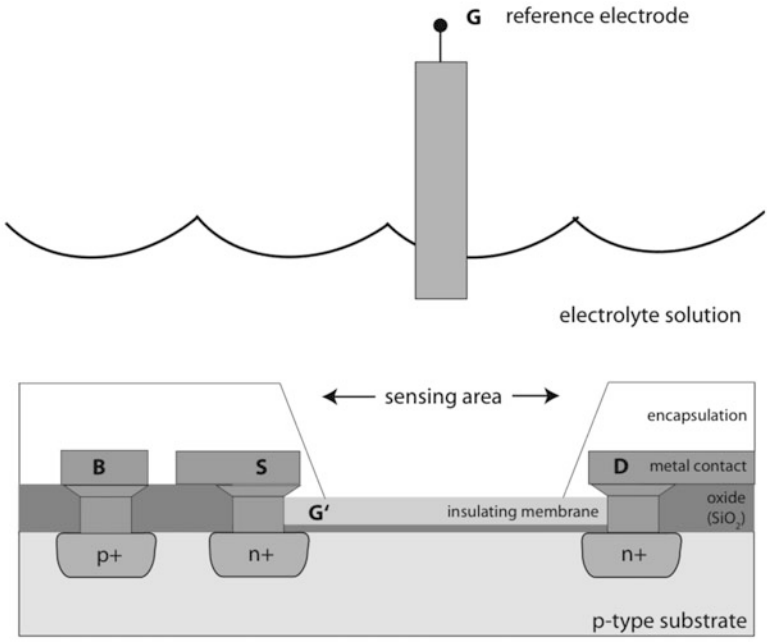


Fig. 7.6 Conventional pH-ISFET fabricated using a custom process flow (non-standard). The MOSFET gate is replaced by the sensing membrane to directly couple ions (H^+) within the electrolyte to the device channel

recent research has been to fabricate FET-based devices with ion sensitivity using a standard commercial CMOS process [31, 32]. These CMOS-based ISFETs circumvent the fact that the CMOS process requires a polysilicon gate for self-alignment of the source and drain diffusions by using the silicon nitride passivation layer on top

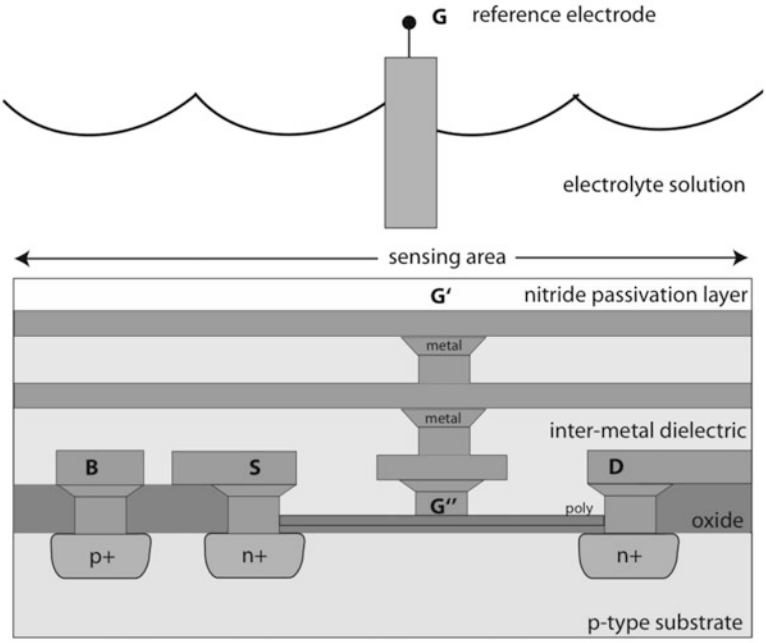


Fig. 7.7 Modified pH-ISFET fabricated in a standard CMOS technology (single poly, three metal layer process illustrated). The polysilicon gate is connected to the top metal layer with silicon nitride passivation acting as sensing membrane

of the polysilicon gate as the sensing membrane as shown in Fig. 7.7 [26]. Charge accumulation on the passivation layer (typically either silicon dioxide or silicon nitride) is capacitively coupled to the ohmic multiconductor "floating gate" structure of metal layers and polysilicon beneath, thus influencing the inversion within the semiconductor channel and therefore device drain-source current.

7.4.2 ISFET-Based Biosensors

Through modification of the sensing membrane, and/or the addition of further layers of sensitive materials, which interact with target species to form ions, ISFETs can be used to sense ions other than hydrogen, as well as gases, antigens and metabolites such as urea and glucose.

7.4.2.1 ChemFET

Through the deposition of various ionophores (ion-selective channels) on top of the insulator sensing membrane, ChemFET sensors for key ions such as sodium, potassium and calcium can be created based on the same principle [33, 34].

7.4.2.2 GasFET

Gas-sensitive FETs or "GasFETs" can be made by taking an ISFET and surrounding it in a thin film of intermediate electrolyte solution enclosed by a gas-permeable membrane. This is often achieved by localising the electrolyte solution to the ISFET surface using a hydrogel, and covering this with a gas-permeable polyimide layer. Thus, the gas of interest diffuses through the membrane and undergoes a chemical reaction with the electrolyte, consuming or forming an ion to be detected by the underlying ISFET. The local activity of this ion is proportional to the amount of gas dissolved in the sample, and the ISFET response is directly related to the concentration of sample. One example of this is the Severinghaus method for the detection of carbon dioxide [35, 36].

7.4.2.3 EnFET

Enzyme-FETs or EnFETs use the specific binding capabilities of enzymes as well as their biocatalytic activity to create FET-based biosensors. Enzymatic action of an enzyme on its substrate gives rise to the production or consumption of ions, which can be detected, by an underlying ISFET or ChemFET if the enzyme is immobilised sufficiently close to its sensing membrane.

The first EnFET was proposed by Janata and Moss in 1976 [37], and was discovered in 1980 with a penicillin-sensitive biosensor using the enzyme penicillinase to catalyse the hydrolysis of penicillin – a reaction which produces hydrogen ions [38]. Since then, a wide range of EnFETs have been reported for the detection of glucose, sucrose, maltose, ethanol, lactose, urea and creatinine among others [39, 40].

EnFET construction involves the attachment of an enzyme or an enzyme-containing layer onto the inorganic gate insulator of the underlying ISFET or ChemFET sensor – a research topic that has generated much investigation. Several methods and protocols have been attempted with varying degrees of success, including physical or chemical adsorption and entrapment within polymeric matrices, covalent bonding, cross-linking by bifunctional agents such as glutaraldehyde and mixed physicochemical methods. The simplest and most frequently used methods are the drop-on technique and the spin coating or dip coating of a mounted sensor chip into an enzyme solution [39]. To improve often-poor adhesion of the enzyme layer, prior surface silinisation of the inorganic gate insulator is often performed.

Specific difficulties associated with EnFETs, other than problems of enzyme adhesion, are their nonlinear and limited dynamic range. Also, the buffer capacity of the sample solution, which is pH-dependent itself, will often counteract the ion-generating or ion-depleting reaction catalysed by the enzyme in a non-linear manner. This, however, is not a major concern if the enzyme layer is immobilised

Table 7.3 Summary of various ISFET sensing targets and recent applications

Year	Ref.	Sensing target	Technology	No. of ISFETs	Application
1997	[41]	Glucose/sucrose	Custom	1	Sugar sensing
1997	[42]	Urease enzyme	Custom	1	Hemodialysis
2001	[43]	H+	Custom	12/20	Cell activity monitoring
2001	[44]	Nitrate	Custom	4	Soil analysis
2004	[45]	K+, H+, penicillin	Custom	4	Physical and chemical detection
2007	[46]	H+	Custom	1	Cell position detection
2007	[47]	H+	Custom	1	Monitoring of yeast fermentation
2008	[48]	H+	0.35 μm CMOS	256	Proton imaging
2008	[49]	K+, Na+, and Cl−	Custom	6	Analysis of mineral water
2010	[50]	H+	0.35 μm CMOS	64	Multifunction analysis
2010	[51]	H+	0.35 μm CMOS	40	DNA SNP detection
2011	[52]	H+	0.35 μm CMOS	13 M	DNA sequencing

close enough to the pH-sensing gate. Another source of non-linearity is the pH-dependency of enzyme kinetics, although this is well modelled and therefore has the potential to be overcome through intelligent local processing.

A summary of various ISFET sensing targets and recent applications is provided in Table 7.3.

7.4.3 Towards Biochemically-Inspired Processing with ISFETs

Despite some difficulties in terms of reliability and reproducibility, the ISFET and its derivatives have been by far the most popular miniaturised potentiometric sensors over recent decades, and the number of ISFET-related publications between 1999 and 2005 approaches 400. Current research directions in this field are the optimisation of the CMOS-based fabrication process; development of on-chip reference electrodes; drift and temperature compensation techniques; cancellation of interference from other ions and modification of the sensing membrane to sense different ions, metabolites and antigens.

Research thus far has not, however, strayed from traditional interface techniques of instrumentation amplifiers and op-amps – the main innovation being that these are now being integrated on-chip. Yet ISFETs, being transistor-based, give plenty of scope for exploration of existing knowledge of device physics and circuit techniques that we are familiar with from the MOSFET design summarised in Tables 7.1 and 7.2 and their application to devices with a chemical input.

7.4.3.1 Weak Inversion Operation

A first step towards exploiting device characteristics of ISFET-based sensors has been made by operating them in the ultra low power current-mode analogue region known as "weak inversion" or "subthreshold", where in digital terms the transistor would be considered "switched off". The current-mode analogue approach [14] allows operations such as addition, subtraction, multiplication, division, scaling, thresholding, power law operations and filtering to be performed at a fraction of the silicon area and power associated with digital processing, as discussed in Sect. 7.3.

In the weak inversion region, MOSFET-based transistors are characterised by the diffusion of electrons across the channel, rather than the drift across an electric field when the device is "switched on" above a given threshold voltage. This operating region is characterised by current levels typically from 1pA to 10 nA, and is generally powered by low (~1 V) power supply voltages, leading readily to the realisation of analogue micro-powered designs. Boltzmann diffusion of electrons in weakly-inverted MOSFET devices dictates an exponential voltage to current relationship:

$$I_D = I_0 \frac{W}{L} e^{V_{GS}/nU_T} \tag{7.1}$$

where I_D is the drain current, I_0 is the pre-exponential multiplier, W and L are gate width and length, VGS is the gate-source voltage, n is the subthreshold slope factor and $U_T = kT/q$ is the thermal voltage.

When an ISFET is operated in weak inversion [53], this equation is modified to include the term V_{chem} that accounts for the linear modulation of ISFET threshold voltage by the pH of the solution:

$$I_D = I_0 \frac{W}{L} e^{V_{GS}/nU_T} e^{-V_{chem}/U_T} \tag{7.2}$$

V_{chem} groups various chemical potentials between the reference electrode and the ISFET surface insulator (e.g. Si_3N_4) such as liquid junction potentials and work functions and, to a first approximation, is linearly proportional to pH with a slightly sub-Nernstian sensitivity of 55 mV/pH at 298 K. This is primarily due to the Boltzmann distribution of hydrogen ions in the electrolyte's diffusion layer, which gives rise to a potential across the electrolyte that varies logarithmically with hydrogen ion concentration and hence linearly with pH.

The dependence of V_{chem} on pH is modelled using a combination of the site-binding theory and the Gouy-Chapman-Stern double-layer theory to model charge distribution across the electrolyte. The ISFET is based on a MOSFET with a remote gate (reference electrode, G), exposing a chemically-sensitive insulator (G′) to an

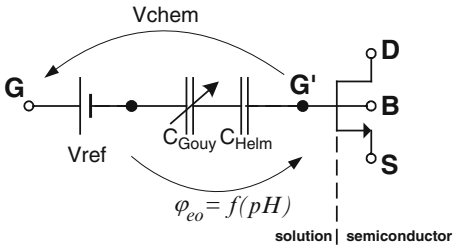


Fig. 7.8 ISFET Behavioural Macromodel consisting of the underlying field-effect transistor (shown on *right*) and pH dependent input network (i.e. V_{ref}, C_{Gouy} and C_{Helm}) generating V_{chem}

electrolyte (Fig. 7.8) and can be represented by a behavioural macromodel such as that of Martinoia et al. [54] in which V_{chem} is given by:

$$V_{chem} = \gamma + 2.303\alpha U_T pH \tag{7.3}$$

where γ is a grouping of pH-independent chemical potentials and α varies between 0 and 1 and relates ISFET sensitivity $S = dV/dpH$ to the ideal Nernstian sensitivity $S_N = 2.303U_T$.

Substituting the sub-Nernstian logarithmic property of the electrolyte (Eq. 7.3) into the Boltzmann exponential distribution of the ISFET (Eq. 7.2):

$$I_D = I_0 exp\left(\frac{V_{GS}}{nU_T}\right) K_{chem}[ionX]^{\alpha/n} \tag{7.4}$$

where $K_{chem} = exp(-\gamma/nU_T)$ is a pH-independent constant, with γ, n and U_T defined as before. The operation of the ISFET in weak inversion thus generates a power-law relation between drain current and ion concentration for a fixed gate-source voltage of the form $I_D = k_1[ion]^{k_2}$ as in Eq. 7.5. This is because the exponential current-voltage characteristic of the transistor in weak inversion is countered by the logarithmic voltage-ionic concentration of the Nernst equation. It is the effective cancellation of the same physical phenomenon – diffusion: one ionic, the other electronic [55].

The advantage of operating in this region is twofold:

- The lower current bias and voltage levels required to operate the transistor in its weak inversion region results in significantly reduced supply power consumption.
- The exponential transconductance characteristic means that these devices can be used as "translinear elements" in the synthesis of static and dynamic translinear circuits – the current-mode analogue design methodology for realising mathematical operations such as power law manipulations, multiplications, correlations and filtering.

7.4.3.2 Translinear Design Methodology

The Translinear Principle was introduced by Barry Gilbert for bipolar transistors in 1975 [20], and is one of the most important circuit theory contributions in the

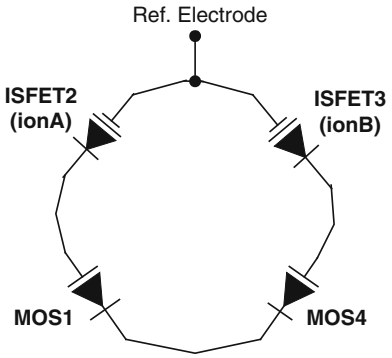


Fig. 7.9 A translinear loop for which the ratio of clockwise current densities to anticlockwise current densities is proportional to the concentration ratio of ion B to ion A (to a known power)

analysis and synthesis of nonlinear circuits. Due to their exponential characteristics, the principle has been extended to MOS transistors in weak inversion [56] for the realization of ultra-low power signal processing circuitry.

The principle has recently been extended to the operation of the ISFET in the weak inversion region to form a "Biochemical Translinear Principle" [57]. The simple "translinear loop" shown in Fig. 7.9 has the property that the product of the clockwise currents is equal to the product of the anticlockwise currents, scaled by the ionic concentration ratio of ion B over ion A to a known power:

$$\frac{I_{D4} \cdot I_{D3}}{I_{D1} \cdot I_{D2}} = \frac{K_{chemB}[ionB]^{\alpha/n}}{K_{chemA}[ionA]^{\alpha/n}} \tag{7.5}$$

This Biochemical Translinear Principle can be manipulated to perform many mathematical operations on biochemical signals. It is immediately apparent from the simple translinear loop shown in Fig. 7.9 that if the enzymes urease and creatinase were bound to ISFETs 2 and 3 respectively, then this circuit would calculate, in real-time, the urea to creatinine concentration ratio to a known power with just four transistors, with no need for digital processing.

A second, purely electronic Translinear loop could then be used to raise this concentration ratio to the power of 1. Plasma urea to creatinine ratio is a key biomarker in the signalling of renal failure and gastrointestinal bleeding, and is an example of a useful parameter in the context of real-time monitoring for BSNs.

As well as real-time concentration ratio calculations, translinear circuits are ideally suited to power law manipulations, be it exponents or roots. The circuit in Fig. 7.10 performs a squaring function designed to linearise Eq. 7.5. Linear sensor output with respect to concentration of species X as opposed to pX, which is logarithmic, is important for subsequent closed-loop feedback applications on-chip, such as drug or hormone delivery, since the control problem is significantly reduced if the sensor system is linear with respect to the real-world biochemical parameter [58]. This circuit shows how easily the sensor output can be linearised in the current-mode analogue domain using translinear circuits. With the inclusion of capacitors into such circuits, log-domain and Dynamic Translinear

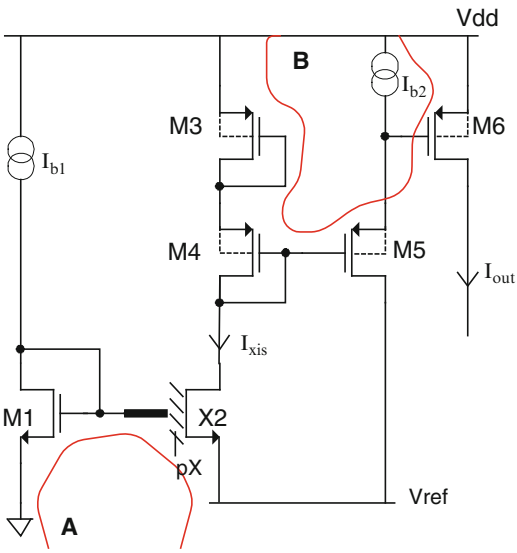


Fig. 7.10 Translinear loop (consisting of devices M1, X1, M3, M4, M5, M6) ensures I_{out} is linearly proportional to the concentration of the ion species detected by ISFET X2 [43]

circuits can be synthesised, which is useful in performing frequency-dependent filtering functions on sensed biochemical signals [59].

Of the operations listed in Tables 7.1 and 7.2, only two (division and squaring) have been demonstrated using the current-mode analogue approach with weak inversion devices. Of the remaining operations, some are trivial using this approach (e.g. addition), but much remains to be explored in terms of applying some of the analogue techniques summarised in these tables, both current- and voltage-mode. For fully optimised, bio-inspired systems, hybrid architectures featuring the interaction between analogue and digital processing should be further investigated.

7.4.4 An ISFET-Based ASIC for Rapid Point-of-Care Gene Detection

Point-of-care diagnostics for detection of genetic sequences require biosensing platforms that are sensitive to the target sequence and are also fast, mass-manufacturable and, ideally, disposable. Conventional lab-based methods of detecting DNA sequences rely on optical methods, typically by the addition of fluorescent tags to the target DNA that in turn latches onto a DNA probe sequence but only if there is a match between the two. These techniques are cumbersome as they require upfront tagging of the DNA with expensive reagents and laboratory equipment to detect the optical signals. Recently, developments have been made in transferring these optical methods to inexpensive CMOS ICs [60], although the requirement for tagging remains. Magnetic beads offer an alternative means of tagging the DNA and their presence can be detected by the shift in resonant frequency of an on-chip

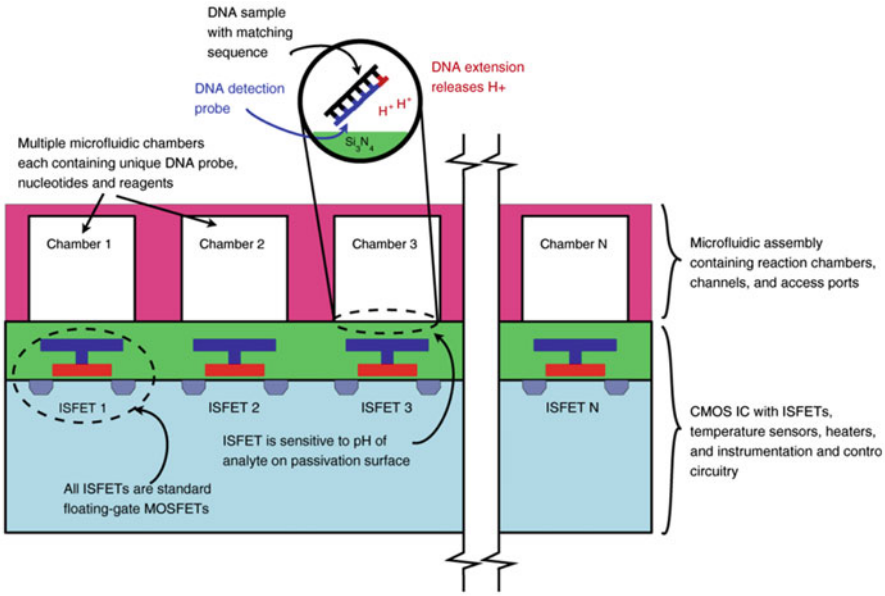


Fig. 7.11 Cross-section through the cartridge containing the Integrated Circuit (*IC*) and the microfluidic assembly. Each microfluidic chamber exposes different ISFETs to different test solutions (each containing unique DNA probe, nucleotides and reagents). Multiple channels implemented on a single ASIC with all associated instrumentation and control

LC tank [61]. There have also been attempts based on "label-free" electrochemical detection using FETs [62], but none of these have been implemented in unmodified standard CMOS.

This example describes an all-electrical approach that does not require tagging of DNA [51]. The principle of label-free electrochemical DNA detection using ISFETs was originally described in [63]. However, this prior art used a custom (i.e., non-CMOS) FET with an exposed Al_2O_3 gate insulator, a macroscopic reference electrode and a heat bath with a control unit to control the temperature of the reagents. Furthermore, it was only able to detect a single DNA sequence, which has limited application. The SOC for DNA detection presented, herein integrates the discrete laboratory apparatus of [63] into a single IC while using standard CMOS manufacturing (unlike the all-electrical approach in [64]), and moreover has scalability to the detection of multiple DNA sequences by the use of multiple on-chip ISFET sensors, temperature sensors and heaters along with ADCs and digital IO. This single-chip solution can be incorporated into disposable cartridges that are interchangeable depending on the DNA sequences to be detected. The structure of the cartridge is shown in Fig. 7.11 and consists of the IC embedded within a microfluidic assembly in which there are microchambers with access channels for the delivery of reagents. The ISFETs are implemented as unmodified floating-gate pMOS devices [27].

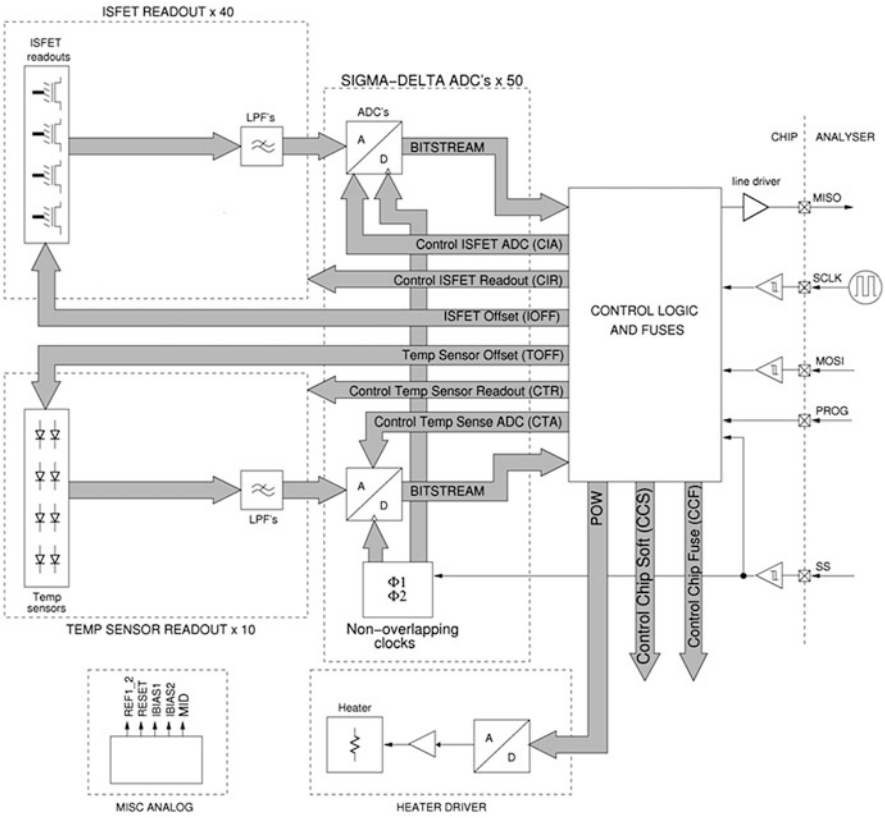


Fig. 7.12 Architecture of the System-on-Chip (SOC) the DNA detection system

When there is a match between target DNA from a sample and the on-chip DNA probes, H+ ions are released, changing the pH of the reagents inside the micro-chamber in which the match took place (Fig. 7.11). ISFETs are sensitive to the pH of liquids on the passivation interface above their floating gates and this manifests itself as a shift in the Vt of the ISFET where Vt is the threshold voltage referred to a reference electrode that sets the potential of the electrolyte [31]. Hence, the IC can detect whether a reaction has taken place in any of the chambers by monitoring the Vt of the ISFETs beneath those chambers. There are many ways to do this, one of which is to pass the drain current through weak-inversion signal processing circuits for the optimum balance of sensitivity and dynamic range [53, 55].

The architecture of the SOC is shown in Fig. 7.12. The IC consists of a bank of 40 ISFETs with associated readout circuitry that allows one of the 40 ISFETs to be used as a reference ISFET (termed a REFET). Thus all ISFET readout signals are outputted differentially as (VISFET-VREFET): this means that all common-mode noise (including electrochemical noise in the reaction chambers, allowing a quasi-reference electrode to be used) is rejected and this, together with up-front gain,

Table 7.4 Key performance metrics of the DNA detection system incorporating the SOC

Parameter	Value	Comments
Reaction chambers	40 per IC	Scalable on-chip or externally by daisy-chaining IC's
Temperature control accuracy	0.5 °C	
Intra-die temp. control precision	0.1 °C	Individual sensor trimming via 8-bit DAC's
Max. no. of daisy-chained IC's	400	Electrical limit
Anti-aliasing filter f3dB	8 Hz	Gm-C using 0.1nS transconductor
DSM OSR	1,024	
Down-sampled sample rate	1 Hz	
Decimation filter	16,384-tap FIR	Implemented as software in the analyser
Peak system SNR	50 dB	Real conditions: reference electrode as input and including electrochemical noise
Input-referred noise voltage	25 μV	Referred to ISFET floating polysilicon gate

gives an input-referred noise voltage of only 25 μV over the bandwidth of interest. Floating-gate transistors have different levels of as deposited charge, which cause large variations in the output voltage: these variations are cancelled using individual 10-bit DACs, thus preventing saturation of the instrumentation circuitry, which is designed to resolve small signals. Therefore, this architecture overcomes the two common performance issues with standard CMOS ISFETs: electrochemical drift and Vt offsets due to charge [31].

Because biochemical reactions are highly sensitive to the temperature of enzymes and reagents, the system controls its temperature to an accuracy of 0.5 °C with an inter-chamber precision of 0.1 °C. This is done with an array of ten temperature sensors distributed across the chip that is outputting the differential signal (VPTAT-VCTAT). These sensors are trimmed using individual 8-bit current-steering DACs fed from registers sitting within an OTP block. The ISFET and temperature signals are filtered by 50 g-C anti-aliasing LPFs with a low f_{3dB} of 8 Hz. The filtered signals are passed through 50 2nd-order ΣΔMs with a peak SNR of 80 dB. The resulting bitstreams are gathered by control logic which implements an SPI interface for full-duplex communication with an external handheld analyser, which also does decimation filtering. The IC as it currently stands, can detect up to 39 different genetic letters (allowing for one channel to be used for the reference chamber). Scalability is a key aspect to the design of the IC because as research into genetics advances, applications will be found that require the identification of more and more genetic letters. To that end, the IC is both scalable internally, in that the architecture allows more ISFETs and temperature sensors to be easily added, and the SPI interface is implemented such that multiple ICs can be daisy-chained together. A table of key performance specifications is given in Table 7.4 in which all values are exact or typical unless otherwise stated.

The sensitivity of the IC to pH of liquids in the microchambers is shown in Fig. 7.13; the output signal has a square-law dependence on pH and the dotted line is a 2nd-order polynomial least squares best fit. Deviations from this best fit are due

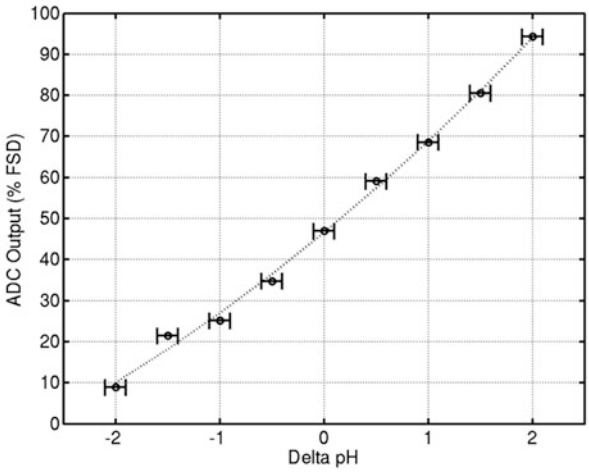


Fig. 7.13 ADC decimation filter output as the pH of the liquid in a reaction chamber is varied

to the difficulty of precisely setting the pH of buffers as shown by the error bars, and do not represent a limitation of the IC itself.

This platform can be applied to many different kinds of genetic testing. One application is the identification of single nucleotide polymorphisms (SNPs), which are a change in a single letter of the genetic code. The IC has sufficient ISFETs to be able to identify up to 19 SNPs but for demonstration purposes in this paper, only one SNP was detected, utilising only three of the ISFETs: one as a REFET described above, and the other two beneath chambers containing reagents suitable for identifying a SNP which reveals information about a person's drug metabolism. This SNP can take the letter A or C only and so each of two chambers was loaded with different DNA detection "probes" suitable for capturing the target region of the DNA and detecting these two letters. In this case, a reaction is deemed to have occurred when the output signal from either or both chambers crosses a preset threshold: if this happens, there is a match between the patient's DNA and the letter of interest. The die is shown in Fig. 7.14. In a 0.35 μm process, it has an area of 26 mm^2.

7.5 Future Outlook

In a BSN with limited bandwidth and power constraints, the conventional method of data acquisition and analogue-to-digital data conversion with signal processing taking place after transmission is not optimal. BSNs are a prime candidate for bio-inspired local processing to take place at the sensor front-end before transmission. This processing could not only include spatial and temporal averaging for drift and failure tolerance, but also for trendspotting and adaptivity. The key principle of bio-inspired engineering in this application area is that biology does not often deal in absolute values, but in relative changes from a given norm.

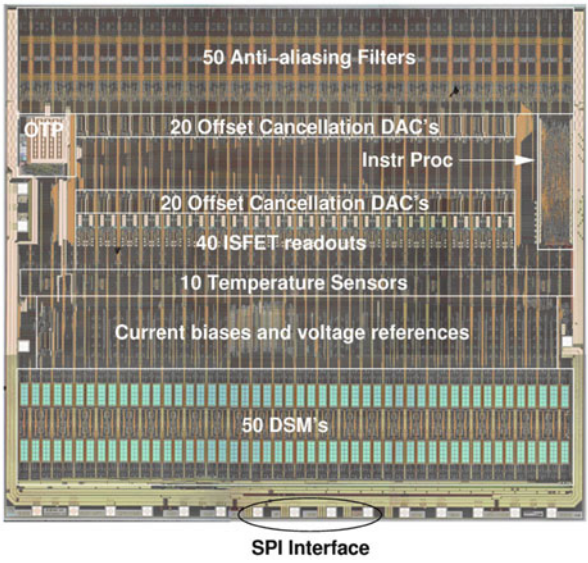


Fig. 7.14 Microphotograph of the ASIC manufactured in a 0.35 μm CMOS technology, with a die size of 5.5 × 4.7 mm. System shown contains 40 independent pH-sensing channels

Ultra-low power signal processing is not limited to ISFET-based devices, and design methodologies such as translinear circuit synthesis do not need to be applied directly to the sensor as they were in Sect. 7.4.3 – indeed, this can only apply to CMOS devices with exponential transconductance properties such as diodes or transistors in weak inversion. The analogue signal processing design methodologies should thus be applied as close to the sensor front-end as possible, thereby building intelligence and adaptivity into the sensor. One example lies in predictive array processing for robustness of sensor output by compensating sensor drift, temperature dependence, sensor failure and interference [50].

Another direction for on-chip intelligence is the implementation of the Hilbert transform for voltammetric sensors as mentioned in Chap. 2 as an analogue filter bank rather than off-chip digital processing, which would give far-reaching instantaneous discrimination capabilities. In optics, interest is developing in the field of integrated optics, which is leading to the development of small chip-sized UV-visual and near-IR spectrometers [65]. A future direction could be the application of expertise from the field of bio-inspired vision chips using hybrid, distributed processing techniques in an embedded photodiode array to facilitate feature extraction [66–68]. Similarly, in ECG sensors, micropower hybrid implementation of low/band pass filters, derivative and template matching functions at the electrode input can extract features such as QRS-wave (heartbeat) complex and detect abnormalities including arrhythmias without the need to continually stream the raw ECG data.

Further scope for bio-inspired design lies in adaptive therapeutics based on neural and metabolic cell modelling. Instead of providing feedback to biological systems using traditional multivariate analysis or PID control, one can use

electronic models of excitable cells. It has already been shown that neural models such as the Hodgkin and Huxley model can be realised in silicon with weak inversion analogue devices [69] and this principle can be extended to metabolic cells in the body such as pancreatic beta cells, whose activity is dominated by action potentials generated by cellular sodium, potassium and calcium ion concentrations. Fully integrated CMOS sensors and processing circuitry will have the ability to measure intra- and extra-cellular ion concentration and to then generate the appropriate signalling to neighbouring cells via electrodes or to a biomimetic hormone delivery unit for example [59]. In addition to metabolic applications, this type of architecture can be useful in neuroprosthetic systems for sensory rehabilitation, for example in restoring balance to individuals with vestibular dysfunction [70].

In summary, the biomedical applications which could benefit from the union of CMOS-based sensors with expertise in optimised hybrid electronics are unlimited if one dares to leave the domain of traditional interface techniques. Two current-mode analogue processing implementations have been proposed for CMOS ISFETs and their derivatives, but there remains an extensive toolkit of circuit techniques and mathematical operations to be explored with biosensors in mind, as well as a plethora of CMOS biosensors each with their own signal processing requirements. These biologically driven intelligent biochemical circuits hold the key to the low power in situ diagnostics and therapeutics in future BSN designs.

References

1. Mead CA. Analog VLSI and neural systems. Reading, Massachusetts: Addison-Wesley, 1989.
2. Eftekhar A, Paraskevopoulou S, Constandinou, TG. Towards Next Generation Neural Interfaces: Optimising Power, Bandwidth and Data Quality. In IEEE Biomedical Circuits and Systems Conference. 2010:122–125.
3. Paraskevopoulou SE, Barsakcioglu D, Saberi M, Eftekhar A, Constandinou TG. Feature Extraction using First and Second Derivative Extrema (FSDE) for Real-time and Hardware-Efficient Spike Sorting. Journal of Neuroscience Methods, 2013; 215(1): 29–37.
4. Constandinou TG, Georgiou J. Micro-Optoelectromechanical Tilt Sensor. Journal of Sensors, 2008:1–7.
5. Gregorian R, Temes GC. Analog MOS integrated circuits for signal processing. John Wiley and Sons, 1986.
6. Constandinou TG, Georgiou J, Toumazou, C. A Micropower Front-end Interface for Differential-Capacitive Sensor Systems. In IEEE International Symposium on Circuits and Systems. 2008:2474–2477.
7. Toumazou C, Hugues JBC, Battersby NC. Switched currents: an analogue technique for digital technology. IEE Publishing, 1993.
8. Hosticka BJ. Performance comparison of analog and digital circuits. Proceedings of the IEEE 1985; 73(1):25–29.
9. Vittoz EA. Future of analog in the VLSI environment. In: Proceedings of the IEEE International Symposium on Circuits and Systems, New Orleans, Louisiana, 1990; 2:1372–1375.
10. Furth PM, Andreou AG. Bit-energy comparison of discrete and continuous signal representations at the circuit level. In: Proceedings of the Fourth Workshop on Physics and Computation 1996.

11. Sarpeshkar R. Efficient precise computation with noisy components: extrapolation from an electronic cochlea to the brain. California Institute of Technology, PhD Thesis, 1997.
12. Sarpeshkar, R. Ultra low power bioelectronics: fundamentals, biomedical applications, and bio-inspired systems. Cambridge University Press, 2010.
13. Sarpeshkar, R. Universal Principles for Ultra Low Power and Energy Efficient Design. IEEE Transactions on Circuits and Systems II 2012; 59(4): 193–198.
14. Gilbert B. A precise four-quadrant multiplier with subnanosecond response. IEEE Journal of Solid-State Circuits 1968; 3(4):365–373.
15. Toumazou C, Lidgey FJ, Haigh DG. Analogue IC design: the current-mode approach. London: Peter Peregrinus Ltd., 1990.
16. Allen PE. Switched Capacitor Circuits. Springer, 2013.
17. Kinniment DJ, Garside JD, Gao B. A comparison of power consumption in some CMOS adder circuits. In: Proceedings of the International Workshop on Power and Timing Modelling optimisation and Simulation, Oldenburg, Germany, 1995; 119–132.
18. Ramirez-Angulo J, et al. Low-voltage CMOS analog four quadrant multiplier based on flipped voltage followers. In: Proceedings of the IEEE International Symposium on Circuits and Systems 2004; 1:681–684.
19. Satyanarayana JH, Parhi KK. A theoretical approach to estimation of bounds on power consumption in digital multipliers. IEEE Transactions on Circuits and Systems II 1997; 44(6):473–481.
20. Weste N, Harris D. Integrated Circuit Design: International Version: A Circuits and Systems Perspective. Pearson Education, 4th Ed., 2010.
21. Gilbert B. Current-mode circuits from a translinear viewpoint: a tutorial. In: Toumazou C, Lidgey FJ, Haigh DG (eds) Analogue IC design: the current-mode approach. London: Peter Peregrinus Ltd., 1990;11–91.
22. Quoc-Hoang D, Trung-Kien N, Sang-Gug L. Ultra low-voltage low-power exponential voltage-mode circuit with tunable output range. In: Proceedings of the IEEE International Symposium on Circuits and Systems, Vancouver, Canada, 2004; 2:729–732.
23. Ercegovac MD, Lang T. Division and square root: digit-recurrence algorithms and implementations. Kluwer Academic Publisher, 1994.
24. Drakakis EM, Payne AJ, Toumazou C. Log-domain filtering and the Bernoulli cell. IEEE Transactions on Circuits and Systems I 1999; 46(5):559–571.
25. Proakis J, Manolakis DG. Digital signal processing: principles, algorithms and applications. Pearson Education, 1995.
26. Hierlemann A, Baltes H. CMOS-based chemical microsensors. Analyst 2003; 128(1):15–28.
27. Bausells J, Carrabina J, Errachid A, Merlos A. Ion-sensitive field-effect transistors fabricated in a commercial CMOS technology. Sensors and Actuators B: Chemical 1999; 57(1–3):56–62.
28. Nam H, et al. Micropotentiometric sensors. Proc. IEEE 2003; 91(6):870–880.
29. Bergveld P. Thirty years of ISFETOLOGY: what happened in the past 30 years and what may happen in the next 30 years. Sensors and Actuators B: Chemical 2003; 88(1):1–20.
30. Bergveld P. Development of an ion-sensitive solid-state device for neurophysiological measurements. IEEE Transactions on Biomedical Engineering 1970; BM17(1):70–71.
31. Liu Y, Georgiou P, Prodromakis T, Constandinou TG, Toumazou C. An Extended CMOS ISFET Model Incorporating the Physical Design Geometry and the Effects on Performance and Offset Variation. IEEE Transactions on Electron Devices 2011; 58(12):4414–4422.
32. Prodromakis T, Liu Y, Constandinou TG, Georgiou P, Toumazou C. Exploiting CMOS Technology to Enhance the Performance of ISFET Sensors. IEEE Electron Device Letters 2010; 31(9):1053–1055.
33. Sibbald A, Covington AK, Carter RF. Online patient-monitoring system for the simultaneous analysis of blood K+, Ca2+, Na+ and pH using a quadruple-function CHEMFET integrated-circuit sensor. Medical and Biological Engineering and Computing 1985; 23(4):329–338.

34. Van der Wal PD, van den Berg A, Derooij NF. Universal approach for the fabrication of Ca(2+)-, K+- and NO3- - sensitive membrane ISFETs. Sensors and Actuators B: Chemical 1994; 18(1–3):200–207.
35. Tsukada K, Miyahara Y, Shibata Y, Miyagi H. An integrated chemical sensor with multiple ion and gas sensors. Sensors and Actuators B: Chemical 1990; 2(4):291–295.
36. Arquint P, van den Berg A, van der Schoot BH, de Rooij NF, Buhler H, Morf WE, et al. Integrated blood-gas sensor for pO2, pCO2 and pH. Sensors and Actuators B: Chemical 1993; 13(1–3):340–344.
37. Janata J, Moss SD. Chemically sensitive field-effect transistors. Biomedical Engineering 1976; 11(7):241–245.
38. Caras S, Janata J. Field-effect transistor sensitive to penicillin. Analytical Chemistry 1980; 52(12):1935–1937.
39. Schoning MJ, Poghossian A. Recent advances in biologically sensitive field-effect transistors (BioFETs). Analyst 2002; 127(9):1137–1151.
40. Dzyadevych SV, et al. Biosensors based on enzyme field-effect transistors for determination of some substrates and inhibitors. Analytical and Bioanalytical Chemistry 2003; 377(3):496–506.
41. Sohn BK, Cho BW, Kim CS, Kwon DH). ISFET glucose and sucrose sensors by using platinum electrode and photo-crosslinkable polymers. Sensors and Actuators B: Chemical. 1997; 41(1):7–11.
42. Pijanowska DG, Torbicz W. pH-ISFET based urea biosensor. Sensors and Actuators B: Chemical. 1997; 44(1):370–376.
43. Martinoia S, et al. Development of ISFET array-based microsystems for bioelectrochemical measurements of cell populations. Biosensors and Bioelectronics. 2001; 16(9):1043–1050.
44. Birrell SJ, Hummel JW. Real-time multi ISFET/FIA soil analysis system with automatic sample extraction. Computers and Electronics in Agriculture, 2001; 32(1):45–67.
45. Poghossian A, Schöning MJ. Detecting Both Physical and Biochemical Parameters by Means of ISFET Dev. Electroanalysis, 2004; 16(22):1863–1872.
46. Castellarnau M, et al. Integrated cell positioning and cell-based ISFET biosensors. Sensors and Actuators B: Chemical. 2007; 120(2):615–620.
47. Krommenhoek EE, et al. Integrated electrochemical sensor array for on-line monitoring of yeast fermentations. Analytical chem., 2007; 79(15):5567–5573.
48. Milgrew MJ, et al. A 16x16 CMOS Proton Camera Array for Direct Extracellular Imaging of Hydrogen-Ion Activity. In IEEE International Solid-State Circuits Conference. 2008; 590–638.
49. Ipatov A, Abramova N, Bratov A, Domínguez C. Integrated multisensor chip with sequential injection technique as a base for “electronic tongue” devices. Sensors and Actuators B: Chemical, 2008; 131(1):48–52.
50. Goh ZDC, Georgiou P, Constandinou TG, Prodromakis T, Toumazou C. A CMOS-based ISFET Chemical Imager with Auto-Calibration Capability. IEEE Sensors Journal, 2011; 11(12):3253–3260.
51. Garner, DM, Bai H, Georgiou P, Constandinou TG, Reed S, Shepherd LM, Wong W, Lim KT, Toumazou C, A Multichannel DNA SoC for Rapid Point-of-Care Gene Detection. In IEEE International Solid-State Circuits Conference. 2010; 27(4):492–493.
52. Rothberg HM, et al. An integrated semiconductor device enabling non-optical genome sequencing. Nature, 2011; 475(7356):348–352.
53. Shepherd L, Toumazou C. Weak Inversion ISFETs for ultra-low power bio-chemical sensing and real-time analysis. Sensors and Actuators B: Chemical 2005; 107(1):468–473.
54. Martinoia S, Massobrio G. A behavioral macromodel of the ISFET in SPICE. Sensors and Actuators B: Chemical 2000; 62(3):182–189.
55. Georgiou P, Toumazou C. ISFET characteristics in CMOS and their application to weak inversion operation. Sensors and Actuators B: Chemical. 2009; 143(1):211–217.
56. Andreou AG, Boahen KA. Translinear circuits in subthreshold MOS. Analog Integrated Circuits and Signal Processing 1996; 9(2):141–166.

57. Shepherd LM, Toumazou C. A biochemical translinear principle with weak inversion ISFETs. IEEE Transactions on Circuits and Systems I. 2005; 52(12):2614–2619.
58. Kalofonou M, Georgiou P, Ou C-P, Toumazou C. An ISFET based translinear sensor for DNA methylation detection. IEEE Transactions on Sensors and Actuators B: Chemical. 2011; 161(1):156–162.
59. Georgiou P, Toumazou C. A silicon pancreatic beta cell for diabetes. IEEE Transactions on Biomedical Circuits and Systems. 2007; 1(1):39–49.
60. Jang B, Cao P, Chevalier A, Ellington A, Hassibi A. A CMOS fluorescent-based biosensor microarray. In IEEE International Solid-State Circuits Conference. 2009:436–437.
61. Wang H, Chen Y, Hassibi A, Scherer A, Hajimiri A. A frequency-shift CMOS magnetic biosensor array with single-bead sensitivity and no external magnet. In IEEE International Solid-State Circuits Conference. 2009:438–439.
62. Sakurai T, Husimi Y. Real-time monitoring of DNA polymerase reactions by a micro ISFET pH sensor. Analytical chemistry. 1992; 64(17):1996–1997.
63. Purushothaman S, Toumazou C, Ou CP. Protons and Single Nucleotide Polymorphism Detection: A Simple Use for the Ion Sensitive Field Effect Transistor. Sensors & Actuators B-Chemical. 2005; 14:964–968.
64. Thewes R, et al., Sensor arrays for fully-electronic DNA detection on CMOS. In IEEE International Solid-State Circuits Conference. 2002; 1:350–473.
65. Diamond D. Principles of chemical and biological sensors. Wiley Interscience, 1998.
66. Constandinou TG, Toumazou C. A micropower centroiding vision processor. IEEE Journal of Solid-State Circuits. 2006; 41(6):1430–1443.
67. Ohta, J. Smart CMOS image sensors and applications. CRC Press, 2010.
68. Constandinou TG, Degenaar P, Toumazou C. An Adaptable Foveating Vision Chip. In: IEEE International Symposium on Circuits and Systems 2006:3566–3569.
69. Yu T, Cauwenberghs G. Analog VLSI biophysical neurons and synapses with programmable membrane channel kinetics. IEEE Transactions on Biomedical Circuits and Systems, 2010; 4(3):139–148.
70. Constandinou TG, Georgiou J, Toumazou C. A Neural Implant ASIC for the Restoration of Balance in Individuals with Vestibular Dysfunction. In IEEE International Symposium on Circuits and Systems. 2009:641–644.

Chapter 8
Multi-sensor Fusion

Guang-Zhong Yang, Javier Andreu-Perez, Xiaopeng Hu, and Surapa Thiemjarus

8.1 Introduction

In the previous chapters, we have discussed issues concerning hardware, communication and network topologies for the practical deployment of *Body Sensor Networks* (BSNs). The pursuit of low power miniaturised distributed sensing under a patient's natural physiological condition has also imposed significant technical challenges on integrating information from what is often heterogeneous, incomplete and error-prone sensor data. For BSNs, the nature of errors can be attributed to a number of sources; but motion artefacts, inherent limitation and possible malfunction of the sensors along with communication errors are the main causes of concern. In practice, it is desirable to rely on sensors with redundant or complementary data to maximise the information content and reduce both systematic errors and random artefacts. This, in essence, is the main drive for multi-sensor fusion, which is concerned with the synergistic use of multiple sources of information.

In cardiac sensing, for example, both ECG and haemodynamic signals, such as blood pressure and pulse wave transmission, have mutually correlated information of the heart due to physiological coupling of electrical and biomechanical function. In situations where the ECG signal is degraded, either due to poor electrode contact

G.-Z. Yang, PhD (✉) • J. Andreu-Perez
The Hamlyn Centre, Imperial College London, London SW7 2BY, UK
e-mail: g.z.yang@imperial.ac.uk

X. Hu
School of Computer Science and Technology, Dalian University of Technology, No 2, Linggong Road, Dalian 116023, China

S. Thiemjarus
National Electronics and Computer Technology Center, Pathumthani 12120, Thailand

G.-Z. Yang (ed.), *Body Sensor Networks*, DOI 10.1007/978-1-4471-6374-9_8,

or patient movement, joint analysis of additional sensors can ensure more reliable cardiac rhythm monitoring. This resolves some of the intrinsic ambiguities involved in rhythm disturbance when assessed by ECG alone [1]. Whilst the use of multiple sensors of the same type for error minimisation is relatively intuitive to understand, the reliance on different sensors in terms of both sensing type and location will require the use of general principles of pattern recognition and machine learning. In practice, the use of multiple sensors with information fusion has the following main advantages [2]:

- Improved *Signal-to-Noise Ratio* (SNR)
- Enhanced robustness and reliability in the event of sensor failure
- Extended parameter coverage
- Integration of independent features and prior knowledge
- Increased dimensionality of the measurement
- Improved resolution, precision and confidence
- Reduced uncertainty

The origin of sensor fusion can be dated back to the 1970s when it was studied extensively for robotics and defence research. To address some of the main issues in data fusion and unify the terminology and procedures involved, a Data Fusion Sub-panel to the *Joint Directors of Laboratories* (JDL) Technical Panel for C3 (command, control, communications) was established by the US Department of Defence in 1986. Under the JDL data fusion framework, five levels of processing have been defined, which include Sub-Object Data Association and Estimation (L0), Object Refinement (L1), Situation Refinement (L2), Significance Estimation or Threat Refinement (L3), and Process Refinement (L4) [3, 4]. The JDL also gave a definition of data fusion, which was subsequently refined as a multilevel, multi-faceted process dealing with the automatic detection, association, correlation, estimation and combination of data and information from single or multiple sources [5]. Although the JDL framework was mainly developed with a strong military emphasis in mind, some of the basic principles provided are still applicable to BSN.

8.1.1 Information Interaction

In general, the nature of information interaction involved in sensor fusion can be classified as *competitive*, *complementary*, and *cooperative* fusion [6–9]. In competitive fusion, each sensor provides equivalent information about the process being monitored. It typically involves the handling of redundant, but sometimes inconsistent, measurements. The nature of competitive sensing means that it is ideally suited for in situ multi-sensor calibration and fault tolerant sensing.

In complementary fusion, on the other hand, sensors do not depend on each other directly as each sensor captures different aspects of the physical process. The measured information is merged to form a more complete picture of the

phenomenon. The example given above on combined ECG and haemodynamic sensing is a typical case of complementary fusion. Another form of complementary fusion is the use of a predefined physical model to combine sensor readings to collectively estimate a higher level of measurement indices. For example, arterial compliance, which is related to aging and diseases such as arteriosclerosis, can be quantified from a pressure-volume relationship by measuring the ability of a vessel to distend with increasing transmural pressure.

In cooperative fusion, sensors work together to provide information that is not obtainable by any of the sensors alone. In stereovision, for example, the measured feature disparity in the image pairs allows the estimation of the depth and shape of the object. Due to the compounding effect, the accuracy and reliability of cooperative fusion is sensitive to any inaccuracies in all of the simple sensor components that are used. For BSNs, the main objective of sensor fusion is to combine information from different sensors to capture data with improved reliability, precision, fault tolerance and inferencing power to a degree that is beyond the capacity of each sensor.

8.1.2 Levels of Processing

From a data processing model point of view, sensor fusion can also be grouped into three different levels of fusion as shown in Fig. 8.1, i.e., *direct data fusion*, *feature-level fusion* and *decision-level fusion* [3].

If the sensors are measuring the same physical parameter with data derived being commensurate, raw sensor data can be directly combined. Otherwise, the data needs to be fused at the feature or decision level. For feature-level fusion, features are first extracted from the sensor data to form multi-dimensional feature vectors so that general pattern recognition methods can be applied. For decision-level fusion, however, the information used has already been abstracted to a certain level through preliminary sensor or feature level processing such that high-level decision can be made. Popular techniques used for this level of fusion include classical inference, Bayesian inference, and Dempster-Shafer's method. Decision level fusion is also an ideal place to incorporate a priori knowledge and high-level domain specific information.

In terms of data communication, the higher levels of data abstraction can help reduce the bandwidth requirement of the sensor network. In a distributed sensor network, feature-level or decision-level fusion allows effective deployment of heterogeneous and independent sensor clusters. For multi-sensor data fusion, it is usually assumed that communication, storage and processing systems are reliable and that the focus is on fusion algorithms that can integrate data from either homogeneous or heterogeneous sources [10].

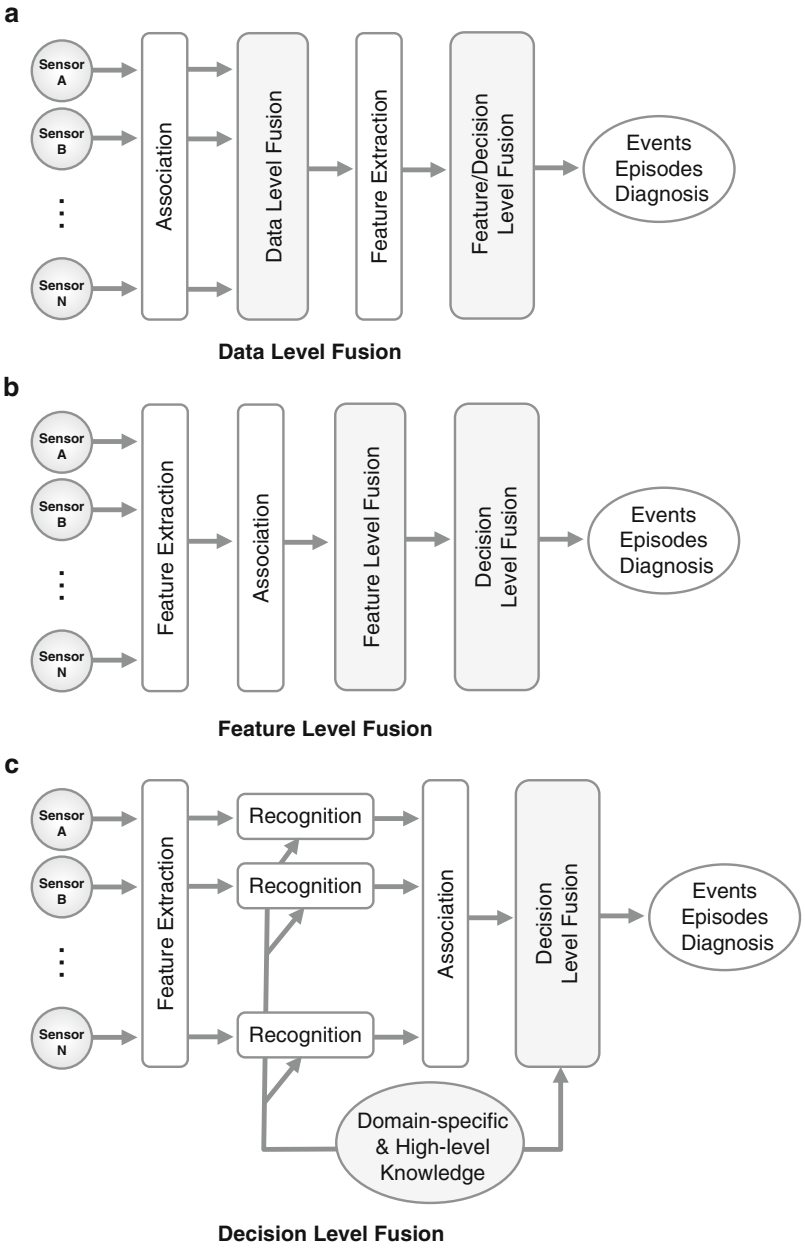


Fig. 8.1 Schematic diagrams showing three different fusion architectures at data, feature and decision levels (Adapted from DL Hall [3])

8.2 Direct Data Fusion

Direct data fusion is useful for sensor arrays, which, unlike the traditional sensor designs, are typically based on the use of redundancy for ensuring sensor accuracy. With sensor fusion, it is possible to overcome some of the inherent limitations of each single element of the ensemble. Another use of direct data fusion for BSN is self-calibration. Traditionally, calibration is performed during the production stage of the sensors and recalibration is necessary periodically for most sensors since factors such as aging, thermal drift, decay and damage can have a detrimental effect on the accuracy of the readings. Frequent recalibration of large-scale sensor networks, however, can be problematic due to the number of sensors involved. Furthermore, MEMS-based sensors such as those motion sensors are usually not calibrated after production and the sensors can be sensitive to offset bias on each axis. This section, will outline two examples of direct data fusion for optimal averaging with outlier detection for sensor arrays and source separation of mixed signals from a set of networked sensors to illustrate the role of direct data fusion.

8.2.1 Optimal Averaging for Sensor Arrays

In its most basic form, sensor fusion can be implemented as a simple average of all sensor readings. This approach, however, is not robust as any error in individual measurements would be included in the final estimate. Weighted averaging reduces the contribution from the worst sensors but erroneous sensor measurements are still included in the final estimate. Numerically, the errors involved in sensors can be attributed to systematic errors or so called *bias*, which is an offset of the mean amplitude of the sensor readings from the true value. This bias can be time-dependent and affected by external factors such as thermal and chemical drifts. Another source of error is random error or noise. This random component can be attributed to hardware noise or other unpredictable transient signals that cause random fluctuations. In sensor fusion, the statistical distribution of the random error can be modelled with a priori knowledge. In the absence of such information, Gaussian distributions are often adopted.

By taking scaling, bias and random noise errors into account, the output $x_i(t)$ of the ith sensor in relation to the original signal $s_0(t)$ can be represented as:

$$x_i(t) = (1 + \gamma_i(t))s_0(t)(t) + b_i(t) + n_i(t) \tag{8.1}$$

where $n_i(t)$ is the measurement noise, and $\gamma_i(t)$ and $b_i(t)$ are the scale and offset biases, respectively. Given a set of N noisy measurements, the recovery of the true signal for the problem formulated above is *ill-posed*. It is only solvable with regard to certain constraints or specific objective functions. For example, Unser and Eden [11]

have shown that the optimal weighting coefficients for N noisy channels can be determined by maximising a quadratic SNR defined by:

$$\mathrm{SNR} = N \frac{\sum_{i=1}^{K} (\overline{s}_i - \overline{s})^2}{\frac{1}{N} \sum_{i=1}^{N} \|\mathbf{s}_i - \overline{\mathbf{s}}\|^2} \tag{8.2}$$

where the temporal signal associated with each sensor is discretised as a K-dimensional vector, i.e., x_i and s_i represent the measured and the source signal for sensor i, respectively, and

$$\overline{\mathbf{s}} = [\overline{s}_1, \ldots, \overline{s}_K]^{\mathrm{T}} \tag{8.3}$$

which represents the ensemble average of $\mathbf{s}_i (i = 1, \ldots, N)$. In the above equations, T denotes vector transpose, and $\|.\|^2$ is the square norm of a vector. In (8.2), $\overline{s}_i$ represents the cross channel average for sample i, and $\overline{s}$ the mean signal given by:

$$\overline{s} = \frac{1}{K} \sum_{i=1}^{K} \overline{s}_i \tag{8.4}$$

The above SNR, in fact, represents the ratio between the rescaled signal and residual noise energies. In (8.2), the relationship between $\mathbf{s}_i$ and $\mathbf{x}_i$ is given by:

$$\mathbf{s}_{i=} w_i [\mathbf{x}_i - b_i \mathbf{I}] \tag{8.5}$$

where $\mathbf{I}$ is a K-dimensional vector of all 1's, and $w_i (i = 1, \ldots, N)$ are the optimal coefficients to be sought. By referring to (8.1) and ignoring n_i, it can be seen that the ideal value of w_i is $(1 + \gamma_i)^{-1}$.

It can be proved that the optimal estimate of the scale and offset biases shown in (8.1) can be derived independently [11]. For any given coefficient $\mathbf{w} = [w_1, \ldots, w_N]^{\mathrm{T}}$, the optimal value of b_i that maximises (8.2) is given by:

$$b_i = \overline{x}_i \ (i = 1, \ldots, N) \tag{8.6}$$

i.e., the average of the measured signal values. Furthermore, it is shown by Unser and Eden [11] that the optimal weighting coefficient $\mathbf{w} = [w_1, \ldots, w_N]^{\mathrm{T}}$ is given by the first generalised eigenvector of the characteristic equation:

$$\mathbf{R}\mathbf{w} = \beta \mathbf{D}\mathbf{w} \tag{8.7}$$

where $\mathbf{R} = [r_{ij}]$ is the $N \times N$ centred inner product matrix defined by

$$r_{ij} = \mathbf{x}_i^{\mathrm{T}} \mathbf{x}_j - K \overline{x}_i \overline{x}_j \tag{8.8}$$

and $\mathbf{D}$ is the corresponding diagonal matrix

$$\mathbf{D} = \left[d_{ij} \right] \text{ with } \begin{cases} d_{ii} = r_{ii} \\ d_{ij} = 0, \quad (i \neq j) \end{cases} \tag{8.9}$$

To demonstrate the effect of the above optimal averaging scheme, Fig. 8.2 shows an example source signal and the corresponding sampled data by an array of five sensors, each with different scale and offset biases. From this figure, it is evident that signals from Channels 1 and 3 are severely corrupted. Direct averaging without discriminating the quality of the data can lead to significant errors. Figure 8.3 gives a comparison of the result with and without the use of the maximum SNR criterion described. It is also interesting to see that after normalising each signal trace with $\|\mathbf{x}_i - \overline{x}_i \mathbf{I}\|^2$, the associated weights for each channel is 0.12, 0.26, 0.10, 0.26 and 0.26 respectively, i.e., Channels 1 and 3 are detected as outliers, thus suggesting the potential use of the method for outlier detection.

8.2.2 Source Recovery

In the previous section, we have assumed that the measured signals are directly at source. For many BSN applications, however, this is not possible. For example, in the case of *Electroencephalography* (EEG) and *Magnetoencephalography* (MEG) measurements, signals associated with spontaneous activities or evoked potentials are mixed, and each sensor measures a different combination of the source signals. In this case, source separation of mixed signals from a set of sensors is required. This aspect of direct data fusion has attracted a significant amount of interest in recent years.

The aim of source separation of blindly mixed signals is to recover unobserved signals or sources from temporally and spatially correlated observations. Generally, a *Blind Source Separation* (BSS) problem can be formulated as finding an inverse system that recovers the original signal sources given an observed number of sensor signals $\mathbf{x}(t) = [x_1(t), \ldots, x_N(t)]^{\mathrm{T}}$ [12]. The mathematical formulation of BSS is typically given in the form of a statistical estimation problem. This model is generative, which means that it describes how the observed data is generated by a process of mixing the source components.

By assuming $\mathbf{s}(t) = [s_1(t), \ldots, s_M(t)]^{\mathrm{T}}$ as the unknown signal sources mixed according to a vector valued non-linear function $\mathbf{f}$ the observations $\mathbf{x}(t)$ can be represented as a non-linear mixture of $\mathbf{s}(t)$ and additive noise $\mathbf{n}(t)$, i.e.,

$$\mathbf{x}(t) = \mathbf{f}(\mathbf{s}(t)) + \mathbf{n}(t) \tag{8.10}$$

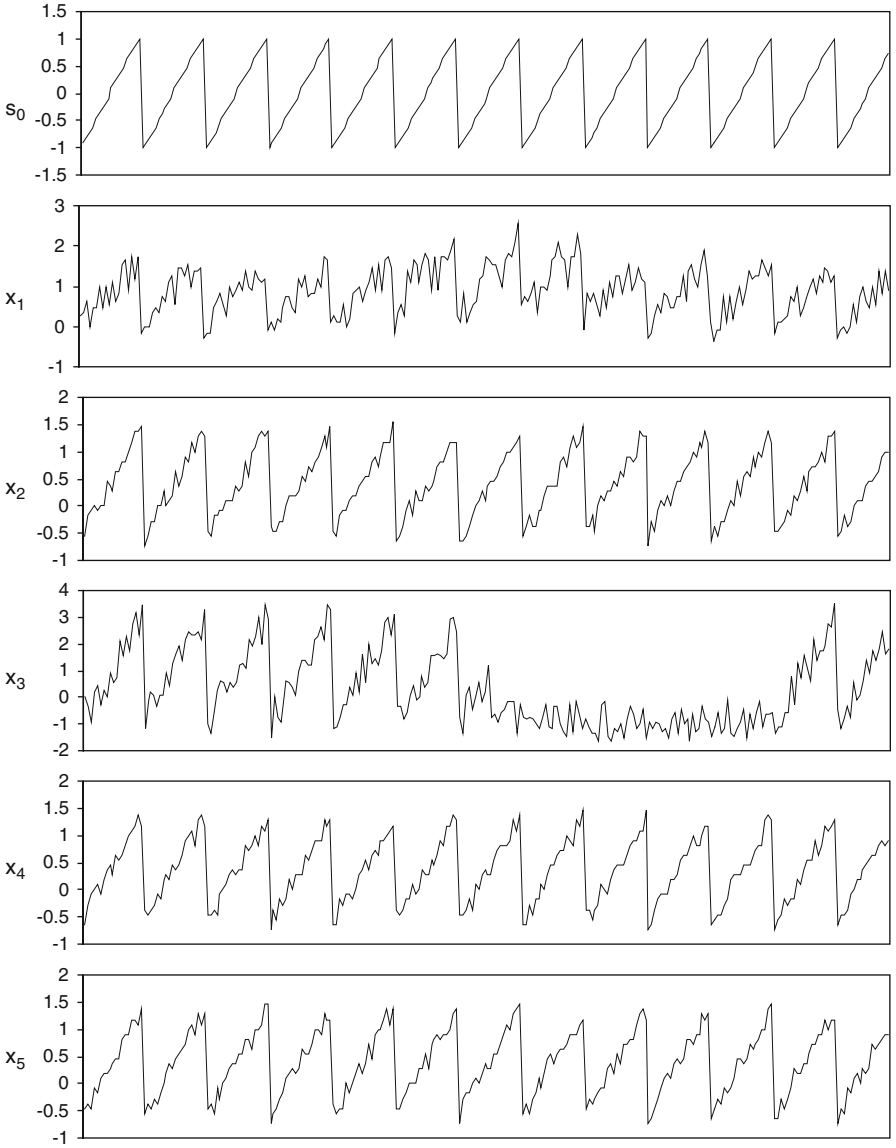


Fig. 8.2 An example source signal (*top*) sampled by an array of five sensors with different scale and offset biases. Each channel is affected by sensor noise but with Channels 1 and 3 being most significant

In most applications, it is desirable to separate the original source signals and to provide information about their spatio-temporal distributions. Ideally, BSS algorithms should make no assumption about the underlying process and the nature of the sources. In practice, however, source separation algorithms range from

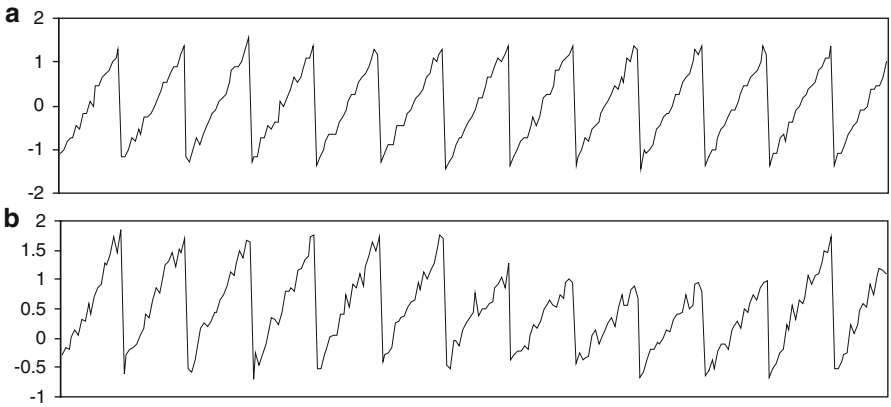


Fig. 8.3 The recovered signal with optimal (**a**) and direct (**b**) averaging

almost blind to highly application specific, where certain characteristics about the sources are available.

For linear mixing models, *Independent Component Analysis* (ICA) is a valuable tool for BSS and it generally conforms to the following main assumptions:

- The sources are linearly mixed and the standard formulation of ICA requires at least as many sensors as sources.
- The sources at each time instant are mutually independent, and at most one source is normally distributed.
- No sensor noise or only low additive noise signals are permitted. However, noise is an independent source itself and if as many sensor outputs are available as the number of sources, the noise signal can be segregated from the mixtures.
- The mixing is assumed to be instantaneous so there is no time-delay between the sources introduced by the mixing medium.
- The mixing process is assumed to be stationary, which implies that the statistics of the signals do not change over time.

The above assumptions ensure that the ICA model is well-defined, which implies the identifiability, separability and uniqueness of the model [13]. It is important to note that the ICA model can be determined up to a scaling factor and a permutation matrix, and these ambiguities are called fundamental indeterminacy. The identifiability stated above suggests the conditions when it is possible to identify the mixing system up to the fundamental indeterminacy. The linear ICA model is identifiable when either all sources are non-Gaussian or the mixing matrix is of full column rank and at most one source is normal. Separability states that the source signals may be recovered up to some ambiguities, whereas the uniqueness ensures that the distribution of the sources can be determined. This is especially important to consider when source separation involves more sources than observations.

The mathematical formulation of the classical ICA is a simplified form of the BSS problem

$$\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t) \tag{8.11}$$

where $\mathbf{A}$ is an $N \times M$ scalar matrix representing the unknown mixing coefficients and it is called transfer or mixing matrix. For most ICA applications, noise is either assumed to be white Gaussian with variance σ^2 or negligible. As stated earlier, noise can also be assumed to be part of the sources. In this case, the noise is assumed to be statistically independent from the other source components. The goal of ICA is to find a linear transformation $\mathbf{W}$ of the dependent sensor signals $\mathbf{x}(t)$ that makes the outputs as independent as possible:

$$\widehat{\mathbf{s}}(t) = \mathbf{W}\mathbf{x}(t) = \mathbf{W}\mathbf{A}\mathbf{s}(t) \tag{8.12}$$

where $\widehat{\mathbf{s}}(t)$ is an estimate of the sources. The sources are recovered exactly when $\mathbf{W}$ is the inverse of $\mathbf{A}$ up to a permutation and scale change. Since both the sources and the mixing coefficients are unknown, it is impossible to determine either the variances or the order of the independent components.

A common way of solving the ICA problem is to use high-order statistics. A classic result in probability theory is the central limit theorem which states that a sum of two independent random variables with finite variances has a distribution that is closer to Gaussian than any of the two original random variables. Since many real world processes yield distributions with finite variances, this explains the ubiquitous nature of the normal distributions. The above principle suggests that a linear combination of the observed mixture variables is maximally non-Gaussian if it is equal to one of the independent components. Therefore, given a function that measures the non-Gaussianity of a signal, each local maximum is an independent component. ICA estimation can be formulated as the search of directions that are maximally non-Gaussian. In practice, non-Gaussianity can be measured by using kurtosis or negentropy. Kurtosis is a higher-order cumulant based on statistical moments defined as:

$$kurt(\mathbf{x}) = E\left(\mathbf{x}^4\right) - 3\left[E\left(\mathbf{x}^2\right)\right]^2 \tag{8.13}$$

where $E(\cdot)$ is the expectation. Alternatively, the normalised kurtosis can also be used:

$$\tilde{\kappa}\left(\mathbf{x}\right) = \frac{E(\mathbf{x}^4)}{\left[E(\mathbf{x}^2)\right]^2} - 3 \tag{8.14}$$

It can be shown that the kurtosis is zero for a Gaussian distribution. In general, a distribution having zero kurtosis is called mesokurtic, whereas a distribution having a positive kurtosis is called super-Gaussian, or leptokurtic in statistics. If the kurtosis is negative, the respective distribution is called sub-Gaussian or platykurtic as a probability densities tend to be flatter than that of the Gaussian.

Prior to ICA estimation, it is usually useful to perform some pre-processing to the measured sensor data to make ICA simpler and better conditioned. One important pre-processing strategy in ICA is to whiten the observed variables, i.e., to transform $\mathbf{x}$ linearly for deriving a new vector $\mathbf{z}$ which is white. This means that the components of $\mathbf{z}$ are uncorrelated and the covariance matrix of $\mathbf{z}$ is equal to an identity matrix.

In practice, a whitening transformation is always possible and a common method of achieving this is through eigendecomposition of the covariance matrix, i.e.,

$$E\left(\mathbf{x}\mathbf{x}^{\mathrm{T}}\right) = \mathbf{\Phi}\mathbf{\Lambda}\mathbf{\Phi}^{\mathbf{T}} \tag{8.15}$$

where $\mathbf{\Phi}$ is the orthogonal matrix of eigenvectors of $E(\mathbf{x}\mathbf{x}^{\mathrm{T}})$ and

$$\Lambda = \mathrm{diag}(\lambda_1, ..., \lambda_N) \tag{8.16}$$

is the diagonal matrix of the corresponding eigenvalues. Therefore, whitening can be represented as:

$$\mathbf{z} = \mathbf{\Phi}\mathbf{\Lambda}^{-1/2}\mathbf{\Phi}^{\mathrm{T}}\mathbf{x} = \mathbf{\Phi}\mathbf{\Lambda}^{-1/2}\mathbf{\Phi}^{\mathrm{T}}\mathbf{A}\mathbf{s} \tag{8.17}$$

where $\mathbf{\Lambda}^{-1/2} = \mathrm{diag}(\lambda_1^{-1/2}, \ldots, \lambda_N^{-1/2})$. It is easy to prove that $E(\mathbf{z}\mathbf{z}^{\mathrm{T}}) = \mathbf{I}$. The effect of whitening is that the new mixing matrix for $\mathbf{s}$ is now orthogonal, and therefore the number of parameters to be estimated for ICA is reduced from N^2 to $N(N-1)/2$.

One popular approach to estimating the ICA model is *Maximum Likelihood* (ML) estimation, which is closely connected to the infomax principle based on maximising the output entropy, or information, of a neural network with nonlinear outputs. The pseudo-code shown in Table 8.1 outlines the infomax learning algorithm for ICA as suggested by Bell and Sejnowski [14] and further extended by Lee et al. [15].

To illustrate how ICA can be used for BSN sensing, Fig. 8.4 illustrates the cardiac surface motion measured from an in vivo experiment showing mixed movement signals along three orthogonal axes. The recovered ICA components are shown in Fig. 8.4b, clearly illustrating the sources of motion due to cardiac, respiratory, and other factors such as noise and jitter.

Since the introduction of the general framework for ICA in the early 80s [16, 17], many new algorithms have been proposed, which lead to a range of successful applications in telecommunication, biomedical signal processing, machine learning, speech recognition, and time-series analysis. Some of the important contributions to ICA include the work of Bell and Sejnowski in developing a fast and efficient ICA based on infomax (a principle introduced by Ralph Linsker in 1992), the introduction of natural gradient by Amari and Cardoso, and Lee and Girolami's work on extending infomax ICA for general non-Gaussian signals [18]. Thus far, a number of different approaches have been pursued for blind source separation, which include maximum likelihood and Bussgang methods based on cumulants, projection pursuit and

Table 8.1 Pseudo code segment for the infomax learning algorithm for ICA as suggested by Bell and Sejnowski [14] and further extended by Lee et al. [15]

1. Initialisation:
   Remove the mean value;
   Perform whitening and $\mathbf{z} = \mathbf{Q}\mathbf{x} = \mathbf{\Phi}\mathbf{\Lambda}^{-1/2}\mathbf{\Phi}^{\mathbf{T}}\mathbf{x}$;
   Initialise the separating matrix so that $\widehat{\mathbf{V}} = \mathbf{I}$.
2. Initialise the optimisation parameters:
   $\boldsymbol{\zeta} = \widehat{\mathbf{V}}\mathbf{z}$
3. Estimate the sign of normalised kurtosis:
   $\tilde{\kappa}(\boldsymbol{\zeta}) = \frac{E(\zeta^4)}{[E(\zeta^2)]^2} - 3$
   where $E(\boldsymbol{\zeta}^4), E(\boldsymbol{\zeta}^2)$ are the fourth and second order moments estimated as the mean of the fourth and second powers of the random variable, respectively.
4. Calculate the contrast based on the infomax principle:
   $\widehat{\mathbf{V}} \leftarrow \mathbf{V} + \eta\left(\mathrm{I} - s \cdot \tanh(\boldsymbol{\zeta}) \cdot \boldsymbol{\zeta}^{\mathrm{T}} - \boldsymbol{\zeta} \cdot \boldsymbol{\zeta}^{\mathrm{T}}\right) \cdot \mathbf{V}$
   where η is the learning rate and s is the sign calculated according to the estimation of kurtosis above.
5. Assess the convergence criteria, ϑ:
   $\delta\mathbf{V} = \widehat{\mathbf{V}} - \mathbf{V}$
   $\vartheta \leftarrow \delta\mathbf{V} \cdot \delta\mathbf{V}^{\mathrm{T}}$
   If the convergence criterion has not reached a predefined constant then return to Step 2.
6. Source separation:
   Estimate the mixing matrix as:
   $\mathbf{A} = \widehat{\mathbf{V}}\mathbf{Q}^{-1}$
   Recover the source components as:
   $\widehat{\mathbf{s}} = \mathbf{A}^{-1}\mathbf{x} = \widehat{\mathbf{V}}^{\mathrm{T}}\mathbf{z}$

negentropy methods. It is expected that BSN will be served as both an important application base for ICA and also as a source of inspiring new algorithms due to the unique features and constraints imposed by BSNs.

8.3 Feature-Level Fusion

Fusion at the feature level is the most important step of sensor fusion and involves the integration of feature sets corresponding to different sensors. These feature vectors are then fused to form joint feature vectors from which the classification is made. Features are an abstraction of the raw data, and the purpose of feature extraction is to find main characteristics of the data that can accurately and concisely represent the original information whilst maximising the discriminative power of the identification process. The first step towards feature-level data fusion is, therefore, effective feature detection. Once the features are selected, the role of feature-level data fusion is to establish decision boundaries in the feature space that can separate patterns belonging to different classes. For general purpose sensing,

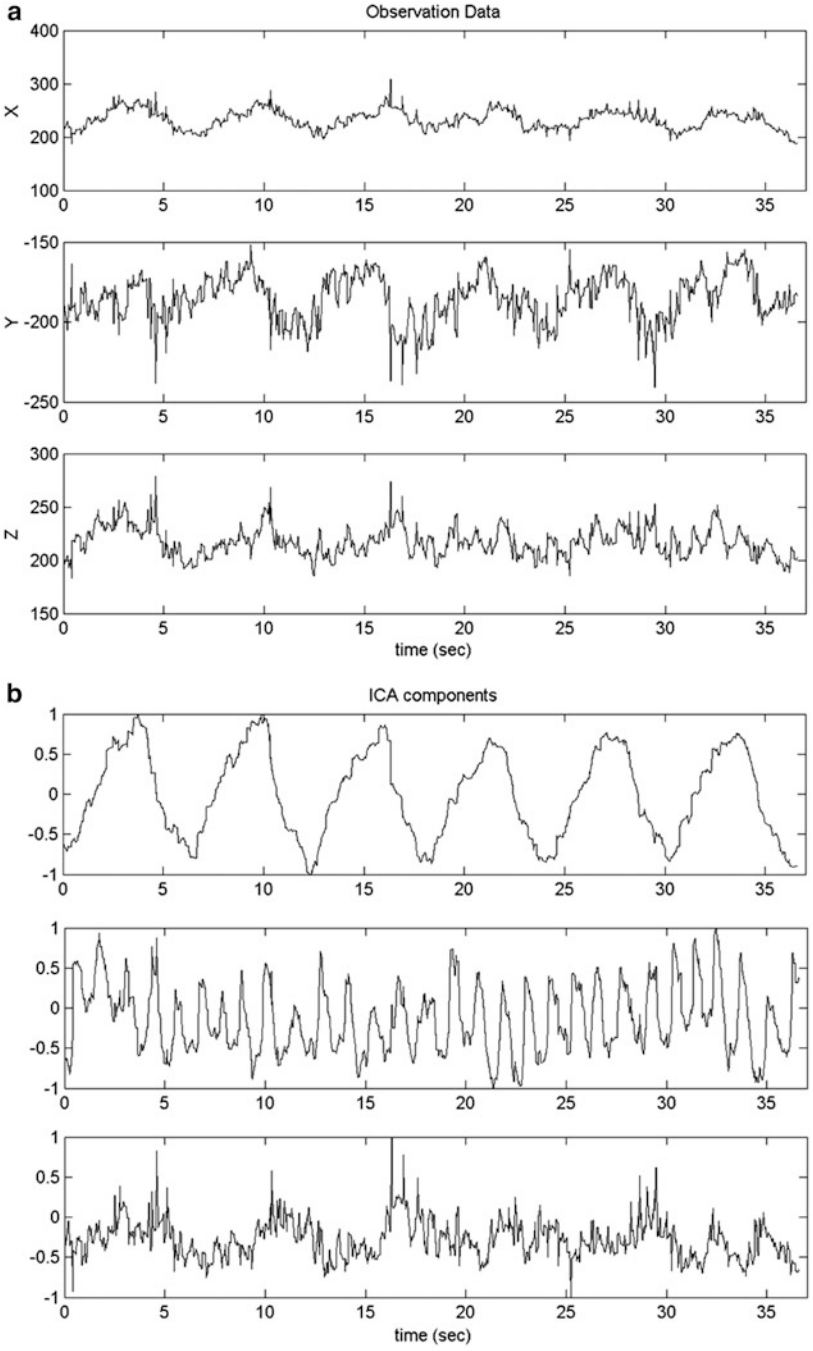


Fig. 8.4 (**a**) The epicardial surface motion measured in vivo showing mixed movement signals along three orthogonal axes. (**b**) The recovered ICA components, illustrating the sources of motion due to respiratory (*top*), cardiac (*middle*) and noise jitter (*bottom*)

Table 8.2 A summary of typical signal features used in general purpose sensing applications (Adapted from Ref. [20])

Feature types		
Time domain	Frequency domain	Hybrid
Waveform characteristics (e.g. slopes, amplitude, envelop, rise time, pulse width, maxima/minima locations, pulse duration, pulse repetition intervals, zero crossing rate)	Periodic structures in the frequency domain	Wavelet representation (e.g. Gabor wavelet features)
Waveform statistics (e.g. mean, standard deviations, mean/standard deviation of signal derivatives, peak-to-valley ratio, average magnitude difference function), energy, kurtosis, entropy and moments	Spectral features (spectral peaks, spectral roll-off, spectral centroid, spectral flux, energy)	Wigner-ville distribution-based analysis
Chaotic models and fractal features	Fourier coefficients	Time-frequency principal component
Ringing, overshoot phenomena, and pulse/ambient noise floor relationship	Chebyshev coefficients	Cyclostationary representations (e.g. cyclic (cross-) correlations, cyclic (cross-) spectra, multi-variate stationary correlations)
	Power spectral density	

there is a wide range of feature extractors that have been developed in the literature. In the following sections, we will provide a brief overview of the main feature detection and classification techniques that are applicable to BSNs.

8.3.1 Feature Detection

In general, signal features can be classified into time-domain, frequency-domain, and hybrid features, as summarised in Table 8.2. Time-domain features include basic waveform characteristics and signal statistics. Conversely, frequency-domain features concentrate on the periodic structures of the signals, which include coefficients such as those derived from Fourier and Chebyshev transforms [19]. For hybrid features, they employ both time and frequency information for complex signals, and a good example of this is the wavelet representation.

Compared to other levels of processing, sensor fusion at the feature level is the most extensively studied yet ongoing research. In terms of BSN, this is also one of the key areas of development. Effective abstraction of the raw sensing data provides important opportunities in using localised processing, such as those described in Chap. 7, for minimising the power and bandwidth utilisation. It also offers the scope of using distributed inferencing for enhancing the reliability and fault tolerance of

BSNs for practical applications. Thus far, the methods developed in this area include deterministic, distance-based, fuzzy logic, neural network, manifold embedding and probabilistic approaches. Due to the large amount of literature in this area, a comprehensive review of the techniques developed is difficult. In the subsequent sections, we will only outline some of the common techniques that are relevant to BSNs. Due to the importance of neural networks for BSNs, particularly the use of *Self-Organising Maps* (SOMs) for context aware sensing and analogue hardware implementation, details concerning this class of techniques will be described in Chap. 9. Similarly, we will dedicate most parts of Chap. 10 to discussing the value of probabilistic approaches in developing autonomic BSNs with distributed inferencing.

8.3.2 Distance Metrics

As mentioned earlier, the goal of feature-level processing is to choose features that allow pattern vectors belonging to different categories to occupy compact and disjoint regions in the feature space. In this case, the decision boundary can be based on either parametric forms or the use of probability distributions specified/learned through training. Thus far, most feature-based clustering techniques are based on the use of distance metrics for measuring similarity or dissimilarity. Let $\mathbf{u}$ and $\mathbf{v}$ be two nonzero K-dimensional feature vectors, so a true distance metric, $d(\mathbf{u},\mathbf{v})$, must conform to the following criteria:

$$\begin{cases} d(\mathbf{u},\mathbf{v}) = 0 & \text{if and only if } \mathbf{u} = \mathbf{v} \\ d(\mathbf{u},\mathbf{v}) \geq 0 & \text{for all } \mathbf{u} \text{ and } \mathbf{v} \\ d(\mathbf{u},\mathbf{v}) = d(\mathbf{v},\mathbf{u}) \\ d(\mathbf{u},\mathbf{v}) \leq d(\mathbf{u},\mathbf{u}') + d(\mathbf{u}',\mathbf{v}) \end{cases} \tag{8.18}$$

In practice, distance metrics may not obey all the properties specified above. Examples of widely used distance metrics include L_p distance (Minkowski distance) and Mahalanobis distance and angle between feature vectors. Mathematically, the L_p distance is defined as:

$$d(\mathbf{u},\mathbf{v}) = \left[\sum_{i=1}^{K} (u_i - v_i)^p\right]^{1/p} \tag{8.19}$$

where $1 \leq p \leq \infty$. In general, the distance is less affected by outliers when p is small. Due to its geometrical and statistical implications, Euclidean distance (L_2) is the most popular distance metric used in practical applications. For instance, it corresponds to the total inter-cluster variance when used to construct the objective function in k-Means clustering. The Euclidean distance is translation and rotation invariant but highly dependent on the scale of each feature. The L_1 distance is known

as the Manhattan (or city block) distance, and counting the number of disagreements is implicitly a Manhattan metric. It is worth noting that the Manhattan distance is invariant to translation or reflection with respect to a coordinate axis, but not rotation. Finally, the L_∞ distance is also known as Chebyshev distance and it corresponds to the maximum of absolute difference in any single dimension.

The above distance metrics are based on the assumption that features are independent. Mahalanobis distance, on the other hand, uses the correlations between variables to remove several of the limitations in Euclidean metrics. It measures the dissimilarity between two random vectors ($\mathbf{u}$ and $\mathbf{v}$) in the same distribution with the covariance matrix $\Sigma = E([\mathbf{u} - E(\mathbf{u})][\mathbf{u} - E(\mathbf{u})]^{\mathrm{T}})$, i.e.,

$$d(\mathbf{u}, \mathbf{v}) = \sqrt{(\mathbf{u} - \mathbf{v})^{\mathrm{T}} \Sigma^{-1} (\mathbf{u} - \mathbf{v})} \tag{8.20}$$

By taking into account the intra-feature correlation, the Mahalanobis distance is scale invariant and it can also provide curved, as well as linear, decision boundaries. However, a robust estimation of the covariance matrix is required, and therefore, it is not suitable for high-dimensional feature vectors. In the special case where the covariance matrix is the identity matrix or features are uncorrelated and the variances in all directions are the same, the above equation is the same as the Euclidean distance. Furthermore, if the covariance matrix is diagonal, the measurement becomes normalised Euclidean distance. In pattern recognition, it is also common to use similarity measures based on the angle between two vectors.

8.3.3 Instance-Based Learning

Once the feature set and the similarity measures are defined, pattern classification techniques can be used to fuse the derived sensor data into meaningful events or episodes as shown in Fig. 8.1b. Instance-based learning [21] is one of the simplest non-parametric statistical learning techniques, in which the decision on how to generalise beyond the training data is deferred until a new sample is encountered [22]. The prototype vectors, in this case, are the samples in the training set but the method can also be generalised for cluster-level prototype vectors. A nearest neighbour classifier [23] separates an unlabelled observation $\mathbf{u}$ by measuring its distances from the labelled training samples $\{(\mathbf{v}_i, c(\mathbf{v}_i))\}_{i=1}^{N}$, where $c(\mathbf{v}_i)$ is the assigned label to $\mathbf{v}_i$, to which the nearest neighbour belongs.

In other words, $\mathbf{u}$ is classified as $c(\mathbf{v}_i) \in \{C_1, C_2, \cdots, C_\ell\}$ if

$$d(\mathbf{u}, \mathbf{v}_i) = \min_{1 \le j \le \mathrm{N}} d(\mathbf{u}, \mathbf{v}_j) \tag{8.21}$$

The decision can also be made based on multiple reference points. In this case, the test sample is classified as belonging to the class with the maximum number of occurrences. The level of confidence in the answer can be measured by using the

number of occurrences for each class in the selected set and the distances between the test sample and the selected reference points. This learning technique is known as *k-Nearest Neighbour* (k-NN) classifier when a set of reference points is selected based on a predefined value k. When the reference points are selected based on their distances from the test sample by using a window of pre-specified size, the learning technique is known as Parzen window or the kernel density estimation method [24].

8.3.4 Distance-Based Clustering

The purpose of distance-based clustering is to group large sets of data $S = \{\mathbf{u}_i\}_{i=1}^{N}$ into clusters, each of which is represented by its mean (centroid) $\boldsymbol{e} = \{\mathbf{c}_i\}_{i=1}^{k}$. In general, distance-based clustering utilises *hard* methods where a data point is assigned to a cluster with a probability either 0 or 1, but the idea can be generalised into *soft* methods by the introduction of fuzzy membership functions. In practice, techniques such as k-Means [25], ISODATA [26] and agglomerative clustering algorithms [27] are most commonly used for sensor data fusion.

The k-Means clustering algorithm is initialised by selecting k initial cluster centres, where k is equal to the final required number of clusters. A common method for selecting the initial points is to assign the centroid of the entire dataset to the first cluster centre and selecting the subsequent centres by the data points farthest form the chosen ones. The k-Means training algorithm is an unsupervised iterative optimising process, aiming to minimise the sum of the distance between the data points and the corresponding centroid:

$$\Delta(\boldsymbol{e}) = \sum_{i=1}^{N} d(\mathbf{u}_i, \boldsymbol{c}(\mathbf{u}_i)) \tag{8.22}$$

where $\boldsymbol{c}(\mathbf{u}_i) \in \boldsymbol{e}$ is the nearest centroid to the data vector $\mathbf{u}_i$. At each step, data points are reallocated to their nearest centroids and the new centroids are recalculated by using the newly assembled clusters. The iteration stops when a stopping criterion is achieved. In other words, when no reassignment occurs or the maximum number of iterations is exceeded. This algorithm is insensitive to data ordering and can be parallelised [28], but that is strongly dependent on the cluster initialisation and the predefined value of k.

The *Iterative Self-Organising Data Analysis* (ISODATA) is a partitioning relocation clustering algorithm similar to k-Means. However, it allows merging and splitting of intermediate clusters according to certain parameters, and thus the number of clusters is not fixed to the predefined value k. An example set of parameters that determines the splitting and merging conditions are as follows:

- The minimum distance between two cluster centroids is d_0,
- The minimum number of data points in a cluster is n_0, and
- The maximum standard deviation allowed for each cluster is σ_0.

Two clusters are merged when the distance between their centroids is below d_0. Each cluster with fewer than n_0 samples are discarded and their elements are distributed amongst the remaining clusters. Clusters in which the maximum coordinate-wise standard deviation exceeds σ_0 are split along that coordinate. Extra conditions can also be introduced to constrain the range of the desired number of clusters. A full implementation of the ISODATA algorithm can be found in [29].

Agglomerative clustering is a type of hierarchical clustering technique. The algorithm is initialised with a set of singleton clusters, each of which is a data point. Clusters are gradually merged based on the distance between clusters until a single big cluster is formed or a stopping criterion is reached. At each level in the hierarchy, clusters are formed from the union of two clusters at the next level down. The distance between clusters can be derived from the distance between individual points. In practice, three common measures of distance between clusters are used and they include single link, complete link and average link metrics. They are equal to the minimum, maximum and average distances from any member of one cluster to any member of the other cluster, respectively, i.e.,

$$\begin{cases} d_{single}(C_1, C_2) &= min\{d(\mathbf{u}, \mathbf{v}) \,|\, \mathbf{u} \in C_1, \mathbf{v} \in C_2\} \\ d_{complete}(C_1, C_2) &= max\{d(\mathbf{u}, \mathbf{v}) \,|\, \mathbf{u} \in C_1, \mathbf{v} \in C_2\} \\ d_{average}(C_1, C_2) &= avg\{d(\mathbf{u}, \mathbf{v}) \,|\, \mathbf{u} \in C_1, \mathbf{v} \in C_2\} \end{cases} \tag{8.23}$$

A single big cluster output from the agglomerative clustering algorithm can be visualised as a tree or a dendrogram. Clusters can be obtained by setting a threshold level across the tree. To evaluate which level in the hierarchy contains the best clusters, the standard measure of within-cluster variance in this case does not apply as the algorithm starts from clusters with no variance at all. Instead, we can use the difference between the level at which it was formed and the level at which it is merged, or compare the average distance within clusters to the average distance between clusters. The disadvantage of this algorithm is decisions made earlier in the process are never revisited. Therefore, if an early agglomeration destroys the structure of a cluster, it will not be detected in the later stages.

As mentioned earlier, feature vectors in the above algorithms are generally partitioned into hard clusters, i.e., each feature vector can be a member of one cluster only. Fuzzy clustering resolves some of the intrinsic problems associated with hard clustering by the introduction of a membership function such that a feature vector can have multiple membership grades to multiple clusters. *Fuzzy c-Means* (FCM) [30] is a clustering technique which allows each data point to be assigned to more than one cluster with different probability or degrees of membership. It is based on an iterative minimisation of the following cost function:

$$\Delta_p(\boldsymbol{e}) = \sum_{i=1}^{N} \sum_{j=1}^{k} \mu_{i,j}^{p} d(\mathbf{u}_i, \boldsymbol{c}_j) \tag{8.24}$$

where p is a weight exponent and $\mu_{i,j}$ indicates the membership or degree that a data point $\mathbf{u}_i$ belongs to cluster C_j. In a fuzzy set, membership must satisfy the following conditions:

$$\begin{aligned} \sum_{j=1}^{k} \mu_{i,j} &= 1, \forall i \\ \sum_{i=1}^{N} \mu_{i,j} &> 0, \forall j \\ \mu_{i,j} &\in [0,1], \forall i,j \end{aligned} \tag{8.25}$$

At each step, the cluster centres $\boldsymbol{c}_j$ and the degree of membership $\mu_{i,j}$ can be updated by:

$$\mu_{i,j}^{p} = \left[\sum_{l=1}^{k} \left(\frac{d(\mathbf{u}_i, \boldsymbol{c}_j)}{d(\mathbf{u}_i, \boldsymbol{c}_l)} \right)^{\frac{2}{p-1}} \right]^{-1} \tag{8.26}$$

$$\boldsymbol{c}_j = \frac{\sum_{i=1}^{N} \mu_{i,j}^{p} \mathbf{u}_i}{\sum_{i=1}^{N} \mu_{i,j}^{p}} \tag{8.27}$$

A common stopping criterion for this algorithm is when the maximum change between the degree of membership at two consecutive steps is less than a pre-defined threshold ϵ.

$$\max_{ij} \left\{ \left| \mu_{i,j}(t+1) - \mu_{i,j}(t) \right| \right\} < \epsilon \tag{8.28}$$

It is also interesting to note that when the degree of membership is constrained to 0 and 1, the FCM becomes the hard k-Means clustering algorithm.

Other techniques for pattern classification include the decision tree method, which is performed by an iterative selection of individual features that are most salient at each node. It therefore implicitly incorporates feature selection during the classification process. The feature selection criteria used include the Fisher's criterion, node purity and the information content. Popular methods in this category include the CART and C4.5 algorithms. Their numerical implementations are both available in the public domain [31, 32]. The main advantage of the method is its speed and the possibility of interpreting the decision rules for each individual feature.

The use of *Support Vector Machines* (SVMs) has also attracted significant research interests for pattern classification. The original idea of SVM was based on Vapnik's method of finding an optimal hyper-plane for dividing two classes, which does not depend on a probability estimation [33, 34]. This optimal hyper-plane is a linear decision boundary that separates the two classes and leaves the

largest margin between the vectors of the two classes. He demonstrated that the optimal hyper-plane is determined by only a small fraction of the data points, the so-called *support vectors*. Cortes and Vapnik extended the method for the case of non-separable classes, and therefore made SVM a general tool for solving general classification problems [35].

One commonly used machine learning technique is *Random Forest* or *Tree Ensemble* [36]. Random forest is a type of machine learning that is within the family of ensemble methods [37, 38]. The idea behind ensemble approaches is to combine the inference provided by "weak learners", such as CART or C4.5 trees, in the case of random forest, to create a "strong learner". Ensemble methods also make use of a sampling procedure so-called *Bootstrap Sampling* (BT). BT samples subsets of the original dataset randomly, uniformly and with replacement. Each model based on weak learners is trained with the generated subsets. The overall result, when the output is a categorical variable, can be either an average or a weighted average from the results of each model. A voting system is commonly employed to obtain the classification decision.

Another important approach towards pattern classification is the neural networks method. The most commonly used techniques include feed-forward networks, such as the multi-layer perceptron and radial basis function networks, and SOM or Kohonen Network [39, 40]. The main advantage of the neural network approach is due to its efficient learning algorithms and the potential for analogue hardware implementation. This is attractive for BSNs, especially for low-power processing requirements. In Chap. 9, we will provide some detailed examples of how the neural network approach, particularly SOM, can be used for pattern classification required for context aware sensing.

8.4 Dimensionality Reduction

In pattern recognition, dimensionality reduction techniques are commonly used when the sample data is assumed to lie on a manifold, which can be non-linear in most general cases. The intrinsic dimensionality is usually related to the number of independent variables that account for most variability within the data. As only intrinsic features are preserved, dimensionality reduction may lead to a better understanding of the data. Thus far, there are a number of techniques in the literature which address the problem of dimensionality reduction. The most commonly used technique is *Principal Component Analysis* (PCA), which provides a reference system for which the variables with small variance are discarded. Thus, the high-dimensional data is projected to the subspace spanned by the most dominant principal components, leading to an approximation of the original data in a least-squares sense. The linear projections of the data are selected according to the maximal variance subject to the orthogonality constraint.

The main disadvantage of PCA is that it is only able to find linear subspaces, and therefore, cannot deal with data lying on non-linear manifolds. Furthermore, the

number of principal components to keep in practice is a complicated issue, although a number of rules of thumb can be applied [41].

Similar to PCA, techniques such as Projection Pursuit can also be used to search for linear projections. It is an unsupervised technique that selects low-dimensional linear orthogonal projections of a high-dimensional point cloud by optimising an objective function called the *projection index*.

Other techniques for dimensionality reduction include *Fisher Projection* (FP) which is based on linear projection of the data to a sub-space where the classes are well-separated. With this technique, however, if the amount of training data is inadequate or the quality of some of the features is poor, then some derived dimensions may be a result of noise rather than the intrinsic differences among feature classes [42].

Although many methods can deal with non-linear dimensionality reduction, most of them rely on local dimensionality reduction. In particular, some locally linear techniques [43, 44] extend PCA to non-linear data by first performing clustering and then applying PCA for each cluster. The main limitation of such techniques is their inability to extract the global structure of the data. Common non-linear techniques leading to a global low-dimensional model of the observations include SOM as mentioned earlier [45] and *Generative Topographic Mapping* (GTM) [46].

8.4.1 Multidimensional Scaling (MDS)

Multidimensional Scaling (MDS) is a technique closely related to PCA, and is based on the definition of a similarity matrix, i.e., a matrix whose elements indicate the degree of similarity between the objects under consideration. MDS has been successfully applied to the visualisation of high-dimensional data in low-dimensional spaces and has been used to discover perceptual representations in psychology by analysing the similarity of stimuli to reveal the underlying structure of the data [47, 48]. The similarity matrix may be defined as a metric distance (metric scaling) but can also be provided as rank ordered information (non-metric scaling) in which case the rank order of the dissimilarities must be preserved. A common way to define the similarity matrix is to consider the stimuli as points in a multidimensional space where similarity is inversely related to the Minkowski distance.

In general, methods that explicitly use a metric are preferred since they enable generalisation from a learned embedding to unseen examples. Generalisation in this case becomes an issue of learning an approximation of the function described by the embedding. The use of a global Minkowski metric leads to a linear reconstruction of manifolds, which, by and large, may not be the most appropriate. In particular, if the Euclidean distance is used, MDS is equivalent to PCA and the method is known as *Classical Multidimensional Scaling* (CMDS).

8.4.2 Locally Linear Embedding (LLE)

Locally Linear Embedding (LLE) exploits the local geometry of the neighbouring points in the high-dimensional space in order to map the input data points onto a global coordinate system of lower dimension while the intrinsic relationships between the points are preserved [49]. Each point in the high-dimensional space is approximated by a linear combination of its neighbours and the coefficients for that combination are selected such that the mapping is invariant to scaling, rotation and translation. LLE consists of the following main steps:

1. ***Computation of the neighbourhood of each data point*** $\mathbf{u}_i$ – The calculations are performed through the selection of the k-nearest neighbours on the basis of the Euclidean metric. However, it is also common to determine a radius, r, defining a ball that encompasses the neighbourhood of each point.
2. ***Computation of the weights*** w_{ij} ***that best reconstruct each data point*** $\mathbf{u}_i$ ***from its neighbours*** – The following function is introduced to measure the reconstruction error that needs to be minimised

$$\varepsilon(w) = \sum_i \left| \mathbf{u}_i - \sum_j w_{ij} \mathbf{u}_j \right|^2 \tag{8.29}$$

   The minimisation is performed subject to two constraints: (a) each point is reconstructed only using its neighbours ($w_{ij} = 0$ *if* $\mathbf{u}_i \notin \Omega$ where Ω represents the neighbourhood of the point), and (b) $\forall i$, $\sum_j w_{ij} = 1$ in order to ensure the invariance to translation. The minimisation problem subject to the above constraints can be solved in a closed form. The reconstruction weights characterise intrinsic geometric properties and provide the invariance to rotation, rescaling and translation of each point and its neighbours.
3. ***Computation of the low-dimensional vector*** $\mathbf{u}'_i$ ***corresponding to each data point*** $\mathbf{u}_i$ ***through the use of the weights*** w_{ij} ***previously calculated*** – Similarly to the previous step, the coordinates of the vectors are found through the minimisation of a cost function. In this case, however, the weights are fixed and the coordinates must be optimised. Therefore, the cost function now becomes:

$$\Psi\left(\mathbf{u}'\right) = \sum_i \left| \mathbf{u}'_i - \sum_j w_{ij} \mathbf{u}'_j \right|^2 \tag{8.30}$$

   To ensure the problem is well-posed, the cost function is also minimised subject to two constraints: (a) the $\mathbf{u}'$ coordinates are centred on the origin as the translation of $\mathbf{u}_i'$ should not alter the cost function and (b) the vectors $\mathbf{u}_i$ have unit covariance, thus avoiding degenerated solutions (i.e., $\mathbf{u}_i' = 0$). As a consequence of the invariance of the cost function to rotation and homogeneous rescaling, there is no loss of generality in imposing the second condition. The cost function can be rewritten in a quadratic form and the minimisation of the above equation can be performed by solving an $N \times N$ eigendecomposition problem.

8.4.3 Laplacian Eigenmaps

Laplacian Eigenmaps [50] is a dimensionality reduction method based on the spectral graph theory. It considers the intrinsic geometry of the data by preserving the local properties of the manifold. For this method, a graph is first developed from the pairwise distances between k-neighbours. The generated graph of local distances serves as a model for the minimisation process. The distance between each node ensures that the low dimensional representation of the data preserves the same geometrical properties from the original manifold. The minimisation process is a spectral graph problem with a cost function that considers the distance between the neighbours' points in the graph, in which the closest neighbour to a point contributes to a higher cost. The creation of a Laplacian Eigenmap involves the following steps:

1. ***Computing the adjacency matrix for k-neighbours*** – Points in a neighbour graph G are connected by an edge $\boldsymbol{i}$ and $\boldsymbol{j}$, for which the opposite sides $\mathbf{u}$ and $\mathbf{v}$ are the nodes. For each node in the graph, the pairwise distances are computed using a Gaussian kernel function:

$$w_{ij} = e^{\frac{\|\mathbf{u}-\mathbf{v}\|^2}{2\sigma^2}} \tag{8.31}$$

   where σ is the scatter of the Gaussian distribution. Each non-linear distance comprises an adjacency matrix or kernel matrix $\mathbf{G}$ of dimension $N \times N$, where N is the total number of points. An edge between $\mathbf{u}$ and $\mathbf{v}$ exists if $||\mathbf{u} - \mathbf{v}||^2 < \epsilon$, where ϵ is a pre-determined threshold. Alternatively, it is also possible to perform this test just for the closest k neighbours. This method has a lower risk of resulting in a disconnected graph.
2. ***Computing eigenmaps from the graph adjacency matrix*** **-** Once the respective weights w of the adjacency matrix or neighbour graph $\mathbf{G}$ has been computed, it is important to check if the graph is connected. If the full graph is disconnected, eigendecomposition is independently performed for each connected component inside the graph.

In order to perform eigendecomposition, two other symmetric matrices derived from graph G have to be computed: (1) the degree matrix $\mathbf{S}$ that is a diagonal matrix whose entries are the row sums of $\mathbf{G}$ (i.e., $S_{ii} = \Sigma_j G_{ij}$), and (2) the Laplacian matrix $\mathbf{L} = \mathbf{G} - \mathbf{S}$. Hence, the eigenvector-generalised problem of these two matrices is defined as:

$$\mathbf{L\Phi} = \lambda \mathbf{S\Phi} \tag{8.32}$$

The resultant eigenvector matrix $\mathbf{\Phi}$ of this decomposition gives rise to different solutions,

$$\begin{aligned}
&\mathbf{L}\boldsymbol{\varphi}_0 = \lambda_0 \mathbf{S}\boldsymbol{\varphi}_0 \\
&\mathbf{L}\boldsymbol{\varphi}_1 = \lambda_1 \mathbf{S}\boldsymbol{\varphi}_1 \\
&\ldots \\
&\mathbf{L}\boldsymbol{\varphi}_{N-1} = \lambda_{N-1} \mathbf{S}\boldsymbol{\varphi}_{N-1} \\
&0 = \lambda_0 \leq \lambda_1 \leq \ldots \leq \lambda_{N-1}
\end{aligned}$$

By discarding the first eigenvector $\boldsymbol{\varphi}_0$ that corresponds to λ_0, the vector of resultant embedding coordinates can be generated as $\mathbf{u}_i' \rightarrow (\boldsymbol{\varphi}_1(i), \ldots, \boldsymbol{\varphi}_N(i))$..

8.4.4 Isometric Mapping (Isomap)

Alternatively, Tenenbaum et al. [51] have proposed the use of geodesic distance measured on the manifold, defined by the data as the basis to calculate the similarity matrix. Non-linear dimensionality reduction is approached as a problem of discovering a Euclidean feature space embedding a set of observations that attempts to preserve the intrinsic metric structure of the data. It should be noted that an *isometry* f:$G \rightarrow G'$ is a distance-preserving map such that $d'(f(\mathbf{u}), f(\mathbf{v})) = d(\mathbf{u}, \mathbf{v})$ for all $\mathbf{u}$, $\mathbf{v}$ in G. Isomap is an isometric feature-mapping procedure that aims to recover low-dimensional non-linear structure and consists of the following three main stages:

1. ***Discrete representation of the manifold*** – Random selection of M points from the N observations to serve as nodes of a topology-preserving network. The neighbourhood is defined through the selection of k-neighbours or a radius r. If M is too small the distance calculation will be a poor approximation to the true manifold distance, whereas if M is too large (relative to N) the graph will miss many appropriate links.
2. ***Manifold distance measure*** – This measure starts with the assignment of a weight, w_{ij}, to each link. Such weight is equal to the Euclidean distance between the nodes $\mathbf{u}$ and $\mathbf{v}$ in the observation space. The geodesic distance is then considered to be the shortest distance along the path by following the previously calculated weights. Should the data be infinite, the graph-based approximation to manifold distance can be made arbitrarily accurate.
3. ***Isometric Euclidean embedding*** – Classical MDS is used to find a low dimensional embedding that preserves as closely as possible the graph distances.

As an example, Fig. 8.5 demonstrates a simple physical exercise sensing experiment where four 2-axis accelerometers were placed on the left and right ankles and legs. The activities of the subject during the exercise routine include: (1) *sitting* (*chair*), (2) *standing*, (3) *steps*, (4) *sitting* (*floor*), (5) *demi-plié*, (6) *galloping left*, (7) *skipping*, (8) *galloping right*, (9) *side kick*, (10) *front kick* and (11) *walking*. Although the dimensionality of the original raw signals is eight, the application of the Isomap reveals that the intrinsic dimensionality of this dataset is in fact

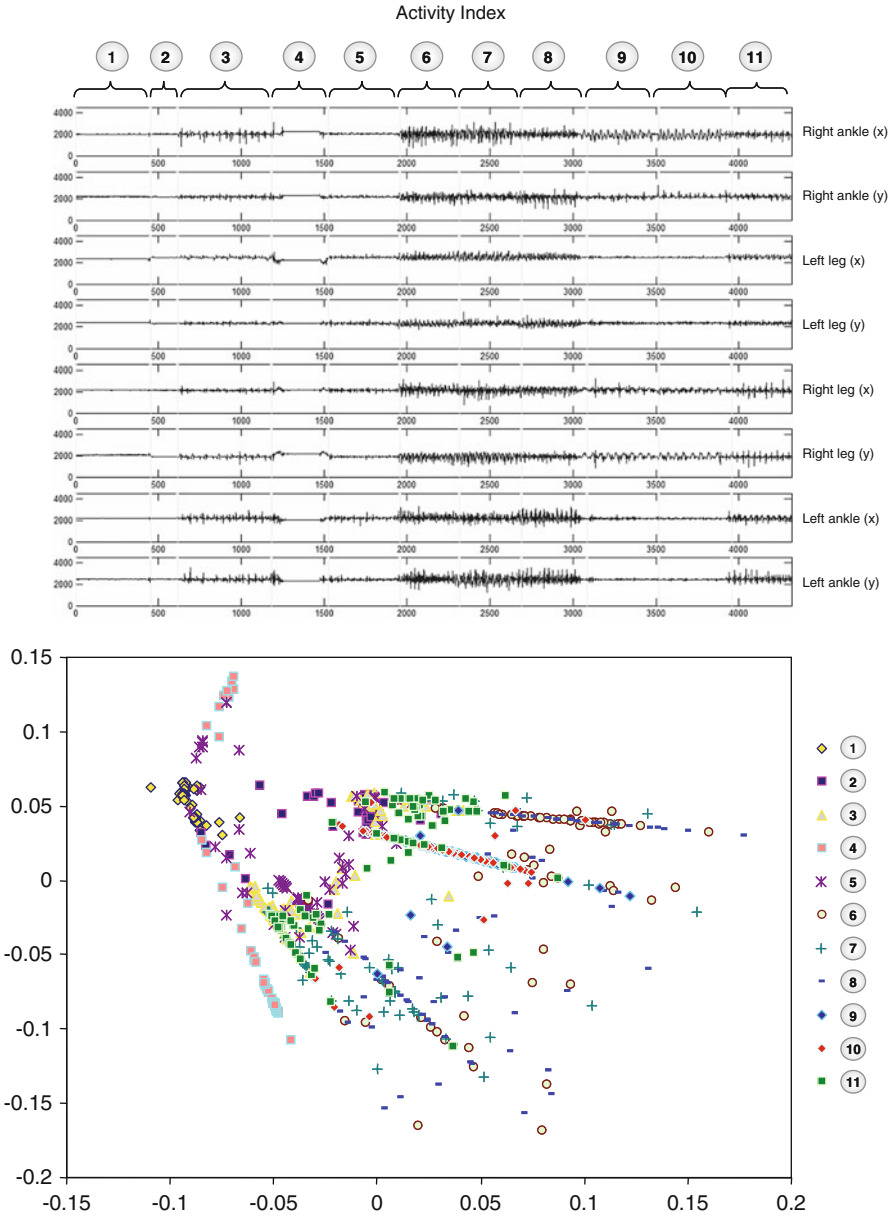


Fig. 8.5 Accelerometer readings from an exercise sequence (*top*) and its corresponding Isomap embedding results (*bottom*). The residual variance of the embedded data decreases rapidly from 0.41 to 0.05, 0.04, 0.03 within the first four Isomap dimensions

only two. This is evident from the residual variance of the embedded data which decreases rapidly from 0.41 to 0.05, 0.04, and 0.03 within the first four Isomap dimensions.

For pattern recognition, the use of dimensionality reduction is advantageous for many classification tasks, as it avoids the over-fitting issue in many high-dimensional problems due to the practical limit of the training sample size. It also offers a convenient way of visualising the intrinsic structure of the data. For example, it is evident from Fig. 8.5 that the separation of the 11 activities involved is good in general, but we may encounter significant challenges when trying to separate activities 6, 7 and 8, (i.e., *galloping left*, *skipping*, and *galloping right*) as with the current sensor placement the distribution of the feature vectors are all mixed together. Another important issue raised by this example is that since the intrinsic dimensionality of the sensor data is only two-dimensional, some of the information collected by these sensors is likely to be redundant. With BSNs, this boils down to the question of how to ensure strategic placement of the sensors to guarantee that the information collected provides the most discriminative power in terms of pattern separation. To achieve this, we introduce in the next section the concept of feature selection and provide an example on how it can be used for optimal sensor placement.

8.5 Feature Selection

In pattern recognition, the aim of feature selection is to reduce the complexity of an induction system by eliminating irrelevant and redundant features. This is essential for reducing computational cost and storage, as well as for improving prediction accuracy. Intuitively, a high-dimensional model is more accurate than a low-dimensional one. However, the computational cost of an inference system increases dramatically with its dimensionality, and therefore we have to balance between accuracy and overall computational cost. It should be noted, however, that the accuracy of a high-dimensional model may deteriorate if the model is built on insufficient training data. The amount of training data required to understand the intrinsic structure of an unknown system increases exponentially with its dimensionality. An imprecise description could lead to over-fitting problems, rendering the inference algorithm poorly generalisable to new data. Furthermore, the use of multi-sensor fusion also requires efficient selection strategies to remove irrelevant and redundant features.

In machine learning, there are three main types of approaches for measuring feature importance:

- *Feature Selection*: From the whole set of features, it finds the minimum optimal set of features with respect to the performance criteria, however, run-time complexity can be high.
- *Feature Ranking*: All features are ranked individually based on the empirical estimates of statistical properties (mutual information, relief, etc.). Features are ranked but the set of features that provides the maximum performance is not provided.

- *Combined Feature Analysis*: Feature selection and feature ranking are not competing approaches but rather complementary. The rank of the features can be used to reduce the complexity of the search for the best subset of features. For example, a sequential feature selection method using a weighted Naïve Bayes algorithm [52] can use the weights, generated through a feature ranking approach, plus the performance of the classifier to select the features.

In order to organise the number of available possibilities of feature importance, a rational division of different methods is proposed in [53]. Feature ranking methods are considered as *filters* as they assess the importance of each feature by analysing statistical properties or information quantities of the data but ignoring the iteration with the machine learning process. Feature selection methods are presented as *wrappers* as they are dependent on the performance of a selected machine learning algorithm that has to be run independently until its performance converges to certain criteria based on a cost function. Combined feature analysis is also regarded as being *embedded* because the machine learning algorithm considers external feature ranks or makes use of the internal feature rank provided by the algorithm itself. The relative merits and potential pitfalls of these methods are provided in Table 8.3.

As a simple demonstration, let us consider a scenario as shown in Fig. 8.6 where there are four input sensor channels used for determining four different activities (C_1, C_2, C_3 and C_4). At a first glance, it is not difficult to see that the information content provided by Channel 4 is limited, while other channels provide different levels of detail for the four classes to be determined. The question is then to what extent each (or a combination) of the sensor channels can provide the right classification. One simple solution is to list out all the possible permutations of the sensor channels and evaluate for each combination the corresponding classification accuracy. This can be shown in Fig. 8.7, where the state of each box represents one combination of the sensor channels (0 or 1 means the sensor channel is or is not selected). In this experiment, a naïve Bayesian classifier has been used. It is evident from the second column of this figure that when each channel is considered on its own, the most relevant feature is from Channel 3, which can achieve an overall accuracy of 80.08 %, whereas for other channels the corresponding accuracy ranges from 23.92 % (Channel 4) to 54.42 % (Channel 2).

When two sensor channels are used, a significant improvement in classification accuracy can be achieved. For example, the classification accuracy achieved by combining Channels 1 and 2 can reach an accuracy of 97.67 %. It is also clear from this figure that from this point the introduction of additional channels can only bring negligible improvement to the classification accuracy, i.e., a maximum accuracy of 99.33 % when three channels are used and 99.42 % when all the four channels are used. Albeit being simplistic, the example highlights some of the challenges in sensor selection. A given feature or sensor channel may have relatively high discriminative power, but its combined use with other sensors may not guarantee the highest performance. Furthermore, irrelevant features or sensor inputs may deteriorate the overall system performance. Therefore, it is important to determine feature relevance, redundancy and interaction given the constraints of the sensing environment.

Table 8.3 A summary of the advantages and disadvantages of different methods used in machine learning techniques for determining feature importance

	Advantages	Disadvantages	Examples
Feature selection	**Exponential search**	**Exponential search**	
	Consider feature dependencies	Complexity, computationally expensive	A*
	Find the best set of values over all possible solutions through an exhaustive search	Dependent on the classifier performance	B*
	No local optima	High risk of overfitting	Depth-first search
			Beam search
			Iterative deepening [54]
	Sequential Search	**Sequential Search**	
	Select next feature and possible successor of the current state, sorted search	Problems of local optima	Forward selection (SFS) [55]
		Dependent on the algorithm performance	Backward selection (SBS) [55]
	Time complexity limited by the number of steps	Some risk of overfitting	
	Tractable complexity		
	Random search	**Random search**	
	Can take into account feature dependencies	Lower risk of local optima	Simulated annealing
	Random selection of features to avoid local optima	Dependent on the algorithm performance	Genetic algorithms
	Conditional tractable complexity	High risk of overfitting	Ant and bee colony optimisation
			Stochastic hill climbing
			Tabu search
Feature ranking	**Bivariate**	**Bivariate**	
	Very fast methods, low complexity	Algorithm independent	Relative entropy
	Scalable to large datasets	Not related to overall performance.	Information gain ratio
	Not related to overall performance	Conditional dependencies between features are not considered	Gini index
			Relief algorithm
			Kruskal-Wallis

	Multivariate	**Multivariate**	
	Take into account conditional dependencies between features	Algorithm independent	Markov blanket filter
	Lower complexity than feature selection methods	Not related to overall performance Higher complexity than bivariate methods	Correlation-based feature selection Maximum-redundancy maximum relevance
Combined analysis	**Weighted**	**Weighted**	
	Use of ranks for weighting the feature relevance in the internal mechanism of machine learning	Necessary to provide the ranks to the algorithm Use of different rank methods may lead to different performances	Weighted Naïve Bayes Weighted Fisher discriminant analysis
	Auto-weighted	**Auto-weighted**	
	Different to the weighted method, external ranks measurements are not needed	Problems related to using a single algorithm, "just one opinion"	Multi-layer perceptron
	Obtain ranks directly from the learning algorithm	Algorithm complexity can be compromised	Linear support vector machines feature weighting Logistic regression feature weights
	Ensemble Methods	**Ensemble Methods**	
	Feature selection is part of the internal mechanism of the machine learning algorithm	High computational complexity	Random forest feature relevance
	Overtake local optima		LPBoost feature relevance
	Problem of overfitting is reduced		AdaBoost feature relevance

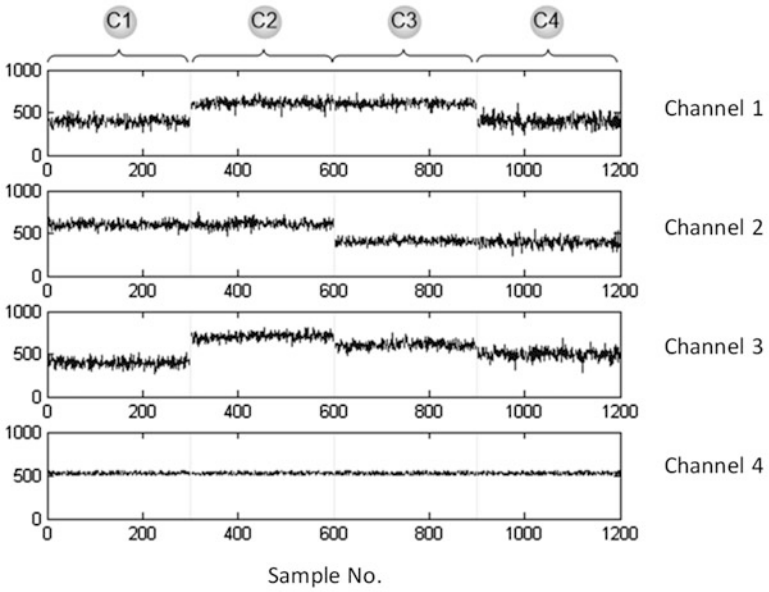


Fig. 8.6 An example dataset showing four sensor channels used to differentiate four different activity classes (From Ref. [56])

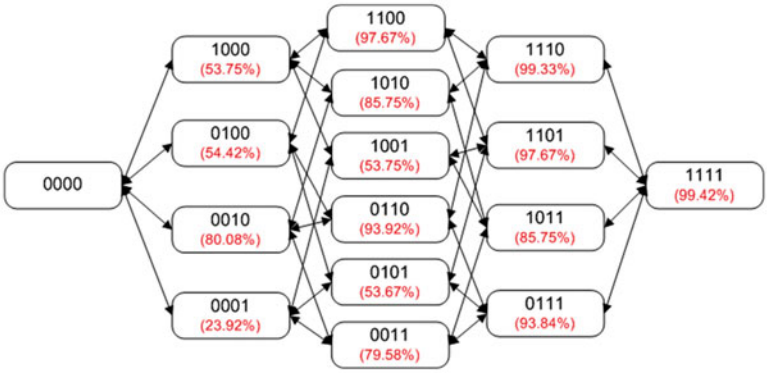


Fig. 8.7 The search space for selecting feature subsets from four input features of Fig. 8.6. Each state represents a candidate feature subset and the corresponding recognition accuracy when a naïve Bayesian classifier is used. For example, when the first and second features are used (state 1,100), an accuracy of 97.67 % can be achieved (From Ref. [56])

8.5.1 Feature Relevance

In practice, it is important to differentiate the aforementioned feature extraction for dimensionality reduction and feature selection. In feature extraction, the given features are transformed into a lower-dimensional space, without much loss of information.

One of the feature extraction techniques is ICA as mentioned earlier, which transforms a number of correlated variables into a number of uncorrelated independent components. For feature selection, however, no new feature is created. The dimensionality is reduced by eliminating irrelevant and redundant features. An irrelevant (or redundant) feature provides no (or no new) information about the target concept [57]. In Bayesian inference, the posterior probability is used for a rational observer to make decisions since it summarises the information available. Consequently, the formal definition of relevance can be based on the conditional independence [58, 59].

Given a set of sensor features $\boldsymbol{\mathcal{G}}^{(1)} = \left\{\boldsymbol{\mathcal{G}}_i^{(1)}, 1 < i < K_1\right\}$, event $\boldsymbol{\mathcal{Y}}$ is conditionally independent of feature set $\boldsymbol{\mathcal{G}}^{(2)} = \left\{\boldsymbol{\mathcal{G}}_i^{(2)}, 1 < i < K_2\right\}$ (i.e., given $\boldsymbol{\mathcal{G}}^{(1)}$, $\boldsymbol{\mathcal{G}}^{(2)}$ provides no further information in detecting $\boldsymbol{\mathcal{Y}}$) if for any assignment of $\boldsymbol{\mathcal{Y}}$

$$P\left(\boldsymbol{\mathcal{Y}}|\boldsymbol{\mathcal{G}}^{(1)}\right) = P\left(\boldsymbol{\mathcal{Y}}|\boldsymbol{\mathcal{G}}^{(1)}, \boldsymbol{\mathcal{G}}^{(2)}\right) \text{ when } P\left(\boldsymbol{\mathcal{G}}^{(1)}, \boldsymbol{\mathcal{G}}^{(2)}\right) \neq 0 \tag{8.33}$$

Optimum selection of feature subset involves two major difficulties: a search strategy to select candidate feature subsets and an evaluation function to assess these candidates. The size of the search space for the candidate subset selection is 2^K, i.e., a feature selection method needs to find the best one amongst 2^K candidate subsets, given K features. The example in Fig. 8.7 also shows the search space for the four features and how forward and backward searches can be used to include/exclude relevant features in the selected feature subset.

Since the size of the search space grows exponentially with the number of input features, an exhaustive search of the space is impractical. As a result, a heuristic search strategy, such as the *greedy search* and the *branch and bound search*, becomes necessary [58–61]. *Forward selection* denotes that the search strategy starts with the empty feature set, while *backward elimination* denotes that the search strategy starts with the full feature set. As an example, Koller and Sahami [59] proposed a sequential greedy backward search algorithm to find '*Markov blankets*' of features based on the expected cross-entropy evaluation.

The feature selection methods can be divided into two groups (*wrapper* and *filter*) according to how the evaluation function is designed. The filters use the information content (typically measured by interclass distance, statistical dependence, and information-theoretic divergences [62]) of the feature subsets for the evaluation. As a result, the filter approaches are independent of any induction algorithm. Kira and Rendell, in their RELIEF [60] algorithm, formulated a weighting method to evaluate the discriminability of each feature by calculating Near-Hit to measure within-concept spread and Near-Miss to measure between-concept separation, whereas the FOCUS [61] algorithm searches for the minimal feature subset via a systematic consistency test – but this algorithm is sensitive to noise.

In contrast, the wrappers directly use the predictive accuracy of some induction algorithms to evaluate candidate feature subsets [63]. Statistical techniques such as cross-validation are employed for the purpose of evaluation [64]. Generally speaking, for a given algorithm, wrappers achieve a better accuracy than filters. However,

compared to filters, they lack generality and can be computationally demanding. This is because they are directly related to specific induction algorithms, and running these algorithms on a large number of features multiple times can be computationally prohibitive. Other popular feature selection techniques include *Sequential Forward Search* (SFS), *Sequential Backward Search* (SBS), and *Sequential Floating Forward Search* (SFFS) methods [65–67].

In this section, we will introduce a *Bayesian Framework for Feature Selection* (BFFS) originally developed by Hu [68] and subsequently extended by Thiemjarus [56] and demonstrate how it can be used for BSNs. In a Bayesian framework, the likelihood probabilities of each class and their priors can be estimated independently through controlled experiments. This is useful in practice to avoid direct random sampling from the whole population, which can be a tedious and costly task.

By using the Bayes rule, an assignment of $\mathcal{Y} = \boldsymbol{a}$, Eq. (8.33) can be rewritten as

$$\begin{aligned} &\frac{P\left(\boldsymbol{\mathcal{G}}^{(1)}\middle|\mathcal{Y}=\boldsymbol{a}\right)P(\mathcal{Y}=\boldsymbol{a})}{P\left(\boldsymbol{\mathcal{G}}^{(1)}\middle|\mathcal{Y}=\boldsymbol{a}\right)P(\mathcal{Y}=\boldsymbol{a})+P\left(\boldsymbol{\mathcal{G}}^{(1)}\middle|\mathcal{Y}\neq\boldsymbol{a}\right)P(\mathcal{Y}\neq\boldsymbol{a})} \\ &=\frac{P\left(\boldsymbol{\mathcal{G}}^{(1)},\boldsymbol{\mathcal{G}}^{(2)}\middle|\mathcal{Y}=\boldsymbol{a}\right)P(\mathcal{Y}=\boldsymbol{a})}{P\left(\boldsymbol{\mathcal{G}}^{(1)},\boldsymbol{\mathcal{G}}^{(2)}\middle|\mathcal{Y}=\boldsymbol{a}\right)P(\mathcal{Y}=\boldsymbol{a})+P\left(\boldsymbol{\mathcal{G}}^{(1)},\boldsymbol{\mathcal{G}}^{(2)}\middle|\mathcal{Y}\neq\boldsymbol{a}\right)P(\mathcal{Y}\neq\boldsymbol{a})} \end{aligned} \tag{8.34}$$

or

$$\frac{P\left(\boldsymbol{\mathcal{G}}^{(1)}\middle|\mathcal{Y}=\boldsymbol{a}\right)}{P\left(\boldsymbol{\mathcal{G}}^{(1)}\middle|\mathcal{Y}\neq\boldsymbol{a}\right)}=\frac{P\left(\boldsymbol{\mathcal{G}}^{(1)},\boldsymbol{\mathcal{G}}^{(2)}\middle|\mathcal{Y}=\boldsymbol{a}\right)}{P\left(\boldsymbol{\mathcal{G}}^{(1)},\boldsymbol{\mathcal{G}}^{(2)}\middle|\mathcal{Y}\neq\boldsymbol{a}\right)} \tag{8.35}$$

Consequently, we can obtain an equivalent definition of relevance by using the likelihood ratio, i.e., given a set of features from the sensor network $\boldsymbol{\mathcal{G}}^{(1)}=\left\{\mathcal{G}_i^{(1)},1<i<K_1\right\}$, event $\mathcal{Y}$ is conditionally independent of feature set $\boldsymbol{\mathcal{G}}^{(2)}=\left\{\mathcal{G}_i^{(2)},1<i<K_2\right\}$ if for any assignment of $\mathcal{Y}=\boldsymbol{a}$ when $P(\boldsymbol{\mathcal{G}}^{(1)},\boldsymbol{\mathcal{G}}^{(2)})\neq 0$.

$$\mathcal{L}\left(\boldsymbol{\mathcal{G}}^{(1)}\parallel \mathcal{Y}=\boldsymbol{a},\mathcal{Y}\neq\boldsymbol{a}\right)=\mathcal{L}\left(\boldsymbol{\mathcal{G}}^{(1)},\boldsymbol{\mathcal{G}}^{(2)}\parallel \mathcal{Y}=\boldsymbol{a},\mathcal{Y}\neq\boldsymbol{a}\right) \tag{8.36}$$

where $\mathcal{L}(\boldsymbol{\mathcal{G}}^{(1)}\|\mathcal{Y}=\boldsymbol{a},\mathcal{Y}\neq\boldsymbol{a})$ is the likelihood ratio:

$$\mathcal{L}(\boldsymbol{\mathcal{G}}\parallel \mathcal{Y}=\boldsymbol{a},\mathcal{Y}\neq\boldsymbol{a})=\frac{P(\boldsymbol{\mathcal{G}}\parallel \mathcal{Y}=\boldsymbol{a})}{P(\boldsymbol{\mathcal{G}}\parallel \mathcal{Y}\neq\boldsymbol{a})} \tag{8.37}$$

Theoretically, Eqs. (8.33) and (8.36) are equivalent. However, Eq. (8.33) is based on the posterior probability, whereas Eq. (8.36) is based on the likelihood ratio. This modification connects feature selection with the performance assessment of decision-making. As will be seen in the following sections, a sufficient performance assessment can be achieved by ROC analysis, in which the likelihood ratio plays an essential role.

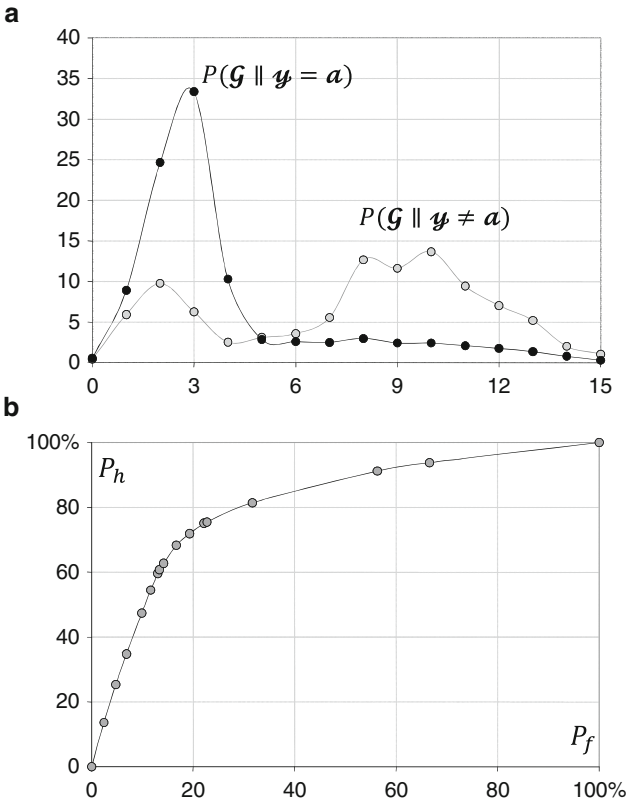


Fig. 8.8 Conditional probabilities (**a**) $P(\mathcal{G} \parallel \mathcal{y} = \mathcal{a})$ and $P(\mathcal{G} \parallel \mathcal{y} \neq \mathcal{a})$ and their ROC curve (**b**). In the ROC space shown in (**b**), hit rate changes from 0 to 100 % when false-alarm rate varies from 0 to 100 %

8.5.2 *Feature Relevance Based on ROC Analysis*

With the above definition, we can now examine the relationship between ROC and feature *irrelevance* to establish the link between feature selection and the performance of decision-making. A proper ROC is generated by using the likelihood ratio or its equivalent as the decision variable [69]. Given a pair of likelihoods, the best possible performance of a classifier can be described by the corresponding ROC, which can be obtained via the Neyman-Pearson ranking procedure by changing the threshold of the likelihood ratio [70]. Given two distributions $P(\mathcal{G} \| \mathcal{y} = \mathcal{a})$ and $P(\mathcal{G} \| \mathcal{y} \neq \mathcal{a})$ as demonstrated in Fig. 8.8a, the hit and false-alarm rates, according to the Neyman-Pearson procedure, are defined as

$$\begin{cases} P_h = \int_{\mathcal{L}(\mathcal{G}\|y=a, y\neq a)>\epsilon} P(\boldsymbol{\mathcal{G}} \parallel \boldsymbol{y} = \boldsymbol{a}) d\boldsymbol{\mathcal{G}} \\ P_f = \int_{\mathcal{L}(\mathcal{G}\|y=a, y\neq a)>\epsilon} P(\boldsymbol{\mathcal{G}} \parallel \boldsymbol{y} \neq \boldsymbol{a}) d\boldsymbol{\mathcal{G}} \end{cases} \tag{8.38}$$

where ϵ is the threshold and $\mathcal{L}(\boldsymbol{\mathcal{G}} \| \boldsymbol{y} = \boldsymbol{a}, \boldsymbol{y} \neq \boldsymbol{a})$ is the likelihood ratio as defined above.

For a given ϵ, a pair of P_h and P_f can be calculated. When ϵ changes from ∞ to 0, P_h and P_f change from 0 to 100 %. Therefore, the ROC curve is obtained by changing the threshold of the likelihood ratio. Figure 8.8b depicts the ROC curve corresponding to the likelihood distributions shown in Fig. 8.8a. In discrete forms, where only some discrete points of (P_h, P_f) are available, straight-line segments are used to connect these points to form a convex hull. Every point on those straight-line segments is realisable by applying a randomised strategy [69, 71]. When misjudgement costs are taken into consideration, minimal maximal-cost points in the ROC space can only be found on the vertices of the convex hull [72].

In a ROC space, the hit rate is the function of the false-alarm rate. The slope at a point on the ROC curve is equal to the associated likelihood ratio. The *Area Under the ROC Curve* (AUC) is an important measure of discriminability of the two classes described by the two likelihood distributions. The equivalent statistical metric to the AUC is the Wilcoxon statistics, which was originally designed to estimate the probability of the rank of two random variables [73, 74].

Based on the above definition and given two pairs of feature distribution, $P(\boldsymbol{\mathcal{G}}^{(1)} \| \boldsymbol{y} = \boldsymbol{a})$, $P(\boldsymbol{\mathcal{G}}^{(1)} \| \boldsymbol{y} \neq \boldsymbol{a})$ and $P(\boldsymbol{\mathcal{G}}^{(2)} \| \boldsymbol{y} = \boldsymbol{a})$, $P(\boldsymbol{\mathcal{G}}^{(2)} \| \boldsymbol{y} \neq \boldsymbol{a})$, we have two corresponding ROC curves obtained from the Neyman-Pearson procedure: $ROC(\boldsymbol{\mathcal{G}}^{(1)} \| \boldsymbol{y} = \boldsymbol{a}, \boldsymbol{y} \neq \boldsymbol{a})$ and $ROC(\boldsymbol{\mathcal{G}}^{(2)} \| \boldsymbol{y} = \boldsymbol{a}, \boldsymbol{y} \neq \boldsymbol{a})$. It can be proven that

$$ROC\left(\boldsymbol{\mathcal{G}}^{(1)} \parallel \boldsymbol{y} = \boldsymbol{a}, \boldsymbol{y} \neq \boldsymbol{a}\right) = ROC\left(\boldsymbol{\mathcal{G}}^{(1)}, \boldsymbol{\mathcal{G}}^{(2)} \parallel \boldsymbol{y} = \boldsymbol{a}, \boldsymbol{y} \neq \boldsymbol{a}\right) \tag{8.39}$$

if, and only if,

$$\mathcal{L}\left(\boldsymbol{\mathcal{G}}^{(1)} \parallel \boldsymbol{y} = \boldsymbol{a}, \boldsymbol{y} \neq \boldsymbol{a}\right) = \mathcal{L}\left(\boldsymbol{\mathcal{G}}^{(1)}, \boldsymbol{\mathcal{G}}^{(2)} \parallel \boldsymbol{y} = \boldsymbol{a}, \boldsymbol{y} \neq \boldsymbol{a}\right) \tag{8.40}$$

and that $ROC(\boldsymbol{\mathcal{G}}^{(1)}, \boldsymbol{\mathcal{G}}^{(2)} \| \boldsymbol{y} = \boldsymbol{a}, \boldsymbol{y} \neq \boldsymbol{a})$ is not under $ROC(\boldsymbol{\mathcal{G}}^{(1)} \| \boldsymbol{y} = \boldsymbol{a}, \boldsymbol{y} \neq \boldsymbol{a})$ at any point in the ROC space. Based on the above equations, an equivalent definition of feature irrelevance based on ROC can be derived; in other words, given a set of features from the sensor network $\boldsymbol{\mathcal{G}}^{(1)} = \left\{\boldsymbol{\mathcal{G}}_i^{(1)}, 1 < i < K_1\right\}$, event $\boldsymbol{y}$ is conditionally independent of feature set $\boldsymbol{\mathcal{G}}^{(2)} = \left\{\boldsymbol{\mathcal{G}}_i^{(2)}, 1 < i < K_2\right\}$ if for any assignment of $\boldsymbol{y} = \boldsymbol{a}$

$$ROC\left(\boldsymbol{\mathcal{G}}^{(1)} \parallel \boldsymbol{y} = \boldsymbol{a}, \boldsymbol{y} \neq \boldsymbol{a}\right) = ROC\left(\boldsymbol{\mathcal{G}}^{(1)}, \boldsymbol{\mathcal{G}}^{(2)} \parallel \boldsymbol{y} = \boldsymbol{a}, \boldsymbol{y} \neq \boldsymbol{a}\right) \tag{8.41}$$

It should be noted however, two ROC curves can be unequal when they have the same AUC. Since $\boldsymbol{\mathcal{G}}^{(1)}$ is a subset of $\{\boldsymbol{\mathcal{G}}^{(1)},\boldsymbol{\mathcal{G}}^{(2)}\}$ we have:

$$AUC\left(\boldsymbol{\mathcal{G}}^{(1)} \parallel \boldsymbol{y} = \boldsymbol{a}, \boldsymbol{y} \neq \boldsymbol{a}\right) = AUC\left(\boldsymbol{\mathcal{G}}^{(1)}, \boldsymbol{\mathcal{G}}^{(2)} \parallel \boldsymbol{y} = \boldsymbol{a}, \boldsymbol{y} \neq \boldsymbol{a}\right) \tag{8.42}$$

In a multiple-class situation, the expected AUC can be used:

$$E_{AUC}(\boldsymbol{\mathcal{G}}) = \sum_{i=1}^{\boldsymbol{\ell}} P(\boldsymbol{y} = \boldsymbol{a}_i) \sum_{j=1\cdots\boldsymbol{\ell}}^{j\neq i} H_{ji} AUC\left(\boldsymbol{\mathcal{G}} \parallel \boldsymbol{y} = \boldsymbol{a}_i, \boldsymbol{y} = \boldsymbol{a}_j\right) \tag{8.43}$$

where $AUC(\boldsymbol{\mathcal{G}} \| \boldsymbol{y} = \boldsymbol{a}_i, \boldsymbol{y} = \boldsymbol{a}_j)$ represents the AUC given two likelihood distributions $P(\boldsymbol{y} = \boldsymbol{a}_j)$ and $P(\boldsymbol{y} = \boldsymbol{a}_i)$, and

$$H_{ji} = \frac{P\left(\boldsymbol{\mathcal{G}} \parallel \boldsymbol{y} = a_j\right)}{\sum\limits_{m=1\cdots\boldsymbol{\ell}}^{m\neq i} P(\boldsymbol{\mathcal{G}} \parallel \boldsymbol{y} = \boldsymbol{a}_m)} \text{ for } i \neq j \tag{8.44}$$

The expected AUC can be calculated from the data distribution independent of a classifier, and thus can be utilised as the evaluation function for a filter algorithm.

8.5.3 Feature Selection Based on ROC Analysis

Since removing or adding an irrelevant feature does not change the expected AUC, both backward and forward greedy selection (filter) algorithms can be designed to use the expected AUC as an evaluation function. A backward elimination approach provides a greedy algorithm for feature selection. It starts with the full feature set and removes one feature at each iteration. A feature $f_j \in \boldsymbol{\mathcal{G}}^{(a)}$ to be removed is determined by using the following equation:

$$f_j = \arg\min_{f_i \notin \mathcal{G}^{(a)}} \left(E_{AUC}\left(\boldsymbol{\mathcal{G}}^{(a)}\right) - E_{AUC}\left(\boldsymbol{\mathcal{G}}^{(a)} - \{f_i\}\right)\right) \tag{8.45}$$

where $\boldsymbol{\mathcal{G}}^{(a)}$ is the temporary feature set after a^{th} iteration and $\boldsymbol{\mathcal{G}}^{(a)} - \{f_i\}$ is the set $\boldsymbol{\mathcal{G}}^{(a)}$ with f_i removed.

Estimating the AUC in a high-dimensional space is time-consuming. The accuracy of the estimated likelihood distribution decreases dramatically with the number of features given limited training samples, which in turn introduces ranking error in the AUC estimation [75]. Therefore, an approximation algorithm is necessary to estimate the AUC in a lower-dimensional space. As explained earlier, the decrease of the total AUC after removal of a feature f_i is related to the overlap of the

Table 8.4 Pseudo-code for BFFS backward elimination algorithm

(a) Let $\boldsymbol{\mathcal{G}}^{(a)}$ be the full feature set and K be the size of the full feature set
(b) Calculate the discriminability differential matrix $\mathcal{M}(f_i, f_j)$

$$\mathcal{M}\left(f_i,f_j\right) = E_{AUC}\left(\left\{f_i,f_j\right\}\right) - E_{AUC}\left(\left\{f_j\right\}\right)$$

for $f_i \in \boldsymbol{\mathcal{G}}^{(a)}, f_j \in \boldsymbol{\mathcal{G}}^{(a)}$ and $f_i \neq f_j$

(c) If $a = K_s$, output $\boldsymbol{\mathcal{G}}^{(a)}$
(d) For $f_i \in \boldsymbol{\mathcal{G}}^{(a)}\ (i = 1 \cdots K)$

Select k_s features from $\boldsymbol{\mathcal{G}}^{(a)}$ to construct a feature subset $\boldsymbol{\mathcal{H}}^{(a_i)}$. The criterion of the selection is to find k_s features f_j, for which $\mathcal{M}(f_i, f_j)$ is the smallest, where $f_j \in \boldsymbol{\mathcal{G}}^{(a)}$ and $f_i \neq f_j$

Calculate $\Psi_{AUC}(f_i) = E_{AUC}\left(\boldsymbol{\mathcal{H}}^{(a_i)} \cup \{f_i\}\right) - E_{AUC}\left(\boldsymbol{\mathcal{H}}^{(a_i)}\right)$

(e) Select feature f_d with the smallest $\Psi_{AUC}(f_i)$ and set $\boldsymbol{\mathcal{G}}^{(a)} = \boldsymbol{\mathcal{G}}^{(a)} - \{f_d\}$
(f) $a = a - 1$; go to (c)

Table 8.5 Pseudo-code for BFFS forward feature selection algorithm

(a) Let $\boldsymbol{\mathcal{G}}^{(a)}$ be empty and a be zero
(b) Calculate the discriminability differential matrix $\mathcal{M}(f_i, f_j)$

$$\mathcal{M}\left(f_i,f_j\right) = E_{AUC}\left(\left\{f_i,f_j\right\}\right) - E_{AUC}\left(\left\{f_j\right\}\right)$$

for $f_i \in \boldsymbol{\mathcal{G}}^{(a)}, f_j \in \boldsymbol{\mathcal{G}}^{(a)}$ and $f_i \neq f_j$

(c) If $a = K_s$, output $\boldsymbol{\mathcal{G}}^{(a)}$
(d) For $f_i \in \boldsymbol{\mathcal{G}}^{(a)}\ (i = 1 \cdots K)$:

Select k_s features from $\boldsymbol{\mathcal{G}}^{(a)}$ to construct a feature subset $\boldsymbol{\mathcal{H}}^{(a_i)}$. The criterion of the selection is to find k_s features f_j, for which $\mathcal{M}(f_i, f_j)$ is the smallest, where $f_j \in \boldsymbol{\mathcal{G}}^{(a)}$ and $f_i \neq f_j$

Calculate $\Psi_{AUC}(f_i) = E_{AUC}\left(\boldsymbol{\mathcal{H}}^{(a_i)} \cup \{f_i\}\right) - E_{AUC}\left(\boldsymbol{\mathcal{H}}^{(a_i)}\right)$

(e) Select feature f_d with the largest $\Psi_{AUC}(f_i)$ and set $\boldsymbol{\mathcal{G}}^{(a)} = \boldsymbol{\mathcal{G}}^{(a)} - \{f_d\}$
(f) $a = a + 1$; go to (c)

discriminability of the feature with other features. In the approximation algorithm, we attempt to construct a feature subset $\boldsymbol{\mathcal{H}}^{(a)}$ from the current feature set $\boldsymbol{\mathcal{G}}^{(a)}$ and use the degree of discriminability overlap in $\boldsymbol{\mathcal{H}}^{(a)}$ to approximate that in $\boldsymbol{\mathcal{G}}^{(a)}$. Similar methods have been reported in [59, 76]. A heuristic approach is designed to select k_s features from $\boldsymbol{\mathcal{G}}^{(a)}$ that have the largest overlap with feature f_i and we assume that the discriminability overlap of feature f_i with other features in $\boldsymbol{\mathcal{G}}^{(a)}$ is dominated by this subset of features. Tables 8.4 and 8.5 summarise the backward elimination and forward selection algorithms for selecting K_s features.

To illustrate how the above algorithm works for BSN applications, Fig. 8.9 illustrates a laboratory activity sequence as sensed by six two-axes accelerometers placed on the ankles, legs, and wrists. The different activities to be detected include *sitting*, *typing*, *writing on paper*, *standing*, *walking*, *writing on white board*, *soldering*, and *drinking*. At a first glance, the placement of the sensors may look rational, as the activities to be differentiated all involve ankles, legs and wrists.

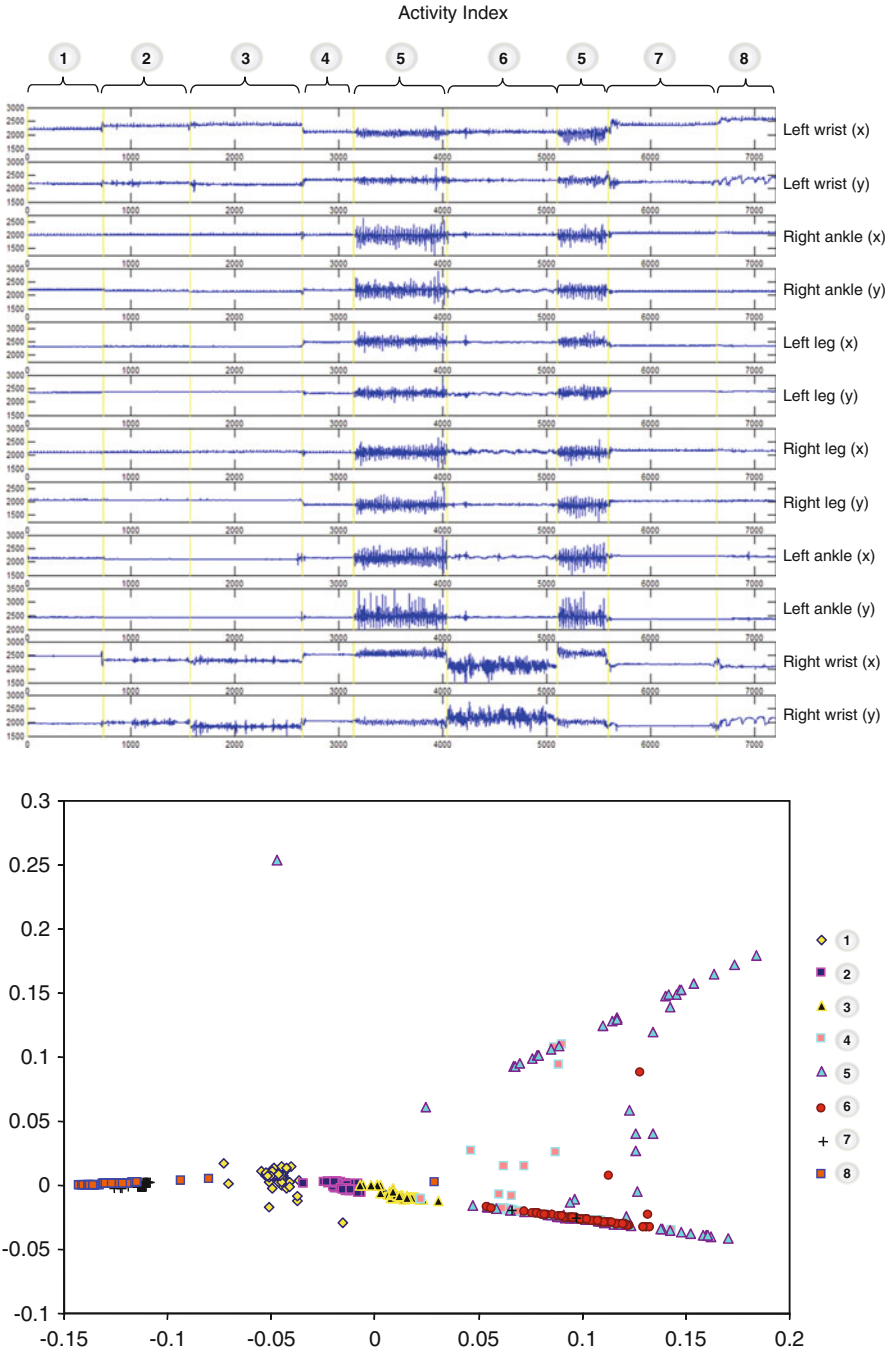


Fig. 8.9 Accelerometer readings from an office activity sequence used for feature selection and the corresponding Isomap embedding result showing the separation of different activity classes

Upon further examination, however, it reveals that the intrinsic dimensionality of the data is low as evident from the Isomap embedding result shown in Fig. 8.8. The Isomap result also demonstrates the good separation of different activities in the embedded feature space. The question now is if we can reduce the number of sensors without affecting the overall sensitivity and specificity of the classification algorithm. By the use of the proposed BFFS algorithm, Table 8.6 illustrates the result of the most discriminative sensor channels as revealed by the backward elimination algorithm. The results in Table 8.6 show that after the incorporation of sensor channels 12, 1, 11, and 2, there is little gain in the AUC by using further sensor channels. In other words, for uniquely separating the eight different activities mentioned above, we only need two accelerometers positioned on the left and right wrists. This result may sound surprising. By careful reasoning, however, the derived sensor placement in fact makes perfect sense. This is because in this problem setting, activities that involve leg and ankle movements are also coupled with unique hand gestures. It is therefore possible in this case to distinguish sitting, standing, and walking from other activities that mainly involve hand motion by the use of wrist sensors only. To demonstrate how the selected sensor channels perform by using different classifiers, Table 8.7 summarises the result of using Naïve Bayes, Pruned C4.5, Instance Based Learning (with $k = 1$ and 3, respectively) and SVM.

It is evident that through the use of only two accelerometers as determined by BFFS, the performance difference is no more than 3 % for all the classifiers concerned. For practical BSN applications, the efficient use of sensor channels plays an important part in the power and communication bandwidth usage. The proposed BFFS algorithm can therefore provide a systematic way of selecting optimal sensing channels for pattern classification.

As described in the original study [68], the overall ability to separate different classes can be represented as the union of the discriminability of each feature set. Similar to [59, 76], BFFS uses the degree of discriminability overlap of a feature f_i to a feature subset $\boldsymbol{\mathcal{H}}^{(a)}$, where $\boldsymbol{\mathcal{H}}^{(a)} \subseteq \boldsymbol{\mathcal{G}}^{(a)}$, to approximate it in $\boldsymbol{\mathcal{G}}^{(a)}$. Table 8.4 describes a simple greedy algorithm for backward BFFS and Fig. 8.10 illustrates the states and the corresponding expected AUC in the search space as calculated by the algorithm when different K_s is used. With sequential search, the search complexity for the BFFS algorithm is bounded by $O(K^2)$. It can be seen from Fig. 8.10 that the algorithm returns the optimal solutions for different K_s values except when K_s is equal to 1. This is because Channel 3 is eliminated in earlier steps of the feature selection. Under this simple elimination scheme, once a feature is eliminated, there will be no backtracking and the solution is not guaranteed to be optimal for all K_s values.

8.5.4 Multi-objective Feature Selection

One of the significant problems of the BFFS algorithm is that it only caters for a single objective function with a simplistic search strategy. It does not differentiate between irrelevant and redundant features. When different criteria are to be

Table 8.6 Changes of AUC during BFFS backward feature elimination

AUC	0.89751	0.95948	0.98649	0.99519	0.99823	0.99938	0.99976	0.99991	0.99994	0.99997	0.99997	0.99997
Sensor Channel ID	11	1	12	2	9	7	5	8	6	10	3	4

Table 8.7 Classification accuracy by using different classification techniques with the selection of sensor channels as determined by Table 8.6

Sensor channels used	12	11	10	9	8	7	6	5	4	3	2	1
Naïve Bayes	95.05	95.06	95.16	95.30	95.19	95.13	95.13	95.33	94.25	92.56	91.30	70.28
Pruned C4.5	97.83	97.44	97.69	97.73	97.48	97.30	96.81	96.86	95.84	95.33	94.28	75.17
Instance based learning (k = 1)	98.83	98.66	98.72	98.64	98.42	98.03	97.45	97.28	96.19	95.20	92.13	64.31
Instance based learning (k = 3)	98.69	98.42	98.44	98.48	98.27	98.05	97.47	97.33	96.47	95.67	93.70	74.83
SVM	92.86	92.61	92.52	92.48	92.03	91.78	91.39	91.41	91.17	89.02	86.48	62.33

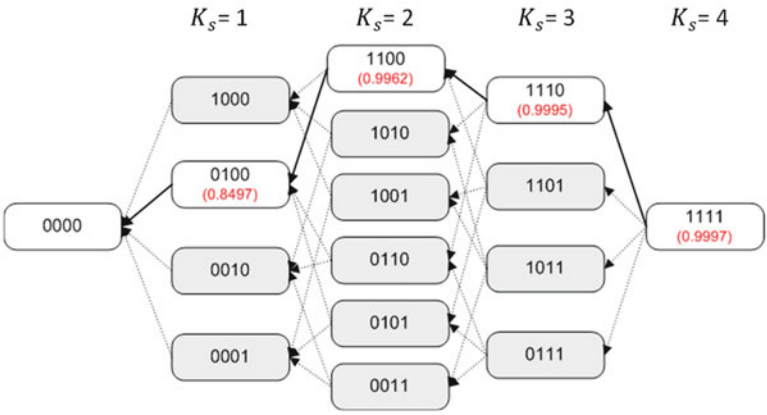


Fig. 8.10 The search space for selecting a feature subset from four input features in Fig. 8.7 with backward BFFS using different number of output features K_s

combined, it is necessary to modify the original objective function to cater for multi-objective constraints. In this case, the feature selection problem can be regarded as a search problem but with a modification to the evaluation function $\Psi(f_i)$. In each step of the backward elimination, for example, a feature which minimises the evaluation function, i.e., $f_d = \text{argmin}_i\ \Psi(f_i)$, will be eliminated from the feature set. A common method for formulating a multi-objective evaluation function is to aggregate the objective vectors into a composite function through linear combination.

$$\Psi(f_i) = \left(1 - \sum_{j=1}^{B-1} w_j\right) O_1(f_i) + \sum_{j=1}^{B-1} w_j O_{j+1}(f_i)$$

where B is the number of objective measures, O_j is the jth objective measure associated with f_i and w_j is the weighting factor ranging between 0 and 1. A backward multi-objective BFFS algorithm can therefore be formulated by replacing the evaluation function $\Psi_{\text{AUC}}(f_i)$ in step (4) of Tables 8.4 and 8.5 with $\Psi(f_i)$.

8.5.5 *Feature Redundancy*

In the previous sections, we have mentioned the importance of differentiating between redundant and irrelevant features. While irrelevant features are uninformative, redundant features are useful features, although their presence may not necessarily increase the discriminative power of the feature set. In the original BFFS algorithm, however, irrelevant and redundant features are treated in a similar manner as both contribute

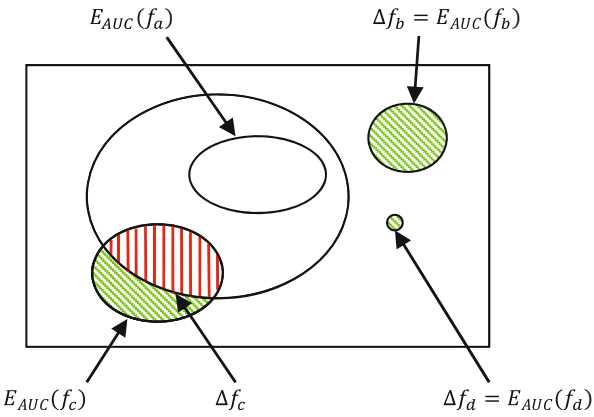


Fig. 8.11 A schematic illustration of the discriminability of features. Δf_i denotes the additional discriminability that f_i adds to the features set $\boldsymbol{\mathcal{G}}^{(a)}$, and $E_{AUC}(f_i)$ that denotes the discriminability of f_i

little to the overall model performance. The order of feature elimination, in this case, is not taken into account as long as the discriminative power of the selected feature set, $E_{AUC}\left(\boldsymbol{\mathcal{G}}^{(a)} - \{f_i\}\right)$, is maintained. In BSN applications, however, redundant features can be used to improve the overall reliability and fault tolerance of the model, whereas the irrelevant features should be removed first during feature elimination.

As an example, Fig. 8.11 depicts the discriminability of several features in the a^{th} iteration of BFFS. If Δf_b is equal to Δf_c and there is no feature interaction, features f_b and f_c will provide the same additional discriminability to the model. However, a higher value of $E_{AUC}(f_c)$ indicates that there is a higher overlap between the discriminability of f_c and the existing feature set than that of f_b. This overlap area (as highlighted in red in Fig. 8.11) can be used as a measure of feature redundancy. In terms of the order with which the features are eliminated for f_a, f_b and f_c, the original BFFS will eliminate f_a first as it provides no additional discriminability to the existing feature set. When it comes to f_b and f_c, the order with which they are removed is arbitrary, although the above analysis clearly demonstrates that it is more preferable to remove f_b first.

It is therefore necessary to construct a new objective function for maximising the discriminability of the selected feature set whilst ensuring the individual discriminative power of each of the selected feature is high. By taking the above discussion into consideration, the revised criterion can therefore be formulated as follows:

$$\Psi_r(f_i) = -(1 - w_1)E_{AUC}\left(\boldsymbol{\mathcal{G}}^{(a)} - \{f_i\}\right) + w_1 E_{AUC}(f_i) \tag{8.46}$$

where w_1 is the weighting factor from 0 to 1. It can be shown that $w_1 = 1/\tau$ indicates that the additional discriminability is τ times more important than feature overlap. When $w_1 = 0$, this is equivalent to the original BFFS objective function.

To approximate the amount of redundancy in the selected feature set $\boldsymbol{\mathcal{G}}_s$, the following equation is used:

$$O_2(\boldsymbol{\mathcal{G}}_s) = \frac{1}{|\boldsymbol{\mathcal{G}}_s|} \sum_{i=1}^{|\boldsymbol{\mathcal{G}}_s|} E_{AUC}(f_i \in |\boldsymbol{\mathcal{G}}_s|) \tag{8.47}$$

which represents the average individual discriminability of all features in $\boldsymbol{\mathcal{G}}_s$. By subtracting the combined discriminability of the solutions from this term, a more accurate measure of the redundancy can be obtained. Examples of the practical use of this feature selection scheme by considering feature redundancy for BSN can be found in [56].

8.6 Decision-Level Fusion

In the previous sections, we have provided detailed explanation of how data- and feature-level sensor fusion techniques can be used for BSN applications. Another important component of multi-sensor fusion as outlined in Fig. 8.1 is decision-level fusion. In general, the decision-level fusion is based on a joint declaration of multiple single source results to achieve an improved event or episode detection result. At this level of fusion, it also provides a unique mechanism for incorporating prior knowledge and domain specific information as shown in Fig. 8.1. For decision-level sensor fusion, common methods include classical inference, Bayesian inference, Dempster-Shafer's method [20] and fuzzy logic [77].

Classical inference seeks to determine the validity of a proposed hypothesis based on empirical probabilities. It computes the joint probability given an assumed hypothesis. In general, classical inference does not take advantage of a priori likelihood assessments and can assess only two hypotheses at a time, i.e., a null hypothesis and its alternative. Although the method can be generalised to include multidimensional data from multiple sensors, complexities arise for multivariate data. In terms of hypothesis acceptance or rejection, the method typically uses maximum a posteriori or maximum likelihood decision rules. For the former, the method accepts the alternative hypothesis C_1 being true if $P(C_1|\mathbf{x})$, i.e., the probability of C_1 given observation $\mathbf{x}$, is greater than $P(C_0|\mathbf{x})$, so the probability of C_1 given $\mathbf{x}$. For the maximum likelihood criterion, it rejects hypothesis C_0 being true if $P(\mathbf{x}|C_1) > P(\mathbf{x}|C_0)$. Other criteria can also be used, which include minimax, Bayes, and Neyman-Pearson decision rules.

Bayesian inference, on the other hand, uses the likelihood of a hypothesis given a previous likelihood estimate and additional observations. Suppose $C_1, C_2, \cdots, C_\ell$ represent mutually exclusive and exhaustive hypotheses that can explain the observation $\mathbf{x}$, then we have

$$P(C_j|\mathbf{x}) = \frac{P(\mathbf{x}|C_j)P(C_j)}{\sum_{i=1}^{\ell} P(\mathbf{x}|C_i)P(C_i)} \tag{8.48}$$

where $P(C_j|\mathbf{x})$ is the a posteriori probability of hypothesis C_j being true given evidence $\mathbf{x}$, $P(C_j)$ the a priori probability of hypothesis C_j being true, and $P(\mathbf{x}|C_j)$ the probability of observing evidence $\mathbf{x}$ given that C_j is true. In contrast to the classical inference method, the Bayesian method provides a way of deriving the probability of a hypothesis being true given the observation. This formulation allows the incorporation of a priori knowledge about the likelihood of the hypothesis and the ability to use subjective probabilities for a priori probabilities for hypotheses when the exact probability density function is not available. In Chap. 10, we will provide more detailed discussion on how to use Bayesian inferencing for BSN applications.

One of the main issues related to Bayesian inferencing is that it requires competing hypotheses being mutually exclusive. In real life, this is not always possible. For humans, we do not generally assign evidence to a set of mutually exclusive and exhaustive hypotheses. Instead, we tend to assign our belief to combinations of hypotheses or propositions. One important fact to note here is that propositions may include overlapping or even conflicting hypotheses. Based on this, Shafer and Dempster provided a generalisation of Bayesian theory that could use probability intervals and uncertainty intervals for determining the likelihood of hypotheses based on multiple evidences [78–81]. In Dempster-Shafer's method, evidence is assigned to both single and general propositions and it uses the concept of a probability mass to represent assigned evidence. When the probability masses are assigned only to an exhaustive and mutually exclusive set of elementary propositions, then the Dempster-Shafer's approach is identical to the Bayesian method. The key advantage of the Dempster-Shafer's method is its ability to establish a general level of uncertainty, and thus provides a means of accounting for possible unknown causes of the observed data. For BSN technology, this is clearly an advantage, as many of the observed events may not have well defined causes.

Inference systems based on *fuzzy logic* also provide a framework for the fusion of different sensors [82–84]. As a difference with Bayesian theory, which deals with *probability*, i.e., the chances that something happens, fuzzy logic deals with *possibility* that is the level of imprecision to determine a specific fact or property (e.g. degree of truth). Fuzzy logic turns sensor readings into fuzzy sets that describe the possibility that the input data describes a particular property. Likewise, defuzzification is the process of obtaining a quantifiable value from a fuzzy set and its membership degrees. It can be achieved through the method called *Center of Gravity* (CoG) of the resulting fuzzy set [85]. This process provides a straight conversion from the fuzzy framework to real values. Fuzzy logic also provides a large number of operators on fuzzy sets, out of which the most used are union, complementation and intersection. Usually, to fuse the information represented by the fuzzy sets, the most common operator is the intersection that describes consensus between sources [86]. The fuzzy inference of the system consists of fuzzy rules

that follow the following form IF antecedent THEN consequent. Each antecedent represents a source and a fuzzy set. Several antecedents can be associated in a rule by using the particle AND (e.g. IF antecendent$_1$ AND antecedent$_2$ THEN consequent). Fuzzy rules can be defined by expert knowledge or an automated learning procedure. It is worth mentioning that the family of fuzzy sets is composed of two types: type-1 and type-2 fuzzy sets, the latter being able to cope with more uncertainty than the former [87, 88].

Other techniques for decision level fusion are based on heuristics. For example, the voting sensor fusion method imitates voting as a means for human decision-making. It combines detection and classification declarations from multiple sensors by treating each sensor declaration as a vote, and the voting process may use majority, plurality or decision-tree rules. Additional methods include ordinal ranking, pair-wise ranking, and *Q*-sort which have been used extensively in assessing the psychometric process by which a human group achieves consensus [20]. To address some of the drawbacks in conventional rule-based schemes, fuzzy logic has also been used extensively to accommodate imprecise states or variables. In this book, however, we will mainly concentrate on the use of Bayesian Inference for practical applications of BSN technology as it provides the basis for understanding some of the advanced techniques in decision-level sensor fusion, including the generalised evidence processing method as proposed by Thomopoulos for addressing the basic assumptions regarding the assignment of evidence to hypotheses or propositions [2].

8.7 Methods for Computing with Large Datasets

With increasing use of continuous sensing, the ability to manage ever-expanding data contributes to the current trend of the *big data* problem. When talking about big data, one of the most well-known computing paradigms is MapReduce [89], which is a programming model that inherits some of the functional ideas of the "map" and "reduce" primitives from the Lisp programming language. As its name suggests, the mechanism consists of two independent functions, *map* and *reduce*. The former gets assigned a set of observations, computes a value and generates a key. The latter takes each of the values computed by the map with the same key and merges them.

MapReduce methods are recommended when the data are dense and variable completeness is required for the calculations, as well as the algorithms implemented on top. Nevertheless, it is possible that the system we are modelling is loosely coupled, e.g., a network with much more nodes than connections, or any data matrix where most elements are 0. The total of null values or zeros that exists in a matrix is called sparsity. For many sparse datasets, compression can be a solution to reduce data storage and ease its manipulation. In the computing literature there are several programming approaches to represent such sparse matrix in terms of sparse vectors or list (compress sparse row, compress sparse column, yale

format etc.). Besides, sparse linear algebra methods provide different solvers to speed up algebraic operations such as factorisation or solving a linear system of equations [90, 91]. When the big dataset is extremely dense and complex, grids methods [92] (which can be structured or unstructured) can be used to split these complex data into simple blocks that discretise the domain in order to enable local manipulation of these datasets.

When considering time complexity and tractability of the algorithm applied to big data, for example when carrying out an exhaustive search, there are some computational paradigms that can be used to ease the burden. One example of this method is *branch and bound*, which defines the problem of optimisation as the minimisation of a function by applying a procedure called bounding, which prunes the candidate solutions en masse. This algorithm is the basis for solving some well-known optimisation processes such as nonlinear programming. Another alternative is dynamic programing, which is for example used for aligning DNA sequences, finding the shortest path of a graph, computing the global distance between two variable time series [93, 94]. Dynamic programming provides solutions to divide a complex problem into simpler sub-problems, enabling the algorithm to go over all possible solutions and select the best choice. Different to MapReduce, dynamic programming does not imply parallelisation. Dynamic programming algorithms aim to solve a complex problem into a sequence of smaller sub-problems, whereas MapReduce focuses on distributing the computation in several machines. In conclusion, there are no general architectures or programming solutions to tackle big data. Hence, researchers must opt for the most suitable solution for the problem at hand and understanding the nature of the data.

8.8 Fusing Datasets in Parallel Using MapReduce

Hitherto, the MapReduce paradigm has been adopted widely in machine learning to parallelise the computation of algorithms across different workers. To speed up the computation, a big dataset is split into several subsets to be individually processed in parallel over multiple processing units. Figure 8.12 shows a schema of the MapReduce computational paradigm. In the map function, each observation that belongs to the same subset is assigned to the same key and transmitted to the reduce function. The reduce function accepts all observations labelled with a specific key and then computes a single but representative value for that key. The main objective of this procedure is to produce a smaller dataset than the original one by assuming that the output values, once reduced and assigned to a specific key, summarise a subset of the original data. Prior to applying this paradigm, it is important to revise if the dataset is big enough for such partitioning. Splitting the dataset into subsets is not recommended when the amount of data is smaller than 1TB as it might be detrimental to the overall performance.

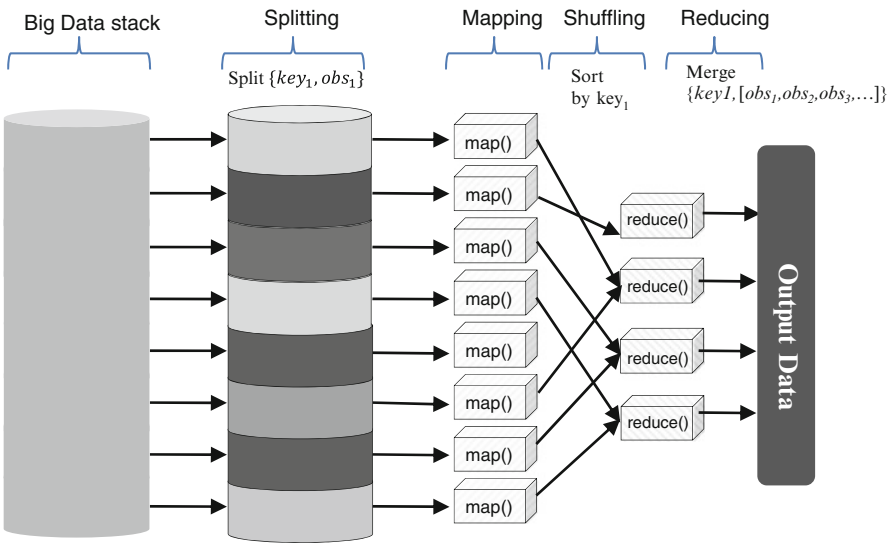


Fig. 8.12 A schematic representation of the map/reduce architecture

8.9 Alternatives and Beyond MapReduce

MapReduce is one the most known paradigms to work with big data. A commonly referred software framework that implements this paradigm is Hadoop. MapReduce implementations have shown good performance when dealing with vast amounts of pre-stored data with many algorithms [95, 96]. However, it also has some inherent pitfalls as described in [97], which makes it difficult to consider MapReduce as a general solution for all bulk processing (mostly batch) problems. Apart from the MapReduce paradigm, there are other architectures and frameworks offering open-source solutions, which improve the concept of MapReduce in terms of processing speed, data caching, querying of nested datasets or other specific features. Table 8.8 below provides a summary of some of these frameworks and their respective properties.

8.10 Conclusions

In this chapter, we have presented the basic concept of multi-sensor data fusion and its implementation at data, feature and decision levels. In essence, sensor fusion is analogous to the sensing and cognitive process used by humans to integrate data from different sources. The advantages of effective sensor fusion include improved SNR, enhanced robustness and reliability in the event of sensor failure, extended

Table 8.8 Alternative frameworks for big data management

Name	Properties	Languages	Machine learning	Reference
Hive	Querying of big data by means of an SQL-like system Build on top of Hadoop	HiveQL C++, Java, Python, PHP, Ruby	Framework does not provide ML specific features	[100]
Dremel	Big data query possible through an SQL-like syntax	DrQl Java	Framework does not provide specific features to implement ML	[101]
Percolator	Incremental updates of data while enabling distributed computation	Java	It does not provide specific features to implement ML	[102]
Spark	Store partial execution of distributed execution in a memory cache to speed computation	Scala Java Python	No specific features for ML, although possible to implement	[103]
Shark	Data caching to speed computation Querying by an SQL-like syntax Fault-tolerance architecture	Hive QL Scala	Specific features for implementing ML algorithms directly on the results of an SQL-query	[104]
Pregel	A specific parallel computation framework for graph theory	Java Erlang	Provides features for computing graph theory operations and related statistic	[105]

parameter coverage, improved resolution, precision, confidence, hypothesis discrimination and reduced uncertainty.

In general, sensor replication imposes the problem of data integration. For multi-sensor data fusion, it is usually assumed that communication, storage, and processing systems are reliable and the research focus is placed on different fusion algorithms for integrating data from either homogeneous or heterogeneous data [10]. Research concerning general wireless networks, on the other hand, assumes that the input data is generally in good quality and sources of error are originated from faults in communication and processing systems.

For direct data fusion, we have mainly concentrated on two examples that are useful for sensor arrays and blind source recovery. In general, the use of direct data fusion can be ad hoc and depends on the application requirement one may have. One of the main applications of direct data fusion is for multi-sensor calibration. Calibration in general refers to the process of correcting systematic errors or bias in sensor readings. It has also been used in reference to the procedure by which the raw outputs of the sensors are mapped to standardised units. Traditionally, calibration is done during production time and recalibration is necessary after an interval for most sensors, as factors such as aging, decay and damage can have a detrimental effect on the accuracy of the readings. Frequent recalibration of large-scale sensor

networks can be problematic for pervasive sensing environments due to the number of sensors involved. For acceleration sensors, calibration is necessary if they are to be used collaboratively. For example, two orthogonally mounted two-axes accelerometers are often used for providing three-axes acceleration measurement [98, 99]. Furthermore, MEMS-based accelerometers are usually not calibrated after production and the sensors can have sensitivity and offset bias on each of their axes.

In this chapter, we have dedicated a significant amount of space on feature-level data fusion because this is one of the most important yet difficult levels of sensor fusion. In general, a feature-based classifier consists of two major components that include feature selection and classification. The goal of defining feature is to preserve the class-discriminative information of the data while ignoring information that is irrelevant. Once a feature is identified, it defines a transformation that maps directly sensed data to the feature space, which typically has a lower dimensionality because of the inherent data abstraction involved [106]. For BSNs, effective sensor fusion and statistical feature reduction is crucial to wireless sensor arrays for both built-in redundancies and tissue heterogeneity.

One of the main theoretical components of the chapter is the development of a filter-based algorithm for feature selection based on Bayesian theory. We have demonstrated the relationship between conditional independence and ideal inference performance, which is described by the ROC curve. We have shown that adding and removing an irrelevant feature will not change the associated AUC. This property, together with the monotonic property of the AUC, defines a theoretical framework that is equivalent to the axiomatic system, consisting of decomposition, weak union and contraction properties [107], for probabilistic dependencies. Based on this framework, the proposed algorithm is designed to provide an accurate, robust and fast feature selection method. The algorithm has shown promising strengths in identifying irrelevant features and improving the accuracy of the classifiers. By using both artificial and real-world datasets for accuracy assessment, we also illustrated the roles of prior and likelihood probabilities in the filter selection.

For a specific induction algorithm, wrapper selection usually achieves a better performance than its filter counterparts in terms of accuracy. However, filter selection is usually faster than wrapper selection, and therefore can serve as a pre-processing step for a wrapper-based method. On the other hand, filter selection provides a more accurate description of the intrinsic structure of an information system, since it is not designed to cater for the bias associated with a specific classifier.

In feature selection, we have attempted to reduce the number of selected features and achieve a high value of AUC. This is a dilemma in practice since the AUC is a monotonically increasing function. Given a small training dataset, a high AUC with a large number of selected features could result in over-fitting. Determining the optimal balance is theoretically and practically important. Further investigation is needed to determine the best way to take the factors, such as misclassification cost information and the standard error of the AUC [108], into consideration.

For direct data and feature level fusion, we have omitted detailed discussion about the use of neural networks in this chapter. The use of neural networks is in fact crucial to the practical deployment of BSNs due to its potential for the

direct hardware implementation of some of the sensor fusion and classification algorithms. This is useful for implementing low-power processing at the source of the sensing environment in order to provide in situ data processing and abstraction.

In the next chapter, we will provide a detailed explanation of the neural networks approach for context-aware sensing and outline the strength of SOMs for effective activity recognition. It is also worth noting that in terms of decision-level sensor fusion, we have only briefly outlined some of the common techniques used without going into extensive details of the techniques. This is because once the data abstraction is reached to this level, the techniques used for BSNs are effectively the same as many other pattern recognition and machine learning techniques, and there is an extensive coverage in the literature of this area. In Chap. 10 we will provide a comprehensive analysis of the use of Bayesian inferencing technique for providing an autonomic sensing environment for BSNs.

Finally, for readers that are new to the field of machine learning and pattern recognition, it is important to differentiate the notion of dimensionality reduction and feature selection. For dimensionality reduction, the independent variables that account for most variability within the data are extracted during the processing stage. As only intrinsic features are preserved, dimensionality reduction may lead to better understanding and classification of the data. For feature selection, however, we are mainly concerned with determining the optimum features both in terms of the number of sensing channels used, and the relevant feature extraction algorithms for effective separation of different events or episodes.

For BSN applications, dimensionality reduction is therefore mainly a post-processing step for re-projecting the data onto a low dimensional space that preserves the internal structure of the data whilst revealing its intrinsic pattern separations. Feature selection, however, is generally used for determining which sensors to be used and how they should be placed on the body, given a set of events or episodes to be monitored. The method is therefore mainly used for determining the sensor architecture with minimal power consumption and data bandwidth during practical deployment of BSNs.

References

1. Hernandez A, Carrault G, Mora F, Thoraval L, Passariello G, Schleich JM. Multisensor fusion for atrial and ventricular activity detection in coronary care monitoring. IEEE Transactions on Biomedical Engineering 1999; 46(10):1186–1190.
2. Thomopoulos SCA. Sensor integration and data fusion. Journal of Robotics Systems 1990; 7:337–372.
3. Hall DL. Handbook of multisensor data fusion. Boca Raton: CRC Press, 2001.
4. Esteban J, Starr A, Willetts R, Hannah P, Bryanston-Cross P. A review of data fusion models and architectures: towards engineering guidelines. Neural Computing and Applications 2005.
5. White FE. Data fusion lexicon. Joint Directors of Laboratories, Technical Panel for C3, Data Fusion Sub-Panel Naval Ocean Systems Center, San Diego 1991.
6. Luo RC, Kay MG. Multisensor integration and fusion in intelligent systems. IEEE Transactions on Systems, Man and Cybernetics 1989; 19(5):901–931.

7. Sadjadi FA. Selected papers on sensor and data fusion. SPIE Milestone, SPIE Optical Engineering Press, 1996.
8. Brooks RR, Iyengar SS. Multi-sensor fusion: fundamentals and applications with software. Upper Saddle River, New Jersey: Prentice Hall, 1998.
9. Luo RC, Yih C-C, Su KL. Multisensor fusion and integration: approaches, applications, and future research directions. IEEE Sensors Journal 2002; 2(2):107–119.
10. Parhami B. Multi-sensor data fusion and reliable multi-channel computation: unifying concepts and techniques. In: Proceedings of the Twenty-Ninth Asilomar Conference on Signals, Systems and Computers, Pacific Grove, California, 1995; 1:745–749.
11. Unser M, Eden M. Weighted averaging of a set of noisy images for maximum signal-to-noise ratio. IEEE Transactions on Acoustics, Speech, and Signal Processing 1990; 38(5):890–895.
12. Cichocki A, Amari S-I. Adaptive blind signal and image processing: learning algorithms and applications. John Wiley and Sons, 2002.
13. Eriksson J, Koivunen V. Identifiability, separability, and uniqueness of linear ICA models. IEEE Signal Processing Letters 2004; 11(7):601–604.
14. Bell AJ, Sejnowski TJ. An information maximisation approach to blind separation and blind deconvolution. Neural Computation 1995; 7(6):1004–1034.
15. Lee T-W, Girolami M, Sejnowski TJ. Independent component analysis using an extended infomax algorithm for mixed subgaussian and supergaussian sources. Neural Computation 1999; 11(2):417–441.
16. Herault J, Jutten C. Space or time adaptive signal processing by neural network models. In: Proceedings of AIP Conference, Neural Networks for Computing, Snowbird, UT, 1986; 151:207–211.
17. Comon P. Independent component analysis, a new concept? Signal Processing 1994; 36(3):287–314.
18. Hyvarinen A, Karhunen J, Oja E. Independent component analysis. John Wiley and Sons, 2001.
19. Boyd JP. Chebyshev and Fourier spectral methods, 1st ed. Springer-Verlag, 1989.
20. Hall DL, McMullen AHS. Mathematical techniques in multisensor data fusion, 2nd ed. Boston: Artech House, 2004.
21. Aha D, Kibler D, Albert M. Instance based learning algorithms. Machine Learning 1991; 6:37–66.
22. Mitchell T. Machine learning. New York: McGraw Hill, 1997.
23. Cover TM, Hart PE. Nearest neighbour pattern classification. IEEE Transactions on Information Theory 1967; 13(1):21–27.
24. Parzen E. On estimation of a probability density function and mode. The Annals of Mathematical Statistics 1962; 33(3):1065–1076.
25. MacQueen B. Some methods for classification and analysis of multivariate observations. In: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability 1967; 1:281–297.
26. Ball GH, Hall DJ. ISODATA, an iterative method of multivariate analysis and pattern classification. In: Proceedings of the IFIPS Congress 1965.
27. Gowda KC, Krishna G. Agglomerative clustering using the concept of mutual nearest neighbourhood. Pattern Recognition 1978; 10(2):105–112.
28. Dhillon I, Modha D. A data clustering algorithm on distributed memory multiprocessor. In: Large-Scale Parallel Data Mining, Lecture Notes in Artificial Intelligence.2000;245–260.
29. Friedman M, Kandel A. Introduction to pattern recognition: statistical, structural, neural and fuzzy logic approaches. London: Imperial College Press, 1999.
30. Bezdek JC. Pattern recognition with fuzzy objective function algorithms. New York: Plenum Press, 1981.
31. Breiman L, Friedman JH, Olshen RA, Stone CJ. Classification and regression trees. Belmont, California: Wadsworth International Group, 1984.

32. Quinlan JR. C4.5: Programs for machine learning. San Mateo, California: Morgan Kaufmann Publishers, 1993.
33. Vapnik V. The nature of statistical learning theory. Berlin: Springer-Verlag, 1995.
34. Cristianini N, Shawe-Taylor J. An introduction to support vector machines. Cambridge University Press, 2000.
35. Cortes C, Vapnik V. Support-vector networks. Machine Learning 1995; 20(3):273–297.
36. Breiman L. Random forests. Machine Learning 2001; 45(1):5–32.
37. Dietterich TG. Ensemble methods in machine learning. In: Multiple Classifier Systems. Springer, 2000;1–15
38. Maclin R, Opitz D. Popular ensemble methods: An empirical study. arXiv preprint arXiv:11060257 2011.
39. Jain AK, Mao J, Mohiuddin KM. Artificial neural networks: a tutorial. Computer 1996; 29(3):31–44.
40. Kohonen T. Self-Organising Maps. Springer Series in Information Sciences, Berlin: Springer-Verlag, 1995.
41. Carreira-Perpinan MA. Continuous latent variable models for dimensionality reduction and sequential data reconstruction. University of Sheffield, PhD Thesis, 2001.
42. Duda RO, Hart PE. Pattern classification and scene analysis. New York: Wiley-Interscience, 1973.
43. Hinton GE, Revow M, Dayan P. Recognizing handwritten digits using mixtures of linear models. Advances in Neural Information Processing Systems 1995; 7:1015–1022.
44. Bregler C, Omoundro SM. Nonlinear image interpolation using manifold learning. Advances in Neural Information Processing Systems 1995; 7:973–980.
45. Kohonen T. Self-organization and associative memory, 2nd ed. Berlin: Springer-Verlag, 1987.
46. Bishop CM, Svensen M, Williams CKI. The generative topographic mapping. Neural Computation 1998; 10(1):215–234.
47. Cox TF, Cox MA. Multidimensional scaling, 2nd ed. London: Chapman and Hall, 2001.
48. Steyvers M. Multidimensional scaling. In: Encyclopedia of Cognitive Science. London: Nature Publishing Group, 2002.
49. Roweis ST, Saul LK. Nonlinear dimensionality reduction by locally linear embedding. Science 2000; 290(5500):2323–2326.
50. Belkin M, Niyogi P. Laplacian eigenmaps and spectral techniques for embedding and clustering. In: Advances in Neural Information Processing Systems. MIT Press, 2001; 585–591.
51. Tenenbaum JB, de Silva V, Langford JC. A global geometric framework for nonlinear dimensionality reduction. Science 2000; 290(5500):2319–2323.
52. Duda RO, Hart PE, Stork DG. Pattern classification. John Wiley and Sons, 2012.
53. Saeys Y, Inza I, Larrañaga P. A review of feature selection techniques in bioinformatics. Bioinformatics 2007; 23(19):2507–2517.
54. Korf RE. Depth-first iterative-deepening: An optimal admissible tree search. Artificial Intelligence 1985; 27(1):97–109.
55. Derksen S, Keselman H. Backward, forward and stepwise automated subset selection algorithms: Frequency of obtaining authentic and noise variables. British Journal of Mathematical and Statistical Psychology 1992; 45(2):265–282.
56. Thiemjarus S. A framework for contextual data fusion in Body Sensor Networks. Imperial College London, PhD Thesis, 2007.
57. John GH, Kohavi R, Pfleger K. Irrelevant features and the subset selection problem. In: Proceedings of the Eleventh International Conference on Machine Learning, New Brunswick, NJ, 1994; 121–129.
58. Blum AL, Langley P. Selection of relevant features and examples in machine learning. Artificial Intelligence 1997; 97(1–2):245–271.

59. Koller D, Sahami M. Towards optimal feature selection. In: Proceedings of the Thirteenth International Conference on Machine Learning, Bari, Italy, 1996; 284–292.
60. Kira K, Rendell LA. The feature selection problem: traditional methods and a new algorithm. In: Proceedings of the Ninth National Conference on Artificial Intelligence, Cambridge, Massachusetts, 1992; 129–134.
61. Almuallim H, Dietterich TG. Learning with many features. In: Proceedings of the Ninth National Conference on Artificial Intelligence, Cambridge, Massachusetts, 1992; 547–552.
62. Dash M, Liu H. Feature selection for classification. Intelligent Data Analysis 1997; 1(3):131–156.
63. Kohavi R, John GH. Wrappers for feature subset selection. Artificial Intelligence 1997; 97(1–2):273–324.
64. Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection. In: Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence, San Mateo, CA, 1995; 1137–1143.
65. Pudil P, Novovicoca J, Kittler J. Floating search methods in feature selection. Pattern Recognition Letters 1994; 15(11):1119–1125.
66. Jain A, Zongker D. Feature selection: evaluation, application, and small sample performance. IEEE Transactions on Pattern Analysis and Machine Intelligence 1997; 19(2):153–158.
67. Pudil P, Novovicova J, Kittler J. Simultaneous learning of decision rules and important attributes for classification problems in image analysis. Image and Vision Computing 1994; 12:193–198.
68. Hu XP. Feature selection and extraction of visual search strategies with eye tracking. Imperial College, PhD Thesis, 2005.
69. Egan JP. Signal detection theory and ROC analysis. New York: Academic Press, 1975.
70. Van-Trees HL. Detection estimation and modulation theory. New York: John Wiley and Sons, 1971.
71. Scott MJJ, Niranjan M, Prager RW. Parcel: feature subset selection in variable cost domains. Department of Engineering, University of Cambridge, England, Technical Report, CUED/F-INFENG/TR, 1998; 323.
72. Srinivasan A. Note on the location of optimal classifiers in n-dimensional ROC space. Oxford University Computing Laboratory, Oxford, England, Technical Report, PRG-TR-2-99, 1999.
73. Wilcoxon F. Individual comparisons by ranking methods. Biometrics 1945; 1:80–83.
74. Hand DJ. Construction and assessment of classification rules. Chichester, UK: John Wiley and Sons, 1997.
75. Coetzee F, Lawrence S, Giles CL. Bayesian classification and feature selection from finite data sets. In: Proceedings of the Sixth Conference on Uncertainty in Artificial Intelligence, Stanford, CA, 2000; 89–97.
76. Singh M, Provan GM. Efficient learning of selective Bayesian network classifiers. In: Proceedings of the Thirteenth International Conference on Machine Learning, Bari, Italy, 1996; 453–461.
77. Zadeh LA. Fuzzy algorithms. Information and Control 1968; 12(2):94–102.
78. Dempster AP. A generalization of Bayesian inference. Journal of the Royal Statistical Society 1968; 30:205–247.
79. Shafer G. A mathematical theory of evidence. Princeton University Press, 1976.
80. Shafer G. Perspectives on the theory and practice of belief functions. International Journal of Approximate Reasoning 1990; 4(5–6):323–362.
81. Shafer G. Readings in uncertain reasoning. San Mateo, California: Morgan Kaufmann, 1990.
82. Conti V, Milici G, Ribino P, Sorbello F, Vitabile S. Fuzzy fusion in multimodal biometric systems. Knowledge-Based Intelligent Information and Engineering Systems 2007; LNCS 4692:108–115.
83. Kenneth E, Rajendra A, Kannathal N, Lim CM. Data fusion of multimodal cardiovascular signals. In: Proceedings of the Annual International Conference of the Engineering in Medicine and Biology Society, Shanghai, China, 2005; 4689–4692.

84. Medjahed H, Istrate D, Boudy J, Dorizzi B. Human activities of daily living recognition using fuzzy logic for elderly home monitoring. In: Proceedings of the IEEE International Conference on Fuzzy Systems, Jeju Island, Korea, 2009; 2001–2006.
85. Zadeh LA. Soft computing and fuzzy logic. IEEE Software 1994; 11(6):48–56.
86. Pirjanian P, Matarić M. Multiple objective vs. fuzzy behavior coordination. In: Fuzzy Logic Techniques for Autonomous Vehicle Navigation. Springer, 2001; 235–253.
87. Mendel JM, John RB. Type-2 fuzzy sets made simple. IEEE Transactions on Fuzzy Systems 2002; 10(2):117–127.
88. Doctor F, Hagras H, Callaghan V. A type-2 fuzzy embedded agent to realise ambient intelligence in ubiquitous computing environments. Information Sciences 2005; 171(4):309–334.
89. Dean J, Ghemawat S. MapReduce: simplified data processing on large clusters. Communications of the ACM 2008; 51(1):107–113.
90. Saad Y. Iterative methods for sparse linear systems. Siam, 2003.
91. Pothen A, Simon HD, Liou K-P. Partitioning sparse matrices with eigenvectors of graphs. SIAM Journal on Matrix Analysis and Applications 1990; 11(3):430–452.
92. Ferziger JH, Perić M. Computational methods for fluid dynamics. Springer Berlin, 1996.
93. Eddy SR. What is dynamic programming? Nature Biotechnology 2004; 22(7):909–910.
94. Denardo EV. Dynamic programming: models and applications. Dover Publications, 1982.
95. Yang H-c, Dasdan A, Hsiao R-L, Parker DS. Map-reduce-merge: simplified relational data processing on large clusters. In: Proceedings of the ACM SIGMOD International Conference on Management of Data, Beijing, China, 2007; 1029–1040.
96. Catanzaro B, Sundaram N, Keutzer K. Fast support vector machine training and classification on graphics processors. In: Proceedings of the Twenty-Fifth International Conference on Machine Learning, Helsinki, Finland, 2008; 104–111.
97. Lee K-H, Lee Y-J, Choi H, Chung YD, Moon B. Parallel data processing with MapReduce: a survey. ACM SIGMOD Record 2012; 40(4):11–20.
98. Krohn A, Beigl M, Decker C, Kochendorfer U, Robinson P, Zimmer T. Inexpensive and automatic calibration for acceleration sensors. In: Proceedings of the Second International Symposium of Ubiquitous Computing Systems, Tokyo, Japan, 2004; Springer LNCS 3598:245–258.
99. Lukowicz P, Junker H, Troster G. Automatic calibration of body worn acceleration sensors. In: Proceedings of the Second International Pervasive Computing Conference, Vienna, Austria, 2004; 176–181.
100. Venner J. Pro Hadoop. Apress, 2009
101. Melnik S, Gubarev A, Long JJ, Romer G, Shivakumar S, Tolton M, et al. Dremel: interactive analysis of web-scale datasets. Proceedings of the VLDB Endowment 2010; 3(1–2):330–339.
102. Peng D, Dabek F. Large-scale incremental processing using distributed transactions and notifications. In: Proceedings of the Ninth USENIX Symposium on Operating Systems Design and Implementation, Vancouver, BC, Canada, 2010; 10:1–15.
103. Zaharia M, Chowdhury M, Franklin MJ, Shenker S, Stoica I. Spark: cluster computing with working sets. In: Proceedings of the Second USENIX Conference on Hot Topics in Cloud Computing, Boston, MA, USA, 2010; 10.
104. Xin R, Rosen J, Zaharia M, Franklin MJ, Shenker S, Stoica I. Shark: SQL and rich analytics at scale. In: Proceedings of the ACM SIGMOD International Conference on Management of Data, NY, USA, 2013; 13–24.
105. Malewicz G, Austern MH, Bik AJ, Dehnert JC, Horn I, Leiser N, et al. Pregel: a system for large-scale graph processing. In: Proceedings of the ACM SIGMOD International Conference on Management of Data, Indianapolis, IN, USA, 2010; 135–146.
106. Liu J, Chang K-C. Feature-based target recognition with a Bayesian network. Optical Engineering 1996; 35(3):701–707.
107. Pearl J. Probabilistic reasoning in intelligent systems: networks of plausible inference. San Mateo, CA: Morgan Kaufmann, 1988.
108. Hanley JA, McNeil BJ. The meaning and use of the area under the receiver operating characteristic (ROC) curve. Diagnostic Radiology 1982; 143:29–36.

Chapter 9
Context Aware Sensing

Surapa Thiemjarus and Guang-Zhong Yang

9.1 Introduction

In recent years, there have been considerable interests in context-aware sensing for pervasive computing. Context can be defined as "*the circumstances in which an event occurs*" and this concept has been successfully used in information processing for over 50 years, particularly for *Natural Language Processing* (NLP) and *Human Computer Interaction* (HCI). The popularity of the context-aware architectures is due to the increasingly ubiquitous nature of the sensors, as well as the diversity of the environment under which the sensed signals are collected. To understand the intrinsic characteristics of the sensed signals and determine how BSNs should react to different events, the contextual information is essential to the adaptation of the monitoring device so as to provide more intelligent support to the users.

One of the earliest examples of context-aware computing was the Active Badge from the Olivetti Research Lab in 1992 [1], and the general term "*context-aware computing*" was first introduced by Schilit and Theimer in 1994 [2]. In this work, they described three important aspects of context-awareness: *where you are*, *who you are with*, and *what resources are nearby*. In other words, they are mainly concerned with location and identity information for context-aware computing. In a mobile distributed computing system named PARCTAB, Schilit et al. considered four different categories of context-aware applications, which included proximate selection, automatic contextual reconfiguration, commands and context-triggered action [3]. Proximate selection provides an interface based on the location or capacity of objects, so that located objects are highlighted or made easier to choose from. Reconfiguration, on the other hand, is the process of adding new components,

S. Thiemjarus
National Electronics and Computer Technology Center, Pathumthani 12120, Thailand
e-mail: surapa.thiemjarus@nectec.or.th

G.-Z. Yang, PhD (✉)
The Hamlyn Centre, Imperial College London, London SW7 2BY, UK
e-mail: g.z.yang@imperial.ac.uk

G.-Z. Yang (ed.), *Body Sensor Networks*, DOI 10.1007/978-1-4471-6374-9_9,

removing existing components or suggesting the connection between components. In the early work of context-aware applications, context triggered actions were simple IF-THEN rules used to specify how the system should adapt to changing environment and user interactions.

Thus far, there has been extensive effort in formally categorising different features of context-aware applications. Whilst the early taxonomy has aimed at identifying different classes of context-aware applications, Pascoe [4] concentrated on the following considerations:

- *Contextual sensing*: the ability to detect contextual information and present it to the user to augment the user's sensory system;
- *Contextual adaptation*: the ability to execute or modify a service automatically based on the current context;
- *Contextual resource discovery*: allows context-aware applications to locate and exploit relevant resources and services; and
- *Contextual augmentation*: the ability to associate digital data with the user's context in such a way that a user can view the data when he is in the associated context.

It is evident that some of the definitions used above are equivalent to the taxonomy proposed by Schilit et al. [3]. For example, contextual sensing can be mapped to Schilit's proximity selection, contextual adaptation is similar to context-triggered action and contextual resource discovery can be regarded as automatic contextual reconfiguration. By considering different categorisations of both context and context-aware applications, Dey and Abowd defined context as [5]:

> Any information that can be used to characterise the situation of an entity. An entity is a person, place, or object that is considered relevant to the interaction between a user and an application, including the user and application themselves.

In addition to location and identity, activity and time were added to their context categorisation, and context-awareness was defined as:

> A system is context-aware if it uses context to provide relevant information and/or services to the user, where relevancy depends on the user's task.

From these studies, the main features for context-aware applications can be considered as the presentation of information and services to a user according to the current context, the automatic execution of a service, and the tagging of context to information for later retrieval. The general definition of a context-aware system is therefore related to the issue of whether the system can extract, interpret and use contextual information and adapt its functionality to the current context of use. Table 9.1 illustrates some of the main considerations for designing context-aware systems [6, 7].

For the purpose of BSNs, the main emphasis of a context-aware design is concerned with the interpretation of physical and biochemical signals acquired from both wearable and implantable sensors and their association with the ambient environment. The contextual information is therefore mainly focussed on the user's activity, physiological states and the physical environment around the user. In this case, the environmental context includes location, proximity, time, social

Table 9.1 Considerations of context-aware systems

Main considerations
• Identity, e.g., user identification
• Spatial information, e.g., location, orientation, speed and acceleration
• Temporal information, e.g., time of the day, date and season of the year
• Environmental information, e.g., temperature, air quality, light or noise levels
• Social interaction, e.g., who you are with and people that are nearby
• Resources that are nearby, e.g., accessible devices and hosts
• Availability of resources, e.g., battery, display, network and bandwidth
• Physiological measurements, e.g., blood pressure, heart rate, respiration rate, tone of voice and emotions
• Activity, e.g., talking, reading, walking and running
• Planned activity, e.g., schedules and agenda
The five W's of context
• Who – The identity of the user or other people in the environment
• What – Human activity and interaction in current systems
• Where – The environment within which the activity is taking place
• When – Timestamp of the capture records
• Why – Person's affective states and intention

interaction, and connectivity information of the healthcare environment. The user-centred context, on the other hand, includes physical action, cognitive/mental activities, and affective states.

From an information processing point of view, context can be regarded as different levels of details linked to physical and perceptual representations. The description of a user's cognitive activities is generally at an abstract level, whereas the recognition of the physical status of a subject is more descriptive and mainly data-driven. For example, movements are the most elementary primitives for motion recognition, which only require local measurements. Activities, on the other hand, involve sequences of movements and require detailed motion modelling. Finally, action is at the highest level of motion understanding which requires the interpretation of the context (e.g., a set of temporal constraints on the relationship between motions) and the interaction of the user with the environment [8]. With the emergence of pervasive sensing technologies, the goal of context sensing is shifting from detecting individual activities to modelling and understanding the complex patterns of a person's activities, which can be transient, interleaved or in parallel.

9.2 Application Scenarios

The above definition outlines the general concepts and considerations for context-aware sensing. Some of the early techniques only involve simple measures such as location for contextual interpretation. These systems include a number of office,

tourist and memory aids such as the *Dynamic Ubiquitous Mobile Meeting Board* (DUMMBO) [9], PARCTAB [3], Cyberguide [10], Forget-Me-Not [11], Remembrance Agent [12], Stick-e Notes [13], and the Olivettit Research Lab's Active Badge [14] mentioned earlier.

Whilst context-awareness in mobile computing is mainly concerned with the adaptation of the application services, its use in BSNs is mainly focused on how to capture signals under varying physical and environmental conditions. This is because there is a growing need clinically for continuous patient monitoring under their natural physiological status so that transient but life-threatening abnormalities can be reliably detected or predicted. For example, Someren et al. [15] illustrated the potential use of motion signal for analysing the effect of medication intake by calculating the average responses during different times of the day for patients with Parkinson's disease. These profiles were then used to evaluate the pharmacological interventions. Bhattacharya et al. [16] investigated the heart rate, oxygen uptake and acceleration profiles during exercise to study the relationship between body movement/acceleration and metabolic rates. Tolkiehn et al. [17] proposed the use of barometric pressure and acceleration sensors for fall and fall direction detection.

The contextual information has also been used to improve the diagnosis accuracy of the acquired physiological signals. This is because similar sensory signals can be interpreted differently depending on the current activities of the patient. For instance, the underlying cause of rapid heartbeats and degenerated ECG can be a result of the vigorous movements of the patient during exercising rather than a genuine cardiac episode. For these reasons, motion signals acquired in situ have been used for recovering biosensor signals corrupted by motion artefacts [18].

Thus far, many of the activity recognition studies are based on the use of motion sensors [19–27], and a record of the daily activities of the patient is used to provide an indication of the general wellbeing of the subject. For patients with disabilities, for example, monitoring tasks that require more effort to accomplish can be used as an objective measure of their functional ability [28, 29]. Tognetti et al. [30] demonstrated the use of limb gesture detection for post-stroke rehabilitation. The use of a wearable system for clinical management of individuals undergoing rehabilitation is attractive since it allows the recording of quantitative measurements in settings other than in hospitals or clinics. Existing research has also investigated the value of continuous patient monitoring for exploring the relationship between activity and disease progression, as demonstrated by the studies of Walker et al. for rheumatoid arthritis [31] and others in detecting changes of posture and gait in patients with neuromuscular diseases and Parkinson's disease [32–36]. In [37], automatic epilepsy seizure detection was conducted based on motion sensors. The results were validated against the ground truth captured by video and *Electroencephalogram* (EEG) recordings associated with cognitions such as, memory, emotion, arousal, fatigues and distraction [38]. The use of EEG patterns for determining different stages of sleep is of significant clinical interest [39]. Eye-gaze is another important information source that can be used for context recognition. Wearable *Electro-oculography* (EOG) and *Video-oculography*

(VOG) have been used for assessing reading patterns [40], as well as the performance of surgical tasks [41]. Other sources of information such as acoustics can also be used for activity recognition [42].

Affective states of depression, anxiety and chronic anger have been shown to impede the immune system and they can potentially be used to assess stress, anger and other emotions that can influence health. Teicher [43] studied the correlations between different activity levels and psychiatric disorders. Myrtek and Brügner [44] investigated the perception of daily emotional events by assessing the correlation between physiological parameters such as heart rate, physical activity and psychological parameters. Picard et al. [45] described a recognition framework for detecting a range of emotional states including anger, hate, grief, platonic love, romantic love, joy and reverence based on signals from *Electromyography* (EMG), blood volume pressure, skin conductance, and Hall-effect respiration sensors. For BSN applications, the emotional states and attention levels are all important contextual information to capture. An example commercial monitoring tool is the SenseWear Armband from BodyMedia [25], which can be worn on the upper arm of the subject to collect data wirelessly from a combination of sensors. Such data includes information about movement, heat flow, skin temperature, ambient temperature, and *Galvanic Skin Response* (GSR). A low-cost wearable *ElectroDermal Activity* (EDA) sensor has been developed for observing patterns of sympathetic arousal [46]. The continuous monitoring of changes in skin conductance has potential clinical applications in autism, epilepsy and sleep disorders.

The use of contextual information has also been used to develop more intelligent healthcare environments. Benditt et al. [47] proposed the concept of rate-variable implantable cardiac pacemakers using motion and physiologic sensors to account for physical exertion or emotional stress. Bardram [48] illustrated the general design principles for context-aware sensing in hospitals and the use of RFID tags for identifying patients and their surrounding clinical team and medical equipment. A context-aware pill container with fingerprint recognition has been proposed to ensure proper dose administration. For introducing context-awareness to BSNs, it is also possible to exploit much of the existing research in pervasive computing, particularly for indoor navigation and tracking [20, 49–52]. Wearable sensors such as accelerometers, magnetometers, temperature and light sensors have been used extensively for location detection and tracking.

Traditionally, the knowledge of the context of the user is acquired through self-reporting based on diaries or questionnaires. This method is both time-consuming and unreliable, especially for the elderly and subjects with memory impairment. Another method of acquiring contextual information is through clinical observation but it requires specialised equipment and a dedicated laboratory set-up. In addition, measurements made in a clinic may not accurately reflect the patient's behaviour in the normal home environment. With the current advances in sensor and wireless technology, it is now possible to provide ubiquitous monitoring of the subjects under their natural physiological status.

9.3 Preprocessing for Context Sensing

Context recognition can be formulated as a general pattern recognition process which consists of data acquisition, feature extraction, model construction and inference, and performance evaluation. Clustering and high-level inferencing techniques as described in Chap. 8 for multi-sensor fusion can all be applied to context recognition. Before reaching to the recognition stage, a number of signal processing issues related to context detection need to be addressed.

9.3.1 Sources of Signal Variations

In context-aware sensing, the acquired signals can be affected by other sources of variations that are detrimental to the classification process. To enhance the overall system sensitivity and specificity, these variations must be carefully considered. Table 9.2 outlines several possible sources of variations that may be encountered by BSN applications.

In general, the variations can be handled either during preprocessing or subsequent data classification stages. In preprocessing, blind source separation, such as the ICA algorithm discussed in Chap. 8, can be used to separate the signal components generated by the real sources. In the event of node failure, the missing data can be assigned with the most common values in the training samples or according to a prescribed probability distribution for assigning the missing feature attribute. Finally, systematic sensor variations can be alleviated by effective online/offline sensor calibration.

Table 9.2 Sources of variation for context sensors

Sources of variation	
Noise	Sensor noise, node failure, and motion artefact can introduce significant errors to data inferencing results
Indirect motion	Indirect motion such as those due to car rides, use of wheelchairs or riding an elevator can contribute to movements without actual physical activities [53]
Intra- and inter-subject variability	The difference in anatomy is a major source of subject specific variations. Motion characteristics can have significant inter- and intra-subject variabilities and this must be taken into account when analysing the motion data. Furthermore, the behaviour of individuals can change due to their emotional status and the surrounding environment
Variation between sensors	Sensors of the same type can be different in terms of characteristic sensitivity, offset, and bias values. These differences can also change over time due to thermal drift [54]. Other subject specific factors such as sensor placement can also introduce sensor variations [53]

9.3.2 Data Normalisation

To account for different ranges of the signals collected, direct data normalisation can be applied. The simplest version of data normalisation is to shift the signal baseline to the same mean and then scale it by its variance so that they share the same data range. This allows a proportional influence of all features whilst standardising the dataset at the same time. Translation or phase shift along the time axis is often applied to eliminate signal drift, whereas scaling or time normalisation is applied to cater for signals acquired with different sampling rates. It is worth noting that many of the data normalisation techniques used so far are ad hoc and require extensive empirical judgements.

As an example, Picard et al. used several methods for handling day-to-day variations in their affective sensing framework [45]. These include the use of day matrix, baseline matrix and day-independent features for the analysis of affective physiological features. For the day matrix, a transform is obtained by applying the Fisher's algorithm to signals appended with a day-dependent vector. The Fisher's algorithm normalises features in the same class so that intra-class difference is minimised whilst interclass differences being maintained. The baseline matrix, on the other hand, is based on the subtraction of the mean of the reference class (baseline) from the respective features of the remaining classes of the same day.

With increasing use of smart phones for activity recognition, the handling of varying device orientation is important. To this end, appropriate signal transformation is applied to data preprocessing [55]. For example, a projection-based method for device coordinate system estimation has been proposed [56] and implemented on an iPhone [57]. The vertical axis of the global coordinate system is derived from the mean of dynamic activities and the forward axis is derived from the principal axis of data on the plane perpendicular to the gravity. Based on the assumption that the first observation is positive, the forward direction can be obtained and the sideward axis is found by cross product between the vertical and the forward axes. By using data acquired from a single device orientation, Henpraserttae et al. have shown that classification with signal transformation performs better than classification without signal transformation by ~42–51 % [56, 58], which is unsurprising. Their results also show that the amount of training dataset can affect the recognition accuracy. Furthermore, based on instance-based learning, when data from all orientations was used for training, classification with signal transformation the accuracy can be improved further [56]. Figure 9.1 illustrates the acceleration signals acquired from an iPhone placed at the waist in 16 different device orientations while a user is performing six activities before and after transformation. Signal truncation due to sensor limitation can be observed during highly dynamic activity.

Another example is the work by Figo et al., who presented a comparison of computational complexity (in terms of storage costs and memory operations) of different preprocessing techniques for accelerometer-based context recognition [59].

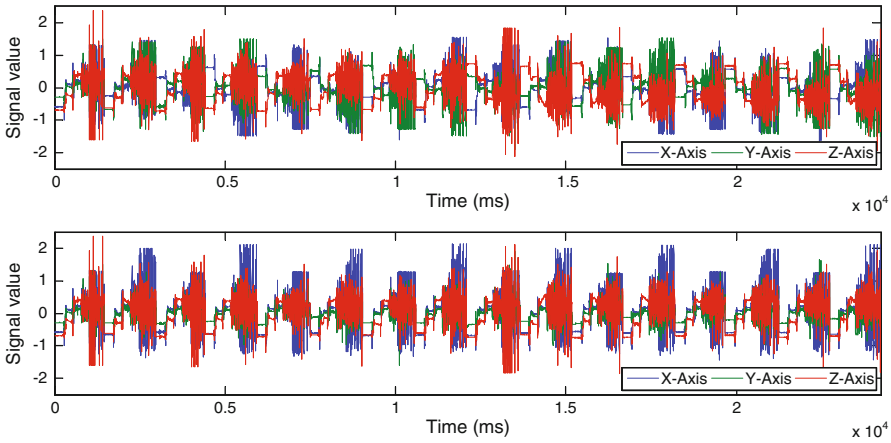


Fig. 9.1 A comparison of a concatenated normalised acceleration signals acquired from 16 different orientations placed on the subject's waist while performing lying, sitting, standing, walking, running and jumping: signals without transformation (*top*) and signals transformed by the transformation matrix (*bottom*)

9.3.3 Information Granularity

For context sensing, we are generally more interested in temporal signal variations in the feature space, as the information derived from the instantaneous signal is usually limited. Short time window analysis is the simplest method for segmenting the input sequence. The basic idea is to divide the time-varying signal into meaningful small segments. This can be done by shifting and multiplying a window function of a chosen width, Ω with the signal. The simplest window function has a rectangular shape but other window functions can also be used to pre-emphasise certain parts of the signal. By the use of a shifting window, the signal can then be divided into a succession of windowed sequences called *frames*, which can be analysed individually. The resolution of the classification process is usually determined by how the feature vectors are constructed. If the window is shifted by a temporal length of $L = 1$, a classification label will be assigned to each sample of the original signal. In most studies, L is set to be equal to Ω. In this case, the derived frames do not overlap and some temporal information may be lost. Signals can also be segmented into varying lengths according to their temporal characteristics. This allows the extraction of information from the entire episode of the signal, which tends to provide a more robust result for the classification process. In practice, however, the performance of this approach is highly dependent on the accuracy of the segmentation algorithm, and in many cases the boundary between episodes is difficult to define.

In addition to simple window-based statistics, other signal characteristics such as peak locations, pulse repetition intervals, and zero crossing rates can also be used [60, 61]. In the peak-based feature extraction method, the 'area of activity' is first detected by applying a thresholding scheme on the running variance followed by

peak localisation. The drawback of the method is that the peak information is not always available and tracking peaks in multiple dimensions is difficult. A more systematic approach to extracting localised signal features is the use of *Discrete Wavelet Transform* (DWT) and it has been applied successfully to a number of context-aware applications [62]. The choice of appropriate mother wavelets and the corresponding scales for different types of activities, however, remains an active research area.

As an example, Loosli et al. [63] proposed an interesting online nonparametric signal segmentation technique with one-class *Support Vector Machines* (SVMs) to detect context changes. The technique is based on the concept of change-detection in signal processing [64], and the nonparametric requirement is achieved by the kernel learning method used in SVMs. A one-class SVM is trained by past data to first learn the current state and then examine the subsequent data sequence. A change in signal characteristics is detected when the proportion of misclassification exceeds a given threshold. The method allows decomposition of a multidimensional time series into stationary (or weakly stationary) segments, thus allowing feature extraction to be specific to each context segment and adaptive to context transitions.

9.4 Context Recognition Techniques

Thus far, most of the context recognition techniques are based on motion sensors and commercially available physiological sensors such as skin conductance, heart rate, and respiratory sensors. Table 9.3 illustrates several examples of the context recognition applications that have been developed up until 2013. In this table, we have also listed out the corresponding processing models that have been used. Of course, these are by no means exhaustive and only provided here as examples. It can be seen that many of the sensor fusion techniques described in Chap. 8 are applicable for this purpose. In this chapter, we will mainly focus on two important approaches for context-aware sensing: *Artificial Neural Networks* (ANNs) – particularly the use of *Self-Organising Maps* (SOMs) and *Probabilistic Graphical Models* (i.e. *Hidden Markov Models* (HMMs) and *Factor Graphs* (FGs)).

9.4.1 Artificial Neural Networks (ANNs)

For context sensing, the use of ANNs offers several important features including nonlinearity (they are suitable for data which is inherently nonlinear), adaptivity (they can be easily retrained to deal with minor changes), evidential response (they can be designed to provide confidence for the decision made) and fault tolerance (their performance degrades gracefully under adverse operating conditions) [81]. Moreover, the relatively small number of operations involved in the combined

Table 9.3 Example applications for context sensing and the recognition techniques used

References	Sensors used	Context detection techniques	Purpose of study
Krause et al. [25]	BodyMedia's Sensewear armband (accelerometers, gyroscope, GSR, heat flux, ambient and skin thermometers)	SOM and 1st order Markov model based on FFT and PCA features; PCA is used for online recalibration	Identification of physiological and activity context
Bao and Intille [19]	Accelerometers	Decision table, instance-based learning, C4.5, and Naïve Bayes using mean, energy, and frequency domain entropy	Activity recognition (20 activities including walking, sitting and relaxing, watching TV, running, stretching, scrubbing, folding laundry, brushing teeth, riding elevator, and eating or drinking, etc.)
Lee and Mase [20]	Accelerometers and gyroscope	Three layers of functional blocks (sensing, motion, and location) incorporating fuzzy logic	Activity and location recognition, where activities involved include sitting, standing, walking level, and walking up/down stairs
Ravi et al. [65]	Accelerometer	Base-level classifier: decision table, decision trees (C4.5), *k*-nearest neighbours, SVM, Naïve Bayes; meta-level classifiers: boosting, bagging, plurality voting, stacking with ordinary-decision trees and stacking with metadecision trees	Activity recognition (standing, walking, running, climbing up stairs, climbing down stairs, sit-ups, vacuuming, and brushing teeth)
Loosli et al. [63]	EMG, blood volume pressure, skin conductivity and respiration sensors	One class SVM and rupture detection algorithm (SVM is trained based on the data in the past window, when the recognition accuracy drops, context change is detected)	Context change detection
Picard et al. [45]	EMG, blood volume pressure, skin conductance sensor, Hall-effect respiration sensor	*K*-nearest neighbour	Emotion recognition (no emotion, anger, hate, grief, platonic love, romantic love, joy, and reverence)

Noguchi et al. [66]	Vivid room	C5.0 decision tree (ID4-based algorithm)	Recognition of human intention (studying, arranging, eating and resting)
Najafi et al. [62]	Accelerometer and gyroscope	A specific rule-based algorithm based on wavelet coefficients	Activity monitoring in the elderly (sleeping posture, sitting, standing and walking)
Patterson et al. [67]	GPS	Dynamic Bayesian Network; Bayes particle filter is learned using EM to model the transportation modes as well as location and velocity	Transport behaviour inference (mode estimation and prediction and location prediction)
Barger et al. [68]	The SmartHouse system with motion sensors and switches	Mixture models trained by EM algorithm	Analysis of work and off day behavioural patterns
Tapia et al. [69]	State-change sensors such as reed switches	Multi-class Naïve Bayesian and binary Naïve Bayesian for each activity	Recognition of daily living activities (such as preparing lunch, toileting, preparing breakfast, bathing, dressing, grooming, preparing a beverage and watching TV, etc.)
Dalton et al. [37]	Accelerometers (with EEG and video as reference)	Incorporates a mass–spring template into the DTW algorithm	Epileptic seizures detection
Chambers et al. [70]	Accelerometers and video	HMMs	Automatic video annotation
Bulling et al. [40]	EOG	SVM	Office activity recognition
Chen et al. [42]	Microphone	Six-state continuous-density HMMs, each state with two Gaussian mixture components	Bathroom activity recognition
Philipose et al. [71]	RFID tags	Activities are modelled as a linear sequence of activity stages and then translated into a Dynamic Bayesian Network incorporating a particle filter based inference engine	Modelling activities of daily living (such as oral hygiene, toileting, washing, housework, safe use of appliances, use of heating, care of clothes and linen, taking medication preparing simple snack, and use of telephone, etc.)
Tanyawiwat and Thiemjarus [72]	Glove sensor (with bend sensors, a accelerometer, conductive fabric contact sensors)	Multivariate Gaussian distribution	Hand sign language gesture recognition

(continued)

Table 9.3 (continued)

References	Sensors used	Context detection techniques	Purpose of study
Poh et al. [46]	Electrodermal activity sensor	Correlation analysis	Understanding of psychological or neurological conditions
Sazonov and Fontana [73]	Piezoelectric strain gauge sensor	SVMs	Food intake detection
Huo and Ghovanloo [74]	Magneto-inductive sensors	Threshold-based comparison	Passage of pill or capsule through the oesophagus upon ingestion
Patel et al. [75]	Accelerometers	SVMs	Estimation of the severity of Parkinsonian symptoms and motor complications
Ghasemzadeh et al. [76]	EMG and accelerometer	*K*-nearest neighbour and neural classifiers	To establish relationship between postural control system and muscular activities and develop a method for assessing human balance
Tolkiehn et al. [17]	Accelerometer and barometric pressure sensor	A threshold-based algorithm	Fall detection and determine the direction of a fall
Bourke and Lyons [77]	Gyroscope	A threshold-based algorithm	Fall detection
Dadashi et al. [78]	Accelerometer and gyroscope	HMMs	Detection of phases in breaststroke swimming
King et al. [79]	Sensor glove with accelerometers and a fibre-optic bend sensor.	HMM clustering framework	Surgical skill assessment
Pradhan et al. [80]	CyberGlove with 22 strip sensors	*K* weight angular similarity (kWAS) comparison	Hand sign language gesture recognition
Mitchell et al. [26]	Smartphone accelerometers	SVM, optimised classification model, a fusion of classifiers (and different Discrete Wavelet Transform features)	Recognition of field sport activities (i.e., stationary, walking, jogging, sprinting, hitting the ball, attempting a standing tackle and dribbling the ball)

learning and classification process makes the model particularly suited for a parallel, on-chip analogue implementation [82]. For a BSN, this means some of the processing steps involved can be performed locally on low-power, miniaturised sensor nodes so as to minimise the communication bandwidth required. This is particularly attractive for distributed inferencing.

The underlying mechanism of an ANN is inspired by the neurobiological system of our brain. A neuron consists of a cell body called a soma which contains the nucleus, and a number of short, branching cellular extensions called dendrites. They form the main information receiving network for the neuron. The axon is a much finer, cable-like structure which carries nerve signals away from the neuron to connect with the dendrites and cell bodies of other neurons. The connection point is called a synapse. Neurons have only one axon, but this axon can undergo extensive branching, enabling communication with many target cells. Each neuron can receive, process and transmit electrochemical signals. Synapses can be either chemical or electrical. In an electrical synapse, the membranes of two neurons are continuous at tiny spots called gap-junctions, making the cells electrically contiguous. In the case of chemical synapses, neurotransmitters are released from a presynaptic neuron and dock with receptor proteins on the postsynaptic neuron. Such binding causes the shape of the protein to change and ion channels to open. The firing of a neuron depends on how many inputs it is receiving as well as the nature of each input signal (excitatory or inhibitory) at each synapse. The net result of these inputs determines whether the neuron will become excited, or depolarised, enough to fire an action potential and release neurotransmitter from its axon terminals.

The history of ANNs begins with the model of the biological neuron introduced by McCulloch and Pitts in 1943 [83]. The *McCulloch-Pitts* (MP) neuron is described as a linear threshold computing unit with multiple inputs and a single binary output. Each input x_i is connected to the j^{th} neuron by a directed synaptic connection with weight w_{ij}. The neuron is activated and returns value 1 when the sum of the weighted inputs exceeds a specified threshold θ_j. Otherwise, the output value is 0. Mathematically, the response of an MP neuron can be written as:

$$y_j = f\left(\sum_i w_{ij} x_i(t) - \theta_j\right) \tag{9.1}$$

where $f(x) = 1$ if $x \geq 0$, otherwise $f(x) = 0$. In 1949, Hebb postulated the first rule for self-organised learning, which states that the effectiveness of a variable synapse between two neurons is increased if the two interconnected neurons are activated at the same time [84]. Based on the McCulloch-Pitts model, the single-layer perceptron was proposed by Rosenblatt in 1958 [85]. The model is considered as the first ANN for supervised learning. A perceptron is a neuron with adjustable weights w_i, for $i = 1, 2, \ldots, d$, and an externally applied threshold bias w_0. Table 9.4 describes the procedure for learning the weights and threshold for a perceptron.

Table 9.4 The perceptron learning algorithm

1. Initialise the weights and threshold to small random values;
2. For each input vector, $\mathbf{x}(t)$ (t is the time step index):
   (a) Evaluate the output y of the neuron by applying the binary step activation function f to the linear combination of inputs and an externally applied bias;
   (b) Update the weights according to
   $$\mathbf{w}(t+1) = \mathbf{w}(t) + \eta(y' - y)\mathbf{x}(t)$$
   where the learning rate η is a constant value between 0 and 1, and y' is the desired output.

The perceptron can be considered as the simplest kind of feed-forward neural network. This model can be generalised by simply replacing the activation function f with a more general nonlinear function. However, the perceptron can only deal with linearly separable patterns as long as a monotonic activation function is used. In 1969, Minsky and Papert [86] demonstrated that a single-layer perceptron was incapable of representing a linearly inseparable function such as the "exclusive or" (XOR). The postperceptron era began with the realisation that adding (hidden) layers to the network could yield significant computational versatility. This stimulated a revival of interest in ANNs especially for multilayered feed-forward structures.

One important type of ANN for context-aware sensing is the *Self-Organising Map* (SOM). SOM is a class of unsupervised competitive neural models with an organised geometrical structure of output neurons. It can be considered as a nonlinear projection of a probability density function $p(\mathbf{x})$ of a high dimensional input onto a discrete, usually two-dimensional output space. In addition to the advantages inherent in ANNs, SOM provides an efficient way of data visualisation and clustering.

SOM can be viewed as a regular lattice of neurons with different areas of the map tuned to different activity patterns as shown in Fig. 9.2. In this figure, the weight vector $\mathbf{w}_j$ associated with each neuron j has an equal dimension to the input vector and is typically initialised with random values. The SOM training algorithm updates the winning node as well as nodes in its topological vicinity. The update rule is formulated so that the node with its weight vector nearest to the input data wins the competition. The most common criterion is based on maximising the inner product $\mathbf{w}_j^T\mathbf{x}$, which is equivalent to minimising the Euclidean distance between the two vectors:

$$i(\mathbf{x}) = \text{argmin}_j \left\| \mathbf{x} - \mathbf{w}_j \right\|, \quad j = 1, 2, \ldots, l \tag{9.2}$$

where $i(\mathbf{x})$ is the winning unit activated by the input vector $\mathbf{x}$ and l is the total number of neurons in the network. The weighting vector of the winning neuron and its neighbours are updated according to the following iterative equation:

$$\begin{aligned} \mathbf{w}_j(t+1) &= \mathbf{w}_j(t) + \Delta\mathbf{w}_j \\ &= \mathbf{w}_j(t) + \eta(t) h_{j,i(\mathbf{x})}(t)\left(\mathbf{x} - \mathbf{w}_j(t)\right) \end{aligned} \tag{9.3}$$

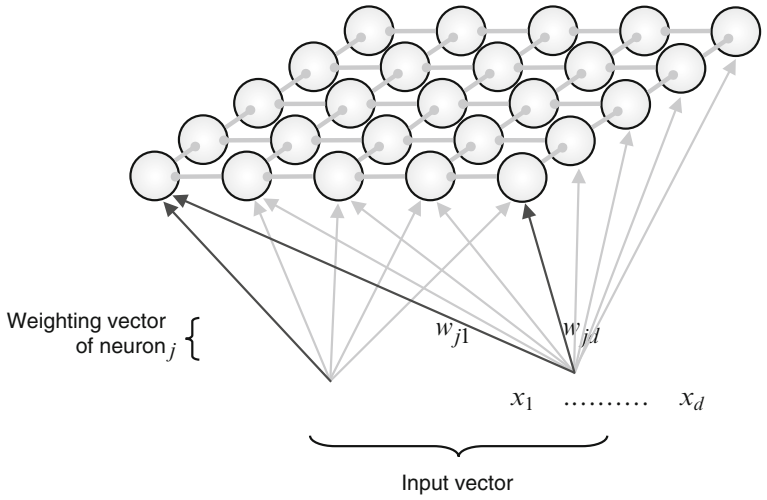


Fig. 9.2 The basic structure of a standard SOM

where $\eta(t)$ is the learning rate and $h_{j,i(\mathbf{x})}(t)$ a neighbourhood function whose value depends on the distance between node j and the wining node $i(\mathbf{x})$. In this way, similar inputs will activate neurons that are close to each other on the SOM map.

A common choice of the neighbourhood function $h_{j,i(\mathbf{x})}(t)$ is a Gaussian function. It has been found that the SOM algorithm converges more quickly with a Gaussian neighbourhood function. The underlying assumptions for the Gaussian neighbourhood function are: (*a*) it is symmetric about the winning node; (*b*) it decreases monotonically in amplitude with increasing lateral distance $d_{j,i}$ and decays to zero as $d_{j,i} \rightarrow \infty$ (a necessary condition for convergence); and (*c*) it is independent of the location of the winning neuron, i.e., it is translational invariant.

In many applications, the quality of the SOM solutions can be improved by using a time-varying neighbourhood function [87]. A time-varying form of the Gaussian function can be described as:

$$h_{j,i(\mathbf{x})}(t) = \exp\left(\frac{d_{j,i}^2}{2\sigma^2(t)}\right) \tag{9.4}$$

where t is the time step index used in training and $\sigma(t)$ is the extent of the neighbourhood. Both the learning parameter $\eta(t)$ and the effective width $\sigma(t)$ should be gradually decreasing over time.

In practical implementations, SOM learning often consists of two different phases of the operation, called ordering and converging phases respectively. The ordering phase involves approximately 1,000 iterations with the learning rate near $\eta(t)$ unity. It is not crucial whether the learning rate decreases linearly, exponentially or inversely proportional to time t. However, after the ordering

Table 9.5 The SOM learning algorithm

1. Initialise the weight vector $\mathbf{w}_j$, learning rate and the "effective width" $\sigma(t)$ of the neighbourhood function $h_{j,i(\mathbf{x})}(t)$.
2. For each input vector, $\mathbf{x}(t)$ (t is the time step index):
   (a) Determine the winning neuron, $i(\mathbf{x})$:
   $i(\mathbf{x}) = \text{argmin}_j \|\mathbf{x} - \mathbf{w}_j\|\ j = 1, 2, \ldots, l$
   (b) Calculate the neighbourhood function:
   $h_{j,i(\mathbf{x})}(t) = \exp\left(\frac{d_{j,i}^2}{2\sigma^2(t)}\right)$
   where $d_{j,i}$ is the distance between weight vectors of node i and j.
   (c) Update the weight vectors of the winning neuron and its neighbours,
   $\mathbf{w}_j(t+1) = \mathbf{w}_j(t) + \eta(t)h_{j,i(\mathbf{x})}(t)(\mathbf{x} - \mathbf{w}_j(t))$
   (d) Reduce the "effective width" $\sigma(t)$ (ordering phase) and the learning rate $\eta(t)$.
3. Repeat step 2. until the convergence condition is satisfied, and reuse the input data if necessary.

phase, $\eta(t)$ should attain a small value (i.e. of the order of or less than 0.01), otherwise the map will lose its adaptive behaviour. The exponential decay function

$$\eta(t) = \eta_0 \exp\left(-\frac{t}{\tau_1}\right) \tag{9.5}$$

provides a way to guarantee the lower bound η_0 of the learning rate, where τ_1 is the time constant. The neighbourhood function $h_{j,i(\mathbf{x})}(t)$ should initially include almost all of the neurons so that the weights will become ordered globally. In the final convergence phase, the number of steps used should be at least 500 times the size of the network in order to achieve a good statistical accuracy [88]. The ordering phase can be omitted if the weight vector is initialised by a linear initialisation scheme. The algorithm in Table 9.5 summarises the main steps involved in the formation of a SOM.

In order to apply a SOM to classification, a class label is assigned to each neuron after convergence. For each neuron that has been activated at least once, it is labelled with the data class that has the highest number of activations for that neuron. For neurons that have not been activated by the training data, they are usually assigned with the label of their nearest neighbours.

To illustrate a simple application of SOM, a map with 100 neurons has been used for clustering different gait patterns based on three-axis accelerometer signals collected using an ear-worn sensor at a sampling rate of 50 Hz. The use of the ear-worn sensor emulates the sensory function of vestibule to sense balance, gait, as well as shock-wave transmission through the human skeleton. It offers remarkable sensitivity and consistency in sensing changes in gait patterns due to lower-limb injuries or surgery [58, 89]. In this example, accelerometer readings of the subject were recorded before and after ankle injury, and when the subject was fully recovered. Average signal energy calculated over a window of 4 s was used as input feature. Figure 9.3 shows the class-specific activation plots of the SOM with 100 neurons for acceleration data acquired from a specific subject before the injury, during the injury and after recovery. Compared to normal gait, distinctive patterns

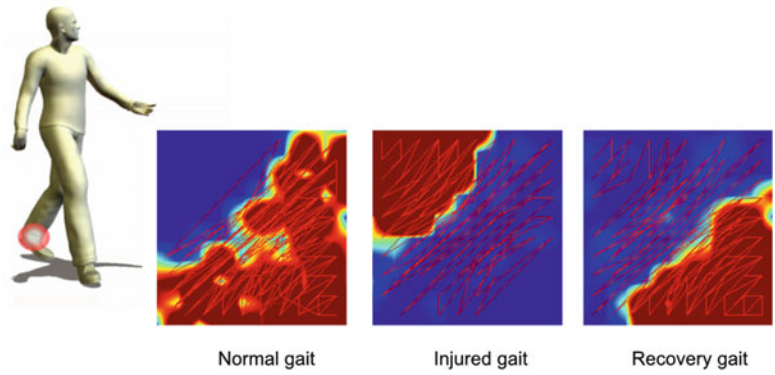


Fig. 9.3 Class-specific activation plots for of a standard SOM with 100 neurons for ankle injury data

were found when the subject was suffering from the ankle injury and different clusters are formed for the different gait patterns. After recovery, the gait patterns became more similar to the normal gait and mostly activate the neuron in the normal gait cluster.

Due to its simplicity, SOM-based architectures have been used in a range of context-aware applications [24, 60, 90–92]. The conventional SOM, however, has a number of limitations. First, it is based on the matching of a snapshot of the input attributes (or features) with the neurons, and its accuracy is influenced by feature variations. In many context-aware applications, each activity can consist of a series of submovements and the resulting activation pattern in this case is no longer restricted to a local cluster of neurons. It tends to span across a large area of the map and overlaps with neuron activations introduced by other activities. These overlaps in neuron excitation can adversely affect the overall recognition accuracy.

Although it is possible to use methods such as the short-term memory model [93] to convert the temporal variation to stable feature vectors, they can significantly increase the dimensionality of the input vector and are not effective when the sub-movements involved have a large temporal variation and poor repeatability. This problem is also compounded by the fact that activities involved in most context-aware applications can have a mixture of stable and dynamic signal features, and it is practically difficult to find low-level feature representations that are effective for both cases. To overcome these problems, existing research has been concentrated on improved feature extraction and selection methods for deriving stable feature vectors suitable for all the activities involved. The SOM in this case is mainly used as a simple classifier at the final stage of the processing steps.

Another problem associated with SOMs is the fact that the neuron activation pattern of the trained map can be highly dependent on the distribution of the training data [94]. If a particular region of the input space contains more frequently occurring stimuli, it will be represented by a larger area of the SOM, and therefore introducing a bias depending on the number of records per class in the training data. Due to the compounding effect of mixing dynamic and static excitations

Table 9.6 The STSOM model learning algorithm

Model learning:

1. Train the static map with the standard SOM training algorithm.
2. Assign the class label to each neuron by:
   (a) Applying the static map on the training set and keep record of activation frequency of each neuron;
   (b) Pruning out the labels of neurons with activation frequency lower than a specified threshold;
   (c) Assigning a label to an unlabelled node with the label of the nearest labelled neighbour.
3. Form sub-clusters of highly confused classes by:
   (a) Applying the static map on the training set;
   (b) Calculating the confusion matrix;
   (c) Creating a list of between-class distances and keep only the elements that have values that are greater than a specified threshold;
   (d) Performing single link clustering based on the distance list;
   (e) Representing each independent spanning tree as a sub-cluster of a confused class.
4. If the distance list is empty, relabel the static map by repeating step 2(a) and 2(c), and output the map and terminate. Otherwise, calculate the index entropy of the classes in the confused subclusters.
5. Extract data samples for dynamic map training
   (a) Partition the data of the confused classes using the index entropy calculated over a fixed window Ω_e*;*
   (b) Based on the number of supporting data decide if a confused class is static or dynamic.
6. Perform feature extraction on the outputs of the static map for the samples that correspond to the dynamic classes and use them to construct the dynamic map.
7. For each subcluster of confused static classes, create a higher layer static map; allocate an integer array to store the class-to-map index.
8. Keep a record of the labelled maps, entropy threshold, window size, features used, and class-to-map index for model inference.

mentioned above, the class discriminability is difficult to control and interclass misclassification (confusion) is inevitable.

To address these issues, *Spatio-Temporal Self-Organising Maps* (STSOMs) [95] have been proposed for integrating temporal excitation pattern and adaptive class-separation into a hierarchical SOM model. The key idea of the method is to rely on class-specific neuron activation patterns and the introduction of an additional temporal layer of the SOM to provide improved class separation. It also incorporates a divide-and-conquer multi-resolution classification scheme to adaptively to remove inter-class overlaps.

The prerequisite of a STSOM is the introduction of both static and dynamic classes of neuron activation. It is worth noting that this should not be confused with the static and dynamic activities mentioned in context-aware sensing. An activity that is dynamic in the physical space can be associated with a static neuron activation given appropriate feature representation. In a STSOM, we consider classes that continuously activate (or in other words, the activation is fixated onto) the same neurons of the map as *static classes*. Other classes that involve activation patterns moving across the map are called *dynamic classes*. Summaries of the STSOM learning and inferencing algorithms are provided in Tables 9.6 and 9.7.

Table 9.7 The STSOM inferencing algorithm

Model inference:
1. For each input vector, $\mathbf{x}_s(t)$ (t is the time step index), determine the winning neuron, $i_s(t)$ of the static map s.
2. Calculate the index entropy over a fixed window Ω_e.
3. If the entropy is higher than a specified threshold,
   (a) Calculate input vector $\mathbf{x}_d(t)$ for the dynamic map d;
   (b) Determine the winning neuron, $i_d(t)$;
   (c) Output the label of the neuron $i_d(t)$
   Otherwise,
   (a) Use the label of the neuron $i_s(t)$ and the class-to-map index to determine the appropriate static map:
   h = class_to_map[label($i_s(t)$)].
   (b) If map h is the same as map s, output the label of the neuron $i_s(t)$, otherwise
   Based on the input vector $i_s(t)$, determine the winning neuron, $i_h(t)$ of the static map h;
   Output the label of the neuron $i_h(t)$.

To illustrate the practical use of the STSOM for context-aware sensing, the method is applied to the same experiment described in Fig. 8.5 of Chap. 8. It features a simple physical exercise sensing experiment where four two-axis accelerometers are placed on the left and right ankles and legs. The activities of the subject during the exercise routine include (1) *sitting* (*chair*), (2) *standing*, (3) *steps*, (4) *sitting* (*floor*), (5) *demi-plié*, (6) *galloping left*, (7) *skipping*, (8) *galloping right*, (9) *side kick*, (10) *front kick* and (11) *walking*. From Fig. 8.5, it is evident that the decision boundaries for the 11 activities involved are highly complex, particularly for some of the dynamic activities involved. Figure 9.4 re-plots the sensor signals collected for this experiment, showing the moving signal energy calculated for the eight sensory channels involved.

For this experiment, the static map of the STSOM involves 100 neurons and the total number of neurons we used after the introduction of the dynamic map and node expansion was 164. The input vector consisted of the raw signal and signal energy calculated over a fixed window of 50 samples (2 s) for each sensor channel.

Figure 9.5 illustrates the class-specific activation plots of the static and dynamic maps of the STSOM for both the training and test data. In order to assess the improvement in model accuracy through the STSOM algorithm, each static map of the STSOM was built from the standard SOM. This enabled a fair comparison since both maps shared the same weight vectors, and so the effect of local minima in different model training was avoided. To compare the performance of the standard SOM and the proposed STSOM, the experiment was repeated 50 times (Fig. 9.6).

9.4.2 Hidden Markov Models (HMMs)

A HMM consists of a finite set of *states*, and the transition between states is governed by a set of probabilities called *transition probabilities*. In a particular

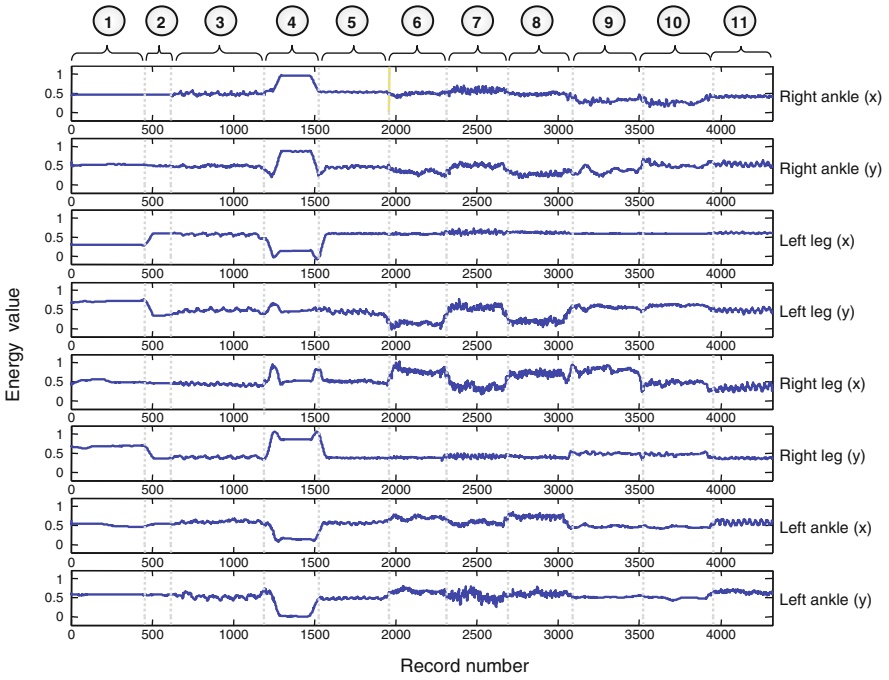


Fig. 9.4 A time series plot of moving signal energy calculated for the eight sensory channels involved for the STSOM example for Fig 9.5

state, an outcome or *observation* can be generated according to the associated probability distribution. It is only the outcome, not the state itself that is visible, hence the name *Hidden Markov Models* because the states are *hidden* from the external world. The basic theory of HMM was published in a series of classic papers by Baum and his colleagues [96] in the late 1960s and since then it has been applied to a wide range of speech processing applications [97, 98]. The key benefit of a HMM is its ability to model temporal statistics of the data by introducing a discrete hidden variable that undergoes a transition from one time step to the next according to a stochastic transition matrix. At each time step, the HMM emits symbols that are dependent on the current state of the hidden variable.

In order to understand HMMs, it is necessary to describe the basic principles of Markov chains. A Markov chain, or first-order Markov model, is a discrete-time stochastic process with a deterministic output function. By describing the evolution of states based on the Markov property (i.e. the probability distribution of the current state depends only on the intermediate previous state and the associated action), it provides a compact representation of all possible paths through the state space. A Markov chain can be described with a triple $\Theta = (Q, \mathbf{A}, \boldsymbol{\pi})$, where Q is a finite set of K states, $\mathbf{A}$ is a matrix of $K \times K$ transition probabilities between the states, and $\boldsymbol{\pi} = \{p(\mathbf{q}_i)\}_{i=1}^{K}$ is a prior probability distribution over the states Q indicating the likelihood of a state $\mathbf{c}$ being the starting point. The prior distribution can sometimes

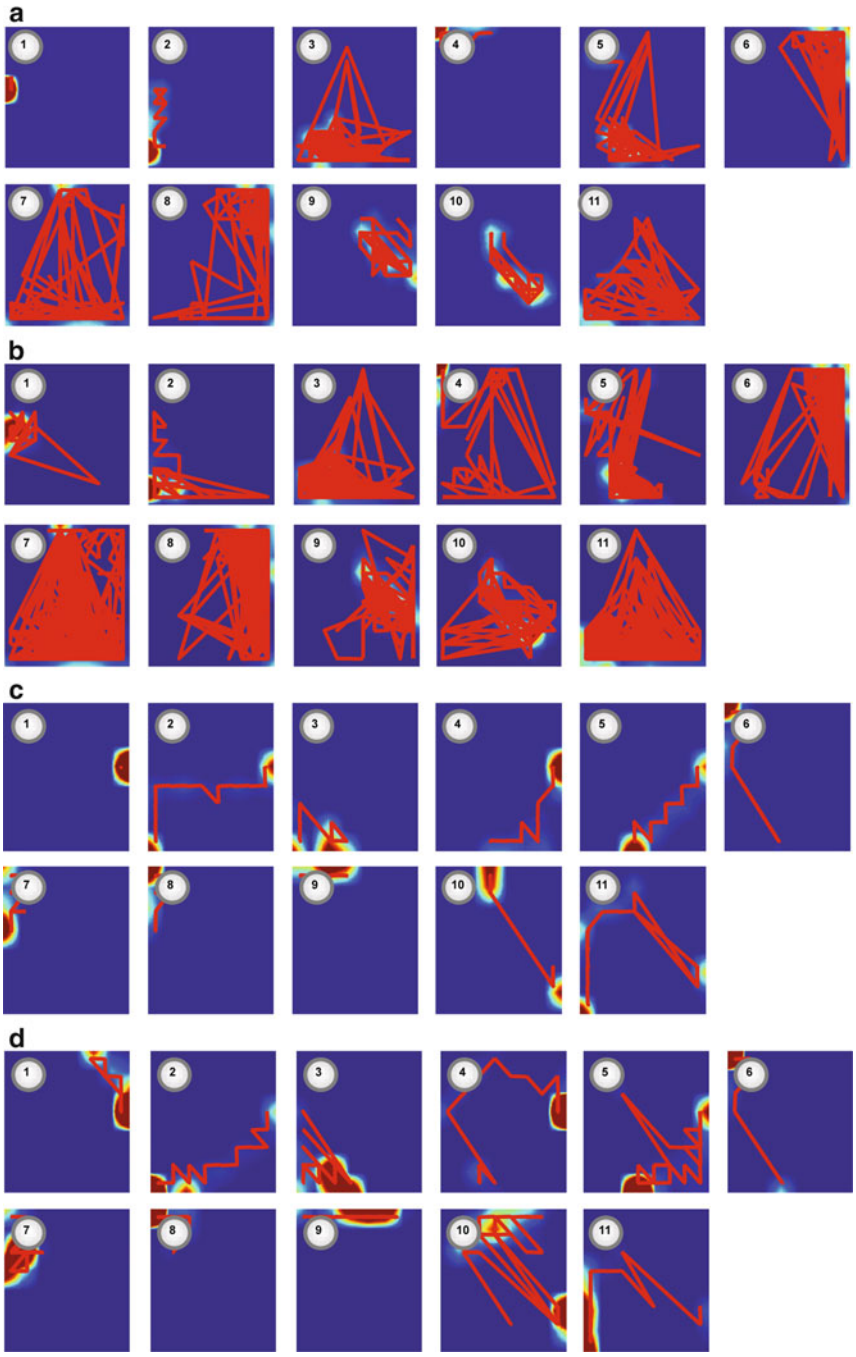


Fig. 9.5 Class-specific activation plots of: (**a**) the static map with training data; (**b**) the static map with test data; (**c**) the dynamic map with training data and (**d**) the dynamic map with test data

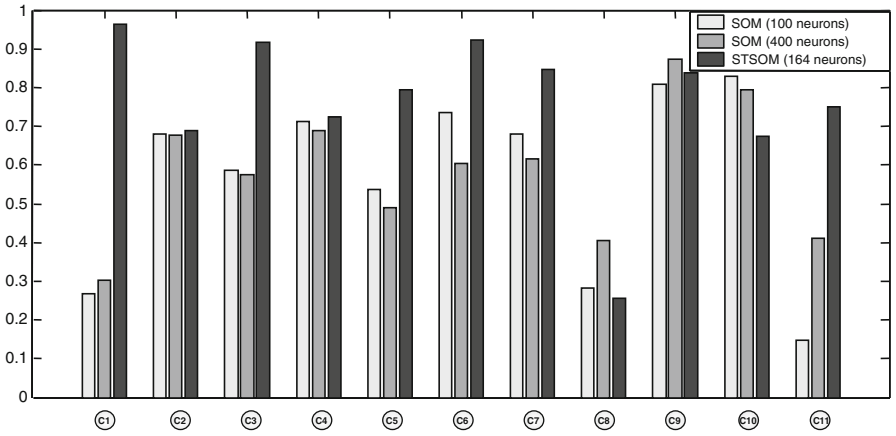


Fig. 9.6 A comparison of the recognition accuracy between a standard SOM with 100 neurons, a standard SOM with 400 neurons, and a STSOM with 164 neurons for the experiment shown in Fig. 9.4

be replaced by the non-emitting (entry and exit) state. In general, the probability of any state sequence $\mathbf{Q}_T = [\mathbf{q}_1, \mathbf{q}_2, \ldots \mathbf{q}_T]$ can be defined as:

$$p(\mathbf{Q}_T|\Theta) = p(\mathbf{q}_1|\Theta)\prod_{t=2}^{T} p(\mathbf{q}_t|\mathbf{q}_{t-1}, \mathbf{q}_{t-2}, \ldots, \mathbf{q}_1, \Theta) \tag{9.6}$$

Based on the first-order Markov assumption, the joint probability $p(\mathbf{Q}_T|\ \Theta)$ can be redefined as a product of the probability of the first state and the probabilities of subsequent transitions in the sequence, i.e.,

$$(\mathbf{Q}_T|\Theta) = p(\mathbf{q}_1|\Theta)\prod_{t=2}^{T} p(\mathbf{q}_t|\mathbf{q}_{t-1}, \Theta) \tag{9.7}$$

where $p(\mathbf{q}_1 \mid \Theta)$ is the initial probability of state $\mathbf{q}_1$ and $p(\mathbf{q}_t|\mathbf{q}_{t-1}, \Theta)$ is the probability of a transition from state $\mathbf{q}_{t-1}$ to $\mathbf{q}_t$. The maximum likelihood estimation of the transition probabilities is defined as:

$$p(\mathbf{q}_t|\mathbf{q}_{t-1}) = \frac{\text{Number of transitions from state } \mathbf{q}_{t-1} \text{ to } \mathbf{q}_t}{\text{Number of transitions from state } \mathbf{q}_{t-1}} \tag{9.8}$$

In context-aware sensing, Markov chains are often used in the supervising layer to extract information about context transitions [24, 25].

The HMM is an extension of the Markov model in which the observation itself is described by a probabilistic output function, but as mentioned earlier, the state sequence in this case is hidden. Each HMM approximates the likelihood that the model generates the observed data based on the assumptions that: (*a*) the signal is stationary over a frame; (*b*) current observations are statistically independent of the previous outputs (observations) and (*c*) transition probabilities are constant

(i.e. independent of observations or previously visited states). A HMM can be defined as a pentuple $\Theta = (Q, X, \mathbf{A}, \mathbf{B}, \boldsymbol{\pi})$, where Q and X are sets of K (hidden) states and L output symbols, respectively. In this definition, $\mathbf{A}$ represents the state transition probabilities, $\mathbf{B}$ is a $K \times L$ output matrix containing the probabilities of emitting observation $\mathbf{x} \in X$ while in state $\mathbf{q} \in Q$, and $\boldsymbol{\pi}$ represents the initial state distribution vector or indicates the non-emitting states. Given an observation sequence $\mathbf{X}_T = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T]$ and model Θ, context recognition can be achieved by determining the best state sequence $\mathbf{Q}_T^* \in Q$ which maximises $p(\mathbf{X}_T|\mathbf{Q}_T, \Theta)$ or equivalently:

$$\mathbf{Q}_T^* = \text{argmax}_{\mathbf{Q}_T \in Q} p(\mathbf{X}_T|\mathbf{Q}_T, \Theta) p(\mathbf{Q}_T|\Theta) \tag{9.9}$$

Since the realisation of $\mathbf{x}_t$ is assumed to be independent of the neighbouring states, the first term of Eq. 9.9 can be simplified as multiplications of the output density, i.e.,

$$p(\mathbf{X}_T|\mathbf{Q}_T, \Theta) = \prod_{t=1}^{T} p(\mathbf{x}_t|\mathbf{q}_t) \tag{9.10}$$

The second term of Eq. 9.9 is used to model the contextual information among states, which can be derived from a first- or higher-order Markov model. The probability of observing a HMM output string $\mathbf{X}_T$ is given by summing the contribution from all the possible state sequences $\mathbf{Q}_T \in Q$ such that:

$$p(\mathbf{X}_T|\Theta) = \sum_{\mathbf{Q}_T \in Q} p(\mathbf{q}_1|\Theta) p(\mathbf{x}_1|\mathbf{q}_1, \Theta) \prod_{t=2}^{T} p(\mathbf{q}_t|\mathbf{q}_{t-1}, \Theta) p(\mathbf{x}_t|\mathbf{q}_t, \Theta) \tag{9.11}$$

This can be efficiently computed by recursively applying the probabilities of the partial state sequences and observations. The most important criterion used for estimating the HMM parameters is to maximise the likelihood of generating the training data in Eq. 9.11. The Viterbi algorithm and Baum-Welch re-estimation method are two of the most commonly used HMM training algorithms. The Viterbi method is a dynamic programming algorithm with which the total likelihood is estimated by the probability of the most likely state sequence $\mathbf{Q}_T^*$, i.e.,

$$p(\mathbf{X}_T|\Theta) \approx p(\mathbf{X}_T, \mathbf{Q}_T^*|\Theta) \tag{9.12}$$

From this, each observation vector can be assigned to exactly one emitting state, and the parameters of each output distribution can be estimated independently based the data segment(s) associated with the state. The Baum-Welch re-estimation algorithm, on the other hand, is based on the popular *Expectation Maximisation* (EM) method. At the E-step, soft alignment is made by estimating the probability of state occupation, whereas at the M-step, the transition probabilities and the output distribution parameters are re-estimated by using the probability of state occupation. At each iteration, an increase in likelihood is guaranteed.

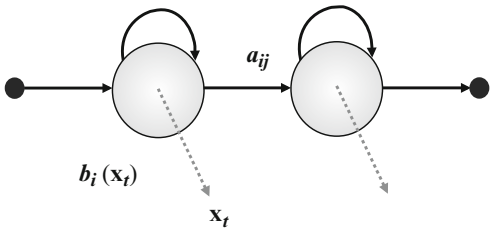


Fig. 9.7 A standard HMM as a finite state machine where the shaded nodes denote an emitting state with output probability $b_i(\mathbf{x}_t)$ and the *small dark nodes* denote nonemitting (entry and exit) states. In this figure, *arcs* represent state transitions with probability a_{ij}

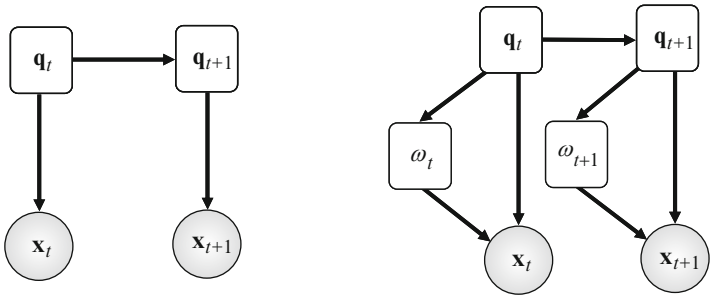


Fig. 9.8 Bayesian network representations of HMMs, where *circles* denote continuous nodes or variables, and *squares* denote discrete nodes. *Unshaded* nodes are hidden nodes

A common way to view a HMM is to regard it as a finite state machine as shown in Fig. 9.7. A HMM can also be viewed as a simple version of a *Dynamic Bayesian Network* (DBN) with one discrete (unobserved) hidden node and one discrete or continuous observed node per time slice, as illustrated in Fig. 9.8. This model provides a compact representation of the joint probability distribution and reveals the underlying independence assumptions among variables in the graph.

To model the output probability distribution for each state in a continuous density HMM, it is common to use a Gaussian distribution with diagonal covariance to reduce the number of parameters required. The distribution of the real-world signal, however, can be non-Gaussian. To overcome this problem, the output distribution can be estimated by using a linear mixture of different models such as the popular mixture of Gaussians represented by the following weighted function:

$$b_j(\mathbf{x}_t) = \sum_{m=1}^{M} \omega_{jm} b_{jm}(\mathbf{x}_t) = \sum_{m=1}^{M} \omega_{jm} N\left(\mathbf{x}_t; \boldsymbol{\mu}_t, \Sigma_{jm}\right) \tag{9.13}$$

where M is the number of Gaussian components in a state, and ω_{jm} is the component weight or prior that is summed to 1.

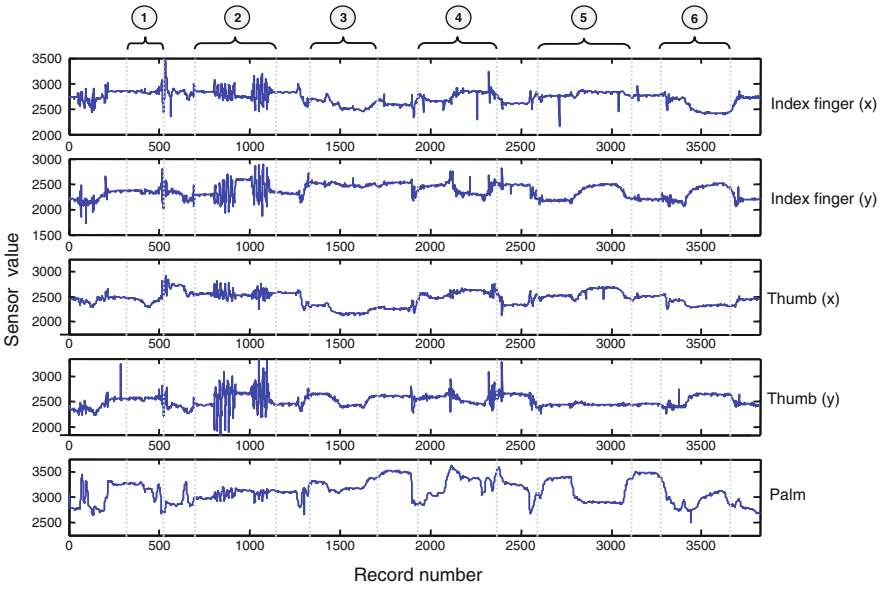


Fig. 9.9 Time series plot of a signal sequence obtained from the sensor glove. The six activities marked represent opening the door, turning on/off the tap, opening/closing the cupboard, making coffee, adding milk and drinking coffee, respectively

Under the Bayesian network representation of HMMs, a general mixture distribution for $b_j(\mathbf{x}_t) = p(\mathbf{x}_t|\mathbf{q}_t = j, \Theta)$ assumes the existence of a hidden variable ω that determines the active mixture component. It follows that:

$$\begin{aligned} p\left(\mathbf{x}_t\middle|\mathbf{q}_t = j, \Theta\right) &= \sum_{m=1}^{M} p\left(\mathbf{x}_t, \omega = m\middle|\mathbf{q}_t = j, \Theta\right) \\ &= \sum_{m=1}^{M} p\left(\omega = m\middle|\mathbf{q}_t = j, \Theta\right) p(\mathbf{x}_t|\omega = m|\mathbf{q}_t = j, \Theta) \end{aligned} \tag{9.14}$$

To illustrate how HMMs can be used for context detection, we present in Fig. 9.9 a simple example of activity detection through the use of a sensor glove mounted with an optical bending sensor and two accelerometers sampled at 50 Hz [99]. The optical bending sensor was placed across the palm, whereas the two accelerometers were positioned on the back of the index finger and thumb, respectively. The dataset consists of six different activities including *opening the door*, *turning on/off the tap*, *opening/closing the cupboard*, *making coffee*, *adding milk* and *drinking coffee*. In this simple example, we used six three-state HMMs with an ergodic fully-connected HMM topology. Baum-Welch re-estimation was used for model training and the output distribution associated with each state is modelled by a simple Gaussian distribution with diagonal covariance matrix so as to keep the number of parameters as small as possible. A total of seven datasets were acquired from the subject; one of which was used for model training and the remaining six were used for evaluating the accuracy of the algorithm.

Table 9.8 HMM classification results for the experiment shown in Fig. 9.9

Class	Rank 1 accuracy (%)	Rank 2 accuracy (%)	Rank 3 accuracy (%)
C1	100	100	100
C2	100	100	100
C3	83.33	83.33	100
C4	33.33	50	100
C5	100	100	100
C6	83.33	100	100
Average	83.33	88.89	100

Since the range of each sensor channel can vary, the overall mean and standard deviation were used for data normalisation for each dataset so that all data shares the same mean and unit variance. Simple noise filtering was applied to the data, and for each channel we also calculated signal energy over a fixed window size of 50 samples. The raw signal and energy were concatenated to form a ten-dimensional vector for the five sensing channels involved. Figure 9.9 illustrates the six different activity segments captured, and Table 9.8 summarises the overall accuracy of the HMM algorithm for the six test datasets used. In this table, rank n accuracy means the correct classification is among the first n highest likelihood models.

Albeit being overly simplistic, the above example demonstrates some of the advantages of HMMs. The method has a sound statistical grounding and its parameter estimation can take into account different sources of uncertainty. Furthermore, it is modular and can be combined into larger models. In general, HMMs are relatively robust with regards to temporal changes and it is also possible to incorporate high-level domain knowledge. The disadvantages of the method, however, include the relatively strong assumptions made about the data and the amount of training data required due to the large number of parameters involved. Another issue related to HMMs are that their training involves maximising the observed probabilities for examples belonging to a certain class but it does not minimise the probability of observation of instances from other classes. In terms of performance, an HMM involves enumerating of all possible paths through the model. Although the search can be efficiently performed by using the token passing algorithm [100], it can be computationally expensive compared to other techniques. Despite these problems, HMMs remain an attractive technique for context-aware sensing.

9.4.3 Factor Graphs (FGs)

A FG is a bipartite graph that represents the global function $g(x_1, \ldots, x_n)$ as a product of J local functions $f_j(X_j)$, which map a subset X_j of $\{x_1, \ldots, x_n\}$ to some range. This graph represents the factorisation:

$$g(x_1, \ldots, x_n) = \prod_{j \in J} f_j(X_j) \tag{9.15}$$

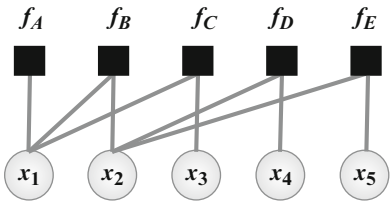


Fig. 9.10 An FG for the product $f_A(x_1)f_B(x_1, x_2)$ $f_c(x_1, x_3)f_D(x_2, x_4)f_E(x_2, x_5)$

In general, a FG consists of two types of nodes: variable nodes and function (or factor) nodes. The presence of an arc from a variable node to a factor node indicates that the variable is an argument of the function. As a simple example, Fig. 9.10 illustrates a FG representation of a function of five variables, expressed as a product of five factors:

$$g(x_1, x_2, x_3, x_4, x_5) = f_A(x_1)f_B(x_1, x_2)f_c(x_1, x_3)f_D(x_2, x_4)f_E(x_2, x_5).$$

Associated with function $g(x_1, \ldots, x_n)$ are n marginal functions $g_i(x_i)$, or the summary for x_i of g, can be obtained using the '*not-sum*' or the summary notation [101]:

$$g_i(x_i) = \sum_{\sim\{x_i\}} g(x_1, \ldots, x_n) \tag{9.16}$$

i.e., summing of the value of $g(x_1,\ldots,x_n)$ over all the configurations (all the possible values) of the variables that have a specific value for x_i.

The sum-product algorithm [101] provides an efficient way to compute all the marginal functions in the factor graph. The algorithm operates according to the following simple rule:

> The message sent from a node v on an edge e is the product of the local function at v (or the unit function if v is a variable node) with all messages received at v on edges other than e, summarised for the variable associated with e

According to this rule, a message from a variable node to a factor node can be calculated by using the following equation:

$$\mu_{x\to f(x)} = \prod_{h\in n(x)\setminus\{f\}} \mu_{h\to x}(x) \tag{9.17}$$

and a message from a factor node to a variable node can be calculated by:

$$\mu_{f(x)\to x}(x) = f(x) \prod_{y\in n(f)\setminus\{x\}} \mu_{y\to f}(y) \tag{9.18}$$

where $n(v)$ denotes the set of neighbours of node v. The algorithm terminates once two messages have been passed in both directions of each edge. The marginal function $g_i(x_i)$ can be computed as the product of all incoming messages of x_i, or as the product of an incoming and an outgoing message that were passed over an incident edge of x_i.

FGs are a generalisation of the '*Tanner graphs*' and have been widely used in coding/decoding problems. In 1981, Tanner [102] introduced bipartite graphs to describe some families of code. In the original formulation, all the variables are visible. Wiberg et al. [103] subsequently introduced hidden state variables into the graph and Kschischang et al. [101] generalised it further by applying them to functions. In behavioural modelling of systems, FGs can be used to represent valid configurations of variables (characteristic function of the system). In probabilistic modelling, FGs are used to represent joint probability mass function of the variable in the system. Compared to MRFs and BNs, FGs provide a more general framework for representing the relationships among variables in the network, i.e., they can describe general functions other than probability distribution. It has been shown in [101] that MRFs and BNs can be converted to a FG without information loss. The Pearl's belief propagation algorithm is basically an example of the sum-product algorithm operating in a BN.

To address link and node failures combined with resource constraints and asynchrony, Chu et al. [104] illustrated the use of FG representation and evolution mechanisms for *Multiple Target Tracking* for sensor networks which is concerned with estimating multiple target trajectories given noisy observations of the target states. The progression of a target entering, moving through, and leaving a sensor field is represented by model evolution, which consists of the following three mechanisms: spawning (hypothesising the existence of new phenomenon), updating (of both distribution and topological changes) and pruning (resolving inconsistencies between expected observations and measured observations by pruning the variable nodes associated with the target from the representation). Such an approach can be particularly useful for integrating BSNs with ambient sensing environments due to the highly mobile nature of the BSNs in entering, traversing through and leaving the ambient sensing environments.

Figure 9.11 illustrates a simple decision tree for detecting the presence of a user context, i.e., the BN model of the problem, the FG representation and the model parameters, respectively. The variables F, S, D, C, E represent values of the three sensors, the presence of the context, and an event detected by F and S.

The example multiply-connected model in Fig. 9.12 describes a cancer diagnosis model developed by Greg Cooper [105]. In this model, it is suggested that the metastatic cancer (A) can cause an increase in total serum calcium (B) and is a cause of brain tumour (C). A brain tumour or an increase in total serum calcium can cause a patient to fall into a coma (D), and the brain tumour can cause papilledema (E). All of the nodes represent binary variables indicating the presence/absence of the event. In the BN model in Fig. 9.12a, Nodes A, B, C and D form a cyclic structure. In the exact inferencing scheme, the model was first transformed into the junction-tree representation and then the corresponding FG representation in Figs. 9.12b and 9.12d, respectively.

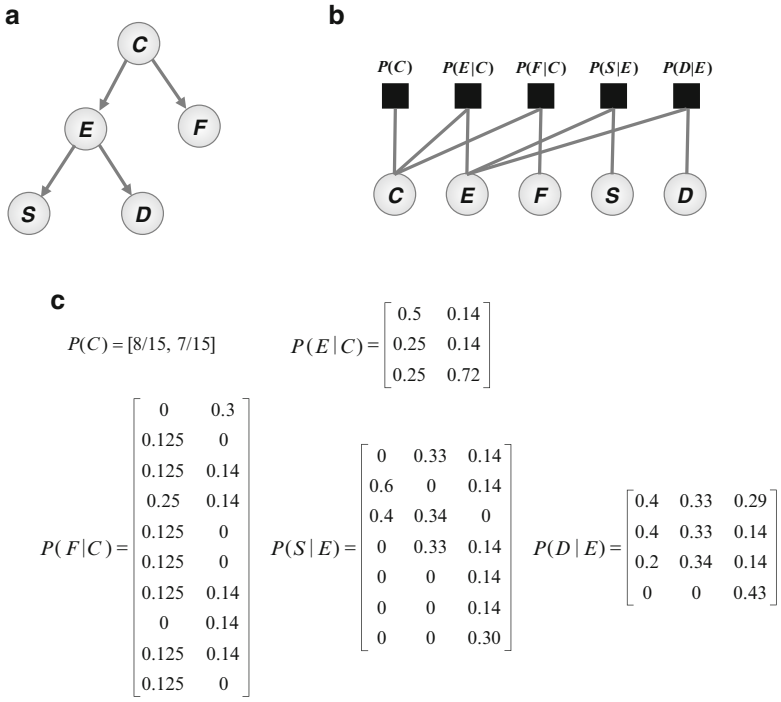
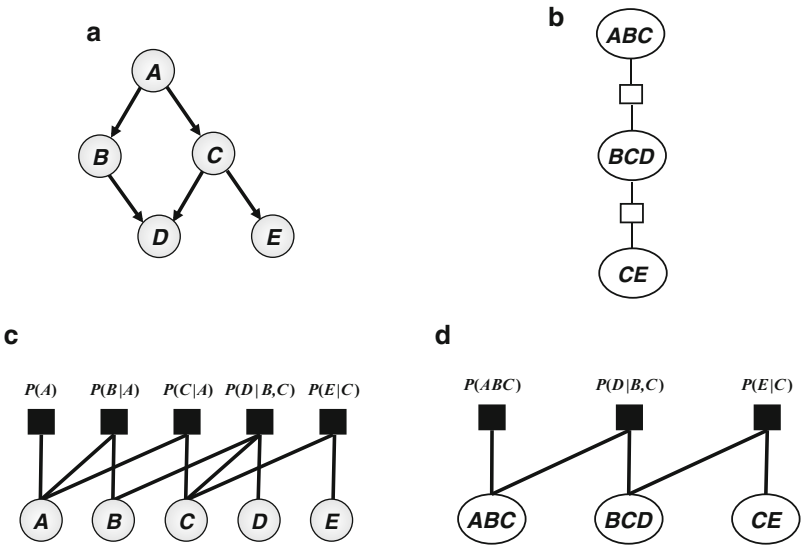


Fig. 9.11 A simple decision tree for detecting the presence of a user context: (**a**) a BN representation, (**b**) a FG representation and (**c**) the prior and conditional probability distributions

9.4.4 Other Techniques

Other techniques for pattern classification include the decision tree method, which is performed by an iterative selection of individual features that are most salient at each node. It therefore implicitly incorporates feature selection during the classification process. The feature selection criteria used include Fisher's criterion, node purity and the information content. Popular methods in this category include the CART and C4.5 algorithms. Their numerical implementations are both available in the public domain [106, 107]. The main advantage of the method is its speed and the possibility of interpreting the decision rules for each individual feature. In the study by Bao and Intille [19], several classifiers were compared and C4.5 was shown the best in recognising everyday activities based on user-annotated acceleration data.

The use of SVMs has also attracted significant research interests for pattern classification. The original idea of SVM was based on Vapnik's method of finding an optimal hyper-plane for dividing two classes, which does not depend on a probability estimation [108, 109]. This optimal hyper-plane is a linear decision boundary that separates the two classes and leaves the largest margin between the vectors of the two classes. He demonstrated that the optimal hyper-plane is



e

$$P(A) = [0.2,\ 0.8] \qquad P(B|A) = \begin{bmatrix} 0.8 & 0.2 \\ 0.2 & 0.8 \end{bmatrix}$$

$$P(C|A) = \begin{bmatrix} 0.2 & 0.05 \\ 0.8 & 0.95 \end{bmatrix} \qquad P(E|C) = \begin{bmatrix} 0.8 & 0.6 \\ 0.2 & 0.4 \end{bmatrix}$$

$$P(D|B,C) = \begin{bmatrix} 0.8 & 0.9 & 0.7 & 0.05 \\ 0.2 & 0.1 & 0.3 & 0.95 \end{bmatrix}$$

Fig. 9.12 A cancer diagnosis model: (**a**) a BN representation, (**b**) a junction-tree representation, (**c**) a FG representation of (**a**), (**d**) a FG representation of (**b**) and (**e**) the prior and conditional probability distributions

determined by only a small fraction of the data points, the so-called support vectors. In 1995, Cortes and Vapnik extended the method for the case of non-separable classes, and therefore made SVM a general tool for solving general classification problems [110]. One-class SVM was proposed for detecting changes in user context. A non-parametric online signal segmentation technique was achieved by training the SVM with the past data and using the misclassification rate to determine the rupture point. In the study on activity recognition by Huynh and Schiele [111], combining a SVM with multiple eigenspaces was shown to outperform the BN regardless of the amount of training data.

Another interesting technique for pattern recognition is to use a combination of multiple classifiers, from which their prediction can be combined by an extra *meta-level classifier* to produce the final classification results. Training different classifiers based on the same data can result in differences in their global performance. Each classifier may have its own region in the feature space where it performs the best. Some classification techniques even yield different results when different

model initialisation or a different order of data records is used. The idea of combining various classification models is to take the advantages of all the attempts in data learning. In order to improve the overall classification accuracy classifiers with difference in the feature set, the training dataset, the classification methods or training sessions can be combined. Ravi et al. [65] demonstrated the use of different types of meta-level classifiers, such as bagging, boosting, plurality voting and stacking, for activity recognition based on the acceleration signals. Choudhury et al. [27] provided a review of their previously proposed inferencing methods for activity recognition, namely, feature selection using adaboost and feed as input to HMMs to estimate "*the most likely activity while providing temporal smoothing*", conditional random fields (CRFs) with virtual evidence boosting to perform "*feature selection and structured prediction in an unified more efficient way*", and semi-supervised CRF to reduce effort in training data labelling.

The use of ontology and other semantic models for pervasive computing has also been increasingly popular. Min and Cho [112], for example, proposed a method for handling with multiple contexts in pervasive computing. The semantic network model is composed of a low-level contexts module that processes sensors data and sends activation signals to a high-level contexts module that incorporates temporal relation and computes activity probability which will be used in probabilistic inference models. Tanantong et al. [113] explored the use of arrhythmia indicator ontology for ECG monitoring with BSNs. A review of other situation identification techniques in pervasive sensing can be found in [114].

9.5 From Context Sensing to Behaviour Profiling

With increasing maturity of wearable sensors, their practical use for long-term monitoring is now a reality. This offers the possibility of capturing richer context patterns associated with user behaviour. Context recognition is therefore no longer limited to inferring a specific context from sensory observation over a short period of time using a single set of models. In this regard, it is necessary to consider complex behavioural patterns involving parallel or interleaved activities. This is because in normal daily life, one is unlikely to rigidly to move from one activity to another, we tend to do several things simultaneously (e.g., talking to someone over the phone while watching TV and having a beer). Furthermore, the transition from one activity to another will no longer have a clear separation. Rather, one activity is morphed gradually into another.

9.5.1 Behaviour Profiling

Changes in behaviour often provide tell-tale signs of the onset or complication of a disease. From long-term behaviour analysis, key indicators of well-being such as sleeping patterns, functional independence, social interaction and risks of isolation

may be inferred. Albeit being mostly surrogate signs, they form an important part of healthcare data analysis [115]. The goal of behaviour profiling is therefore more towards understanding, rather than just detecting, the intrinsic patterns of an individual's daily routines. To this end, probabilistic models are often used.

In [116], a survey of probabilistic models used for behaviour profiling is presented. These include HMMs and their extensions such as *Conditional Random Fields* (CRFs) and *Dynamic Bayesian Networks* (DBNs). Hierarchical extensions of HMMs mirror ethological research, where hierarchy has been shown to be underlying certain behaviour [117]. Examples of this include *Layered HMM* [118], *Factorial HMM* [119] and *Abstract Hidden Markov Model* (AHMM) [120]. A *Hidden Semi-Markov Model* (HSMM) extends HMM by incorporating an explicit model of state duration, allowing each state to emit a sequence of observations. For example, Duong et al. proposed the use of switching semi-Markov model, a two-layered extension of the HSMM, for activity recognition and abnormality detection [121]. With this approach, the bottom layer is a sequence of concatenated HSMMs representing atomic activities. The top layer, which represents a sequence of high-level activities, is a Markov sequence of switching variables which determine the parameters of HSMMs in the bottom layer. By introducing the conditional probabilities between hidden states of two or more HMMs, coupled HMMs provide a way of modelling interactions between multiple processes. Each chain has its own observation sequence and is allowed to evolve independently. The model has been successfully applied on vision-based modelling of interactive activities involving two-handed gestures [122] and a surveillance task [123]. In [124], Natarajan and Nevatia combined the concept of semi-Markov model and coupled HMMs and proposed *Coupled Hidden Semi Markov Models* (CHSMM) for activity recognition.

While behaviour modelling allows a detailed analysis of these patterns, it may be difficult to be applied to pervasive sensing scenarios, given the volume of the data and the lack of access to detailed annotations. This is because behaviour is manifested over a period of time, and therefore methods designed to analyse large datasets, such as those developed in the field of data mining, can be utilised.

One typical method is called *frequent pattern mining* [125]. Frequent pattern mining relates to searching for frequently repeating patterns in a database, where a pattern is considered to be frequent if its occurrence in the database is above a user-specified threshold. Lühr et al. used an extension of frequent pattern mining in a smart home to model temporal relationships in activities [126]. Activations on object sensors are mined to learn associations between them over a long period, and the extended mining algorithm allows these associations to span database transactions. This, for instance, allows the modelling of toilet activity through successive activations of activity sensors. A data structure for representing daily routines from mined activity data was proposed in [127]. This hierarchical data structure, called *Routine Tree*, is generated through recursive application of pattern mining on activity data, associating activity patterns with times of the day.

Routine activity is an important aspect of behaviour. The main biologically driver of daily routine is the circadian rhythm, a natural regulation of hormones determining rest and activity cycles, centred around the 24-h motion of the Earth

around the Sun. Much of pervasive sensing research in routine analysis is in the smart-home area, where ambient sensor statistics are associated with the circadian rhythm. Anomalies are detected for significantly large variations in the expected activity [68, 128]. A state-occupancy time approach was taken by Li and Parker [129], utilising a Fuzzy-ART (adaptive resonance theory) neural network supplemented with a Markov model to detect abnormal events generated by ambient sensors. The neural network categorises raw sensor data, with each category representing a state in the Markov model, which then learns the state transitions during normal behaviour. Detection of anomalous behaviour is based on state occupancy time by recording the average time spent in each state during normal behaviour. Similarly, elderly patients residing in a smart home are profiled based on room-occupancy times [130]. Behaviour is tracked by maintaining a *Gaussian Mixture Model* (GMM) of room occupancy duration at different time periods. Changes are detected at two scales. The first is at the level of "local anomaly" where an outlier detection algorithm is used to detect behaviours with unlikely timing or duration. For example, an inordinate time spent in the bedroom would be classified as an anomaly. The second is a "global anomaly" where changes in the model are tracked over longer time periods. Daily differences in behaviour are computed and an anomaly is flagged if a sudden change in 1 day exceeds a threshold after discounting for seasonal variations.

Recently, there have been extensive interests in tracking behaviour using consumer electronic devices such as smartphones. Examples of these include smartphone based activity recognition and energy expenditure estimation [131]. As devices become more powerful, the potential for continuous monitoring by apps executing in the background increases, allowing for long-term profiling of behaviour. This was explored in [132] indicating potential for clustering types of behaviour. The *Routine Tree* data structure proposed in [127] was extended by introducing algorithms to calculate distances between different routines, thereby allowing a quantitative analysis of daily routines.

A study relying more on social behaviour was conducted in [133], which used mobile-phone data including Bluetooth use, phone usage and location data. Principal components were extracted from the dataset. A weighted sum of these components approximated an individual's behaviour. The study used this measure, termed *Eigenbehaviours*, to cluster people and analyse social groups, for instance to analyse friendship and work group affiliations.

9.5.2 Transitional Activities

Another important direction of research in context sensing has been to analyse the transition between activities. Joint diseases such as arthritis, trauma or other conditions such as obesity, can impair the ability of patients to fluidly transition between activity states. For example, many elderly people face difficulties in rising from a chair [134]. As functional mobility is impaired with age, disease, disability

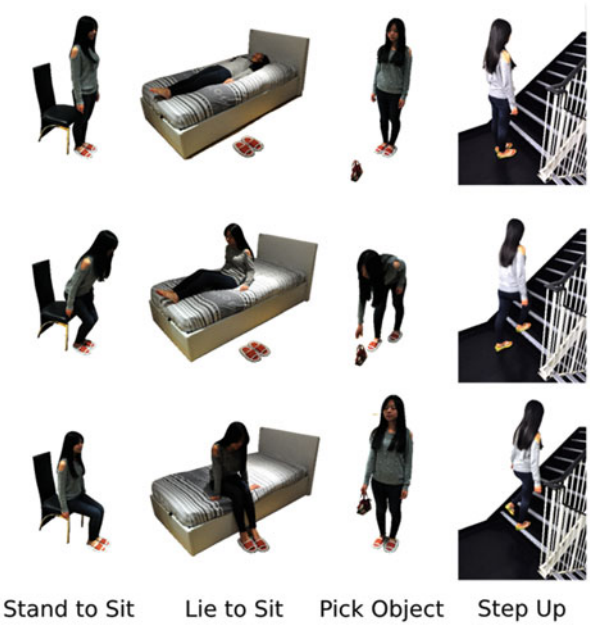


Fig. 9.13 Example transitional activities demonstrating the gradual postural change between different stable states

or injury, transitional movements such as these can become more laboured and distinct. The ability to perform transitions with ease has been suggested as a target for rehabilitation [135, 136] and as an indicator for musculoskeletal strength and motion coordination [137]. Transitions from sitting to standing have been shown to be indicative of the likelihood of falls [138, 139]. They have also been studied in association with stroke [140, 141], neuromuscular conditions [142], and found to be an indicator of difficulty in movement due to obesity [143]. Examples of some transitional activities are shown in Fig. 9.13.

While there is abundant clinical research in laboratory settings, there has been limited work in developing pervasive systems to detect transitions. As an example, a decision-tree classifier is trained to discriminate between sit-to-stand strategies based on single and multiple camera sensors [144]. Najafi et al. [145] used a gyroscope to record sit-to-stand and stand-to-sit transitions extracting the average and standard deviation of transition durations and the occurrence of abnormal successive transitions. In both of these studies, participants performed exclusively sit-to-stand transitions in controlled settings. Figure 9.14 shows signals of an accelerometer worn by a participant while performing stand-to-sit transition.

A challenge in translating laboratory research into home environments lies in detecting transitions in activities of daily living, and analysing transitions performed at patient's own home environments. It is important to note that transitions between activities in natural behaviour may not be sharp (or instantaneous) with distinctive boundaries. Instead, one activity may morph into another.

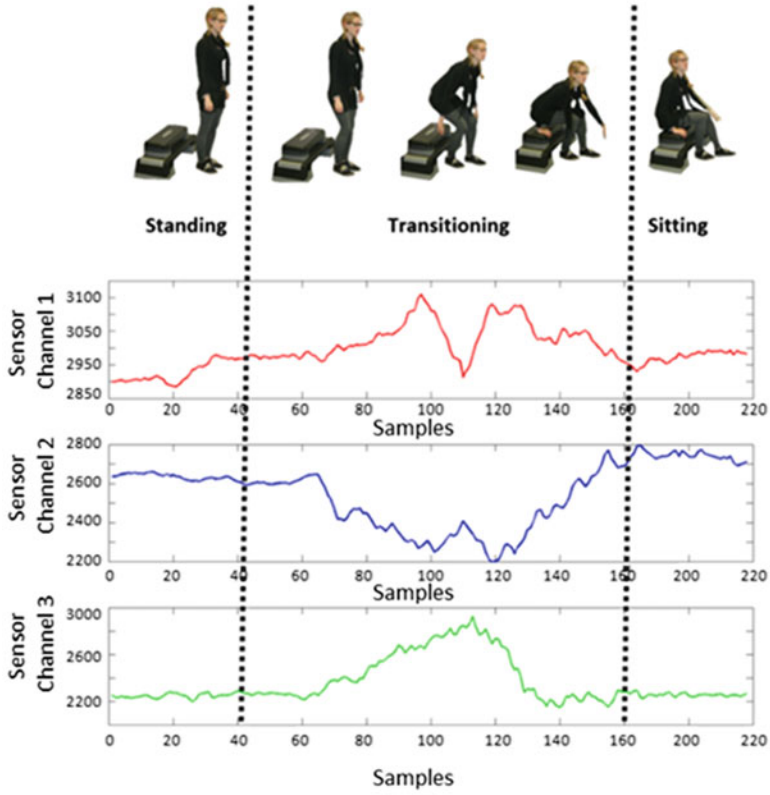


Fig. 9.14 Time series plot of a signal sequence obtained from a BSN node while the subject is performing the stand-to-sit transition (From Ref. [147])

For prolonged transitions, the marking of the boundary, particularly in real time can be subjective and error-prone. In [146], a framework for transitional activity detection and analysis in pervasive sensing contexts was proposed. Reliance on manually placed class labels is reduced by modelling the structure of the data through the geometric concept of manifolds. The neighbourhoods of a manifold can be represented as a graph, where data points correspond to vertices, edges correspond to neighbourhood relationships, and edge weights correspond to distances. Strongly connected regions of the manifold can be found by partitioning this graph. A transition is specified as when a participant moves from activity in one partition to the other. Most of the work presented in this section is from the work of Raza et al. In his thesis [147], the importance of transitional activity analysis is discussed in details, along with developing methods for mining temporal associations to activity patterns and visual quantitative analysis of user's routines. The methodologies have been adopted in a smartphone application and validated in both laboratory and free-living scenarios.

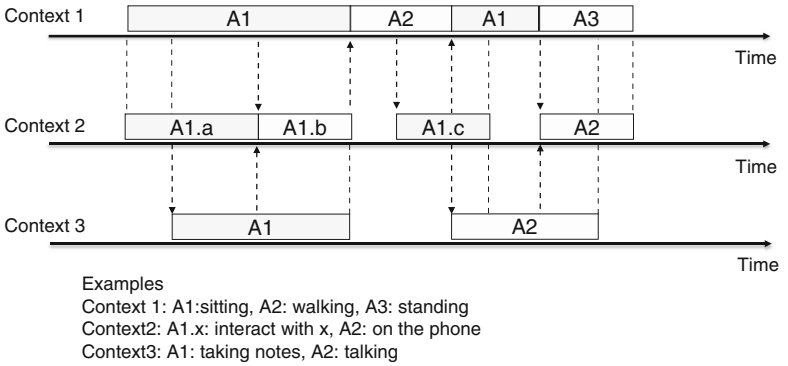


Fig. 9.15 A demonstration of different complex activity scenarios (Adapted from Ref. [116])

9.5.3 Concurrent and Interleaving Contexts

As mentioned earlier in this chapter, modelling of human behaviour may involve more than one context recognition problems. As demonstrated in Fig. 9.15, the context may occur in parallel or interleaved with other contexts, and/or comprises sub-contexts. Solving each of the context recognition tasks may require overlapped input information from shared sources of sensory signals. One complication in fusing information from multiple heterogeneous data sources is the alignment of sensor sequences. Furthermore, a piece of derived information can also affect the outcome of another context recognition process and decisions can be synchronously or asynchronously made. In general, all problems with multiple goals need to consider two important features: concurrency and interleaving. Concurrent or parallel goals are usually pursued at the same time whereas for interleaving goals, one activity can be paused and then resumed after executing some other activities for a different goal.

For dealing with concurrent and interleaving contexts, two general approaches can be used: logic-based approach and probabilistic-based approach. A logic-based approach is based on the use of human expert knowledge and provides a precise means of understanding the problem domain and defining problem specification. Furthermore, it can be implemented in a hierarchical manner, making knowledge refinement and modification very convenient. The major limitation of logic-based approaches, however, is the incapability to handle uncertainty and noise in sensor data. Probabilistic approaches, on the other hand, can provide inference under uncertainty. Bayesian network as explained in Chap. 8, for example, is a probabilistic model that can incorporate both logic and probability information.

More complex activity and behaviour modelling, including that of parallel and interleaving activities, are often achieved through extensions of HMM. Coupled HMM [122], allows the modelling of concurrent activities using more than one HMM. In this case, the model can "switch" between independently evolving Markov models. An example of a usage scenario would be when a person was talking on the phone while preparing food. In this case, the observations received by an agent would allow mapping to activity components switching between the two

activities. Without making specific assumptions about the dependence structure between observations, CRFs are capable of modelling more complex relationship than HMMs. In [148], hierarchical CRFs were deployed to extract a person's activities and significant locations from GPS data. Hu and Yang [149] proposed a two-level probabilistic framework for handling both concurrent and interleaving activities in a unified CRF model. That is, interleaving goals are modelled by skip-chain CRFs, and concurrent goals are modelled through correlation graph. Constructing a probabilistic or statistical model requires a sufficient amount of representative data for model training. Learning an accurate recognition model usually involves a laborious and time-consuming task of data collection and data annotation.

Practically, there has been extensive effort to incorporate the strengths of the two approaches. Riboni and Berttini [150], for example, used ontology to model context data and enhance the activity recognition result of a statistical classifier, such as *Multiclass Logistic Regression (MLR) with a ridge estimator*, with semantic relationships expressing the feasibility of performing an activity in a given location context. Helaoui et al. [151] proposed the use of Markov logic, a set of first-order formulae with weights, for recognizing interleave and concurrent activities. Markov logic incorporates both *hard* and *soft* logical statements. In the study, the soft formulas were used to capture the temporal relationship between individual activities. Their weights were learned by voted perceptron. In both studies, algorithm validation was performed offline on pre-collected datasets. In [152], a hybrid framework for pervasive sensing was proposed to facilitate organization of context recognition tasks into multiple dependent layers. While derivation of each piece of contextual information is achieved with a suitable statistical/numerical model, interdependencies between different levels of contexts, along with expert knowledge, can be integrated into the system using the traditional rule-based approach. In the study, *Java Expert System Shell* (JESS) was deployed as the inference engine to illustrate the framework concept. For validation, an experiment on real-time concurrent recognition of device location, orientation and user activity was conducted.

9.5.4 A Distributed Inferencing Model for Context Recognition

From a machine-learning perspective, parallel and interleaving context recognition can be viewed as a multi-label classification problem [153], where multiple output labels can be assigned to an observation. According to Tsoumakas et al. [153], mining multiple labelled data can be classified into two board categories: *problem transformation* and *algorithm adaptation*. The former involves transforming the learning tasks into one or more single-label classification tasks and is algorithm independent, whereas the latter group directly extends specific algorithms to solve the multi-label classification problem.

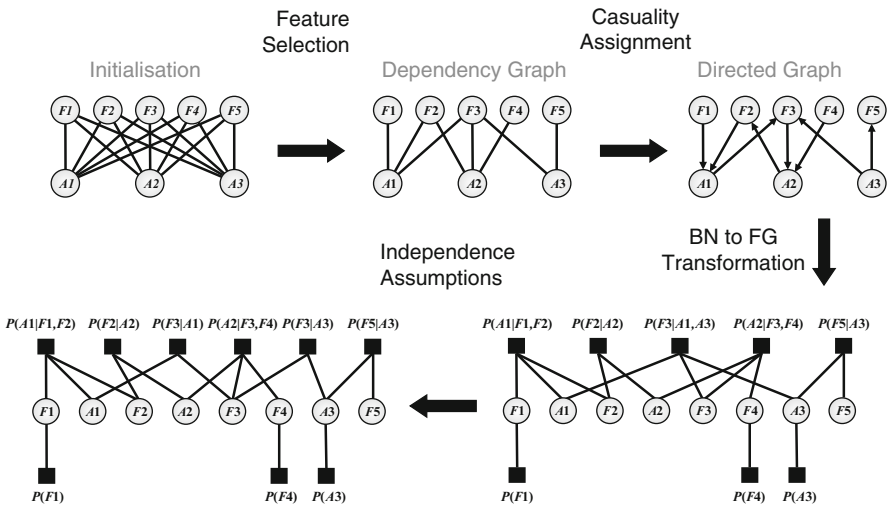


Fig. 9.16 An illustration of distributed model construction

In pervasive sensing, several context recognition problems can be formulated based on subsets of shared input sensory signals. It is also common that for a given context recognition problem, some sensors are more informative than others. Correct reasoning about a user's context may therefore rely on different sensors at different time and locations. Based on these basic concepts, an autonomic sensing framework for distributed inferencing was proposed [154, 155].

Figure 9.16 shows the key elements of the proposed framework, namely, dependency graph construction by using feature selection, causal direction assignment based on dependency analysis, creation of the FG representation, resolving inappropriate factorisation, computation of model parameters, and model inference by the use of the sum-product algorithm. In this example, a *multi-objective Bayesian Framework for Feature Selection* (multi-objective BFFS), which extends the basic algorithm in Chap. 8 and as described in [41], is used for learning the model structure. Feature redundancy and network complexity measures are used to cater for fault tolerance and minimal resource utilisation. After the dependency links between variables are obtained, causality assignment is performed. From the feature analysis perspective, a multiple parent configuration is required when there exists some correlation or interaction between the observable features. Otherwise, a more compact version of link matrix with a single parent should be used. By using the above factorisation strategy, it is also possible that a cyclic graph structure is obtained. The cyclic graph will violate the standard definition of a BN, i.e., it has to be a *Direct Acyclic Graph* (DAG). Compared to BN, FG representation is more expressive. In addition, the separation in the actual measurements and the function storages in FG facilitates the mappings of the logical model onto the physical network. The learned directed graph is therefore transformed into FG graph, after which certain assumptions can be asserted for more efficient use of

computational/network resources. After obtaining the information on what variables should be included in the model and how the joint probability of the variables is to be factorised, the value of each local conditional distribution can be learned from the data.

To demonstrate how this works, an experiment on parallel activity recognition has been conducted. Three sets of activities were used in this experiment. The activities involving legs are (1) *lying*, (2) *sitting*, (3) *standing*, (4) *walking* and (5) *running*. The activities involving arms are (1) *doing nothing*, (2) *reading newspaper*, (3) *eating food*, (4) *waiving arms in the air* and (5) *talking on the phone*. The mouth activities are (1) *doing nothing*, (2) *chewing* and (3) *speaking*. In the experiment, a subject was asked to perform predefined train and test routines of activities. The training routine consists of (1) *lying*, (2) *sitting*, (3) *walking*, (4) *standing*, (5) *running*, (6) *sitting and reading newspaper*, (7) *sitting and waving arm in the air*, (8) *sitting and talking on the phone*, (9) *sitting and speaking*, (10) *sitting and eating food*, (11) *standing and chewing* and (12) *running and waiving arm in the air*. In the test routine, the activities in the training routine was repeated, followed by a set of unseen combination of parallel activities, namely, (13) *lying and speaking*, (14) *standing and speaking*, (15) *walking and speaking*, (16) *standing and reading newspaper*, (17) *standing and talking on the phone*, (18) *walking and talking on the phone* and (19) *running and waiving arms in the air and speaking*.

Five BSN nodes, each equipped with a tri-axial accelerometer, were placed on the subject's waist, left wrist, right wrist, left ankle and right ankle, respectively. An ear-worn vision sensor embedded with sound recording was used for each subject [156]. The acceleration signals were synchronised with the video and sound data for further processing. Acceleration signals were acquired at 30 Hz and the sound at 44.1 kHz. Figure 9.17 illustrates the audio and acceleration signals captured during a performance of a test routine.

A fixed window of size of 1,000 ms and a shifted window of size of 500 ms were used for feature computation for both data sources. For audio feature extraction, a total of 76 features as detailed in [156] were extracted. These are statistical factors, such as maximum max $(\cdot)$, minimum min $(\cdot)$, standard deviation $\sigma(\cdot)$ and mean values $\mu(\cdot)$, of three classes of audio features, i.e., (1) energy features (e.g., energy entropy EE(i) and short time energy STE (i)); (2) spectral features (e.g., spectral roll-off SRO(i), spectral centroid SC(i), spectral flux SF(i), an spectral average of sub-bands SA(i, j)); and (3) temporal features (e.g., zero crossing rate ZCR(i) and peak gap between two neighboured local maximal energy peak PG(i)), where i is the time index and j is the frequency sub-band index. For acceleration features, mean and standard deviation of the tri-axial acceleration signals were computed. A total of 106 audio and acceleration features are described in Table 9.9.

Based on the multi-objective BFFS algorithm, the number of features associated to the activities was reduced to 3, 7 and 26, respectively. For leg activities, features from right ankle and left wrist were selected. For arm activities, most selected features are from left and right wrists with one feature from the right ankle and one feature from the waist. For the last set of activities, 16 audio features and 10 acceleration

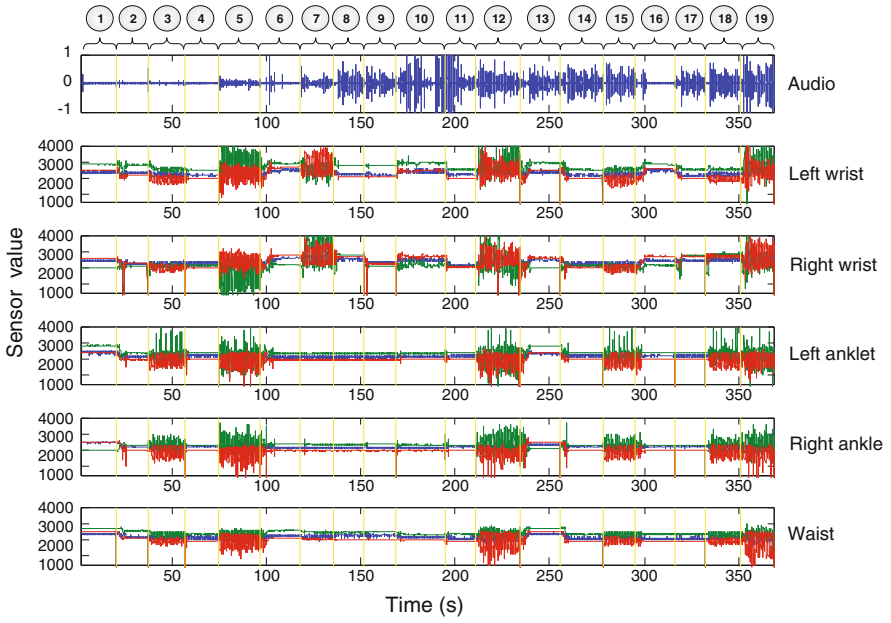


Fig. 9.17 Time series plot of the synchronised signal sequence during a test routine of the parallel activity recognition experiment

Table 9.9 Description of features used in the parallel activity recognition experiment

No.	Feature description
1–6	Energy entropy, EE(i) (max, min, σ, μ, σ/μ, $a1$)
7–12	Zero crossing rate, ZCR(i) (max, min, σ, μ, σ/μ, $a1$)
13–18	Spectral roll-off, SRO(i) (max, min, σ, μ, σ/μ, $a1$)
19–24	Spectral centroid, SC(i) (max, min, σ, μ, σ/μ, $a1$)
25–30	Spectral flux, SFi) (max, min, σ, μ, σ/μ, $a1$)
31–36	Short time energy, STE(i) (max, min, σ, μ, σ/μ, $a1$)
37–38	Peak gap, PG(i) (μ, σ)
39–47	Spectral power, SA(i, j) (μ over 3 channels, i.e.,1: 3, 4: 6, 7: 9, 10, 13, 14: 17, 18: 21, 22, 24, 25, 27)
48–56	Spectral power, SA(i, j) (σ over 3 channels, i.e.,1: 3, 4: 6, 7: 9, 10, 13, 14: 17, 18: 21, 22, 24, 25, 27)
57–66	Spectral power on frequency channels 7–16 (μ)
67–76	Spectral power on frequency channels 7–16 (σ)
77–82	$\mu(a_x)$, $\mu(a_y)$, $\mu(a_z)$, $\sigma(a_x)$, $\sigma(a_y)$, $\sigma(a_z)$ at left wrist
83–88	$\mu(a_x)$, $\mu(a_y)$, $\mu(a_z)$, $\sigma(a_x)$, $\sigma(a_y)$, $\sigma(a_z)$ at right wrist
89–94	$\mu(a_x)$, $\mu(a_y)$, $\mu(a_z)$, $\sigma(a_x)$, $\sigma(a_y)$, $\sigma(a_z)$ at left ankle
95–100	$\mu(a_x)$, $\mu(a_y)$, $\mu(a_z)$, $\sigma(a_x)$, $\sigma(a_y)$, $\sigma(a_z)$ at right ankle
101–106	$\mu(a_x)$, $\mu(a_y)$, $\mu(a_z)$, $\sigma(a_x)$, $\sigma(a_y)$, $\sigma(a_z)$ at waist

features were selected. For simplicity, only single parent structure is used. On the training dataset, a recognition accuracy of 92.40, 90.78 and 88.98 % were obtained for the three sets of activities. On the test dataset, the recognition accuracy is reduced to 86.11, 66.40 and 70.3 %.

From the experiment, it can be seen that for a multi-label classification problem, issues occur when the class variables are dependent. For example, if a leg activity always co-occurs with an arm activity in the training set, the features that are discriminatory for the arm activity can be statistically associated to that of the leg. In general, by capturing the dependencies among class labels, a multi-classification learning algorithm is expected to lead to improved classification performance. However, if the leg activity co-occurs with another arm activity in the test dataset, the learned recognition model will fail to recognise the leg activity due to wrong sensor association. In general, if the dependency of class variables in the training set is different from that in the testing dataset. It is likely that the learned model will fail for some unseen cases. When we can control the training data collection protocol, a carefully designed protocol for training data collection can alleviate some of the problems. In this study, for collecting training data, the activities in the second-class variable will be collected with fixed activity (state) of the first class variable and like-wise for the activities in the third class variable. If activity in the third class variable can occur in different contexts of the first two class variables, we must make sure the training data for that activity is collected in at least two contexts. In cases when the data already exists, some close analysis on dependency between class variables and manual intervention may be required to fill in the missing information.

9.6 Conclusions

In this chapter, we have described the use of context awareness for more accurate and intelligent pervasive sensing. The use of contextual information, however, is not new and it has been widely used in many pattern recognition applications including NLP, HCI, image processing, and computer vision. Its use for pervasive sensing, however, has introduced some interesting new challenges. The popularity of context-aware architectures is due to the increasingly ubiquitous nature of the sensors as well as the diversity of the environment under which the sensed signals are collected. For the purpose of BSNs, the main emphasis of a context-aware design is concerned with the interpretation of physical and biochemical signals acquired from both wearable and implantable sensors and their association with the ambient environment. The contextual information is therefore mainly focussed on the user's activity, cognitive/mental activities, physiological and affective states, and the environment context such as location, proximity, time, social interaction, and connectivity to the general healthcare environment.

The use of contextual information is important to the improvement of diagnosis accuracy because in a BSN, similar sensory signals can be interpreted differently depending on the current activities of the patient. To understand the intrinsic

characteristics of the sensed signal and determine how BSNs should react to different events, the contextual information is essential to the adaptation of the monitoring device so as to provide more intelligent support to the users.

The use of ANNs, and SOMs in particular, offers the advantages of nonlinearity, adaptivity, evidential response, and fault tolerance. Moreover, the relatively small number of operations involved in the combined learning and classification process makes the model particularly suited for parallel, on-chip analogue implementations. For BSNs, this means some of the processing steps involved can be performed locally on low-power, miniaturised sensor nodes so as to minimise the communication bandwidth required based on effective distributed inferencing. In this chapter, we have also briefly mentioned the extension of SOMs to STSOMs. The main features of the STSOM architecture proposed in this chapter are the introduction of the dynamic layer and an adaptive mechanism for class separation and node expansion. It has been shown that the overall number of the neurons involved in the proposed STSOM is relatively small compared to traditional approaches. This is essential for BSN nodes, which have limited computational and storage resources. For low power analogue implementation, the implications for hardware design are expected to be even greater.

Our discussion in this chapter has also covered the use of HMMs and factor graphs for context-aware sensing. The main advantage of the HMM is that it has a sound statistical grounding and its parameter estimation can take into account different sources of uncertainty. Furthermore, it is modular, relatively robust, and can be combined into larger models. It is also possible to incorporate high-level domain knowledge. The disadvantage of the method, however, originates in the relatively strong assumptions about the data and the amount of training data required due to the large number of parameters involved.

With the maturity and ease of use of BSNs for activities of daily living, the amount of data involved is increasing rapidly. How to model and discover the underlying patterns is a significant research challenge. Behaviour profiling provides a means of understanding of a person's routines or activities, which can potentially be an indicator of an onset of adverse events. In this chapter, we have provided some simple examples of how to perform behaviour profiling and detect transitional activities, as well as concurrent and interleaving events. A distributed inferencing model has been introduced by combining the use of multi-objective feature selection, dependency analysis and FG representation. In this way, multiple context recognition problems can be concurrently solved with minimal use of computational resources and the ease of logical-to-physical model mapping.

It should be noted that in this chapter, we have paid little attention to signal feature extraction and selection. These are, in fact, essential to the overall performance of context detection techniques. In Chap. 8, we have summarised a number of different approaches for extracting intrinsic signal characteristics. The effective use of these features can greatly enhance the accuracy of the context detection algorithm. Although we have only mentioned in this chapter the use of HMMs and SOMs for context recognition, other techniques such as clustering and high-level inferencing techniques as described earlier in the book are equally applicable for this purpose.

Acknowledgements The authors would like to thank Dr Jindong Liu and Mr. Krittameth Teachasrisakul for their assistance in data collection, audio feature extraction and sensor data analysis for the parallel activity experiment in Sect. 9.5.4. We also thank Dr Raza Ali for his contribution to the contents of Sect. 9.5 on behaviour profiling and transient activity recognition.

References

1. Want R, Hopper A. Active badges and personal interactive computing objects. IEEE Transactions on Consumer Electronics 1992; 38(1):10–20.
2. Schilit B, Theimer M. Disseminating active map information to mobile hosts. IEEE Network 1994; 8(5):22–32.
3. Schilit BN, Adams N, Gold R, Tso MM, Want R. The PARCTAB mobile computing system. In: Proceedings of the Fourth Workshop on Workstation Operating Systems 1993; 34–39.
4. Pascoe J. Adding generic contextual capabilities to wearable computers. In: Proceedings of the Second IEEE International Symposium on Wearable Computers 1998; 92–99.
5. Dey AK, Abowd G. Towards a better understanding of context and context-awareness. In: Proceedings of the CHI 2000 Workshop on "The What, Who, Where, When, and How of Context-Awareness" 2000.
6. Korkea-aho M, "Context-aware applications survey", http://users.tkk.fi/~mkorkeaa/doc/context-aware.html, 2000.
7. Abowd GD, Mynatt ED. Charting past, present, and future research in ubiquitous computing. ACM Transactions on Computer-Human Interaction 2000; 7(1):29–58.
8. Bobick AF. Movement, activity and action: the role of knowledge in the perception of motion. Philosophical Transactions of the Royal Society of London B: Biological Sciences 1997; 352(1358):1257–1265.
9. Brotherton JA, Abowd GD, Truong KN. Supporting capture and access interfaces for informal and opportunistic meetings. GVU Center, Georgia Institute of Technology, Technical Report, GIT-GVU-99-06, 1999.
10. Long S, Aust D, Abowd GD, Atkeson CG. Rapid prototyping of mobile context-aware applications: the Cyberguide case study. In: Proceedings of the Second ACM International Conference on Mobile Computing and Networking, NY, USA 1996; 97–107.
11. Lamming M, Flynn M. Forget-me-not: intimate computing in support of human memory. In: Proceedings of FRIEND21: International Symposium on Next Generation Human Interfaces, Meguro Gajoen, Japan, 1994; 125–128.
12. Rhodes BJ. The wearable remembrance agent: a system for augmented memory. In: Proceedings of the First International Symposium on Wearable Computers, Cambridge, MA, USA, 1997; 123–128.
13. Brown PJ, Bovey JD, Chen X. Context-aware applications: from the laboratory to the marketplace. IEEE Personal Communications 1997; 4(5):58–64.
14. Want R, Hopper A, Falcao V, Gibbons J. The active badge location system. ACM Transactions on Information Systems 1992; 10(1):91–102.
15. Van Someren EJ, Vonk BF, Thijssen WA, Speelman JD, Schuurman PR, Mirmiran M, et al. A new actigraph for long-term registration of the duration and intensity of tremor and movement. IEEE Transactions on Biomedical Engineering 1998; 45(3):386–395.
16. Bhattacharya A, McCutcheon EP, Shvartz E, Greenleaf JE. Body acceleration distribution and O2 uptake in humans during running and jumping. Journal of Applied Physiology 1980; 49(5):881–887.
17. Tolkiehn M, Atallah L, Lo B, Yang G-Z. Direction sensitive fall detection using a triaxial accelerometer and a barometric pressure sensor. In: Proceedings of the Thirty-Third Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Boston, USA, 2011.

18. Asada HH, Jiang H-H, Gibbs P. Active noise cancellation using MEMS accelerometers for motion-tolerant wearable bio-sensors. In: Proceedings of the Twenty-Sixth Annual International Conference of Engineering in Medicine and Biology Society 2004; 1:2157–2160.
19. Bao L, Intille SS. Activity recognition from user-annotated acceleration data. In: Proceedings of the Second International Conference on Pervasive Computing, Vienna, Austria, 2004; 1–17.
20. Lee S-W, Mase K. Activity and location recognition using wearable sensors. IEEE Pervasive Computing 2002; 1(3):24–32.
21. Thiemjarus S, Lo BPL, Yang GZ. Feature selection for wireless sensor networks. In: Proceedings of the First International Workshop on Wearable and Implantable Body Sensor Networks, Imperial College, London, 2004.
22. Thiemjarus S, Lo BPL, Yang GZ. A distributed Bayesian framework for body sensor networks. In: Proceedings of the Second International Workshop on Body Sensor Networks, Imperial College, London, 2005.
23. Thiemjarus S, Lo BPL, Yang GZ. A noise resilient distributed inference framework for body sensor networks. In: Proceedings of the Third International Conference on Pervasive Computing, Munich, Germany, 2005; 13–18.
24. Van Laerhoven K, Cakmakci O. What shall we teach our pants? In: Proceedings of the Fourth IEEE International Symposium on Wearable Computers 2000.
25. Krause A, Siewiorek DP, Smailagic A, Farringdon J. Unsupervised, dynamic identification of physiological and activity context in wearable computing. In: Proceedings of the Seventh IEEE International Symposium on Wearable Computers 2003; 88–97.
26. Mitchell E, Monaghan D, O'Connor NE. Classification of sport activities using smartphone accelerometers. Sensors 2013; 13(4):5317–5337.
27. Choudhury T, Consolvo S, Harrison B, Hightower J, LaMarca A, LeGrand L, et al. The mobile sensing platform: an embedded activity recognition system. IEEE Pervasive Magazine – Special Issue on Activity-Based Computing 2008; 7(2):32–41.
28. Inzitari D, Basile AM. Activities of daily living and global functioning. International Psychogeriatrics 2003; 15(Supplement 1):225–229
29. Senanarong V, Harnphadungkit K, Prayoonwiwat N, Poungvarin N, Sivasariyanonds N, Printarakul T, et al. A new measurement of activities of daily living for Thai elderly with dementia. International Psychogeriatrics 2003; 15(2):135–148.
30. Tognetti A, Lorussi F, Bartalesi R, Quaglini S, Tesconi M, Zupone G, et al. Wearable kinesthetic system for capturing and classifying upper limb gesture in post-stroke rehabilitation. Journal of NeuroEngineering and Rehabilitation 2005; 2(8):1–16.
31. Walker DJ, Heslop PS, Plummer CJ, Essex T, Chandler S. A continuous patient activity monitor: validation and relation to disability. Physiological Measurement 1997; 18(1):49–59.
32. White R, Agouris I, Selbie RD, Kirkpatrick M. The variability of force platform data in normal and cerebral palsy gait. Clinical Biomechanics (Bristol, Avon) 1999; 14(3):185–192.
33. Chang WN, Tsirikos AI, Miller F, Schuyler J, Glutting J. Impact of changing foot progression angle on foot pressure measurement in children with neuromuscular diseases. Gait Posture 2004; 20(1):14–19.
34. Verghese J, Lipton RB, Hall CB, Kuslansky G, Katz MJ, Buschke H. Abnormality of gait as a predictor of non-Alzheimer's dementia. The New England Journal of Medicine 2002; 347(22):1761–1768.
35. Mueller MJ, Salsich GB, Bastian AJ. Differences in the gait characteristics of people with diabetes and transmetatarsal amputation compared with age-matched controls. Gait Posture 1998; 7(3):200–206.
36. Zijlstra W, Rutgers AW, Van Weerden TW. Voluntary and involuntary adaptation of gait in Parkinson's disease. Gait Posture 1998; 7(1):53–63.
37. Dalto A, Patel S, Chowdhury AR, Welsh M, Pang T, Schachter S, et al. Development of a Body Sensor Network to detect motor patterns of epileptic seizures. IEEE Transactions of Information Technology in Biomedicine 2012; 59(11):3204–3211.

38. Dahal N, Nandgopal N, Nafalski A, Z. N. Modeling of cognition using EEG: a review and a new approach. In: Proceedings of IEEE Region 10 Conference, Bali, Indonesia, 2011; 1045–1049.
39. Rodenbeck A, Binder R, Geisler P, Danker-Hopfe H, Lund R, Raschke F, et al. A review of sleep EEG patterns. part I: a compilation of amended rules for their visual recognition according to Rechtschaffen and Kales. Somnologie 2006; 10(4):159–175.
40. Bulling A, Roggen D, Troster G. What's in the eyes for context-awareness? Pervasive Computing 2011; 10(2):48–57.
41. Thiemjarus S, James A, Yang G-Z. An eye hand data fusion framework for pervasive sensing of surgical activities Pattern Recognition 2012; 45(8):2855–2867.
42. Chen J, Kam AH, Zhang J, Liu N, Shue L. Bathroom activity monitoring based on sound. In: Proceedings of the Third International Conference on Pervasive Computing, Munich, Germany, 2005; 47–61.
43. Teicher MH. Actigraphy and motion analysis: new tools for psychiatry. Harvard Review of Psychiatry 1995; 3(1):18–35.
44. Myrtek M, Brugner G. Perception of emotions in everyday life: studies with patients and normals. Biological Psychology 1996; 42(1–2):147–164.
45. Picard RW, Vyzas E, Healey J. Toward machine emotional intelligence: analysis of affective physiological state. IEEE Transactions on Pattern Analysis and Machine Intelligence 2001; 23(10):1175–1191.
46. Poh MZ, Swenson NC, Picard RW. A wearable sensor for unobtrusive, long-term assessment of electrodermal activity. IEEE Transactions on Biomedical Engineering 2010; 57 (5):1243–1252
47. Benditt DG, Milstein S, Buetikofer J, Gornick CC, Mianulli M, Fetter J. Sensor-triggered, rate-variable cardiac pacing: current technologies and clinical implications. Annals of Internal Medicine 1987; 107(5):714–724.
48. Bardram JE. Applications of context-aware computing in hospital work: examples and design principles. In: Proceedings of the 2004 ACM Symposium on Applied Computing, Nicosia, Cyprus, 2004; 1574–1579.
49. Golding AR, Lesh N. Indoor navigation using a diverse set of cheap, wearable sensors. In: Proceedings of the Third IEEE International Symposium on Wearable Computers 1999; 29–36.
50. Wilson D, Atkeson C. The Narrator: a daily activity summarizer using simple sensors in an instrumented environment. In: Proceedings of the Fifth International Conference on Ubiquitous Computing, Seattle, Washington, 2003.
51. Patawari N, Arbor A, Ash JN, Kyperountas S, Hero AO. Locating the nodes: cooperative localization in wireless sensor networks IEEE Signal Processing Magazine 2005; 22 (4):54–69.
52. Wymeersch H, Lien J, Win MZ. Cooperative localization in wireless networks. Proceedings of the IEEE 2009; 97(2):427–250.
53. Steele BG, Belza B, Cain K, Warms C, Coppersmith J, Howard J. Bodies in motion: monitoring daily activity and exercise with motion sensors in people with chronic pulmonary disease. Journal of Rehabilitation Research and Development 2003; 40(5):45–58.
54. Lukowicz P, Junker H, Troster G. Automatic calibration of body worn acceleration sensors. In: Proceedings of the Second International Conference on Pervasive Computing, Linz/Vienna, Austria, 2004; 176–181.
55. Thiemjarus S. A device-orientation independent method for activity recognition. In: Proceedings of the International Conference on Wearable and Implantable Body Sensor Networks, Biopolis, Singapore, 2010; 19–23.
56. Henpraserttae A, Thiemjarus S, Marukatat S. Accurate activity recognition using a mobile phone regardless of device orientation and location. In: Proceedings of the International Conference on Wearable and Implantable Body Sensor Networks, Dallas, TX, USA, 2011.

57. Thiemjarus S, Henpraserttae A, Marukatat S. A study on instance-based learning with reduced training prototypes for device-context-independent activity recognition on a mobile phone In: Proceedings of the Ninth International Conference on Wearable and Implantable Body Sensor Networks, MA, USA, 2013.
58. Lo B, Thiemjarus S, Panousopoulou A, Yang G-Z. Bioinspired design for body sensor networks. IEEE Signal Processing Magazine 2013; 30(1):165–170.
59. Figo D, Diniz PC, Ferrerira DR, Cardoso JMP. Preprocessing technique for context recognition from accelerometer data. Personal and Ubiquitous Computing 2010; 14(7):645–662.
60. Van Laerhoven K, Kern N, Gellersen HW, Schiele B. Towards a wearable inertial sensor network. In: Proceedings of IEE Eurowearable 2003; 125–130.
61. Van Laerhoven K, Gellersen H-W. Spine versus porcupine: a study in distributed wearable activity recognition. In: Proceedings of the Eighth International Symposium on Wearable Computers 2004; 1:142–149.
62. Najafi B, Aminian K, Paraschiv-Ionescu A, Loew F, Bula CJ, Robert P. Ambulatory system for human motion analysis using a kinematic sensor: monitoring of daily physical activity in the elderly. IEEE Transactions on Biomedical Engineering 2003; 50(6):711–723.
63. Loosli G, Lee SG, Canu S. Context changes detection by one-class SVMs. In: Proceedings of Workshop on Machine Learning for User Modeling: Challenges, the Tenth International Conference on User Modelling, Edinburgh, Scotland, 2005.
64. Basseville M, Nikiforov IV. Detection of abrupt changes – theory and application. Prentice-Hall, 1993.
65. Ravi N, Dandekar N, Mysore P, Littman ML. Activity recognition from accelerometer data. In: Proceedings of the Seventeenth Annual Conference on Innovative Applications of Artificial 2005.
66. Noguchi K, Somwong P, Matsubara T, Nakauchi Y. Human intention detection and activity support system for ubiquitous autonomy. In: Proceedings of IEEE International Symposium on Computational Intelligence in Robotics and Automation 2003; 2:906–911.
67. Patterson D, Fox D, Kautz H, Philipose M. Expressive, tractable and scalable techniques for modeling activities of daily living. In: Proceedings of the Second International Workshop on Ubiquitous Computing for Pervasive Healthcare Applications, Seattle, Washington, USA, 2003.
68. Barger TS, Brown DE, Alwan M. Health-status monitoring through analysis of behavioral patterns. IEEE Transactions on Systems, Man and Cybernetics: Part A 2005; 35(1):22–27.
69. Tapia EM, Intille SS, Larson K. Activity recognition in the home setting using simple and ubiquitous sensors. In: Proceedings of the Second International Conference on Pervasive Computing 2004; Springer LNCS 3001:158–175.
70. Chambers G, Venkatesh S, West G, Bui H. Hierarchical recognition of intentional human gestures for sports video annotation. In: Proceedings of the Sixteenth IEEE Conference on Pattern Recognition 2002; 6.
71. Philipose M, Fishkin KP, Perkowitz M, Patterson DJ, Fox D, Kautz H, et al. Inferring activities from interactions with objects. IEEE Pervasive Computing 2004; 3(4):50–57.
72. Tanyawiwat N, Thiemjarus S. Design of an assistive communication glove using combined sensory channels. In: Proceedings of the International Conference on Wearable and Implantable Body Sensor Networks, London, UK, 2012.
73. Sazonov ES, Fontana JM. A sensor system for automatic detection of food intake through non-invasive monitoring of chewing. IEEE Sensors Journal 2012; 12(5):1340–1348.
74. Huo X, Ghovanloo M. A wireless pharmaceutical compliance monitoring system based on magneto-inductive sensors. IEEE Sensors Journal 2007; 7(12):1711–1719.
75. Patel S, Lorincz K, Hughes R, Huggins N, Growdon J, Standaert D, et al. Monitoring motor fluctuations in patients with Parkinson's disease using wearable sensors. IEEE Transactions on Information Technology in Biomedicine 2009; 13(6):864–872.

76. Ghasemzadeh H, Jafari R, Prabhakaran B. A body sensor network with electromyogram and inertial sensors: multimodal interpretation of muscular activities. IEEE Transactions on Information Technology in Biomedicine 2010; 14:198–206.
77. Bourke AK, Lyons GM. A threshold-based fall-detection algorithm using a bi-axial gyroscope sensor. Medical Engineering and Physics 2008; 30:84–90.
78. Dadashi F, Arami A, Crettenand F, Millet GP, Komar J, Seifert L, et al. A hidden Markov model of the breaststroke swimming temporal phases using wearable inertial measurement units In: Proceedings of the Tenth International Conference on Wearable and Implantable Body Sensor Networks, MIT, Cambridge, USA, 2013.
79. King R, Atallah L, Lo B, Yang GZ. Development of a wireless sensor glove for surgical skills assessment. IEEE Transactions of Information Technology in Biomedicine 2009; 13 (5):673–679.
80. Pradhan G, Prabhakaran B, Li C. Hand-gesture computing for the hearing and speech impaired. IEEE Multimedia 2008; 15(2):20–27.
81. Haykin S. Neural networks: a comprehensive foundation, 2nd ed. Upper Saddle River, New Jersey: Prentice Hall, 1994.
82. Macq D, Verleysen M, Jespers P, Legat J-D. Analog implementation of a Kohonen map with on-chip learning. IEEE Transactions on Neural Networks 1993; 4(3):456–461.
83. McCulloch WS, Pitts W. A logical calculus of the ideas immanent in nervous activity. Bulletin Mathematical Biophysics 1943; 5:115–133.
84. Hebb DO. The organization of behavior. New York: John Wiley and Sons, 1949.
85. Rosenblatt F. The perceptron: a probabilistic model for information storage and organization in the brain. Psychological Review 1958; 65(6):386–408.
86. Minsky ML, Papert S. Perceptrons; an introduction to computational geometry. MIT Press, Cambridge, Massachusetts: MIT Press, 1969.
87. Mehrotra K, Mohan CK, Ranka S. Elements of artificial neural networks. Cambridge, Massachusetts: MIT Press, 1997.
88. Kohonen T. The self-organizing map. Proceedings of the IEEE 1990; 78(9):1464–1480.
89. Attallah L, Wiik A, Jones GG, Lo B, Cobb J, Amis A, et al. Validation of an ear-worn sensor for gait monitoring using a force-plate instrumented treadmill. Gait and Posture 2011; 35 (4):674–676.
90. Himberg J, Flanagan JA, Mäntyjärvi J. Towards context awareness using symbol clustering map. In: Proceedings of the Workshop for Self-Organizing Maps, Kitakyushu, Japan, 2003; 249–254.
91. Varsta M, Heikkonen J, Millan JdR. Context learning with the self-organizing map. In: Proceedings of the Workshop on Self-Organizing Maps, Helsinki University of Technology, Finland, 1997.
92. Van Laerhoven K. Combining the self-organizing map and *k*-means clustering or on-line classification of sensor data. In: Proceedings of the International Conference on Artificial Neural Networks 2001; 464–469.
93. Barreto G, Araujo A, Ritter H. Time in self-organizing maps: an overview of models. International Journal of Computer Research, Special Issue on Neural Networks: Past, Present and Future 2001; 10(2):139–179.
94. Yin H, Allinson NM. Towards the optimal Bayes classifier using an extended self-organising map. In: Proceedings of the International Conference on Artificial Neural Networks 1995; 2:45–49.
95. Thiemjarus S, Lo BPL, Yang G-Z. A spatio-temporal architecture for context-aware sensing. In: Proceedings of the Third International Workshop on Wearable and Implantable Body Sensor Networks, MIT Media Lab, 2006; 191–194.
96. Baum LE, Pertrie T, Soules G, Weiss N. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. Annals of Mathematical Statistics 1970; 41:164–171.

97. Rabiner L. A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of the IEEE 1989; 77.
98. Gold B, Morgan N. Speech and audio signal processing: processing and perception of speech and music. New York: John Wiley and Sons, 2000.
99. King R, Lo BPL, Yang GZ. Hand gesture recognition with body sensor networks. In: Proceedings of the Second International Workshop on Body Sensor Networks, Imperial College, London, 2005.
100. Young SJ, Rusell NH, Thornton JHS. Token passing: a conceptual model for connected speech recognition. Cambridge University Engineering Department, Technical Report, CUED/F-INFENG/TR38, 1989.
101. Kschischang FR, Frey BJ, Loeliger H-A. Factor graphs and the sum-product algorithm. IEEE Transactions on Information Theory, Special Issue on Codes on Graphs and Iterative Algorithms 2001; 47(2):498–519.
102. Tanner RM. A recursive approach to low complexity codes. IEEE Transactions on Information Theory 1981; 27(5):533–547.
103. Wiberg N, Loeliger H-A, Kotter R. Codes and iterative decoding on general graphs: iterative and turbo decoding. European Transactions on Telecommunications and Related Technologies 1995; 6:513–525.
104. Chu M, Mitter SK, Zhao F. An information architecture for distributed inference on ad hoc sensor networks. In: Proceedings of the Forty-First Annual Allerton Conference on Communication, Control, and Computing, Monticello, IL, USA, 2003.
105. Cooper GF. NESTOR: A computer-based medical diagnostic that integrates causal and probabilistic knowledge. Stanford University, Stanford, California, PhD thesis 1984.
106. Breiman L, Friedman JH, Olshen RA, Stone CJ. Classification and regression trees. Belmont, California: Wadsworth International Group, 1984.
107. Quinlan JR. C4.5: Programs for machine learning. San Mateo, California: Morgan Kaufmann Publishers, 1993.
108. Vapnik V. The nature of statistical learning theory. Berlin: Springer-Verlag, 1995.
109. Cristianini N, Shawe-Taylor J. An introduction to support vector machines. Cambridge: University Press, 2000.
110. Cortes C, Vapnik V. Support-vector networks. Machine Learning 1995; 20(3):273–297.
111. Huynh T, Schiele B. Towards less supervision in activity recognition from wearable sensors. In: Proceedings of the Tenth IEEE International Symposium on Wearable Computing, Montreux, Switzerland, 2006.
112. Min J-K, Cho S-B. Semantic management of multiple contexts in a pervasive computing framework. Expert Systems with Applications 2012; 39:8655–8664.
113. Tanantong T, Nantajeewarawat E, Thiemjarus S. Towards ECG healthcare monitoring through body sensor networks using an arrhythmia indicator ontology. In: Proceedings of the Sixth International Conference on Knowledge, Information and Creativity Support Systems, Beijing, China, 2011.
114. Ye J, Dobson S, McKeever S. Situation identification techniques in pervasive computing. Personal and Ubiquitous Computing 2012; 8(1):36–66.
115. Atallah L, ElHelw MA, Pansiot J, Stoyanov D, Wang L, Lo BPL, et al. Behaviour profiling with ambient and wearable sensing. In: Proceedings of the International Workshop on Wearable and Implantable Body Sensor Networks, Aachen, Germany, 2007; 133–138.
116. Atallah L, Yang GZ. The use of pervasive sensing for behaviour profiling – survey. Pervasive and Mobile Computing 2009; 5:449–464.
117. Dawkins R. The selfish gene. Oxford University Press, 1989.
118. Glodek M, Bigalke L, Schels M, Schwenker F. Incorporating uncertainty in a layered HMM architecture for human activity recognition. In: Proceedings of the Joint ACM Workshop on Human Gesture and Behavior Understanding, Scottsdale, Arizona, USA, 2011; 33–34.
119. Chen C, Liang J, Zhao H, Hu H, Tian J. Factorial HMM and parallel HMM for gait recognition IEEE Transactions on Systems, Man, and Cybernetics 2009; 39(1):114–123.

120. Bui HH, Venkatesh S, West G. Policy recognition in the abstract hidden Markov model. International Journal of Artificial Life Research 2002; 17(1):451–499.
121. Duong TV, Bui HH, Phung DQ, Venkatesh S. Activity recognition and abnormality detection with the switching hidden semi-Markov model. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CA, USA, 2005; 838–845.
122. Brand M, Oliver N, Pentland A. Coupled hidden Markov models for complex action recognition. In: Proceedings of the Conference on Computer Vision and Pattern Recognition, San Juan, Puerto Rico, 1997; 994–999
123. Oliver N, Rosario B, Pentland A. A Bayesian computer vision system for modeling human. IEEE Transactions on Pattern Analysis and Machine Intelligence 2000; 22(8):831–843.
124. Natarajan P, Nevatia R. Coupled hidden semi Markov models for activity recognition. In: Proceedings of the IEEE Workshop on Motion and Video Computing, Texas, USA, 2007.
125. Han J, Cheng H, Xin D, Yan X. Frequent pattern mining: current status and future directions. Data Mining and Knowledge Discovery 2007; 15(1):55–86.
126. Luhr S, West G, Venkatesh S. Recognition of emergent human behaviour in a smart home: a data mining approach. Pervasive Mobile Computing 2007; 3(2):95–116.
127. Ali R, ElHelw M, Atallah L, Lo B, Yang GZ. Pattern mining for routine behaviour discovery in pervasive healthcare environments. In: Proceedings of the International Conference on Information Technology and Applications in Biomedicine, Shenzhen, China, 2008; 241–244.
128. Virone G, Noury N, Demongeot J. A system for automatic measurement of circadian activity deviations in telemedicine. Transactions on Biomedical Engineering 2002; 49(12):1463–1469.
129. Li YY, Parker LE. Detecting and monitoring time-related abnormal events using a wireless sensor network and mobile robot. In: Proceedings of the International Conference on Intelligent Robots and Systems, Nice, France, 2008; 3292–3298.
130. Mori T, Urushibata R, Shimosaka M, Noguchi H, Sato T. Anomaly detection algorithm based on life pattern extraction from accumulated pyroelectric sensor data. In: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems, Nice, France, 2008; 2545–2552.
131. Nanami R, Kawahawa Y, Asami T. A calorie count application for a mobile phone based on METS value. In: Proceedings of the IEEE Communications Society Conference on Sensor, Mesh and Ad Hoc Communications and Networks, San Francisco, CA, USA, 2008; 583–584.
132. Ali R, Lo B, Guang-Zhong Y. Unsupervised routine profiling in free-living conditions – can smartphone apps provide insights? In: Proceedings of the Tenth International Workshop on Wearable and Implantable Body Sensor Networks, MIT, Cambridge, USA, 2013.
133. Eagle N, Pentland A. Eigenbehaviors: identifying structure in routine. Behavioral Ecology and Sociobiology 2009; 63(7):1057–1066.
134. Dawson D, Hendershot G, Fulton J. Aging in the eighties: The prevalence of comorbidity and its association with disability. National Center for Health Statistics, Hyattsville, MD, 1987.
135. Britton E, Harris N, Turton A. An exploratory randomized controlled trial of assisted practice for improving sit-to-stand in stroke patients in the hospital setting. Clinical Rehabilitation 2008; 22(5):458–468.
136. Dall PM, Kerr A. Frequency of the sit to stand task: An observational study of free-living adults. Applied Ergonomics 2010; 41(1):58–61.
137. Janssen W, Bussmann J, Selles R, Koudstaal P, Ribbers G, Stam H. Recovery of the sit-to-stand movement after stroke: a longitudinal cohort study. Neurorehabilitation and Neural Repair 2010; 24(8):763–769.
138. Topper AK, Maki BE, Holliday PJ. Are activity-based assessments of balance and gait in the elderly predictive of risk of falling and/or type of fall? Journal of the American Geriatrics Society 1993; 41(5):479–487.
139. Cheng PT, Liaw MY, Wong MK, Tang FT, Lee MY, Lin PS. The sit-to-stand movement in stroke patients and its correlation with falling. Archives of Physical Medicine and Rehabilitation 1998; 79(9):1043–1046.

140. Janssen WGM, "The sit-to-stand movement recovery after stroke and objective assessment", Erasmus University Rotterdam, http://hdl.handle.net/1765/13555, 2008.
141. Frykberg GE, Åberg AC, Halvorsen K, Borg J, Hirschfeld H. Temporal coordination of the sit-to-walk task in subjects with stroke and in controls. Archives of Physical Medicine and Rehabilitation 2009; 90(6):1009–1017.
142. Nitz JC, Burns YR, Jackson RV. Sit-to-stand and walking ability in patients with neuromuscular conditions. Physiotherapy 1997; 83(5):223–227.
143. Galli M, Crivellini M, Sibella F, Montesano A, Bertocco P, Parisio C. Sit-to-stand movement analysis in obese subjects. International Journal of Obesity and Related Metabolic Disorders 2000; 24(11):1488–1492.
144. Allin S, Mihailidis A. Automated sit-to-stand detection and analysis. In: Proceedings of Symposium on AI in Eldercare: New Solutions to Old Problems, Arlington, Virginia, 2008.
145. Najafi B, Aminian K, Loew F, Blanc Y, Robert PA. Measurement of stand-sit and sit-stand transitions using a miniature gyroscope and its application in fall risk evaluation in the elderly. IEEE Transactions on Biomedical Engineering 2002; 49(8):843–851.
146. Ali R, Atallah L, Lo B, Yang G-Z. Detection and analysis of transitional activity in manifold space. IEEE Transactions on Information Technology in Biomedicine 2012; 16(1):119–128.
147. Ali R. Behavior profiling using wearable sensors for pervasive healthcare. Imperial College London, PhD Thesis, 2012.
148. Liao L, Fox D, Kautz H. Extracting places and activities from GPS traces using hierarchical conditional random fields. The International Journal of Robotics Research 2007; 26 (1):119–134.
149. Hu DH, Yang Q. CIGAR: concurrent and interleaving goal and activity recognition. In: Proceedings of the Twenty-Third Conference on Artificial Intelligence Chicago, Illinois, USA, 2008; 3:1363–1368.
150. Riboni D, Bettini C. COSAR: hybrid reasoning for context-aware activity recognition. Personal and Ubiquitous Computing 2011; 5:271–289.
151. Helaoui R, Niepert M, Stuckenschmidt H. Recognizing interleaved and concurrent activities: a statistical-relational approach. In: Proceedings of the International Conference on Pervasive Computing and Communications Seattle, WA, USA, 2011; 1–9.
152. Theekakul P, Thiemjarus S, Nantajeewarawat E. A hybrid framework for real-time continuous monitoring with body sensor networks. International Journal of Innovative Computing Information and Control (IJICIC) 2013; 8(6):2377–2392.
153. Tsoumakas G, Katakis I, Vlahavas I. Mining multi-label data, 2nd ed. Data Mining and Knowledge Discovery Handbook, Springer, 2010.
154. Thiemjarus S, Yang G-Z. An autonomic sensing framework for body sensor networks. In: Proceedings of the Second International Conference on Body Area Networks, Florence, Italy, 2007.
155. Thiemjarus S. A framework for contextual data fusion in Body Sensor Networks. Imperial College London, PhD Thesis, 2007.
156. Liu J, Johns E, Attallah L, Pettitt C, Lo B, Frost G, et al. An intelligent food-intake monitoring system using wearable sensors. In: Proceedings of the Ninth International Conference on Wearable and Implantable Body Sensor Networks, London, UK, 2012; 154–160

Chapter 10
Autonomic Sensing

Benny Lo, Athanasia Panousopoulou, Surapa Thiemjarus, and Guang-Zhong Yang

10.1 Introduction

In most engineering problems, our main concern is the exact specification and modelling of a system's architecture and its associated responses. In this way, we can discover whether the analytical solution is tractable and practical. For complex systems, however, this is not always possible and the use of bio-inspired design provides a way of imitating how biological systems adapt to complex, dynamic and rapidly changing environments.

Natural selection has shown the weakness of distinct creatures and strengths of surviving species. To survive, animals, insects and even microorganisms are constantly competing for their lives. Despite the innate abilities of hunting for food and avoidance from predators, natural behaviours, such as social interactions, and the formation of societies are all effective survival tools. Through millions years of evolution, species have evolved to adapt to complex, competitive and dynamic environments. From anatomical and immunological properties, to those that are neurological and behavioural, nature and biology have inspired numerous innovations including the musculoskeletal humanoid, the immune network theory, artificial intelligence and quorum sensing.

B. Lo • G.-Z. Yang, PhD (✉)
The Hamlyn Centre, Imperial College London, London SW7 2BY, UK
e-mail: g.z.yang@imperial.ac.uk

A. Panousopoulou
Institute of Computer Science, Foundation for Research and Technology – Hellas (FORTH), N. Plastira 100, Vassilika Vouton, GR-700 13 Heraklion, Crete, Greece

S. Thiemjarus
National Electronics and Computer Technology Center, Pathumthani 12120, Thailand

G.-Z. Yang (ed.), *Body Sensor Networks*, DOI 10.1007/978-1-4471-6374-9_10,

In searching for biological inspiration for sensor network design, two biological systems are regarded to be of particular importance to the BSN – the *Autonomic Nervous System* (ANS) and the *Biological Immune System* (BIS).

The Autonomic Nervous System (ANS) comprises of autonomic ganglia and nerves. Also known as the involuntary nervous system, it is concerned primarily with the control of the body's internal environment. The ANS controls the human body's internal environment through innervation of the non-skeletal muscle of the heart, blood vessels, bronchial tree, gut, pupils and secretomotor supply of many glands, namely those in the gastrointestinal tract, its embryological outgrowths, sweat glands and the adrenal medulla [1]. The higher centres of this system are located in the brain and spinal cord, with peripheral nerve fibres connecting these to both sensor and effector organs. Functionally, the ANS can be divided into two groups: sympathetic and parasympathetic. The sympathetic nervous system is primarily concerned with stress reactions of the body and when stimulated, results in what is commonly termed as the "*fight or flight*" response. The effect of stimulation of the sympathetic nervous system can, in essence, be related to the preparation for a stressful situation such as battle. In the eyes, pupils dilate allowing more light onto the retina at the back of the orbit, thereby enhancing vision. Peripheral blood vessels constrict resulting in a reduction of heat loss from the skin and a diversion of blood to vital organs. The force of contraction of the heart, along with its rate and oxygen consumption increases, also in an attempt to maintain a good blood pressure and perfusion to vital organs. The diameter of airways in the lungs is increased, allowing more oxygen into the lungs and thereby allowing for better oxygenation of blood. The motility and contraction of bowel normally required to digest food is inhibited and the tone of sphincters increased, slowing the transit of faeces through the alimentary system. In a similar way, bladder contraction is inhibited, and the sphincter controlling the flow of urine out of the bladder stimulated to contract. Blood glucose levels are increased by inhibition of insulin release and increased breakdown of the liver glycogen stores into glucose, providing fuel for the increased activity with the body. The adrenal gland is stimulated to secrete hormones such as epinephrine, further stimulating the increased cardiac output. Finally there is an increase in the production of sweat along with an elevation of hairs on the surface of the body. Together all these complex actions prepare the human body for "*fight or flight*" action.

The opposite of this occurs with the stimulation of the parasympathetic system, with, for example, a reduction in heart rate, conduction and excitability accompanied by a reduction in pupil size. There is also an increase in gut and bladder motility with sphincter relaxation, resulting in a quicker transit of faeces and urine than with sympathetic stimulation. Although the actions of both systems seem antagonistic, it is important to appreciate that both do work in synergy, with many organs, such as the heart, having both a parasympathetic and sympathetic innervation. For these organs, it is the proportion of sympathetic and parasympathetic stimulation they receive that is important in how they behave.

Figure 10.1 is a representation of the autonomic system, highlighting important differences between its sympathetic and parasympathetic components.

Whilst the ANS regulates the internal environment of the body, the Biological Immune System (BIS), is the defence system of the body. Living in a world full of

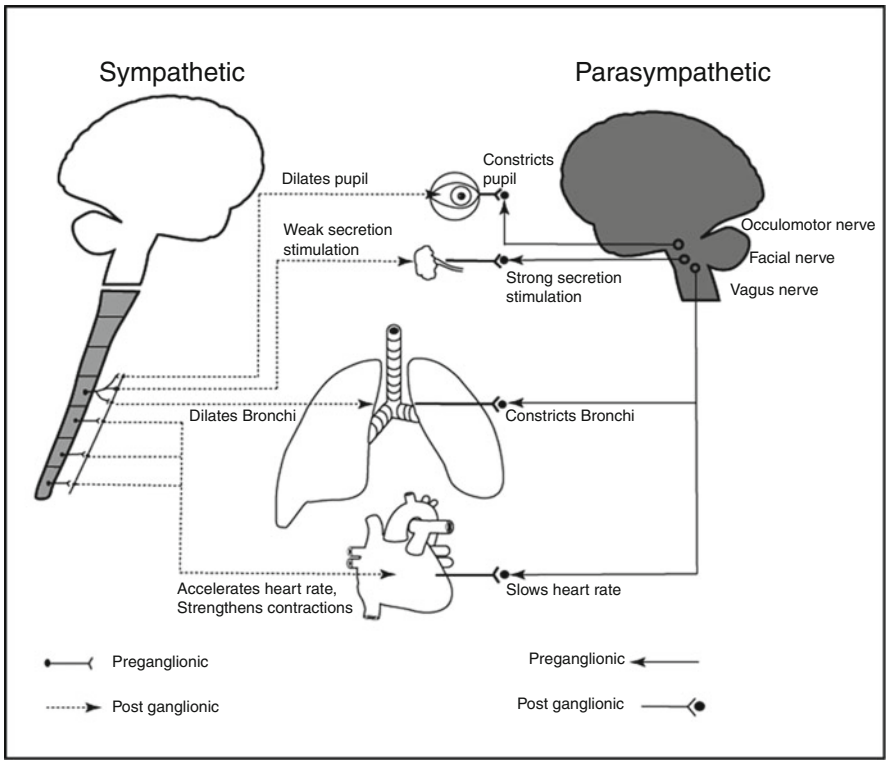


Fig. 10.1 A diagrammatic representation of the autonomic nervous system with both sympathetic (*left*) and parasympathetic components (*right*)

microbes, toxic and allergenic substances, our lives are constantly threatened by countless pathogens and toxic substances that we encounter every day. It is our highly effective immune systems, which protect us against pathogens and toxins. In general, the human immune system can be broken down into two main systems as shown in Fig. 10.2 [2–4]:

1. ***The innate immune system*** consists of a range of defence mechanisms ranging from external barriers to tissue fluids containing antimicrobial agents and leucocytes:

   - Physical barriers include epithelial cell layers, the mucus membranes (in respiratory, gastrointestinal, urogenital, visual, auditory and genitourinary tracts), and the epithelial cilia which sweep mucus or debris out of the body;
   - Innate leucocytes (white blood cells) which include phagocytes and auxiliary cells:

     – Phagocytes, which include macrophages, neutrophil, eosinophil and dendritic cells, can ingest and destroy pathogens, antigens and cell debris. Macrophages are large phagocytic cells that reside in most tissues. They can act as antigen presenting cells that activate the adaptive immune

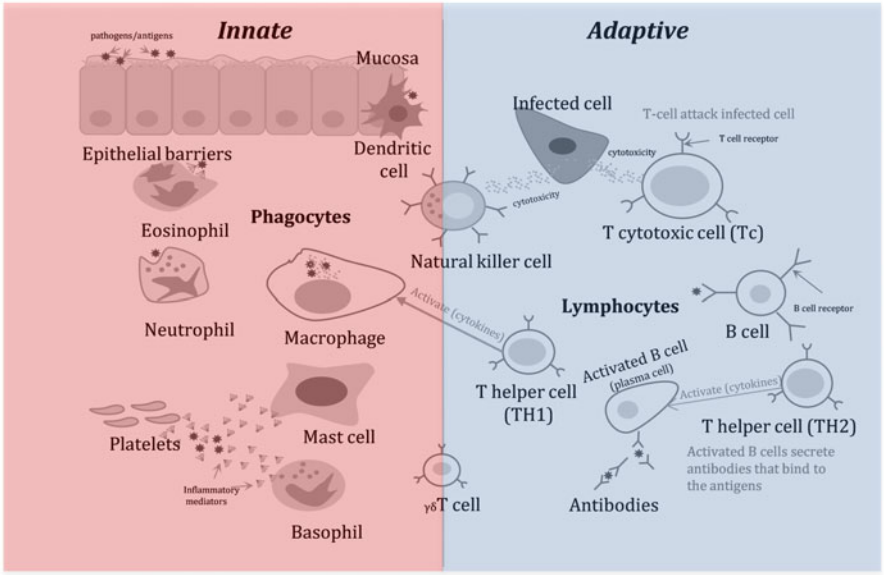


Fig. 10.2 Human immune system can be divided into innate (*left*) and adaptive (*right*) immune systems. The innate immune system provides rapid responses against antigens. Adaptive immune system, on the other hand, often takes much longer to response but becomes more powerful in repeated encounters with the same antigens

system. Neutrophils are the most abundant type of leucocytes that are normally found in the bloodstream. Eosinophils represent another type of leucocytes that can engulf large extracellular parasites and release protein to damage the parasites. Dendritic cells reside in tissues that are in contact with external environment, such as skin, noise and lungs. Similarly to macrophages, dendritic cells can function as antigen-presenting cells, which can take up the microbes and present them to lymphocytes in a form that they can recognise.

– Auxiliary cells, which include mast cells, basophils and platelets, can induce inflammatory responses against antigens. Both mast cells and basophils have similar functions and contain numerous mediators which can trigger inflammation in surrounding tissues. Mast cells are present close to blood vessels in all tissues. Basophils, on the other hand, are mobile and circulated around the body. Platelets are small disk shaped cellular fragments that are essential for blood clotting, but they can also release inflammatory mediators during an immune response.

2. ***The adaptive immune system*** consists mainly of lymphocytes, which can recognise antigens. There are two main types of lymphocytes, namely B cells and T cells. There is also a minority population of large granular lymphocytes (LGLs), such as Natural Killer (NK), in the adaptive immune system.

- B cells express specific surface immunoglobulin as their antigen receptors. When activated, B cells secrete antibody molecules that bind to antigens. B cells mainly act against extracellular pathogens.
- T cells, on the other hand, mainly response to intracellular pathogens, such as viruses. There are three main different types of T cells:
  - Type 1 helper T cells (TH1) interact with macrophages and help them to destroy pathogens;
  - Type 2 helper T cells (TH2) help B cells to divide, differentiate and make antibodies;
  - Cytotoxic T lymphocytes (Tc) cells identify and kill infected cells;
  - γδT cells are a small population of T cells that can recognise and respond to antigens without the normal constraint of antigen processing.
- Natural Killer (NK) cells can recognise cells with low levels of the cell surface marker MHC (Major Histocompatibility Complex) Class I, which indicate that the cells are virally infected or are tumour cells. Once such cells are recognised, cytotoxic substances will be secreted by the NK cells to destroy the targeted cells.

The innate immune system is often regarded as our first line of defence which provide immediate responses against pathogenic microbes, toxic and allergic substances. The adaptive immune system, on the other hand, takes much longer to react to antigens, but it is able to memorise the characteristics of pathogens and provide an enhanced response against the same pathogens in subsequent encounters. Although the mechanisms against pathogens are different, both innate and adaptive immune systems are usually act together to eliminate pathogens.

Having understood the anatomical makeup of the ANS and BIS, one can begin to explore how some of the principles can be adopted to overcome the challenges faced by BSNs. For ANS, some of these features can be summarised as the so-called *self-** properties such as self-management, self-organisation and self-healing. The BIS, on the other hand, provides us with important design principles in developing sensor networks with effective self-protection by exploiting the adaptivity and versatility of the biological systems in dealing with bacterial attack and viral infection. In the subsequent sections of this chapter, we will illustrate some of the design principles of autonomic sensing based on the features derived from the ANS and BIS, and highlight some of the computational considerations involved in implementing these concepts.

10.2 Autonomic Sensing

The term *Autonomic Sensing* follows the concept of *Autonomic Computing,* coined by IBM in their manifesto responding to the looming crisis of software complexity facing the IT industry. The spiralling cost of managing increasingly complex

Table 10.1 The eight defining characteristics of an autonomic system

Characteristics of autonomic systems	
Self-management	An autonomic system needs to have detailed knowledge about its components, current status, ultimate capacity and all connections to other systems to govern itself through effective resource management, utilisation and sharing
Self-configuration	An autonomic system can automatically and dynamically configure and reconfigure itself under varying conditions and changing environments
Self-optimisation	An autonomic system can constantly optimise its performance and resource utilisation by monitoring its constituent components, and fine-tune workflow to achieve predetermined performance and resource utilisation goals
Self-healing	An autonomic computing system can gracefully recover from routine and extraordinary events that may cause component malfunction. It is able to discover problems and establish means of using alternative resources or configuration to maintain system functionality
Self-protection	An autonomic computing system must be able to exert self-protection by automatically detecting and identifying different types of attacks to maintain overall system security and integrity
Self-adaptation	An autonomic system must be context-aware and adapt itself for improved interaction and performance under changing working environments and user requirements
Self-integration	An autonomic system can fully function under heterogeneous infrastructures and be seamlessly and securely integrated with other systems
Self-scaling	An autonomic system will anticipate the optimised resources required and scale its functionality while keeping its complexity hidden from the user

Adapted from Kephart and Chess [6]

systems is becoming a significant obstacle that is undermining the future growth and societal benefits of IT technology. As stated by Kephart and Chess of IBM in their article on the vision of autonomic computing [5]:

> Computing systems' complexity appears to be approaching the limits of human capability, yet the march toward increased interconnectivity and integration rushes ahead unabated. This march could turn the dream of pervasive computing – trillions of computing devices connected to the Internet –into a nightmare.

The biological connotations of autonomic sensing are not coincidental. It reflects our inspiration by biological systems, which manage networks that are so complex, and yet are handled so effectively and gracefully. It also echoes our desire to develop self-management systems that can pull us through the present complexity crisis and free us from the system administration nightmare. It has been expected that IT systems, particularly considered within the context of pervasive computing, will become so massive that even the most skilled system integrators will find them too complex to install, configure, optimise and maintain [5]. All of these considerations have motivated the investigation of an alternate paradigm based on the strategies used by biological systems to deal with the challenges of scale, complexity, heterogeneity and uncertainty involved in pervasive sensing.

The overall goal of an autonomic system is to provide self-management in accordance with high-level guidance from humans. Table 10.1 outlines the eight

defining characteristics of an autonomic system advocated by IBM [6, 7]. While the definition of autonomic sensing is likely to evolve as the contributing technologies mature and become established, these eight *self-** properties have highlighted some of the major requirements as well as challenges faced by the pervasive sensing community.

In this chapter, we will focus on three of the *self-** properties listed in Table 10.1: self-healing, self-organisation, and self-protection for an improved design of BSNs. It must be pointed out that our aim is not to reverse-engineer biological systems but only to follow some of the same design principles, so that plausible computational architectures suitable for the BSN can be developed.

10.3 Fault Detection and Self-Healing

Effective fault detection and recovery is one of the major concerns of wireless sensing. Although the basic principles of fault localisation and root cause analysis have been addressed by the control and process engineering communities for many years [8, 9], fault detection and self-healing for BSNs present a number of unique challenges. For a BSN, we need to examine both hard and soft failures by considering the mobile, ad hoc nature of the underlying network, the presence of both transient and permanent abnormalities, and the possibility of multiple and correlated failures. The *hard-failure* mentioned above includes node failures due to faulty sensors, loss of wireless communication or depleted battery, whereas *soft-failure* can be caused by excessive noise artefact due to poor sensor contact and motion. The isolation of these problems is also compounded by the complexity of the BSNs interacting with the heterogeneous ambient sensing environment.

Figure 10.3 illustrates a typical system architecture for fault diagnosis. Many of the traditional systems take a global approach and assume that only one fault may occur in the system at a given time. In addition, these systems frequently use deterministic models that assume that all dependencies and causal relationships are known. More recently, the use of *Finite State Machines* (FSMs) and probabilistic models have gained considerable interest. The latter is particularly relevant to BSNs as it provides an effective means of managing heterogeneous networks with non-deterministic factors. For example, uncertainty about the dependencies amongst the sensor nodes can be represented by assigning probabilities to the links in the dependency or causality graph that can be transformed into a belief network.

Given an evidence set, belief networks can be used for queries including: (1) belief assessment, (2) *Most Probable Explanation* (MPE), (3) *Maximum a posteriori* (MAP) hypothesis and (4) maximum expected utility [10]. As an example, MPE can be used to find a complete assignment of values to variables in a way that best explains the observed evidence, and therefore we are able to see its value in fault diagnosis. The main advantage of belief networks is due to not only their ability to handle uncertainties, but also the possibility of their implementation in a distributed framework. They are, therefore, particularly suited for wireless body

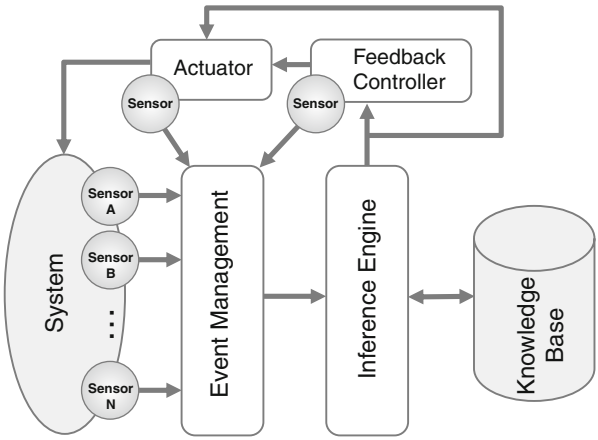


Fig. 10.3 Architecture for a typical fault diagnosis system

sensor networks, due to their similarity to biological systems in using simple local processing and messaging to form an effective global behaviour with fault detection and self-healing properties.

10.3.1 Belief Networks

To understand how belief networks work, it is necessary to explain the concept of probabilistic graphical models, which are the outcomes of fusing graph theory and probability theory. Qualitatively, their structures are graphs in which the nodes represent random variables and the arcs represent dependencies. Generally, there are two types of graphs: undirected and directed graphs. Undirected graphical models are known as *Markov Random Fields* (MRFs) or Markov networks. Directed graphs, particularly *Directed Acyclic Graphs* (DAGs), are known as Bayesian networks or Belief Networks. The latter have a more complicated notion of dependency by taking into account the directionality of the arcs which connote the causality. Another commonly used probabilistic graphical model is called *Factor Graphs* (FGs) [11, 12], which is a bipartite graph. It subsumes both Bayesian networks and MRFs. All of the independency relationships in a Bayesian network or MRF can be expressed in an FG, so that a single unified belief propagation algorithm can be used for data inference.

Quantitatively, probabilistic graphical models provide an economical way to encode a complete *Joint Probability Distribution* (JPD) over a large set of variables. The basic decomposition scheme offered by the probabilistic graphical models relies on the chain rule of probability calculus, i.e.,

$$P(x_1, \ldots, x_n) = \prod_j^n P\left(x_j \middle| x_1, \ldots, x_{j-1}\right) \tag{10.1}$$

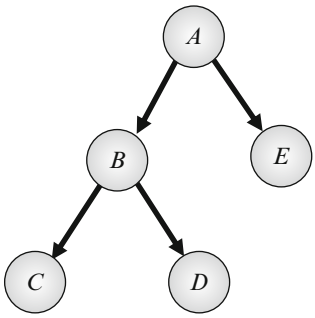


Fig. 10.4 An example Bayesian network showing how the JPD can be calculated

which permits the decomposition of a joint distribution $P(x_1, \ldots, x_n)$ as a product of n local conditional distributions. Under certain conditional independence assumptions, large distribution functions can be decomposed into several small distributions while the global nature of the problem domain is preserved.

In a Bayesian network, the direction of the arc denotes a direct child-parent relationship between two variables. That is, the arc is directed from a Markovian parent to a child node. The conditional probability of x_1 is sensitive only to the Markovian parents when:

$$P\left(x_i \middle| x_1, \ldots, x_{i-1}\right) = P\left(x_i \middle| a_i\right) \tag{10.2}$$

where a_i represents a set of parent nodes, $a_i \subseteq \{x_1, \ldots, x_{i-1}\}$, and x_i and $\{x_1, \ldots, x_{i-1}\}$ are conditionally independent given a_i. As such, the joint distribution can be simplified into the product decomposition:

$$P(x_1, \ldots, x_n) = \prod_{i=1, i \neq a}^{n} P\left(x_i \middle| a_i\right) \tag{10.3}$$

In other words, under the conditional independence assumptions denoted by the network structure, the product of the local conditional distributions of all nodes is equal to their JPD. This is essential to the calculation of JPD because as the number of variables grows, their JPD is not readily accessible. Figure 10.4 illustrates an example Bayesian network, using the chain rule, the JPD can be calculated by:

$$P(A, B, C, D, E) = P\left(E \middle| A, B, C, D\right) P\left(D \middle| A, B, C\right) P\left(C \middle| A, B\right) P\left(B \middle| A\right) P(A) \tag{10.4}$$

According to the directed graph as shown in Fig. 10.4, node E is connected only to its parent node A, and is conditionally independent of others. Therefore, $P(E|A, B, C, D)$ can be simplified as: $P(E|A)$. In addition, both nodes C and D are only connected to its parent node B; hence, $P(D|A, B, C)P(C|A, B)$ can be simplified as: $P(C|B)P(D|B)$. Therefore, the JPD of the Bayesian network can be simplified as:

$$P(A, B, C, D, E) = P(A) P\left(B \middle| A\right) P\left(C \middle| B\right) P\left(D \middle| B\right) P\left(E \middle| A\right) \tag{10.5}$$

In real-life problems, it is often possible through observation that some consequences are more likely to occur than the other, although not with absolute certainty. By summing up all the possible values of the irrelevant variables (called marginalisation), all possible inference queries can be answered accordingly. In practice, however, more efficient inference methods are used, as the direct method yields exponential time complexity.

Bayesian networks were first introduced during the mid-1980s, largely through the work of Judea Pearl when he developed a belief propagation algorithm for inferencing in a singly-connected network [13–15]. This algorithm is also commonly called the polytree algorithm. A singly-connected network is a network that contains no closed loops, i.e., there is only one single path between any two nodes in the network. Inferencing in a general Bayesian network has been proven to be NP-Hard [16], but for graphs that are singly-connected, several algorithms exist that can provide exact solutions run in polynomial time.

The inference tasks in Bayesian networks include:

- *Belief updating vs belief revision*: belief updating involves the determination of the probabilities of a set of query variables given the evidence. Belief revision (MAP explanation) consists of determining the most probable instantiations of a set of variables given the evidence [17].
- *Diagnostic vs causal reasoning*: diagnostic reasoning infers the most likely cause from the obtained evidence, while causal or "top-down" reasoning infers how the cause generates effects [18].

In general, learning in Bayesian networks involves two parts: *structure learning* and *parameter learning*. Murphy [18], for example, classified structure and parameter learning in Bayesian networks based on prior knowledge of the network structure and the availability of the data. When the structure of the network is known (e.g. from prior knowledge) and the data is fully observed, maximum likelihood estimation can be used. That is, the parameters can be derived from data distribution. Due to the potential sparse data involved, small priors are usually added to the calculation of the conditional probabilities.

When the structure is known but the data is partially observed, the probabilities associated with the unobservable or hidden nodes can be learned from optimisation methods based on *Expectation Maximisation* (EM) or gradient descent.

Finally, when the structure is unknown, model selection can be formulated as a search problem. For example, a spanning tree is an objective approach in constructing a Bayesian network structure for observable cases, and the *Maximum Weight Spanning Trees* (MWST) algorithm proposed by Chow and Liu [19] is a well-known algorithm for constructing Bayesian networks with singly-connected structures.

10.3.2 Belief Propagation Through Message Passing

Belief Propagation is a decentralised iterative algorithm that operates by transmitting messages between nearby nodes in the network. Each node acts as a processing

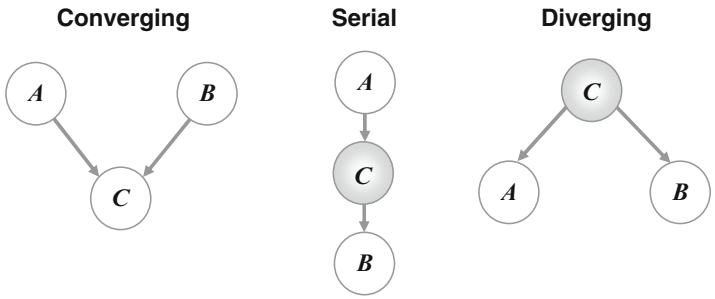


Fig. 10.5 Three cases of blocked paths. In the converging case, the path between A and B is blocked when C is not instantiated and has no evidence. In serial and diverging cases, the paths are blocked when C is instantiated

unit which can communicate only to its direct neighbours via local message passing. Associated with each node in the network are the conditional probability distributions that quantify the node and its parents.

In local message propagation, when a message is received, the node will pass the messages to all of its neighbours, with the exception of the node from which the receiving message originated. The actual mechanism also depends on its instantiation status and how it interconnects with other nodes, as shown in Fig. 10.5. The belief propagation can be accomplished in the following three steps of any order: (1) belief updating; (2) bottom-up propagation and (3) top-down propagation.

Consider node X in a typical fragment of a singly connected network, as shown in Fig. 10.6. Let e be the total evidence available, $e^-$the evidence connected to X via its children, $\mathbf{Y} = \{Y_1, \ldots, Y_m\}$ and e^+ the evidence connected to X via its parents $\mathbf{U} = \{Y_1, \ldots, U_m\}$. The belief distribution of a particular node X is updated based on the evidence received from its child- and parent-nodes. This can be derived as:

$$\begin{aligned} BEL(X) &= P(x|e) = P(x|e^-, e^+) = \alpha P(e^-|x) P(x|e^+) \\ &= \alpha P\left(e^-_{XY_1}, \ldots, e^-_{XY_m}|x\right) P\left(x|e^+_{U_1X}, \ldots, e^+_{U_nX}\right) \\ &= \alpha\lambda(x)\pi(x) \\ &= \alpha\left[\prod_{j=1}^{m}\lambda_{Y_j}(x)\right]\left[\sum_{u_1,\ldots,u_n} P(x|u_1, \ldots, u_n)\prod_{i=1}^{n}\pi_x(u_i)\right] \end{aligned} \tag{10.6}$$

where $\alpha = P(e^-|e^+)^{-1}$ is a normalisation constant; $e^-_{XY_j}$ denotes the evidence contained in the subnetwork that connects to node X via the link $X \rightarrow Y_j$ and $e^+_{U_iX}$ denotes the evidence contained in the sub-network that connects to node X via the link $U_i \rightarrow X$. The belief update in node X can be performed by inspecting the evidence in the λ messages from its children, $\lambda_{Y_j}(x), j = 1, \ldots, m$, and the evidence in the π messages from its parents, $\pi_X(u_i)$, $i = 1, \ldots, n$. The accumulation of the

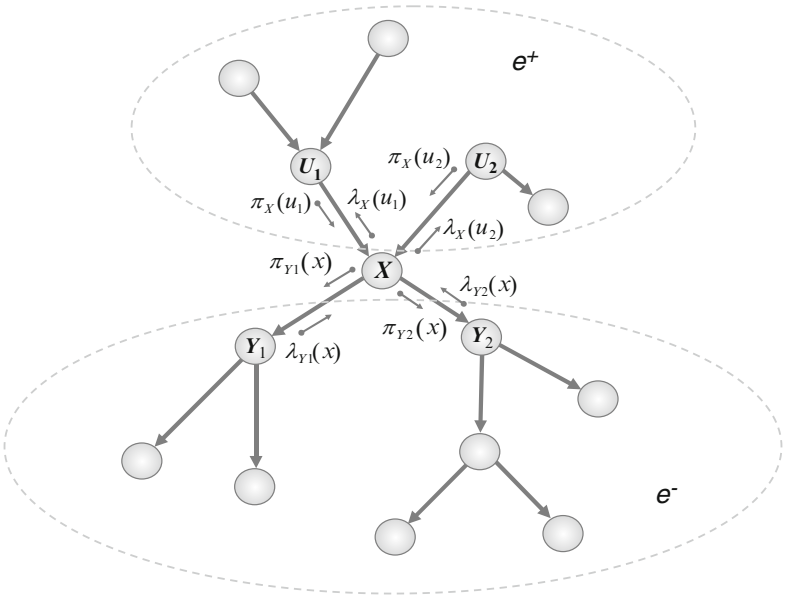


Fig. 10.6 A fragment of a singly connected network to illustrate the message passing to/from parents and children of node X

evidence from child nodes is also known as diagnostic support or λ evidence, $\lambda(x)$. The accumulation of the evidence from the parent-nodes is also known as causal support or π evidence, $\pi(x)$.

During bottom-up belief propagation, each node updates its parents' λ message array. For example, the new $\lambda_x(u_i)$ message from node X for parent U_i is calculated by:

$$\lambda_X(u_i) = \beta \sum_x \lambda(x) \sum_{u_k : k \neq i} P\left(x \middle| u_1, \ldots, u_n\right) \prod_{k \neq i} \pi_x(u_k) \tag{10.7}$$

which is the accumulation of evidence in messages from all children and parents of X apart from U_i, and β is a normalising constant. The summations indicate marginalisation over all possible values of the variables given below the Σ sign.

During top-down belief propagation, each node updates its children's π message array. For example, the new $\pi_{Y_j}(x)$ message from node X to the child Y_j is calculated by:

$$\pi_{Y_j}(x) = \gamma \frac{BEL(X)}{\lambda_{Y_j}(x)} \tag{10.8}$$

which is the accumulation of evidence in messages from all children and parents of X apart from Y_j, and γ is a normalising constant. At the boundary, exceptions are applied. For root nodes, for example, the π evidence is set to be the prior probability. If X has no children and has not been instantiated, all of the elements in

the λ evidence vector are set to be 1. Finally, if X is instantiated for x_k, the elements of the λ evidence vector at the kth position is set to 1, but 0 otherwise [17].

In Fig. 10.6, $\pi_x(u_1)$ and $\pi_x(u_2)$ are π messages from the parent-nodes U_1 and U_2 to X, $\pi_{Y1}(x)$ and $\pi_{Y2}(x)$ are π messages from X to the child-nodes. Y_1 and Y_2, $\lambda_{Y1}(x)$ and $\lambda_{Y2}(x)$ are λ messages from its child-nodes Y_1 and Y_2 to X, and $\lambda_X(u_1)$ and $\lambda_X(u_2)$ are λ messages from X to the parent-nodes U_1 and U_2. In a multiply-connected network where loops exist, messages may circulate infinitely and may not converge. To avoid evidence being counted twice in a multiple-connection network, alternative inference mechanisms have been proposed. A list of currently used inference mechanisms in Bayesian networks is summarised in Table 10.2, which categorises these techniques into exact (by modelling all the dependencies in the data) and approximate inference algorithms as suggested by Guo and Hsu [39].

It has been shown that for all exact inference methods, the time complexity is exponential to the induced width of the graph (i.e., the size of the largest clique in the triangulated moral graph). For networks with many loops and a large induced width, the exact solutions become intractable. In this case, approximation that trades the complexity of the running time with the accuracy of the results can be used. In addition to the techniques listed in Table 10.2, other approaches are also available. For example, it is possible to use orthogonal transformation of variables; for instance, using the Principal Component Analysis (PCA) to remap the data so that the correlations between variables are minimised [40]. It is also possible to use the hidden node insertion algorithm to represent the arc between two or more child-nodes of the same parent from which the link matrices associated with the hidden node can be calculated by an error minimisation scheme [41].

The message passing mechanism can be summarised as follows:

1. ***Initialisation***:
   - $BEL(X)$ of each node X should be initialised to its prior probability π_X, and
   - Set all λ evidence to 1, i.e. $\lambda = \{1, \ldots, 1\}$
2. ***Instantiation***:
   - When node X is instantiated with a value j. The jth λ evidence should be set to 1, i.e. $\lambda_{X,k} = \begin{cases} 1 & k = j \\ 0 & otherwise \end{cases}$
3. ***Belief update***:
   - The posterior probability or the belief of node X is updated by the π messages from its parents and λ messages from its children, as shown in Eq. 10.6:

$$BEL(X) = \alpha\lambda(x)\pi(x)$$

$$\lambda(x) = \prod_{j=1}^{m} \lambda_{Y_j}(x)$$

$$\pi(x) = \sum_{u_1, \ldots, u_n} P\left(x \middle| u_1, \ldots, u_n\right) \prod_{i=1}^{n} \pi_x(u_i)$$

Table 10.2 A summary of the exact and approximate belief network inference methods

Exact inference	Polytree algorithm	A generalisation of the forwards-backwards algorithm for HMMs, which is only for singly-connected networks. For multiply-connected models this becomes the loopy belief propagation algorithm [17]
	Clustering	The algorithm involves transforming the graph structure into clique tree by node clustering. Local message-passing is used for data inference, and the time complexity of the clique tree propagation algorithm is exponential to the size of the largest clique [20–24]
	Cutset conditioning	This method is based on breaking the loops in the multiply-connected network by instantiating a selected group of variables called a loop cutset [17]
	Variable elimination	The key idea of this method is to rewrite the operations in a manner which minimises the number of numerical operations required, typically through iteratively moving all irrelevant terms outside the innermost sum so that the variable associated to the innermost sum can be eliminated by marginalisation [25, 26]
	Arc reversal	This method is based on applying a sequence of operators to the network that reverses the links using Bayes' rule so that the network is reduced to just the query nodes with the evidence nodes as immediate predecessors [27, 28]
	Symbolic probabilistic inference	This method treats probabilistic inference as a combinatorial optimisation problem of finding an optimal factoring given a set of probabilistic distributions [29]
	Differential method	This method formulates a Bayesian network as a multivariate polynomial and computes its partial derivatives with respect to each variable so that answers to a very large class of probabilistic queries can be computed in a constant time [30]
Approximate inference	Stochastic simulation algorithm	Uses random instantiations from the Bayesian network and only keeps those that are consistent with the values of the observations. They can be divided into two main categories: importance sampling algorithms [31, 32] and *Markov Chain Monte Carlo* (MCMC) methods [33]
	Model simplification	Simplifies the model until an exact method becomes feasible. It includes partial evaluation methods such as the bounded cutset conditioning algorithm (which works by instantiating subsets of the variables and break loops in the graph) [34] and Mini-buckets [35]
	Search based method	Assumes the majority of the probability mass is contained in small regions of the probability space, and uses high probability instantiations to approximate the true distribution
	Loopy belief propagation	The use of Pearl's polytree propagation algorithm in a Bayesian network with loops, the marginals are often good approximations to the true marginals found through the junction tree algorithm [36]. Loopy belief propagation is equivalent to using the Bethe approximation in the variational methods [37]
	Variational methods	A class of inference methods is based on reformulating the belief propagation inference problem as an energy minimisation method. The idea is to approximate the true posterior distribution with a simpler but similar distribution. Searching for the approximate distribution can be performed by minimising the distance, such as the Kullback-Leibler divergence, between the two [38]

4. ***Bottom-up propagation***:

- When a λ message is received, the node X will update and forward λ messages to its parents (using Eq. 10.7):

$$\lambda_X(u_i) = \beta \sum_x \lambda(x) \sum_{u_k : k \neq i} P\left(x \middle| u_1, \ldots, u_n\right) \prod_{k \neq i} \pi_x(u_k)$$

5. ***Top-down propagation***:

- Likewise, when a π message is received, the node X will update and forward π messages to its parents (using Eq. 10.8):

$$\pi_{Y_j}(x) = \gamma \frac{BEL(X)}{\lambda_{Y_j}(x)}$$

When evidence is observed (or a node is instantiated), messages will be propagated throughout the networks, and the belief of each node will be updated accordingly, until an equilibrium state is reached [15].

The use of belief networks with distributed inferencing for WSNs and BSNs has attracted significant interest in recent years. For instance, to address link and node failures combined with resource constraints and asynchrony, Chu et al. [42] illustrated the use of FG representation and evolution mechanisms for *Multiple Target Tracking* which is concerned with estimating multiple target trajectories given noisy observations of the target states. The progression of a target entering, moving through and leaving a sensor field is represented by model evolution, which consists of the following three mechanisms: spawning (hypothesising the existence of new phenomenon), updating (of both distribution and topological changes) and pruning (resolving inconsistencies between expected observations and measured observations by pruning the variable nodes associated with the target from the representation).

A distributed implementation can be derived from this architecture by assigning nodes of the graph to sensors based on the so-called agent assignment algorithm. Such an approach can be particularly useful for integrating BSNs with ambient sensing environments due to the highly mobile nature of BSNs in entering, traversing through, and leaving the ambient sensing environments.

To enhance the understanding of how network topologies may affect the consensus of the MAP estimation, Alanyali et al. [43] studied the behaviour of Pearl's belief propagation algorithm in different predefined network topologies. The objective of the work is to identify communication schemes which guarantee each sensor eventually being able to identify a MAP estimate. Other research includes the use of *Nonparametric Belief Propagation* combined with Monte Carlo stochastic approximation for self-localisation under noisy sensor measurements [44] and the application of loopy belief propagation for asynchronous, rapidly changing environments troubled with node failures [45]. Paskin and Guestrin [46, 47] have also developed an inference architecture and *Robust Message Passing* algorithm for

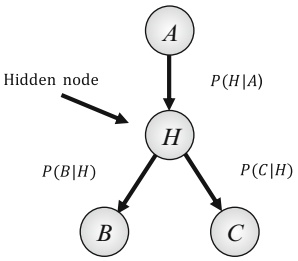


Fig. 10.7 Introduction of hidden node for redundant sensing nodes

probabilistic inferencing in distributed systems. The inference architecture is tailored for distributed implementation and consists of a spanning tree formation, optimised junction tree formation and message passing. Compared to other approaches, the algorithm is more robust when it comes to dealing with node failure and missing messages.

10.3.3 Self-Healing with Hidden Node

To enable self-healing, redundant sensors can be integrated into the network and modelled using a Bayesian network. However, with highly correlated signals, the conditional independency among children (sensing) nodes given their parents will no longer hold and lead to misclassifications [46]. Pearl introduced the use of hidden variables to maintain the conditional independency among children nodes [47]. Figure 10.7 shows a hidden node H is inserted between the parent node A and its children, nodes B and C. Nodes B and C are assumed to be redundant sensing nodes and which will be conditionally dependent given the parent node A, and the insertion of the hidden node H will maintain the conditional independency of the children nodes in the Bayesian network.

As the hidden node is a virtual variable, there is no sensing data for extracting the conditional probabilities required for the Bayesian network. To estimate the conditional probabilities, Kwoh and Gillies proposed the backward propagation with gradient descent algorithm to learn the probabilities from a set of training data [46]. The learning process can be summarised into the following steps:

1. Initialisation
   - Initialise the conditional probabilities with random values
   - Acquire a set of data for training the network
2. Classification
   - Use the current Bayesian network to perform classification with the training dataset as described in the previous section.

- Find the error of the classification result by:

$$E_n = \sum_{i}^{|T|}\sum_{k}^{|A|}(D_i(A_k) - BEL_i(A_k))^2 \tag{10.9}$$

where E_n is the classification error of the learning iteration n, $|T|$ is the number of training records, $|A|$ is the number of states of node A, $D_i(A_k)$ represents the desired value of A_k for training record $\underline{i}$ (i.e. the kth states of node A), and $BEL_i(A_k)$ is the posterior probability of A_k.

3. Propagate the error backwards and using gradient descent to update the conditional probabilities

- The conditional probability P(H|A) can be updated by:

$$P^{new}(H|A) = P(H|A) - \eta\frac{\partial E_n}{\partial P(H|A)} \tag{10.10}$$

where P^{new}(H|A) is the new P(H|A), η is the learning rate (with value between 0 and 1). With a small value of η (i.e. $\eta \cong 0$), P(H|A) will be updated marginally (at a very slow rate), and with large value of η (i.e. $\eta \cong 1$), the P(H|A) will be updated drastically (at a very high rate). In addition, the partial differential variable is defined as:

$$\frac{\partial E_n}{\partial P(H_r|A_j)} = \sum_{i}^{|T|}\sum_{k}^{|A|}\frac{\partial E_n}{\partial P'(A_k)}\frac{\partial P'(A_k)}{\partial \lambda(A_j)}\frac{\partial \lambda(A_j)}{\partial P(H_r|A_j)} \tag{10.11}$$

$$\frac{\partial E_n}{\partial P'(A_k)} = -2(D_i(A_k) - BEL_i(A_k))$$

$$\frac{\partial P'(A_k)}{\partial \lambda(A_j)} = \alpha\pi_i(A_j)(\delta(j,k) - BEL_i(A_k)) \tag{10.12}$$

$$\frac{\partial \lambda(A_j)}{\partial P(H_r|A_j)} = \lambda_i(H_r)$$

where α is the normalising constant, $\pi_i(A_j)$ is the π evidence of node A and $\lambda_i(H_r)$ is the λ evidence of hidden node H with state r at the ith training record, and

$$\delta(j,k) = \begin{cases} 1 \; if \;\; j = k \\ 0 \; if \;\; j \neq k \end{cases} \tag{10.13}$$

- The conditional probability P(B|H) can be updated as follows:

$$P^{new}(B|H) = P(B|H) - \eta \frac{\partial E_n}{\partial P(B|H)}$$

where

$$\frac{\partial E_n}{\partial P(B_s|H_r)} = \sum_{i}^{|T|} \sum_{k}^{|A|} \frac{\partial E_n}{\partial P'(A_k)} \frac{\partial P'(A_k)}{\partial \lambda(A_k)} \frac{\partial \lambda(A_k)}{\partial \lambda(H_r)} \frac{\partial \lambda(H_r)}{\partial P(B_s|H_r)} \tag{10.14}$$

$$\frac{\partial P_i'(A_k)}{\partial \lambda(A_k)} = \alpha \pi(A_k)\left(1 - P_i'(A_k)\right)$$

$$\frac{\partial \lambda(A_k)}{\partial \lambda(H_r)} = P(H_r|A_k) \tag{10.15}$$

$$\frac{\partial \lambda(H_r)}{\partial P(B_s|H_r)} = \sum_{j}^{|C|} P'(C_j) P(C_j|H_r) P'(B_s)$$

$|C|$ is the number of states of node C, $P'(C_j)$ is the posterior probability of node C with state j, and $P'(B_s)$ is the posterior probability of node B with state s.

- The conditional probability P(C|H) can be updated as follows:

$$P^{new}(C|H) = P(C|H) - \eta \frac{\partial E_n}{\partial P(C|H)}$$

where

$$\frac{\partial E_n}{\partial P(C_s|H_r)} = \sum_{i}^{|T|} \sum_{k}^{|A|} \frac{\partial E_n}{\partial P'(A_k)} \frac{\partial P'(A_k)}{\partial \lambda(A_k)} \frac{\partial \lambda(A_k)}{\partial \lambda(H_r)} \frac{\partial \lambda(H_r)}{\partial P(C_s|H_r)} \tag{10.16}$$

$$\frac{\partial \lambda(H_r)}{\partial P(C_s|H_r)} = \sum_{j}^{|B|} P'(B_j) P(B_j|H_r) P'(C_s) \tag{10.17}$$

$|B|$ is the number of states of node B, $P'(B_j)$ is the posterior probabilities of node B with state j, and $P'(C_s)$ is the posterior probability of node C with state s.

4. If the difference between the classification error obtained in this iteration n and the error in the previous iteration $(n - 1)$ is larger than a threshold [i.e. $(E_n - E_{n-1}) > threshold$], go back to step 2 and repeat the learning processes.

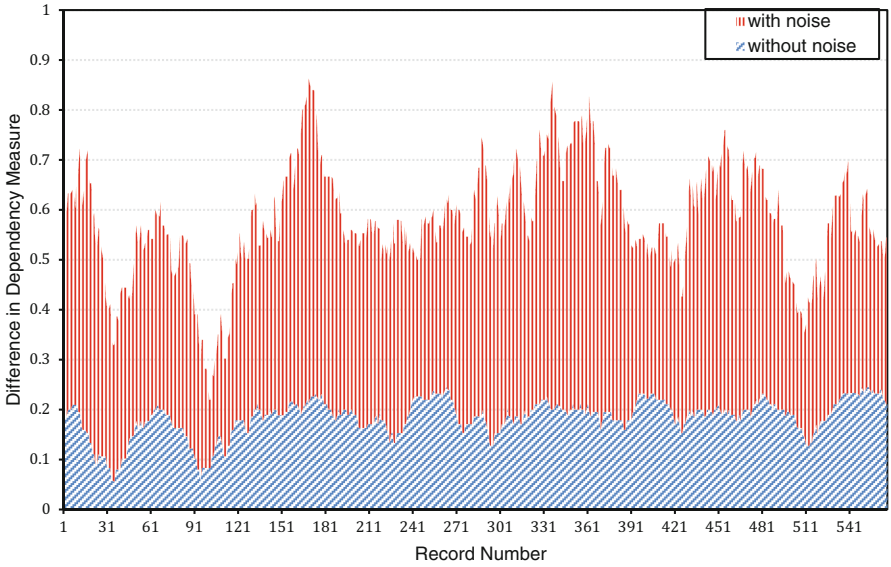


Fig. 10.8 Difference in child-parent dependency between a normal BSN node and a node corrupted by Gaussian noise with 0.5 standard deviation of the signal

Using redundant nodes, the child-parent dependency can be used to detect faulty sensor responses. For instance, L1 dependency measure was proposed to detect faulty sensors [48, 49].

$$Dep(A,B) = \sum_{i}^{|A|}\sum_{j}^{|B|} P\left(A_i, B_j\right) - P\left(B_j\right)P(A_i) \tag{10.18}$$

Since node C and node B are highly correlated, the absolute difference between the L1 dependency measures (i.e. $Dif = |Dep\ (A, C) - Dep(A, B)|$) will be very small. Figure 10.8 demonstrates one example of how the difference in L1 dependency measures (*Dif*) is used to indicate noise interface before and after Gaussian noise is introduced to one of the sensor nodes (i.e. node B) [50, 51]. As shown on the graph, the absolute difference (*Dif*) value before the noise is introduce is relatively small compare to the difference after the noise is introduced.

Once the faulty sensor node is identified, the noisy node and the hidden node can be removed from the Bayesian network, and the classification accuracy can be maintained using the remaining node (i.e. node C). By dynamically reconfiguring the structural of the Bayesian network, self-healing can be achieved. Figure 10.9 shows an example of how self-healing can be enabled using the Bayesian network. To simulate sensor failure, Gaussian noise is introduced in one of the sensor node (i.e. node B) while the network is in operation (i.e. starting from record number 200). As shown in the figure, after the noise is introduced, the classification error is elevated. However, if self-healing is enabled, (i.e. reconfiguring the Bayesian network), the classification error will be able to maintain at a relatively low level.

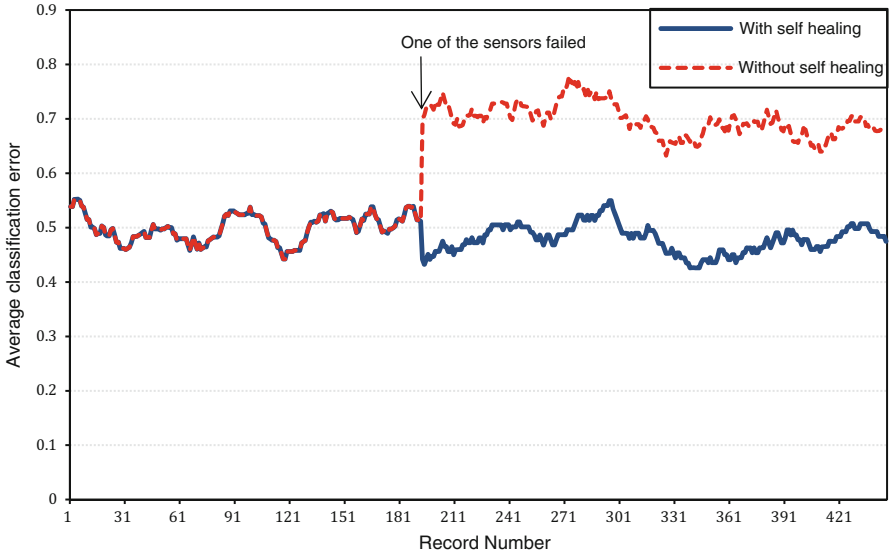


Fig. 10.9 An example of using Bayesian Network to enable self-healing. Noise is introduced in one of the sensors at record number 200, and the classification accuracy deteriorated as shown as the *red* trace. However, by removing the noisy sensor, the classification error remains at a relatively low level (as shown as the *blue* trace)

10.4 Networking and Self-Organisation

Driven by the diversity of the application-specific requirements, networking for BSN can be challenging. A typical BSN setup is comprised of heterogeneous network characteristics, regarding to the desired data rates and topology. In addition, while each sensing mode corresponds to a different set of applications, the network needs to simultaneously support different monitoring schemes, ranging from real-time data streaming to long-term trend analysis. In Chap. 5, different network topologies, protocols and standards for BSN applications have been discussed. In this section, we will discuss how self-organisation approaches can be used for optimising the structure of BSNs.

Vital sign monitoring requires "always-on" links between the body-worn/implanted sensors and off-body network infrastructure. Continuous information streaming is considered as a prerequisite for obtaining the contextual information from data acquired by a wireless BSN [48]. On the other hand, the body-centric platform should be operated with very low power and must be highly energy efficient in order to prolong monitoring period and improve user compliance. In terms of networking, the on-body/off-body links should be seamlessly configured, without user inputs or restricting activities. This requires, for example, the

sensors to gracefully detect and react to the presence or absence of the fixed tier network infrastructure, and therefore minimising the user interaction.

The aforementioned requirements on delivering heterogeneous, power-efficient and unobtrusive sensor network are constrained by the behaviour of typical wireless propagation media around the human body. Considering the cases for both wearable and implantable wireless BSNs, the human body as a medium can have a significant impact on wireless performance as discussed in Chap. 4. In addition, the operation of a wireless BSN is prone to motion artefacts and interferences, thereby generating a strong dependency of the network performance on antenna characteristics, body composition, and even body postures [49]. A systematic study by Reusens et al. [52] highlights that the behaviour of a typical wireless BSN operating at the ISM (Industrial-Scientific-Medical) band is affected by the position of the transmitter-receiver pairs on the body surface. Natarajan et al. in [53–55] further addressed the impact of the surrounding environment on the packet delivery ratio of on-body networked platforms. These studies indicate the dependency of the topology on the persistence of multipath effects. As such, while star topologies are more efficient for indoor environments for which multipath is anticipated, multi-hopping is more efficient for open spaces, where the multipath effects on the signal propagation are constrained.

As surveyed in Ullah et al. [56] and Latré et al. [57], several approaches have addressed basic networking issues by considering technology- and protocol-oriented solutions for the physical, data link and network layers. In the majority of the cases examined, static and centralised mechanisms have been adopted. However, as the authors in [57] have highlighted, the resulting protocols have limited consideration for the dynamic topology posed by the human body and its interaction with the environment. As such, despite their moderate size, BSNs can benefit from self-organised networking schemes, which are applicable across the main layers of the OSI (Open Systems Interconnection) protocol stack.

10.4.1 Medium Access Control Sub-layer

As extensively described in Chap. 5, the standardisation efforts for personal and body area networks are focused on the specifications for the Physical Layer and the Medium Access Control (MAC) sub-layer. For the latter, the dominating approaches for accessing the transmission medium are (a) contention-based, (b) contention-free. More specifically, contention-based techniques allow the operational nodes to compete for accessing the common transmission medium. This competition can either rely on assessing the traffic on the channel and prevent frames from colliding with each other (e.g. CSMA/CA (Carrier Sense Multiple Access with Collision Avoidance)) or simply wait for a random amount of time before attempting a transmission (e.g. slotted ALOHA). In contrast, during contention-free medium access, nodes agree on the exclusive use of the channel

for a limited amount of time, which is often expressed in slots (e.g. TDMA (Time Division Multiple Access) and round-robin).

The standards for personal and body area networks provide both the design guidelines for switching from contention-based to contention-free periods and vice versa, as well as the operational range of the parameters for each medium access policy. Nevertheless, the coordination of these operational parameters with respect to the varying topologies and demands in bandwidth is not part of the standardisation efforts.

Introducing self-organisation in medium access techniques can therefore provide the desired parameters coordination in a distributed manner. This is achieved by allowing nodes to converge to global decisions based on local network interactions. These interactions usually take place in the 1-hop or 2-hop neighbourhood and can, for instance, rely on the formation of autonomic feedback loops. These autonomic loops are inspired by the reaction-diffusion mechanism, which, in its most primitive form, is responsible for the pattern formation on the body surface of animals. Reaction-diffusion relies on the interaction of two basic components: (a) the activator, which activates itself with (b) the inhibitor, which restrains the activator. The principle of operation of this biological mechanism has been described by Alan Turing in [58] in the form of the following 2nd order partial differential equations:

$$\begin{aligned} \frac{\partial a}{\partial t} &= f(a,h) + D_a \nabla^2 a \\ \frac{\partial h}{\partial t} &= g(a,h) + D_h \nabla^2 h \end{aligned} \tag{10.19}$$

where functions f and g represent the non-linear chemical reactions, D_a, D_h are the diffusion rate of the activator and the inhibitor respectively, and ∇^2 is the Laplacian operator.

A representative example that highlights how reaction-diffusion can contribute to the distributed parameter coordination of a medium access control policy is described in Durvy and Thiran [59], for the case of probabilistic protocols, such as the slotted ALOHA. More specifically, the connection probability $p_i(t)$, which is assigned to each connection i at the tth time instant, can be instantly reconfigured by considering the remaining connections that are located in close proximity. This is achieved by modelling each connected as the feedback loop shown in Fig. 10.10.

According to this scheme, at time instant t, an active connection i, which is established by the nodes, will inhibit its closest connections (Inhibitor Area, Δ_I), formed by their 1-hop neighbourhoods, to become active during t. This will encourage the connections that are not in close proximity to i to become activated (Activation Area, Δ_A). At node level, these reaction-diffusion rules are translated into adopting the following rule for updating the connection probability for the next time instant $t + 1$:

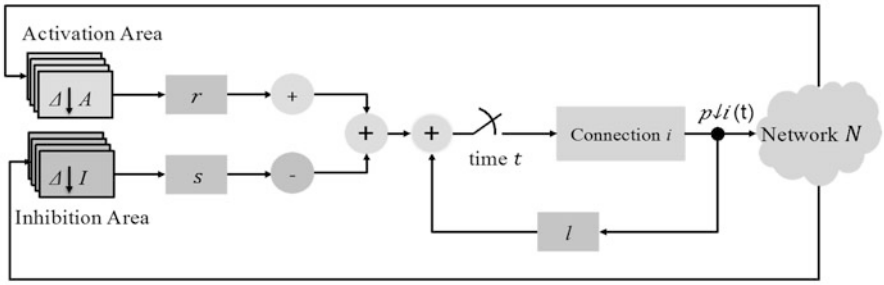


Fig. 10.10 The feedback loop per connection inspired by the reaction-diffusion mechanism

$$p_i(t+1) = \begin{cases} 1 & if \ \ \pi_i > 1 \\ \pi_i & if \ \ 0 \leq \pi_i \leq 1 \\ 0 & otherwise \end{cases} \tag{10.20}$$

where the term π_i expresses the interaction between self-activation(l), inhibition (s) and activation (r) from the network:

$$\pi_i = lp_i(t) - s\sum_{j \in \Delta_I} p_j(t) + r\sum_{j \in \Delta_A} p_j(t) \tag{10.21}$$

While at a local perspective, the nodes update their connection probabilities in an asynchronous manner, at a global perspective this results in the formation of dense and collision-free transmission patterns. In addition, it is proven that all trajectories of the non-linear system described by Eqs. 10.20 and 10.21 always converge to a stable state. Finally, the range of values for the activation-inhibition parameters that enhance the contention conditions is proven to be the following:

$$\begin{aligned} &l > 1 \\ &s > l - 1 \\ &0 < r < \frac{1}{\Delta}(1 - l + s) \end{aligned} \tag{10.22}$$

where Δ is the average number of nodes in Δ_A of a connection.

The aforementioned approach highlights how self-organised feedback loops can improve the performance of contention-based medium access control techniques. Nevertheless, the majority of the modern network techniques for BSN-oriented applications require scheduling of collective actions based on a common time base. In terms of MAC policies, this is translated in synchronisation of all devices with respect to a central node, which coordinates the operation of the entire network. The objective of synchronisation is thereby the allocation of resources on a contention-free basis. Achieving synchronisation in a self-organised manner has a two-fold advantage when compared to centralised schemes: (a) it can compensate for single-point failures resulting from packet losses and poor-quality links; and (b) it can deal with on-demand bandwidth management of each node.

A representative example of self-organised synchronisation with direct extensions to MAC design protocol for BSNs is the operation of pulse-coupled oscillators (PCO) [62]. Pulse coupled oscillators are elements in nature that change the period of their pulsing depending on whether they are operating individually or in groups. More specifically, when separated, the oscillators pulse periodically, while when combined, they alter their pulsing pattern in response to the signals heard from the remaining oscillators. Consequently, in order to achieve global synchronisation, the recipient of a pulse will desynchronise itself from its coupled oscillator, by moving earlier or later in time for its own firing.

The concept of pulse-coupled oscillators has been extensively used for achieving self-synchronisation [60–63] and relies on the update of a local state s_x of each node x based on the following equation [64]:

$$s_x = f(\phi) \tag{10.23}$$

Function f:[0, 1] $\rightarrow$ [0, 1] is a smooth and monotonically increasing function, and it is called the PCO dynamics. ϕ is the phase variable of the node with the following properties:

- $\phi \in [0, 1]$;
- $\frac{d\phi}{dt} = \frac{1}{T}$, where T is the cycle period (*i.e.* the node emits a pulse every T);
- $\phi = 0$ when $s_x = 0$ (*i.e.* the oscillator is at its lower state) and $\phi = 1$ when $s_x = 1$, (*i.e.* the oscillator reaches its threshold; hence, the node 'fires').

The combination of the properties of function f and the phase variable ϕ leads to the observation that $f(0) = 0$, and $f(1) = 1$.

While Eq. 10.23 expresses the local update on the node, the reception of a pulse from any other node can change the phase. This variation is quantified by the coupling variable ε, and

$$\phi(t^+) = f^{-1}(f(\phi(t)) + \varepsilon) \tag{10.24}$$

As proven in Mirollo and Strogatz [64], if function f is concave and the coupling is negative, the nodes will be synchronised by scheduling their firing in asymptotically equal distances.

This concept is exploited in Pagliari et al. [65, 66] and Tannious and Scaglione [67] for designing contention-free medium-access control protocols for UWB-based BSN. The basic difference with the previous approaches is that the coupling variable ε is considered positive and the function f is concave up. In particular, considering an initial offset ϕ_o, the phase variable ϕ is modelled as:

$$\phi(t) = \frac{t}{T} + \phi_o \; mod \; 1 \tag{10.25}$$

The function f considered is $f(\phi) = -\log(\phi)$ and the coupling variable equals to $\varepsilon = -\log(1-\delta)$, where $\delta \in (0,1)$. As such, the update of the phase variable is according to Eq. 10.24, where $\phi(t^+)$ equals to:

$$\begin{aligned} \phi(t^+) &= f^{-1}(f(\phi(t)) + \varepsilon) \\ f(\phi(t^+)) &= f(\phi(t)) + \varepsilon \\ f(\phi(t^+)) &= -\log(\phi(t)) - \log(1-\delta) \\ -\log(\phi(t^+)) &= -\log(\phi(t)) - \log(1-\delta) \\ \phi(t^+) &= \phi(t)(1-\delta) \end{aligned} \tag{10.26}$$

Based on this update, inhibitory coupling between operational nodes can be achieved. This implies that when a message is received from a neighbour, the node will decrease its local clock, instead of increasing it. This approach achieves weak synchronisation, defined as the case when a constant spacing between the pulsing times of consecutive nodes is achieved, but the difference between their local phases drifts over time. This shortcoming can be compensated for by alternating Eq. 10.26 to accommodate only a subset of neighbouring nodes. Consequently, firing of a node should push away only the nodes whose phase is too close. The resulting scheme can therefore achieve strict synchronisation, while retaining the time complexity to $O(N/logN)$, where N is the network size.

Based on this concept, contention-free medium access protocols for BSN with self-organised bandwidth allocation are achievable. At the level of each node x, the inhibitory coupling is integrated with the preservation of two local clocks: $\phi_{1,x}$, $\phi_{2,x}$, instead of one. These clocks serve as the boundaries of the time needed for the x node to transmit its frames (Fig. 10.11). The update of $\phi_{1,x}$, $\phi_{2,x}$ is made according to: (a) the firing of the succeeding node x_{next} in the firing cycle; (b) the bandwidth demands K_x of the x-th node and (c) the phase of $\phi_{2,x_{prev}}$ of the preceding node x_{prev} in the firing cycle [65, 66].

The firing of each clock triggers a beacon transmission, which is used to mark the start and the end of the allocated time slot respectively. In order to achieve synchronisation, the beacon associated with the first clock contains the current value of the second clock. Subsequently, upon the reception of the beacon associated with the firing of $\phi_{1,xprev}$, node x will record its distance from node x_{prev}:

$$d < x_{prev}, x >= mod\left(\varphi_{2,x_{prev}} - \phi_{1,x}, 1\right) \tag{10.27}$$

Node x will use this information to update its local clocks, when node x_{next} fires its first clock. This is determined by the following set of equations:

$$\begin{aligned} \phi_{1,x} &= \delta \min\left(\Phi_{1,x}, \frac{f\left(\phi_{2,x_{prev}}\right) + \phi_{1,x}}{2}\right) + (1-\delta)\phi_{1,x} \\ \phi_{2,x} &= \delta \max\left(\Phi_{2,x}, \frac{\phi_{2,x}}{2}\right) + (1-\delta)\phi_{2,x} \end{aligned} \tag{10.28}$$

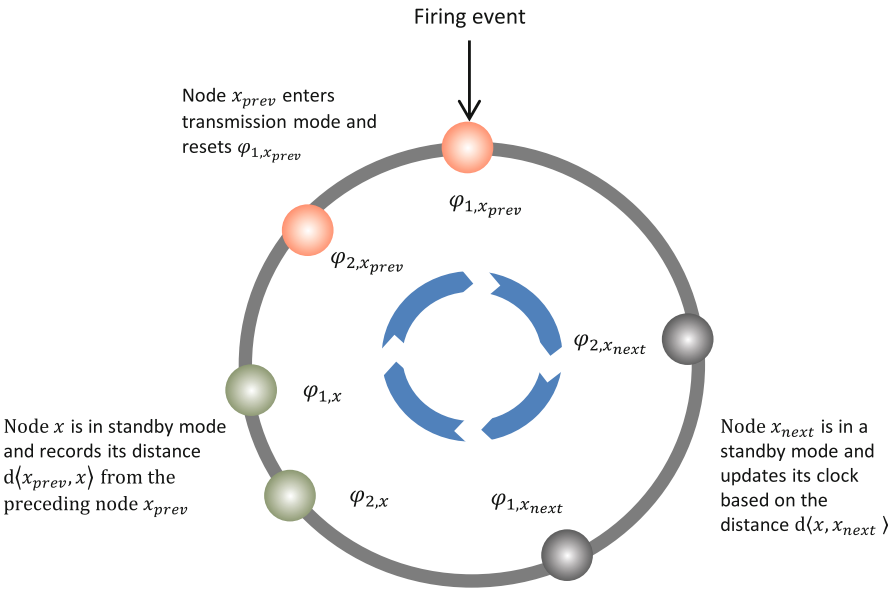


Fig. 10.11 The inhibitory coupling with two local clocks per node. The firing event stimulates the previous node to record its distance from the firing node and the next node to update its local clocks

where

$$\Phi_{j,x} = \frac{(2-j)K_x + \beta}{K_x + 2\beta} f\left(\phi_{2,x_{prev}}\right), \text{with } j = 1, 2, \beta > 0,$$

and

$$f\left(\phi_{2,x_{prev}}\right) = \phi_{1,x} + d < x_{prev}, x >.$$

As elaborated by Pagliari et al. [65, 66], convergence conditions represent an important aspect of this approach. More specifically, it is therein proven that the phase difference between $\boldsymbol{\phi_{1,x}}$ and $\boldsymbol{\phi_{2,x}}$, as well as between the extreme clock phases on adjacent nodes always converges to the same value. Nevertheless, the speed of convergence depends on the selection of parameter $\boldsymbol{\delta}$.

Figure 10.12 illustrates the significance of the value of δ, by demonstrating the convergence of the phase difference between $\phi_{2,2}$, $\phi_{1,3}$ of nodes 2 and 3 for a network comprised by 5 nodes for $\delta = \{0.009, 0.09, 0.9\}$. More specifically, the convergence speed increases when the selected value for parameter δ increases, similar to that described in Degesys et al. [61].

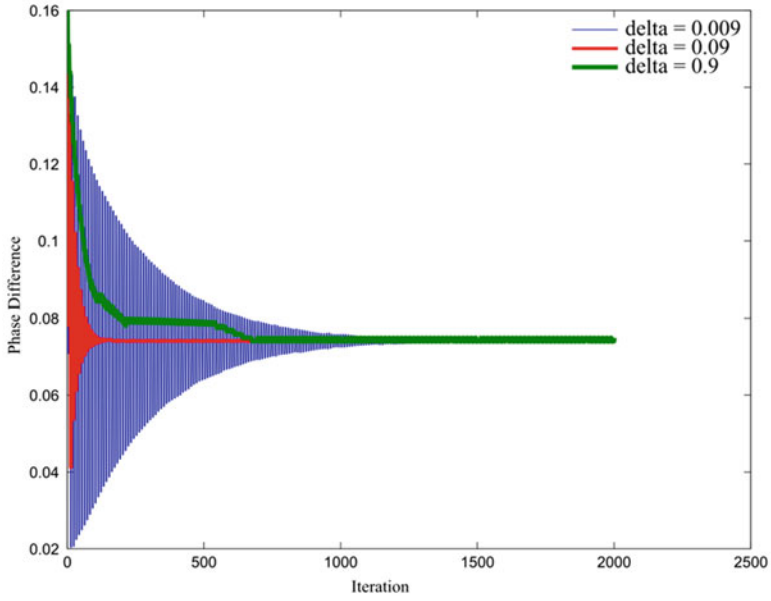


Fig. 10.12 The convergence of the phase difference during PCO-based de-synchronisation for two adjacent nodes, and different values of δ. The network is comprised by five nodes, while the remaining parameters are derived from Pagliari et al. [65] ($\beta = 2$ and $\boldsymbol{K} = \{5, 5, 1, 1, 5\}$)

10.4.2 Network Layer

The network layer is responsible for data flow control and deriving routes for packet transmission. Routing for BSN differs significantly from techniques employed by conventional network, mainly due to energy and resource constraints. Despite extensive research in routing for large scale WSN [68–72], not all energy-aware approaches are applicable to BSNs. This is due to the fact that BSN has a higher demand on data rates and Quality of Service (QoS), often under un-predictable mobile situations. The majority of routing protocols for WSNs, however, only consider networks with homogeneous sensors, while the topology of the network is often considered static. Routing techniques for BSN typically adopt a hierarchical flow [73], which is also compliant to the ZigBee standard (Chap. 5).

Hierarchical routing schemes essentially group nodes intro clusters. A cluster is a group of nodes that share a similar set of characteristics (for instance, proximity or similar sensing mode), and is represented to the sink of the BSN by its leader. A cluster leader typically acquires data from all members and transmits them to the sink. The leader is either elected from its members, based on some specific criteria, or is predetermined. Likewise, the member of the cluster can either be predetermined or dynamically updated according to the operational characteristics of the BSN.

A representative example of static cluster formation and leader selection for body-centric monitoring is typically based on 1-hop and 2-hop neighbourhood during the formation of the network. As explained in Watteyne et al. [74], this clustering scheme for routing purposes between body-centric networking and the send-tier remote infrastructure can be based on the exchange of HELLO messages between the operating nodes. Leaders are elected based on their popularity, which is expressed in terms of the range of their 1-hop neighbourhood.

A technique for self-organised BSN clustering relies on the relative positions of the sensors. As the human body is characterised by almost continuous movement, efficient approaches for mapping the quality of the link to distance need to be considered. One method is to use *Multi-Dimensional Scaling* (MDS) [75]. MDS is responsible for finding a low-dimensional representation of a set of high-dimensional data for which the distances are preserved. The starting point of MDS is a matrix consisting of the pair-wise dissimilarities (or distances as mentioned in Chap. 8) of the entities. In a simple BSN deployment example, let's assume that the RF (Radio Frequency) attenuation model is given by:

$$P_{\text{receive}} \propto \frac{P_{send}}{r^{\alpha}} \tag{10.29}$$

where r is the transmission distance between the nodes and α is the RF attenuation exponent. Given the deployment of the sensor nodes with unknown relative sensor locations, a pairwise distance d_{ij}^{*} between nodes i, j can therefore be derived. From these pairwise distance measures, the main idea of MDS is to find a configuration of the nodes in a low dimensional space so that the mapped distance in this low dimensional space d_{ij} is as close as possible to d_{ij}^{*}. When a square-error cost is used, the objective function to be minimised can be written as:

$$E = \sum_{i \neq j} \left(d_{ij}^{*} - d_{ij} \right)^2 \tag{10.30}$$

An effective projection method closely related to MDS is through the use of Sammon's mapping [76] where the errors in distance preservation are normalised with the distance in the original space, i.e.,

$$E_s = \frac{1}{\sum_{i<j} \left[d_{ij}^{*} \right]} \sum_{i<j} \frac{\left(d_{ij}^{*} - d_{ij} \right)^2}{d_{ij}^{*}} \tag{10.31}$$

It can be shown that the problem stated above can be solved iteratively by using the error measure in (10.30). By defining $E_s(m)$ and $d_{ij}(m)$ as the mapping error and the mapped low-dimensional distance after the mth-iteration, respectively, the newly estimated coordinates of the sensor nodes at iteration $m + 1$ is given by

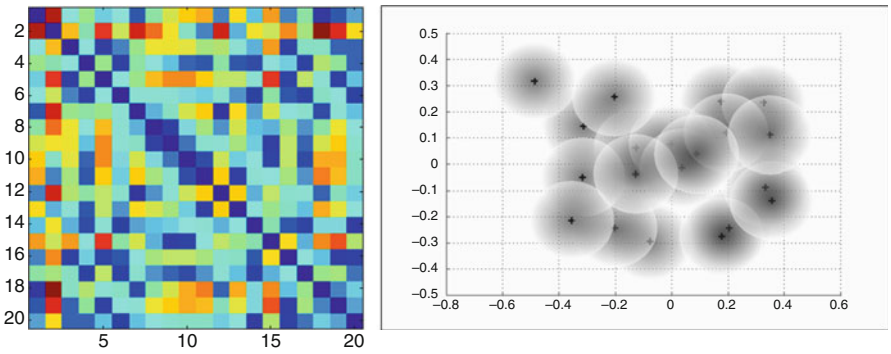


Fig. 10.13 Result of nonlinear mapping based on the relative distance measures between the 20 sensors. (*Left*) The pairwise signal attenuation matrix. (*Right*) The reconstructed relative spatial configuration of the sensors represented in 2D

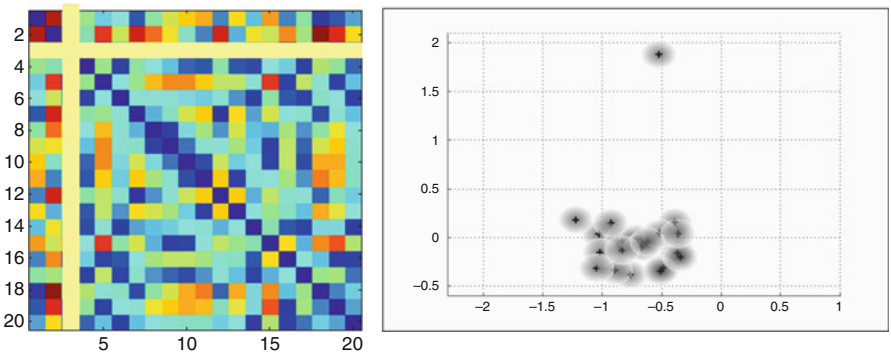


Fig. 10.14 Result of nonlinear mapping based on the relative distance measures between the 20 sensors when one of the sensor nodes is faulty, illustrating the ability of the algorithm to single out the defective node (node 3) while the relative geometrical configuration of the other sensors remains intact. (*Left*) Pair-wise signal attenuation matrix. (*Right*) The reconstructed relative spatial configuration of the sensors represented in 2D

$$x_{pq}(m+1) = x_{pq}(m) - \alpha \left[\frac{\partial E(m)}{\partial x_{pq}(m)} \right] \left| \frac{\partial^2 E(m)}{\partial x_{pq}(m)^2} \right|^{-1} \tag{10.32}$$

where x_{pq} is the qth coordinate component of sensor p, and α is the step size (referred to as the *magic factor* by Sammon).

To demonstrate how this nonlinear mapping can be used for the self-organisation of BSNs, Fig. 10.13 illustrates a group of 20 sensors with their pairwise distances (a) based on the signal attenuation model given in (10.29). In this figure, the dark (blue) cells represent short distances whereas bright (orange) cells signify large distances between the sensor nodes. Figure 10.13(b) is the corresponding nonlinear embedded result, showing a near perfect reconstruction of the co-locations of the sensors. To further illustrate the ability of the algorithm to engage in self-organisation in the presence of sensor failures, Fig. 10.14 shows the nonlinear

mapped result of how the defective sensor (c) is singled out from the sensor cluster, while the relative geometrical configuration of the remaining sensors is kept intact.

It has been shown that performing localised data processing and information passing is an energy efficient way of sensor routing [77]. As the volume of information increases, methods relying on *Multiply Sectioned Bayesian Networks* (MSBN) [78, 79] can be used for efficient partitioning and energy-aware clustering [80].

10.4.3 Application Layer

In practice, user and application specific parameters can also determine the nature of communication of self-organising networks. Self-organising methods in the application layer are often application specific. Examples include on-the-fly reconfiguration of the sensing mode with respect to low-level interpretation of the raw data (change of priorities during alarm events) [81], and on-node processing for feature extraction [82], which may additionally encourage or prevent neighbouring nodes from similar actions. The principles of self-organisation can also be exploited for evaluating the performance of BSNs. This can be achieved, for example, by assessing the trade-off between the desired transmission rate and the actual capacity of the BSN. This is essential for addressing Quality of Service (QoS) challenges of the BSN involving continuous sampling under dynamic movement.

To highlight how self-organising principles can be exploited in the application layer, we will show in this section an example that relates application-dependent parameters to QoS characteristics of a BSN. The problem considered here involves distributed self-regulation of the transmission period of a BSN with respect to the desired QoS. In particular, we consider a set of N on-body sensors that transmit their sampled information towards a sink node in a single-hop setup. The objective of the network is to maximise the volume of the information to be transmitted to the sink node (or the network throughput). More specifically, the objective for each node is to self-minimise its sampling period (or maximise the sampling frequency), such that the percentage of successfully received packets PDR (Packet Delivery Ratio) remains higher than a predefined threshold. The bio-inspired approach adopted for addressing this problem is closely related to the reaction-diffusion mechanism as explained in Sect. 10.4.1.

The self-regulation of the sampling period T_i at each node i relies on the network activity monitoring during short evaluation periods, henceforth called Network Monitoring Periods (NMP). Under the assumption of a typical medium access mechanism that supports the transmission of acknowledgement (ACK) packets, the i-th node records the number of successfully transmitted data packets within each NMP. When the NMP expires, node i then locally evaluates the quality of its communication with the sink by calculating the value of PDR_i. If $PDR_i \geq \gamma$, then the sampling period T_i is decreased for the next NMP. Otherwise, the sampling period T_i is increased.

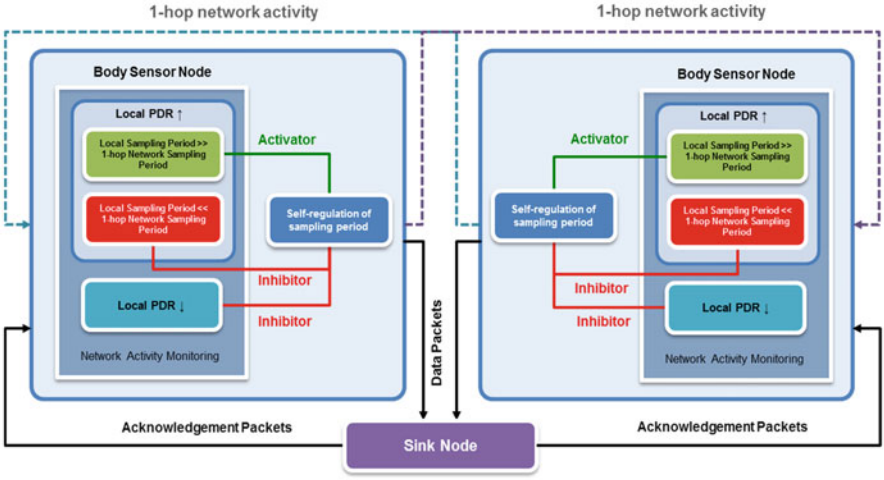


Fig. 10.15 Self-organisation of the sampling period for a set of BSN nodes. The activator-inhibitor based mechanism that considers both the local PDR (Packet Delivery Ratio) and the 1-hop network activity

Note that the activator refers to the positive feedback; the local estimation of the QoS allows for a further decrease of the sampling period, therefore allowing the evaluation of the network performance at more competitive network conditions. In contrast, the inhibitor is interpreted as the negative feedback, since the reduction of the PDR below the predefined threshold indicates the fact that the requested bandwidth is not available at the given network conditions. Therefore, the sampling period needs to be increased.

The self-regulation scheme described above considers only the QoS of the isolated links established between a BSN node and the sink node. Depending on the medium access mechanism, this may result in highly unbalanced allocation of the common network resources, therefore introducing highly varied sampling periods between different nodes. In order to make the activator-inhibitor based self-regulation scheme more resilient to these network conditions, a second network feedback is introduced as shown in Fig. 10.15. This feedback is associated with the sink-related activities of the BSN nodes, where the i-th node can passively receive within the vicinity of its 1-hop neighbourhood. More specifically, during a NMP, each node records the mean value of the sampling period $\tilde{T}_i$ within its 1-hop neighborhood of body sensor nodes. When the NMP expires with a satisfactory value of $PDR_i \geq \gamma$, then the node will examine the difference between T_i and $\tilde{T}_i$. If the deviation of T_i from $\tilde{T}_i$ is not significant ($\leq D_T$), then the activator will be stimulated. Otherwise, in order to balance the access of the common propagation channel from all operational nodes, T_i is set equal to $\tilde{T}_i$ for the following NMP.

This scheme has been experimentally evaluated in a bio-motion estimation application. A number of BSN sensors are placed on joint segments of a volunteer,

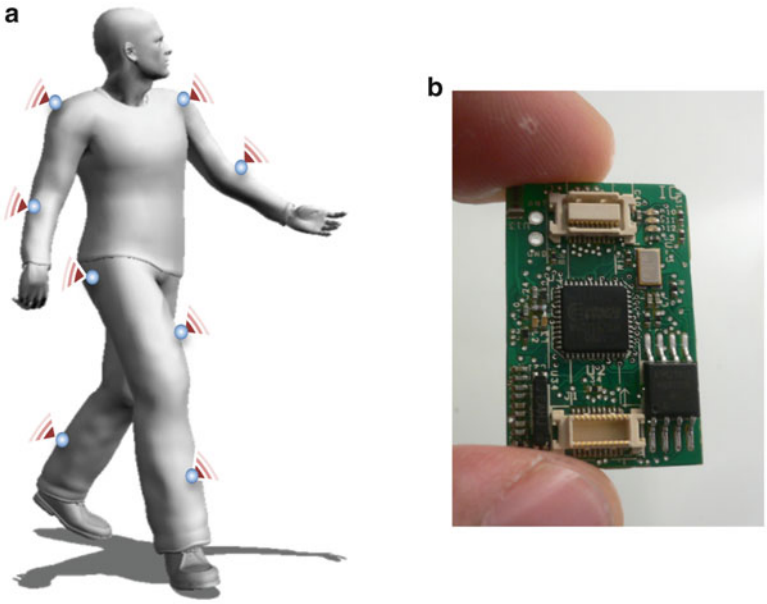


Fig. 10.16 The experimental setup for evaluating the self-regulation of the transmission period with respect to the performance of the network. (**a**) The on-body deployment of the BSN nodes, (**b**) the BSN platform used for the experiment, designed and developed by Imperial College London [83]

as shown in Fig. 10.16. Each sensing node is equipped with 9-axes inertial and magneto sensors for estimating attitude of the person. Details on using such wearable sensors for bio-motion reconstruction will be discussed in Chap. 12 and we only demonstrate here some of the autonomic principles involved.

In this example, the sensor nodes are networked with a sink node, from which the acquired data is collected for further processing. Each node records the value of the locally calculated PDR during fixed-length NMPs, which includes the related information ($T_i, \tilde{T}_i$ and PDR_i) to the packets disseminated towards the sink node. The on-body sensor network was tested with an indoor cycling activity to allow repetitive and controlled experiments. In the experiment, the total number N of sensor nodes was gradually increased from 3 to 8. The BSN nodes were networked with the sink using an IEEE 802.15.4 network using an un-slotted CSMA/CA medium access mechanism and a star topology. The self-regulation parameters are set as follows: (a) the duration of NMP is set to 2,000 ms; (b) the upper and lower thresholds for T_i are 30 ms and 100 ms respectively; (c) the activator/inhibitor regulation step is $T_i/10$; (d) the value of D_T is 15 m and (e) the PDR threshold γ is set to 85 %.

An example on the self-regulation procedure implemented on the body sensor nodes is shown in Fig. 10.17. The inhibitor is stimulated either when the local value of the PDR drops below the value of threshold γ (for instance,

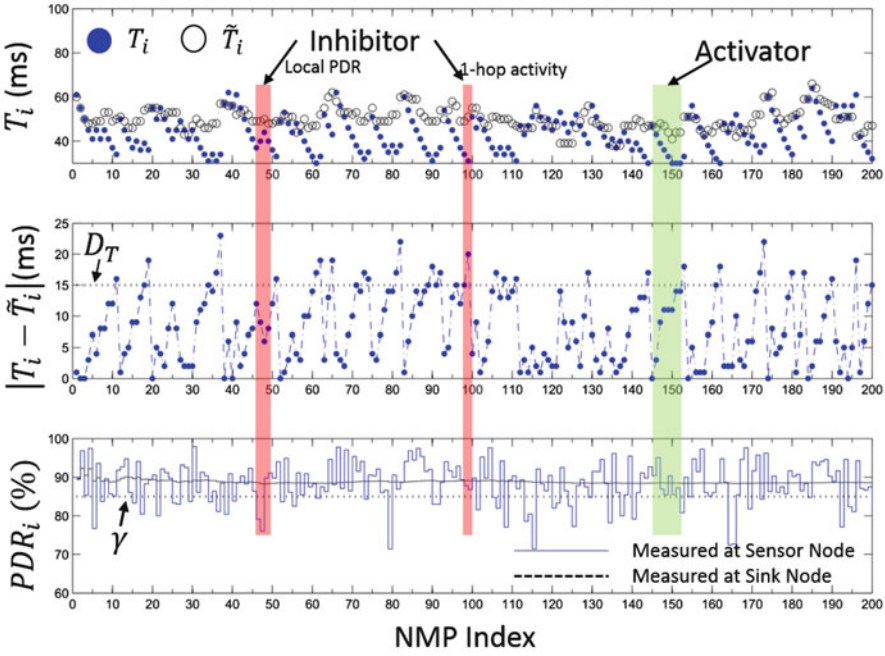


Fig. 10.17 A snapshot on the self-regulation of the sampling period $\boldsymbol{T_i}$ (*top diagram*), accompanied by the difference between $\boldsymbol{T_i}$ and $\tilde{T}_i$ (*middle diagram*) and the corresponding (*bottom diagram*) PDR_i, for the sensor node placed on the right ankle

when NMP = 46–48), or when the 1-hop network activity indicates high deviation between T_i and $\tilde{T}_i$ (for instance, when NMP = 99–100). The activator is triggered according to local PDR and 1-hop network activities. As shown in the top diagram of Fig. 10.17, this may result in a successive decrease of the sampling period, until either the minimum threshold of the sampling period is achieved, or the inhibitor mechanism forces the relaxation of the transmission rate.

Figure 10.18 highlights the benefits of the sampling period self-regulation when the inhibitor-activator mechanism combines local estimation of the PDR with the monitoring of the network activity within the 1-hop neighbourhood. More specifically, both activator-inhibitor schemes preserve the PDR above 80 %. It can be seen that the difference between the mean values of T_i and $\tilde{T}_i$ exceeds 32 ms ($N \geq 6$) when only the local PDR is considered for the self-regulation, and the mean value of $\left|T_i - \tilde{T}_i\right|$ is less than 3 ms in this fully integrated activator-inhibitor mechanism.

The overall performance of the body sensor network based on the bio-inspired self-regulation mechanism is shown in Table 10.3. The activator-inhibitor scheme that combines the local estimation of *PDR* and the monitoring of the sink-related activities highlights the limitations posed by the network size to the

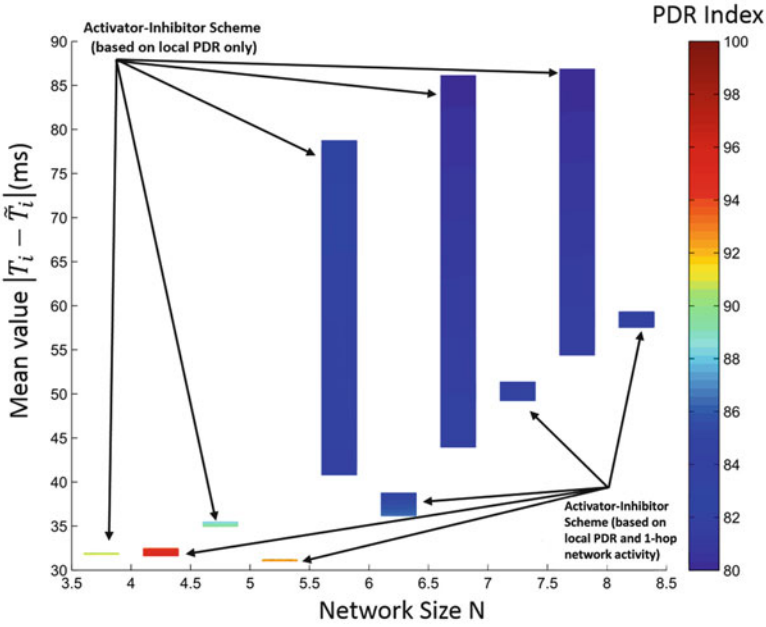


Fig. 10.18 The difference between $\boldsymbol{T_i}$ and $\tilde{T}_i$, for the body sensor placed at the left arm versus the network size, when the activator-inhibitor scheme is either based only on the local value of PDR or combines both the local value of PDR and the 1-hop network activity

Table 10.3 Results of the distributed self-regulation of the sampling period

Network size	Sampling period T_i(ms)	Packet delivery ratio PDR_i (%) measured at sink node	Packet delivery ratio PDR_i (%) measured at sensor nodes
3	34.12 ± 0.12	94.25 ± 1.15	91.45 ± 0.58
4	31.35 ± 0.24	93.92 ± 2.20	93.28 ± 1.92
5	31.27 ± 0.42	93.08 ± 1.63	92.33 ± 1.41
6	36.20 ± 4.00	88.62 ± 3.60	87.57 ± 3.76
7	48.92 ± 4.55	86.78 ± 2.94	85.62 ± 3.37
8	57.31 ± 6.36	86.23 ± 3.61	85.48 ± 3.96

data-transmission rate. As such, when $N = 4$, the mean value of the *PDR* remains higher than 90 %, while the self-regulated sampling period is below 32 ms. In contrast, when $N = 8$, the activator-inhibitor scheme preserves the mean value of *PDR* above the predefined threshold at the expense of increasing the mean value of the sampling period to approximately 57 ms.

This example highlights how network performance can be optimised based on local interactions between the operational nodes. This self-regulation has direct influence on the sensing quality and contextualisation of the raw information, as it allows a more balanced access of the BSN sink. These self-organising attributes can be beneficial as the operational environment imposes significant changes to the topology of the network.

The discussion thus far on networking and self-organisation can address a range of network-related challenges for body-centric applications. It should be noted that practical deployment of BSNs should also incorporate resilience in terms of privacy and security. In the following section, the self-protection issues are addressed with respect to their applicability to body-centric platforms.

10.5 Security and Self-Protection

The nature of BSN in handling patient information and coordinating real-time data for both wearable and implantable sensors, can potentially become an ideal target for malicious intervention. It is not difficult to appreciate that the potential impact involved can far exceed the damage caused to desktop computers. For this reason, security and self-protection are an integral part of a BSN design.

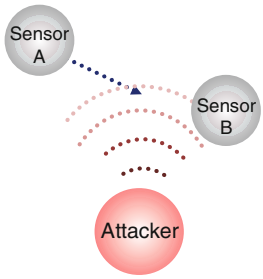
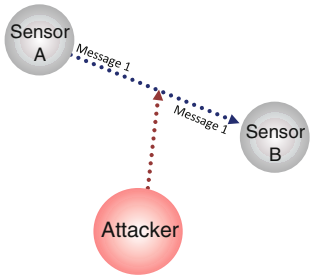
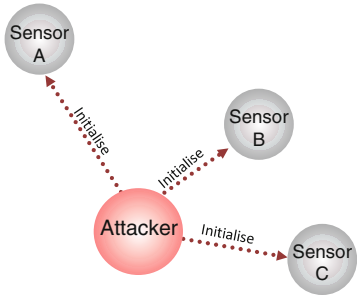
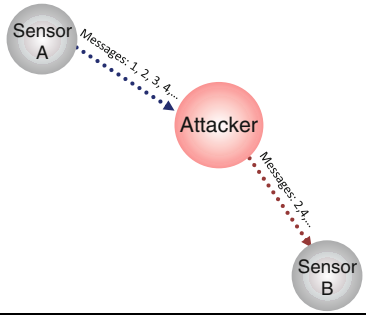
An analogy is often drawn between biological systems and computer systems. In terms of security, "virus" is a well-known term for security attacks on computer systems, as computer viruses share certain characteristics with the real viruses that lead to diseases. In biological systems, infectious disease is caused by a biological agent, called a pathogen. There are four main types of pathogen; including viruses, bacteria, fungi and parasites [84]. A virus is a microscopic parasite [85], which contains only a limited genetic blueprint and is incapable of ordinary reproduction. It can only replicate this by hijacking other biological cells and injecting its RNA (Ribonucleic acid) into the host cell, which will then allow it to reproduce and create more virus cells. A bacterium, on the other hand, is a cellular organism that can reproduce itself. It attacks the host by releasing toxins which could damage a cell or block the transmission of cellular signals. In a BSN, bacteria can be seen as compromised sensor nodes that attack the sensor network integrity, whereas a virus can be regarded as data packets or malicious programs injected into a sensor node with the intent to damage the sensor and thus the network.

To outline some of the main security considerations for BSNs, we will investigate in this section some of the identifiable threats to BSNs by following a bio-inspired framework. We will also discuss related protocols in the WSNs and draw a parallel between the AIS (Artificial Immune System) and the future design of BSNs with effective security and self-protection measures.

10.5.1 Bacterial Attacks

In this subsection, we will use the analogy of bacterial attack to describe attacks that require the subordination of at least one sensor node (or a sensor node emulated by a more powerful device, such as a laptop computer) to overcome the self-protective system and lead to system failure. Thirteen possible bacterial attacks summarised in Table 10.4.

Table 10.4 The bacterial attacks which are likely to compromise the privacy and security of a BSN

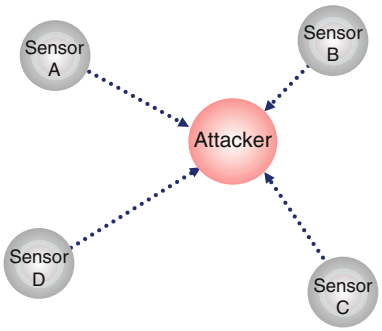
Jamming: A low level technique to disrupt the service of a wireless network is through jamming the frequency band of the wireless devices. However, in the case of BSNs, due to the ubiquitous nature of the application and the short range radio design, jamming can only cause localised interruption to the service of the system. In addition, the use of frequency hopping schemes, such as the *Direct Sequence Spread Spectrum* (DSSS), means that the effect of jamming is minimal	
Collision: Similarly to physical layer attack, attackers can jam the data transmission path by corrupting the packets at the link layer [86, 87]. For example, an attacker can disrupt the checksum of a packet causing retransmission, thus leading to collisions in the network	
Exhaustion and interrogation: For critical messages, sensor node has to retransmit the message repeatedly until it reaches the sink, if the message itself or the acknowledgement message is corrupted. Through corrupting the acknowledgement messages or send false acknowledgement messages, an attacker can cause endless retransmission and eventually bring the network down. In addition, as in most wireless network protocol designs, there are a number of power consuming commands, such as network initialisation and time synchronisation. If an attacker continuously broadcasts those commands, the battery power of the sensors will soon be exhausted. This repeated solicitation of energy-draining responses is called interrogation [87]	
Selective forwarding: In a multi-hop network, messages are expected to be forwarded to the sink through multiple hops. In a selective forwarding attack, a compromised node may selectively reject certain messages, or may randomly drop messages, thus causing data loss and triggering costly network recovery mechanisms [87]	

(continued)

Table 10.4 (continued)

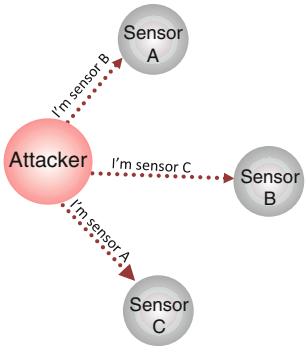
Sinkhole attacks:

In a sinkhole attack, an attacker attempts to lure the network traffic through a compromised node, and does this by making the compromised node look especially attractive with respect to the routing algorithm [88]. Once the sinkhole is created, the network is opened to other attacks, such as selective forwarding or eavesdropping



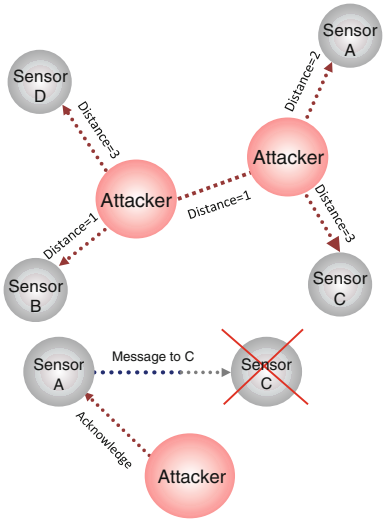
Sybil attacks:

To form a network and create routing tables, every sensor node has to have a unique identity. In a Sybil attack [89], a malicious node creates fake identities, in order that it can appear in multiple places at the same time. It will therefore be more likely to be selected as part of a hop for forwarding messages, thus opening the gate for selective forwarding attacks



Wormholes:

In a wormhole attack, attackers cooperate to simulate a low-latency communication link [87, 90]. Messages received by one of the attackers will be replayed by the other attacker, so that the message appears to be forwarded from a nearby hop. As such, the neighbouring sensor nodes will favour the attacker for routing [91], similarly to a sinkhole. In addition, the network will be severely disrupted when the wormhole is removed



Acknowledgment spoofing:

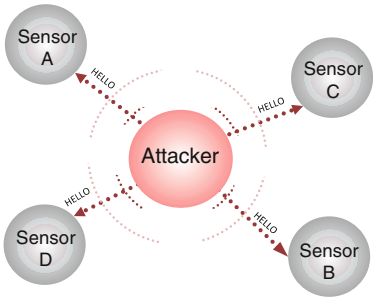
To ensure the integrity of the link and improve the reliability of the data transmission, acknowledgement messages are often required from the receiver. In an acknowledgement spoofing attack, the malicious node spoofs acknowledgement messages aiming to trick the sender into believing that a weak link is strong or a dead node is alive [88]. As typical routing protocols select hops based on the reliability of the link, the attacker can effectively launch a selective forwarding attack by encouraging the target node to transmit packets through those weak links

(continued)

Table 10.4 (continued)

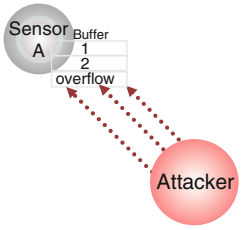
Hello flood attacks:

To introduce a new sensor and update the routing table dynamically, many wireless network protocols require the new sensor to broadcast an announcement message, the HELLO message, to notify its neighbours of its request to join the network. In a HELLO flood attack, an adversary broadcasts powerful HELLO messages to many nodes in the network, so that every node thinks the attacker is within a short radio range [88]. This will cause a large number of nodes to attempt to use the attacker as a hop to route messages, thus confusing the entire routing system



Buffer overflow:

As limited storage is available in each sensor node, flooding a sensor node with messages will lead to buffer overflow and subsequently crash the node



Network scanning:

Through scanning the network and obtaining the network topology, the attacker can locate a particular sensor node or even the individual wearing the sensors

Traffic analysis:

As the network traffic pattern is often used to detect and defend attacks on the network; through observing the traffic patterns, the attacker can potentially compromise the security defence of the system

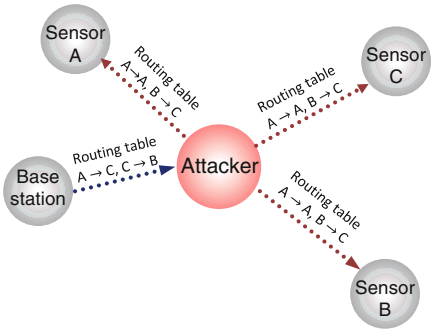
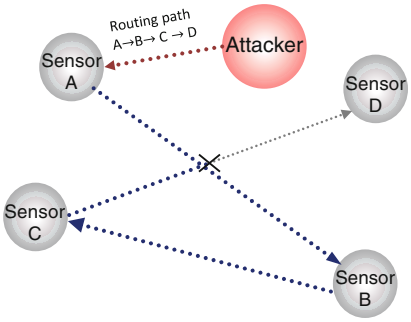
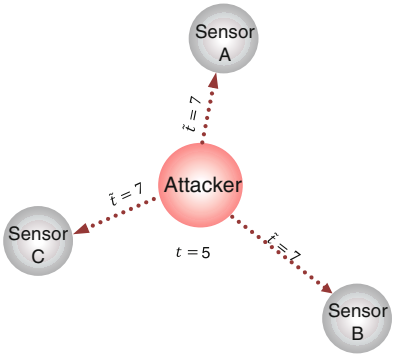
False alarms:

Being a pervasive monitoring system, BSNs are required to accurately capture abnormal events and transmit alarm messages to the system. By generating numerous false alarm messages, an attacker can effectively undermine the system's ability to detect genuine abnormal events

10.5.2 Viral Infection

In a biological viral infection, the virus hijacks a cell by injecting its RNA into the cell. Similarly, an attacker can use a compromised sensor node like a virus and inject malicious data packets or programs into sensor nodes. Unlike bacterial attacks, which do not alter the internal properties of the sensors, viral infections damage the system by altering the parameters or programs of the nodes. Table 10.5 outlines seven possible viral infections which are likely to compromise the security and privacy of BSN.

Table 10.5 Infections which are likely to compromise the privacy and security of a BSN

Corrupting the routing table: The most direct and effective approach to attacking a network is to corrupt the routing table. By spoofing, altering or replaying the routing table, attackers can severely damage the sensor network by creating routing loops, attracting or repelling network traffic, and making false partitions in the network [88]	
Misdirection: In a misdirection attack, an attacker corrupts the message forwarding paths by advertising false routing updates which could isolate sensor nodes or flood a victim	
Time synchronisation corruption: Due to the high bandwidth requirement and limited battery power available to miniaturised sensors, one approach to developing an energy efficient network is to implement a scheduling or *Time Division Multiple Access* (TDMA) scheme for data transmission, such as the TD-DES protocol [92]. With TDMA, the power consumption can be optimised by switching on the radio only when it is required. Collision is also avoided with this scheme. As the timing information of the TDMA protocol is crucial for scheduling, in order to launch an attack on such a network, an attacker can simply corrupt the time synchronisation. One method of achieving this is to broadcast invalid synchronisation commands to the sensors, where a slight offset of the time could potentially lead to collisions. In addition, if dynamic timeslot allocation is enabled on the network, malicious sensor nodes can request an excessive amount of bandwidth and cause the scheduling to fail	

(continued)

Table 10.5 (continued)

Worms:
Although very limited resources are available on a wireless sensor, a network worm attack is possible on a sensor node. Dynamic reprogramming of a sensor node is possible via software tools, such as Deluge [93]. Based on its source dissemination approach, worms can be easily created and spread across the network based on the program distribution framework, thus causing severe damage to the system
Trojan horse:
As with worms, Trojan horse programs can be transmitted to sensor nodes to corrupt sensor data, override control of the sensor node, or damage the network by using the compromised node
Backdoor:
Programs or routines can be uploaded to the sensor nodes by the attackers to open up a backdoor for enabling future access to the network and data
Hoaxes:
An attacker can compromise network security by sending false warning messages regarding security attacks, which can not only trigger energy consuming recovery or protection processes, but also reduce the ability of an adaptive protective system to prevent further virus attacks. As the system may disregard the hoax attacks after identifying them as being hoaxes, the system may fail to capture them when they become real attacks rather than hoaxes

10.5.3 Secured Protocols

Unlike traditional computer systems, very limited resources are available for BSN nodes, and traditional cytological algorithms, encryption techniques and secure protocols cannot be directly applied to BSN networks. There are a number of secured protocols that have been developed within the Sensor Networks community. One example of a secured protocol design for WSN is the *Security Protocols for Sensor Networks* (SPINS) introduced by Berkeley. It mainly consists of two components: μTESLA (*micro version* of the *Timed, Efficient, Streaming, Loss-tolerant Authentication Protocol*) and SNEP (*Secure Network Encryption Protocol*) [94] for unicast and broadcast messages. In addition, to cater for the ad hoc nature of sensor networks, different key distribution schemes, such as the pairwise key distribution method, are proposed, accompanied by approaches that exploit the inherited randomness of the BSN operational space, in the form of biometric-based information.

10.5.3.1 SNEP

The SNEP protocol utilises the RC5 block cipher [95] to encrypt unicast messages. To fit the RC5 onto the severely constrained sensor node, SNEP implements only a subset of functions of RC5 [94]. To ensure the confidentiality of the data, a semantic security mechanism is imposed whereby a shared counter is used to encrypt the data with the block cipher. However, instead of transmitting the counter, each sensor of the node pairs keeps a monotonically increasing counter to minimise the overhead imposed. As shown in Fig. 10.19, the block cipher is applied to the monotonically

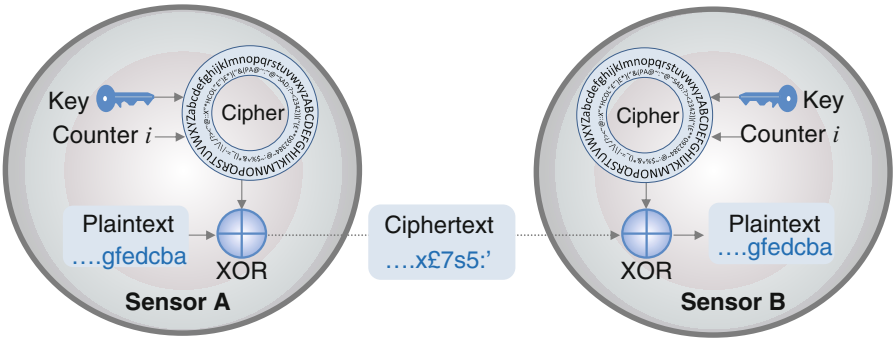


Fig. 10.19 SNEP: counter mode encryption and decryption

increasing counter which will then be XORed with the plaintext to generate the ciphertext. The same process is also applied to decrypt the message [94]. In this way, an eavesdropper will not be able to reconstruct the messages and the counter used can also guarantee the freshness of the messages. In addition, a separate master key is used in each sensor node, so that the authentication of the messages can be ensured. To authenticate the message, the Cipher Block Chaining Message Authentication Code (CBC-MAC) technique is adopted and the same block cipher is used to compute the Message Authentication Code (MAC).

10.5.3.2 µTESLA

Although SNEP provides secure communication between a pair of sensor nodes, SNEP does not support message broadcasting due to the use of different master keys in each sensor node. The µTESLA, which is a micro-version of the TESLA protocol [96], is proposed for secured data broadcast [94]. In µTESLA, the base station generates a chain of secret keys where the last key of the chain is picked randomly and all the other keys are then generated by applying a one-way function F, as shown in the following equations:

$$\begin{aligned} K &= \{K_0, K_1, K_2, \ldots, K_n\} \\ K_n &= random \\ K_i &= F(K_{i+1}) \;\; where \;\; i = 0, \ldots, n-1 \end{aligned} \tag{10.33}$$

As function F is a one-way function, any key from 0 to $i-1$ can be derived by the key K_i, but none of the other keys (K_i, to K_n) can be determined. To securely broadcast messages, the µTESLA utilises a loose time synchronisation scheme where different keys are used to encrypt the packets as shown in Fig. 10.20, where key K_i is used to encrypt the packets sent in time slot t_i. In addition, the initial key K_0 is first disclosed to the sensor nodes by using a secured link provided by SNEP. Subsequent key K_i is not disclosed until a certain interval after the time

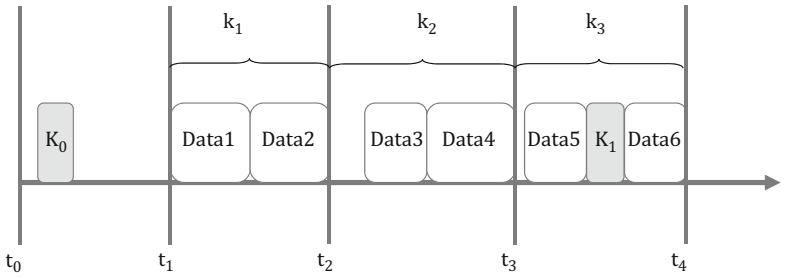


Fig. 10.20 μTESLA data authentication

interval t_i. As shown in the figure, K_1 is disclosed at t_3, and once K_1 is disclosed, the receiver can then authenticate the data packet 1 and 2 by verifying that $K_0 = F(K_1)$.

As the initial key has to be distributed as a unicast message individually to each sensor node, this process can be energy inefficient. Liu et al. proposed encoding the initial key K_0. onto the sensor nodes before deployment instead of distributing the key wirelessly [97]. In addition, Liu et al. proposed a hierarchical organisation of the keys to reduce the memory required in the base station to store the key chain.

10.5.3.3 Cryptography for Ad Hoc Links

The SPINS protocols described above are mainly designed for sensor networks with a static architecture. For BSN applications, however, sensors are often distributed in an ad hoc fashion where sensors may join or disconnect from sensor clusters dynamically. For instance, when a patient with on-body sensors is walking through a care home equipped with environmental sensors, the body sensors may attach to certain clusters of the care home network depending on the location of the patient. In addition, instead of routing all of the messages back to the server, certain sensors in an ad hoc network could request messages, and sensors can also broadcast messages to other nodes. In the case of a patient in a homecare environment, ambient sensor data can be requested by the on-body sensors to validate the findings of wearable sensors, e.g., to confirm if the patient has had a fall. In such scenarios, SPINS-like approaches will not be able to provide the necessary secured ad hoc links.

Due to the limited resources of the sensor nodes, a symmetric key system is often adopted for ad hoc sensor networks where a key is shared between the sensor nodes for data encryption and authentication. One efficient key distribution approach is the pairwise key management system, as proposed in Liu et al. and Du et al. [98–100]. In a pairwise key management system, a selection of keys are initially assigned to each sensor node [101]. To form secured links, sensor nodes announce and compare their keys, and a connection is established if the same key is found on both sensor nodes.

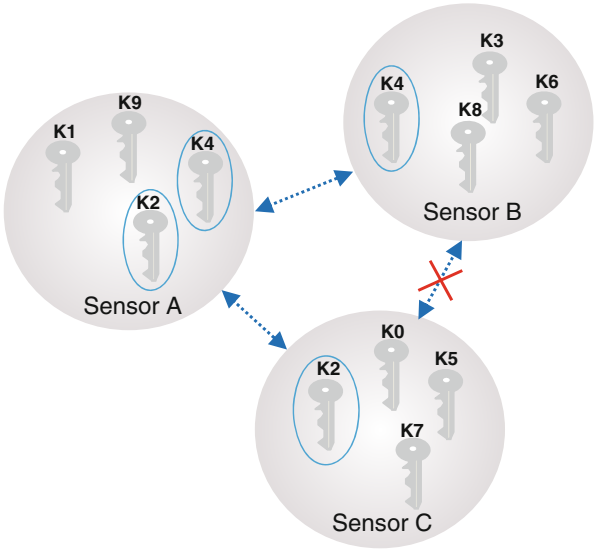


Fig. 10.21 A schematic illustration of the pair-wise key management scheme

Figure 10.21 illustrates how secured links are formed by using the pairwise key management scheme. As shown in the diagram, sensors A and B share the same key K4. Therefore, a secured link is established between the sensor nodes. On the other hand, as no common key is found between sensors B and C, no direct connection can be established. However, since sensors A and C share the same key K2, a secured link can be established between A and C. Once secured links are formed, new keys can be generated to connect disjointed nodes, such as nodes B and C in Fig. 10.21.

With the above key-sharing scheme, a subset of nodes can potentially be opened to security threats as each sensor node contains a subset of keys. To improve the resilience against attack on the subnet, an improved pairwise scheme is proposed by Chan et al. [102]. Instead of connecting nodes if only one common key is found, the proposed method requires the nodes to share a certain number (q) of keys in order to form a direct link. Figure 10.22 illustrates the improved keying scheme with the q-composite key distribution scheme. In this case, $q = 2$ and a secured link is formed between sensors A and C as they both share K_2 and K_4.

Although an asymmetric key scheme is often considered to be too expensive for miniaturised sensor nodes [103], it is argued that public key cryptographic key generation is necessary for sensor networks [104]. In addition, Malan et al. demonstrated the possibility of using asymmetric key distribution schemes for sensor networks [105], where the Elliptic Curve Cryptography key generation was implemented on a MICA-2 sensor node.

10.5.3.4 Biometrics-Based Cryptography

While the aforementioned approaches on distributed schemes for key management compensates for the problems related to computational constraints for the general

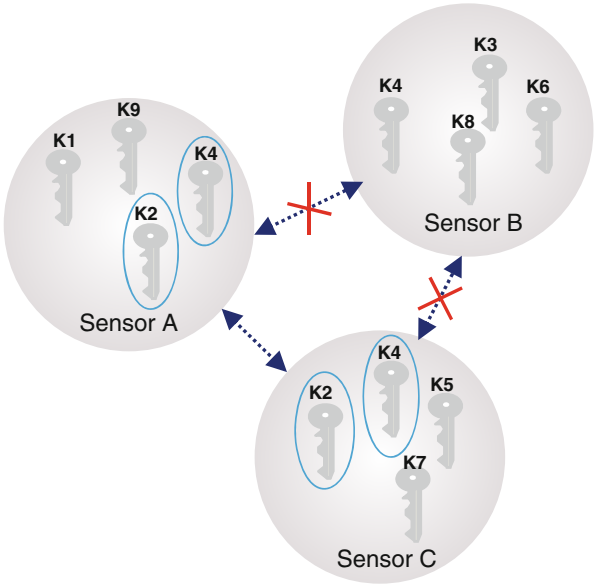


Fig. 10.22 *Q*-composite key distribution scheme

case of sensor networks, they may not be practical for BSN applications. In particular, driven by their moderate size, the majority of the approaches found in recent literature, considers centralised symmetric cryptography schemes [106] by also assuming "always-on" on body – off body connectivity. These can further be extended to include features for linking between customised users and sensors [107].

With regard to security and privacy, BSNs have a great advantage when compared to remaining types of sensor networks; they monitor biometric information, which is characterised by an inherent randomness and uniqueness that sophisticated cryptography for computer networks attempt to articulate with heavy computational schemes. As such, research has exploited the biometric information for encryption purposes. Studies by Uludag et al. [108], Cherukuri et al. [109], and Szczepanski et al. [110], for example, are in favour of encrypting information based on the biometric data, as it additionally promotes highly personalised schemes. Based on these observations, the work presented in Raazi et al. [111] describes a centralised architecture for wireless body-area networks that uses biometric information for key management and refreshment. Although this approach indicates an example of utilising biometrics for cryptography, it does not promote self-protection for two reasons: (a) biometric information utilised for keys management is a-priori acquired and (b) this scheme heavily relies on human intervention in case of attack or network failures. Thus, biometric-based cryptography does not imply self-protection. More specifically, self-protection based on biometrics is achieved when the corresponding scheme relies on fresh, self-generated and evolving information for data encryption, signature and authentication. Based on the recent

trends, there are two main categories of biometric-based cryptography rely on extracting biometric:

Biometrics – Time Domain: The time elapsed between successive nerve impulses is called *Inter-Pulse Interval* (IPI). It consists of a versatile biometric feature, as it can be obtained with different types of sensors and from different physiological signals, such as ECG, PPG, heart sounds, blood pressure wave and blood flow. As such, IPI is available to a wide range of biosensors. In addition, it can be measured with both wearable and implantable devices.

Due to its simplicity and low computational cost, the use of IPI as the means of cryptography based on biometrics has been a favourable approach. Poon et al. experimentally proved in [112] that the IPI-related information, derived from two different types of physiological signals, delivers trait with solid cryptographic qualities for a body-area network. More specifically, it is sufficiently different between different individuals and between different capture slots. In addition, it is time-variant, thus increasing the level of security.

Based on these observations, IPI-related information that is extracted from a heartbeat can be used to generate a unique entity identifier per sensor node in a distributed way [113]. An example of use of these entity identifiers is the synthesis of protocols for establishing symmetric keys, resulting in secure protocols [114]. More specifically, the concept relies on securing the data by using an arbitrary key. The secured data is then transmitted along with the key, which is blurred by the IPI value. The receiver uses its own IPI information for revealing the key, and, subsequently, for decrypting the data.

The concept of keys being locally derived, yet globally converging to the same value is appealing. This forgoes the need for distributing and pre-deploying keys, especially in the case of implanted networks. However, the noise and variance in the measurement of biometric data may degrade the quality of the security provided. The elimination of pre-deployed keys can again be based on the self-emergent value of the IPI, measured at different locations of the body. The associated error compensation relies on a fuzzy commitment scheme, which extends the communication between two nodes A and B in order to increase the network throughput needed for robust authentication [115]. This further implies that node A will send to node B: (a) multiple measurements of the locally calculated IPI and (b) the value of the bit-coded hash function, for comparison against the value of the hash function calculated at node B.

IPI-based cryptography is accompanied by synchronisation constraints. The resulting problems can be compensated by summarizing and verifying a key that is employed at the extremes of a communication link [116]. This is achieved by using *Gaussian Mixture Models* (GMM). The IPI is summarised as the authentication key by using a GMM. Prior to transmission, data is scrambled with the IPI. The outcome is then modelled with the GMM. The resulting message is augmented by the characteristics of the GMM and sent to the receiver. At the receiver end, the local likelihood is compared against the one received, and if the difference is below

a predefined threshold, the verification is considered successful. Otherwise, the received information is regarded as contaminated and discarded.

Biometrics – Frequency Domain: It is also possible to use frequency domain information for biometric-based encryption, for example the *Photoplethysmograph* (PPG) signal, which describes the volumetric change in the distension of arteries, due to the perfusion of blood though them, during a cardiac cycle. PPG can be measured with a pulse oximeter, attached on fingers, earlobes, forehead, etc.

Utilising PPG as the means of cryptography implies that features can be extracted from it for encryption purposes. This feature extraction can be based on Fast Fourier Transform (FFT)), which is further utilised for encrypting a random number [117, 118]. In a way similar to the IPI case, PPG signal is locally utilised by each sensor node in order to extract features in the frequency domain. These features are further exploited for hiding a randomly generated key, which is shared between the transmitter and the receiver. Utilising locally calculated value of the PPG-based features then reveals the key. The procedure for hiding and concealing the key relies on the fuzzy vault approach, and therefore on polynomial operations on the bit sequence. The analysis on the distinctiveness and randomness of the resulting symmetric keys highlights the viability of using PPG for biometric based cryptography.

The approaches discussed above indicate the benefits of using biometric-based information for securing BSN communications. The feature extraction both in time and frequency domain eliminates the need for key pre-deployment and maintenance. In addition, the IPI itself can also solve some of the hardware problems that are responsible for increasing the underlying complexity [119]. However, IPI-based cryptography relies on the assumption of a synchronised network during the sampling of the related biometric data. Schemes that are based on communication between the nodes for resolving authentication issues are efficient, however they can also introduce network overhead. This can be compensated by the GMM-based processing, which, however, is accompanied by the computational overhead introduced.

The advantage of the feature extraction on the frequency domain is that the frequency components of physiological signals can have similar value regardless of the position of sampling. This is in contrast to the time domain analysis, which is characterised by similar trends but higher deviations at different body positions. Therefore, using frequency domain analysis relaxes the synchronisation constraints, which can be met when time domain analysis is used. As such, after features extraction, the cryptography scheme has less demand on the robustness against noise and de-synchronisation effects, than a scheme that relies on IPI. However, as the authors explain in [120], extracting frequency-domain PPG- features can be computational demanding for conventional low power micro-controllers. Consequently, characterising the trade-off between computational power and the robustness of features extraction is necessary.

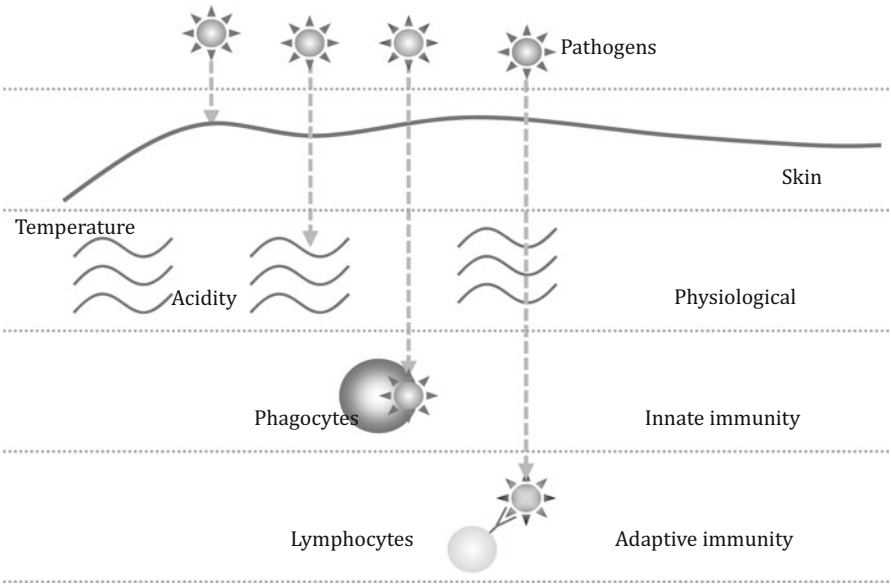


Fig. 10.23 The basic architecture of human immune system

10.5.4 Self-Protection

An ideal model of a self-protection system is the human immune system, as it is an extremely effective defence mechanism that is capable of preventing the onset of infection from approximately 1,016 different molecules [121]. In addition to being able to identify and destroy antigens autonomously, the immune system is able to adapt to virus mutation. From an architectural perspective, the BIS can be viewed as a multiple layered system where each layer is independently equipped with different defence mechanisms [84, 122–124] as schematically illustrated in Fig. 10.23. They include:

- *Physical barrier* – the first line of defence of the immune system is the skin, and also mucus coating of the gut and airway, which physically blocks pathogens from entering the host. In addition, the respiration system also helps in keeping pathogens out of the system by trapping irritants in nasal hairs, coughing and sneezing.
- *Physiological barrier* – the physiological properties of the human body, such as temperature and acidity, actually present a hostile environment for many pathogens.
- *Innate immune system* – this is composed of the build-up of phagocyte cells which can engulf pathogens, and the complement system which is made up of several plasma proteins. The plasma proteins normally circulate in an inactive form and are sequentially activated when an antigen is detected in order to

Table 10.6 The analogies between BIS and AIS

Mechanism	BIS	AIS
Recognising antigens	The BIS recognises antigens by using the bio-receptors of the immunity cells, and it can recognise known antigens with only partial matches, meaning that it can detect marginally mutated antigens	In order to incorporate this very effective defence mechanism, AIS often implements a certain fuzziness in recognising viruses and only a certain short sequence of the virus is examined [125]. Various matching rules for virus identification are proposed by Harmer et al. [128]
Eliminating antigens	In the BIS, antibodies neutralise antigens by binding to the microorganisms, and T-cells kill infected host cells to prevent spreading of the virus	In AIS, once a known virus is detected, conventional virus recovery processes will be used to remove the infected entity from the system, such as destroying the infected file. To prevent self-replication, a "kill signal" mechanism for AIS isolates the infected host and thus prevents it from passing its infected message to its neighbours [125]
Adapting to new antigens	BIS uses a training mechanism to train T-cells to recognise antigens	A negative selection mechanism enables the system to identify antigens while not attacking its own cells [126]. In addition, Kephart proposed the use of virtual environments and "decoy" programs to analyse the characteristics of the virus and develop antibodies accordingly [125]

eliminate the microorganism. This leads to cytolysis, inflammation, and other immune responses. During an inflammatory response, the body temperature will rise and the blood flow increase, fever being a possible consequence.

- *Adaptive immune system* – this consists of two systems: (a) Humoral immune system: the production of antibodies by B cells in response to antigens and (b) Cellular immune system: this recognises and destroys infected cells with T-cells.

One possible approach to introducing self-protection to BSNs is to follow the general principles of the human immune system as illustrated in the concept of a bio-inspired *Artificial Immune System* (AIS) [125, 126], which has been successfully utilised for the case of computer networks [127]. Based on the human immune system, AIS mainly implements three main protective mechanisms, summarised in Table 10.6.

Since the architecture of BSNs are much closer to the biological system than that of personal computers, the self-protection systems of BSNs could be modelled closer to the BIS. For instance, instead of the centralised approach in certain AIS methods, the distributed characteristics of the BIS can be adopted by the BSN. To explain the immune system concept for BSNs, the layered architecture of the BIS can be used, as shown in Fig. 10.24. This includes mechanisms, which are

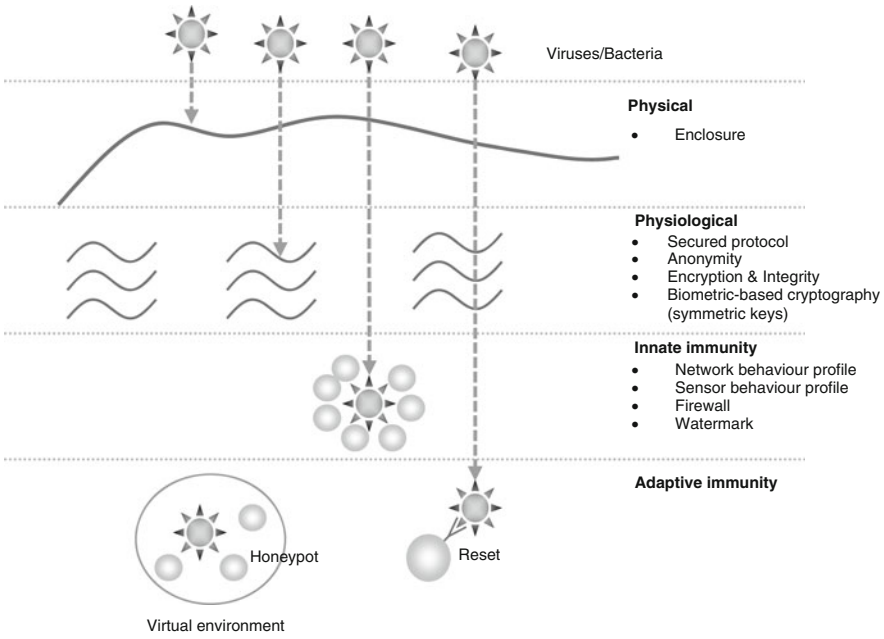


Fig. 10.24 The architecture of a BSN immune system

located at the Physical and Physiological Barrier, and the Innate and Adaptive Immune system and are summarised in Table 10.7.

A representative example of self-protected sensor networks is based on the analogy between innate immunity and forward and backward secrecy [129]. Considering a set of sensors that operate in an autonomous and unattended environment, similar to the case of unobtrusive BSN, the objective is to preserve the secrecy of the collected information either before, or after the potential compromise by an adversary. Considering the case of read-only adversary model, an adversary can simultaneously occupy a subset of nodes (occupied nodes) and compute the secret keys of a previously occupied nodes (sick nodes). The proposed self-healing approach, referred to as DISH, relies on public-key cryptography and collaboration between nodes. The idea is to let healthy sensors, which are neither occupied nor sick, to assist sick sensors to regain their health. Affected sensors ask for contributions from healthy sensors, which in return provide a share for a new key. An affected sensor uses contributions from healthy sensors, along with its current key, as input to a one-way function to generate a new key. As long as there is at least one contribution from a healthy sensor, the adversary is unable to learn the new key. Consequently, a previously sick sensor will become healthy after a key update.

This scheme has a strong relevance to the innate immune system, and the way that an infected cell self-destructs by its interaction with the T-cells. As such, sick

Table 10.7 The mechanisms per BIS layer for developing self-protected BSN

Physical barrier	Sensor enclosures can provide physical protection against tampering, and for implantable sensors, biocompatible casing will be required to protect the sensor from the human immune system
	The short range radio design of a BSN will limit the spread of virus and effectively lower the probability of being infected
Physiological barrier	A secured network protocol will be able to identify and prevent certain service attacks. This can create an uninhabitable environment for antigens, like the physiological barrier of in the BIS
	The use of anonymity and transmitting raw sensor data could discourage eavesdroppers, because deriving the identity of the subject and the context of the data will require significant amounts of effort and time
Innate immune system	To engulf an antigen, neighbouring sensor nodes can form a guard to isolate the attacker. For instance, the sensor nodes can act as sinkholes to the antigen, so that malicious packets or programs cannot be distributed in the network
	The BIS inflammation concept can be used in BSNs to slow down the spread of antigens
Adaptive immune system	In order to recognise antigens, instead of keeping a record of all known antigens at a central storage, each sensor node could hold one or more sets of antigen information, similarly to the B-cells in the BIS. To investigate a suspicious packet or program entity, the signature of the packet or program (such as the checksum), could be broadcast to the B-cells to let them compare it with the known antigen records
	In eliminating infected hosts, the infected node can be reset to a trusted program or even switched off upon receipt of the self-destruct message from a T-cell, as there will be redundant nodes around that would serve its same function, similarly to biological cells
	To adapt to new antigens, a virtual environment can be created by allocating a small group of sensors to extract the characteristics of the virus [6]. In the virtual environment, a decoy sensor node (honeypot) can be assigned to lure the virus to infect the node, and then the characteristics of the virus can be extracted by monitoring the infected node's interaction with other nodes in the virtual environment

nodes represent the infected cells and healthy nodes represent T-cells. In addition, the contribution of each healthy node to the reconstruction of a new secure key for affected nodes represents the plasma proteins model. Consequently, in a way similar to the immune system described above, the infected host returns to a safe condition, by resetting its key upon the reception of corresponding messages originated by healthy nodes.

Finally, it should be noted that, although having an immune system for BSNs can provide protection against attacks, the nature of the immune system could cause adverse effects to the network itself. Overreaction to certain stimuli, analogously to the case of human allergies, could bring certain parts or the whole BSN system into disarray. Nevertheless, these problems represent an interesting research issue for the BSN community in the coming years.

10.6 Conclusions

In this chapter, we have discussed the use of autonomic principles for self-healing, self-organisation, and self-protection in developing BSNs with effective fault tolerance and self-protection. Due to the inherent complexities involved in managing a large number of wireless sensors, bio-inspired sensing and networking is an important area of study and most of the work we discussed in this chapter only represents the tip of the iceberg.

One of the key principles derived from biological systems is the effective coordination that is possible through some very simple mechanisms in information exchange and message passing; and how this can be used to achieve a global behaviour that is adaptive and self-governing. In sensor networks, in addition to sensor noise, bias and node failures, uncertainties due to an imperfect understanding of the system to be monitored, as well as incomplete knowledge of the state of the environment at the time of the interaction are important factors to consider. For self-organisation, as well as fault detection and self-healing, belief propagation represents an attractive method for implementing some bio-inspired concepts in sensor networks. This is because the method has a compact representation, is distributed, and robustness to noise and network degradation. It has also been shown that belief propagation is highly effective for asynchronous communication and is suitable for heterogeneous networks.

In this chapter, we have also outlined some of the basic principles of message passing in a belief network. The main benefits of probabilistic graphical representation to model sensor networks also include the ease of integrating heterogeneous data from different sensors, the possibility of continuously improving the accuracy of the system by learning from available data and the advantage of reasoning under uncertainty, thus permitting the incorporation of high-level and domain specific knowledge into the distributed inferencing framework.

It is worth noting that the self-healing mechanism that we have discussed in relation to autonomic sensing is often referred to as the survivability of the system. One of the key emphases of self-healing is how the system can perform gracefully under deteriorating environmental conditions and system hardware. As mentioned in previous chapters, whilst the perpetual powering of the sensor nodes through energy scavenging remains an active research topic, our current focus should be directed at the effective processing and routing strategies that can maximise overall power efficiency. For the practical deployment of BSNs, it is also important to consider a proactive approach towards fault tolerance, a strategy that is often relied upon by biological systems to improve their immunity to system failures.

As mentioned in the introduction to this chapter, the focus of our discussion has been mainly restricted to the illustration of some of the design principles of autonomic sensing based on the features derived from the ANS and BIS. Our aim here is not to reverse-engineer these biological systems but only to investigate some of their design principles; so that plausible computational architectures suitable for the BSN can be developed. In this regard, the discussion about security and self-protection is only intended to outline the challenges, as well as the opportunities, faced by the BSN research community.

References

1. H. Ellis, Clinical anatomy: a revision and applied anatomy for clinical students, 10th ed., 2002.
2. D. Male, J. Brostoff, D. B. Roth, and I. M. Roitt, Immunology, 8 ed.: Elsevier Ltd, 2012.
3. D. Male, Immunology – an illustrated outline, 4 ed.: Mosby, Elsevier Science Limited, 2004.
4. D. D. Chaplin, "Overview of the immune response," Journal of Allergy and Clinical Immunology, vol. 125, pp. S3–S23, 2010.
5. J. O. Kephart and D. M. Chess. (January 2003) The vision of autonomic computing. IEEE Computer Magazine. 41–50.
6. J. O. Kephart and D. M. Chess, "The vision of autonomic computing," Computer, vol. 36, pp. 41–50, 2003.
7. R. Murch, Autonomic Computing, 1 ed.: IBM Press, 2004.
8. V. Venkatasubramanian, R. Rengaswamy, K. Yin, and S. N. Kavuri, "A review of process fault detection and diagnosis. Part I: Quantitative model-based methods," Computers and Chemical Engineering, vol. 27, pp. 293–311, 2003.
9. A. A. Lazar, W. Wang, and R. H. Deng, "Models and algorithms for network fault detection and identification: a review," in Singapore International Conference on Communications/ International Symposium on Information Theory and Its Applications, ed, 1992, pp. 9999–1003.
10. L. Kant, A. S. Sethi, and M. Steinder, "Fault localization and self-healing mechanisms for FCS networks," in the Twenty-Third Army Science Conference, Orlando, Florida, 2002.
11. F. R. Kschischang, B. J. Frey, and H.-A. Loeliger, "Factor graphs and the sum-product algorithm," IEEE Transactions on Information Theory, Special Issue on Codes on Graphs and Iterative Algorithms, vol. 47, pp. 498–519, 2001.
12. B. J. Frey, "Extending factor graphs so as to unify directed and undirected graphical models," in the Nineteenth Conference on Uncertainty in Artificial Intelligence, Acapulco, Mexico, 2003.
13. J. H. Kim and J. Pearl, "A computational model for causal and diagnostic reasoning in inference engines," in the Eighth International Joint Conference on Artificial Intelligence, Karlsruhe, West Germany, 1983, pp. 190–193.
14. J. Pearl, "A constraint-propagation approach to probabilistic reasoning," in the First Annual Conference on Uncertainty in Artificial Intelligence, Rome, New York, USA, 1985, pp. 357–370.
15. J. Pearl, Probabilistic reasoning in intelligent systems: networks of plausible inference: Morgan Kaufmann, 1988.
16. G. F. Cooper, "The computational complexity of probabilistic inferences," Artificial Intelligence, vol. 42, pp. 393–405, 1990.
17. J. Pearl, Probabilistic reasoning in intelligent systems: networks of plausible inference. San Mateo, CA: Morgan Kaufmann, 1988.
18. K. Murphy. (1998). A brief introduction to graphical models and Bayesian networks. Available: http://www.cs.ubc.ca/~murphyk/Bayes/bayes_tutorial.pdf
19. C. K. Chow and C. N. Liu, "Approximating discrete probability distribution with dependence trees," IEEE Transactions on Information Theory, vol. IT-14, pp. 462–267, 1968.
20. P. P. Shenoy and G. Shafer, "Propagating belief functions using local computations," IEEE Expert, vol. 1, pp. 43–52, 1986.
21. G. Shafer and P. P. Shenoy, "Probability propagation," Annals of Mathematics and Artificial Intelligence, vol. 2, pp. 327–352, 1990.
22. P. P. Shenoy and G. Shafer, Axioms for probability and belief-function propagation. San Francisco, CA: Morgan Kaufmann, 1990.
23. F. V. Jensen, S. L. Lauritzen, and K. G. Olesen, "Bayesian updating in causal probabilistic networks by local computation," Computational Statistics Quarterly, vol. 4, 1990.

24. S. Lauritzen and D. Spiegelhalter, "Local computations with probabilities on graphical structures and their application to expert systems," Journal of the Royal Statistical Society B, 1988.
25. R. Dechter, "Bucket elimination: a unifying framework for several probabilistic inference algorithms," in the Twelfth Conference on Uncertainty in Artificial Intelligence, Portland, Oregon, USA, 1996.
26. F. Cozman, "Generalizing variable elimination in Bayesian networks," in Workshop on Probabilistic Reasoning in Artificial Intelligence, ed. Atibaia, Brazil, 2000.
27. R. Shachter, "Intelligent probabilistic inference," in the First Annual Conference on Uncertainty in Artificial Intelligence, ed. UCLA, Los Angeles, CA, 1985.
28. R. Shachter, "Evidence absorption and propagation through evidence reversals," in the Fifth Conference on Uncertainty in Artificial Intelligence, ed. University of Windsor, Ontario, 1989.
29. Z. Li and B. D'Ambrosio, "Efficient inference in Bayes network as a combinatorial optimization problem," International Journal of Approximate Reasoning, vol. 11, 1994.
30. A. Darwiche, "A differential approach to inference in Bayesian networks," in the Sixteenth Conference on Uncertainty in Artificial Intelligence, ed. Stanford, Palo Alto, CA, 2000.
31. R. D. Shachter and M. A. Peot, "Simulation approaches to general probabilistic inference on belief networks," in the Fifth Conference on Uncertainty in Artificial Intelligence, University of Windsor, Ontario, 1989.
32. S. Geman and D. Geman, "Stochastic relaxation, Gibbs distribution and the Bayesian restoration of images," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 6, pp. 721–741, 1984.
33. D. MacKay, "Introduction to Monte Carlo methods," in Learning in Graphical Models, M. I. Jordan, Ed., ed: MIT Press, 1998.
34. H. J. S. H. Horvitz, and G. F. Cooper, "Bounded conditioning: flexible inference for decisions under scarce resources," in the Fifth Conference on Uncertainty in Artificial Intelligence, ed. Windsor, Ontario, 1989.
35. R. Dechter, "Mini-buckets: a general scheme for generating approximations in automated reasoning in probabilistic inference," in the Fifteenth International Joint Conference on Artificial Intelligence, Nagoya, Japan, 1997.
36. K. Murphy, Y. Weiss, and M. I. Jordan, "Loopy belief propagation for approximate inference: an empirical study," in the Fifteenth Conference on Uncertainty in Artificial Intelligence, ed, 1999, pp. 467–475.
37. Y. Weiss, "Correctness of belief propagation in Gaussian graphical models of arbitrary topology," Neural Computation, vol. 13, 2001.
38. M. J. Wainwrght and M. I. Jordan, "Graphical models, exponential families and variational inference," University of California, Berkeley, Technical Report 649, 2003.
39. H. Guo and W. Hsu, "A survey of algorithms for real-time Bayesian network inference," in Joint Workshop on Real-Time Decision Support and Diagnosis Systems, ed. Edmonton, Canada, 2002.
40. C.-K. Kwoh and D. F. Gillies, "Probabilistic reasoning and multiple-expert methodology for correlated objective data," Artificial Intelligence in Engineering, vol. 12, pp. 21–33, 1998.
41. C.-K. Kwoh and D. F. Gillies, "Using hidden nodes in Bayesian network," Artificial Intelligence, vol. 88, pp. 1–38, 1996.
42. M. Chu, S. K. Mitter, and F. Zhao, "An information architecture for distributed inference on ad hoc sensor networks," in the Forty-First Annual Allerton Conference on Communication, Control, and Computing, ed. Monticello, IL, USA, 2003.
43. M. Alanyali, S. Venkatesh, O. Savas, and S. Aeron, "Distributed Bayesian hypothesis testing in sensor networks," in the 2004 American Control Conference, 2004, pp. 5369–5374.
44. A. T. Ihler, J. W. Fisher, R. L. Moses, and A. S. Willsky, "Nonparametric belief propagation for sensor network self-calibration," IEEE Journal on Selected Areas in Communications, 2005.

45. C. Crick and A. Pfeffer, “Loopy belief propagation as a basis for communication in sensor networks,” in the Nineteenth Conference on Uncertainty in Artificial Intelligence, Acapulco, Mexico, 2003.
46. M. A. Paskin, C. E. Guestrin, and J. McFadden, “A robust architecture for distributed inference in sensor networks,” in the Fourth International Symposium on Information Processing in Sensor Networks, 2005.
47. M. A. Paskin and C. E. Guestrin, “Robust probabilistic inference in distributed systems,” in the Twentieth ACM International Conference on Uncertainty in Artificial Intelligence, ed. Banff, Canada, 2004, pp. 436–445.
48. L. Atallah, D. McIlwraith, S. Thiemjarus, B. Lo, and G. Z. Yang, “Distributed inferencing with ambient and wearable sensors,” Wireless Communications and Mobile Computing, vol. 12, pp. 117–131, 2012.
49. P. S. Hall and Y. Hao, Antennas and Propagation for Body-Centric Wireless Communications: Artech House Inc, 2006.
50. S. Thiemjarus, B. P. L. Lo, and G. Z. Yang, “A distributed Bayesian framework for body sensor networks,” in the Second International Workshop on Body Sensor Networks, Imperial College, London, 2005.
51. S. Thiemjarus, B. P. L. Lo, and G. Z. Yang, “A noise resilient distributed inference framework for body sensor networks,” in the Third International Conference on Pervasive Computing, ed. Munich, Germany, 2005, pp. 13–18.
52. E. Reusens, W. Joseph, B. Latré, B. Braem, G. Vermeeren, E. Tanghe, L. Martens, I. Moerman, and C. Blondia, “Characterization of on-body communication channel and energy efficient topology design for wireless body area networks,” Information Technology in Biomedicine, IEEE Transactions on, vol. 13, pp. 933–945, 2009.
53. A. Natarajan, M. Motani, B. de Silva, K.-K. Yap, and K. C. Chua, “Investigating network architectures for body sensor networks,” in Proceedings of the 1st ACM SIGMOBILE international workshop on Systems and networking support for healthcare and assisted living environments, 2007, pp. 19–24.
54. A. Natarajan, B. de Silva, K.-K. Yap, and M. Motani, “Link layer behavior of body area networks at 2.4 ghz,” in Proceedings of the 15th annual international conference on Mobile computing and networking, 2009, pp. 241–252.
55. A. Natarajan, B. De Silva, K.-K. Yap, and M. Motani, “To hop or not to hop: Network architecture for body sensor networks,” in Sensor, Mesh and Ad Hoc Communications and Networks, 2009. SECON’09. 6th Annual IEEE Communications Society Conference on, 2009, pp. 1–9.
56. S. Ullah, H. Higgins, B. Braem, B. Latre, C. Blondia, I. Moerman, S. Saleem, Z. Rahman, and K. S. Kwak, “A comprehensive survey of wireless body area networks,” Journal of Medical Systems, vol. 36, pp. 1065–1094, 2012.
57. B. Latré, B. Braem, I. Moerman, C. Blondia, and P. Demeester, “A survey on wireless body area networks,” Wireless Networks, vol. 17, pp. 1–18, 2011.
58. A. M. Turing, “The chemical basis of morphogenesis,” Bulletin of mathematical biology, vol. 52, pp. 153–197, 1990.
59. M. Durvy and P. Thiran, “Reaction-diffusion based transmission patterns for ad hoc networks,” in INFOCOM 2005. 24th Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings IEEE, 2005, pp. 2195–2205.
60. A. Tyrrell, G. Auer, and C. Bettstetter, “Biologically inspired synchronization for wireless networks,” in Advances in Biologically Inspired Information Systems, ed: Springer, 2007, pp. 47–62.
61. J. Degesys, I. Rose, A. Patel, and R. Nagpal, “DESYNC: self-organizing desynchronization and TDMA on wireless sensor networks,” in Proceedings of the 6th international conference on Information processing in sensor networks, 2007, pp. 11–20.

62. J. Degesys and R. Nagpal, "Towards desynchronization of multi-hop topologies," in Self-Adaptive and Self-Organizing Systems, 2008. SASO'08. Second IEEE International Conference on, 2008, pp. 129–138.
63. H. Hernández and C. Blum, "Distributed graph coloring: an approach based on the calling behavior of Japanese tree frogs," Swarm Intelligence, vol. 6, pp. 117–150, 2012.
64. R. E. Mirollo and S. H. Strogatz, "Synchronization of pulse-coupled biological oscillators," SIAM Journal on Applied Mathematics, vol. 50, pp. 1645–1662, 1990.
65. R. Pagliari, Y.-W. Hong, and A. Scaglione, "Bio-inspired algorithms for decentralized round-robin and proportional fair scheduling," Selected Areas in Communications, IEEE Journal on, vol. 28, pp. 564–575, 2010.
66. R. Pagliari, A. Scaglione, and R. A. Tannious, "Design of a distributed protocol for proportional fairness in wireless body area networks," in Global Telecommunications Conference (GLOBECOM 2010), 2010 IEEE, 2010, pp. 1–6.
67. R. A. Tannious and A. Scaglione, "Physical layer design of distributed proportional fair scheduling in wireless body area networks," in Personal, Indoor and Mobile Radio Communications Workshops (PIMRC Workshops), 2010 I.E. 21st International Symposium on, 2010, pp. 78–83.
68. W. R. Heinzelman, A. Chandrakasan, and H. Balakrishnan, "Energy-efficient communication protocol for wireless microsensor networks," in System Sciences, 2000. Proceedings of the 33rd Annual Hawaii International Conference on, 2000, p. 10 pp. vol. 2.
69. S. Lindsey and C. S. Raghavendra, "PEGASIS: Power-efficient gathering in sensor information systems," in Aerospace conference proceedings, 2002. IEEE, 2002, pp. 3-1125–3-1130 vol. 3.
70. R. C. Shah and J. M. Rabaey, "Energy aware routing for low energy ad hoc sensor networks," in Wireless Communications and Networking Conference, 2002. WCNC2002. 2002 IEEE, 2002, pp. 350–355.
71. A. Manjeshwar and D. P. Agrawal, "TEEN: A Routing Protocol for Enhanced Efficiency in Wireless Sensor Networks," in IPDPS, 2001, p. 189.
72. A. Manjeshwar and D. P. Agrawal, "APTEEN: A Hybrid Protocol for Efficient Routing and Comprehensive Information Retrieval in Wireless Sensor Networks," in ipdps, 2002, p. 48.
73. B. Braem, B. Latre, I. Moerman, C. Blondia, and P. Demeester, "The wireless autonomous spanning tree protocol for multihop wireless body area networks," in Mobile and Ubiquitous Systems: Networking & Services, 2006 Third Annual International Conference on, 2006, pp. 1–8.
74. T. Watteyne, I. Augé-Blum, M. Dohler, and D. Barthel, "Anybody: a self-organization protocol for body area networks," in Proceedings of the ICST 2nd international conference on Body area networks, 2007, p. 6.
75. X. Ji and H. Zha, "Sensor positioning in wireless ad hoc sensor networks using multidimensional scaling," in the IEEE INFOCOM, 2004.
76. J. W. Sammon, "A nonlinear mapping for data structure analysis," IEEE Transactions on Computers, vol. C-18, pp. 401–409, 1969.
77. R. Yan, D. Ball, A. Deshmukh, and R. X. Gao, "A Bayesian network approach to energy-aware distributed sensing," in Sensors, 2004. Proceedings of IEEE, 2004, pp. 44–47.
78. Y. Xiang, D. Poole, and M. P. Beddoes, "Multiply sectioned Bayesian networks and junction forests for large knowledge based systems," Computational Intelligence, vol. 9, pp. 171–220, 1993.
79. Y. Xiang and V. Lesser, "Justifying multiply sectioned Bayesian networks," in the Fourth International Conference on Multi-Agent, Boston, 2000, pp. 349–356.
80. R. Yan, D. Ball, A. Deshmukh, and R. X. Gao, "A Bayesian network approach to energy-aware distributed sensing," in IEEE Sensors Conference, Vienna, Austria, 2004, pp. 44–47.
81. F. Dressler, "Self-organized event detection in sensor networks using bio-inspired promoters and inhibitors," in Proceedings of the 3rd International Conference on Bio-Inspired Models of Network, Information and Computing Systems, 2008, p. 10.

82. K. Lorincz, B.-r. Chen, G. W. Challen, A. R. Chowdhury, S. Patel, P. Bonato, and M. Welsh, "Mercury: a wearable sensor network platform for high-fidelity motion analysis," in SenSys, 2009, pp. 183–196.
83. The BSN node. Available: http://ubimon.doc.ic.ac.uk/bsn/index.php?article=167
84. L. N. De Castro and J. Timmis, Artificial immune systems: a new computational intelligence approach: Springer Verlag, 2002.
85. K. Akkaya and M. Younis, "A survey on routing protocols for wireless sensor networks," Ad hoc networks, vol. 3, pp. 325–349, 2005.
86. H. Karl and A. Willig, Protocols and architectures for wireless sensor networks. Chichester, West Sussex: John Wiley and Sons, 2005.
87. A. Wood and J. Stankovic, "A taxonomy for denial-of-service attacks in wireless sensor networks," in Handbook of Sensor Networks, M. Ilyas and I. Mahgoub, Eds., ed Boca Raton, Florida: CRC Press, 2005, pp. 32.1-32.20.
88. C. Karlof and D. Wagner, "Secure routing in wireless sensor networks: attacks an countermeasures," in the First IEEE International Workshop on Sensor Network Protocols Applications, 2003.
89. J. R. Douceur, "The Sybil attack," in the First International Workshop on Peer-to-Peer Systems, 2002.
90. Y. C. Hu, A. Perrig, and D. B. Johnson, "Wormhole detection in wireless ad hoc networks," Department of Computer Science, Rice University, Technical Report TR01-384, 2002.
91. A. D. Wood and J. A. Stankovic, "A taxonomy for denial-of-service attacks in wireless sensor networks," Handbook of Sensor Networks: Compact Wireless and Wired Sensing Systems, pp. 739–763, 2004.
92. U. Cetintemel, A. Flinders, and Y. Sun, "Power-efficient data dissemination in wireless sensor networks," in the Third International ACM Workshop on Data Engineering for Wireless and Mobile Access, San Diego, California, 2003.
93. J. W. Hui and D. Culler, "The dynamic behavior of a data dissemination protocol for network programming at scale," in the Second ACM Conference on Embedded Networked Sensor Systems, ed. Baltimore, Maryland, USA, 2004.
94. A. Perrig, R. Szewczyk, V. Wen, S. Culler, and J. D. Tygar, "SPINS: security protocols for sensor networks," in the Seventh Annual International Conference on Mobile Computing and Networking, Rome, Italy, 2001.
95. R. L. Rivest, "The RC5 encryption algorithm," in the Second Workshop on Fast Software Encryption, ed, 1995, pp. 86–96.
96. A. Perrig, R. Canetti, J. D. Tygar, and D. Song, "Efficient authentication and signing of multicast streams over lossy channels," in Security and Privacy, 2000. S&P 2000. Proceedings. 2000 I.E. Symposium on, 2000, pp. 56–73.
97. D. Liu and P. Ning, "Efficient distribution of key chain commitments for broadcast authentication in distributed sensor networks," in the Tenth Symposium of Network Distributed System, Security, Internet Society, ed. Reston, VA, 2003.
98. W. Du, J. Deng, Y. Han, S. Chen, and P. Varshney, "A key management scheme for wireless sensor networks using deployment knowledge," in IEEE INFOCOM, ed. Hong Kong, 2004.
99. D. Liu and P. Ning, "Establishing pairwise keys in distributed sensor networks," in The Tenth ACM Conference on Computer and Communications Security, Washington D.C., 2003.
100. D. Liu and P. Ning, "Location-based pairwise key establishments for relatively static sensor networks," in the 2003 ACM Workshop on Security of Ad Hoc and Sensor networks, ed. George W. Johnson Center at George Mason University, Fairfax., V.A., USA, 2003.
101. G. Eschenauer and D. Gligor, "A key management scheme for distributed sensor networks," in the Ninth ACM Conference on Computer Communication Security, ed. New York, USA, 2002.
102. H. Chan, A. Perrig, and D. Song, "Random key predistribution schemes for sensor networks," in IEEE Symposium of Security and Privacy, Los Alamitos, CA, 2003.

103. S. Slijepcevic, J. Wong, and M. Potkonjak, "Security and privacy protection in wireless sensor networks," in Handbook of Sensor Networks: Compact Wireless and Wired Sensing System, M. Ilyas and I. Mahgoub, Eds., ed Florida, USA: CRC Press, 2004.
104. B. Arazi, I. Elhanany, O. Arazi, and H. Qi, "Revising public-key cryptography for wireless sensor networks," IEEE Computer, vol. 3, pp. 103–105, 2005.
105. D. Malan, M. Welsh, and M. D. Smith, "A public-key infrastructure for key distribution in TinyOS based on elliptic curve cryptography," in the First Annual IEEE Communication Society Conference on Sensor and Ad Hoc Communications and Networks, 2004, pp. 71–80.
106. C. S. Jang, D. G. Lee, J.-w. Han, and J. H. Park, "Hybrid security protocol for wireless body area networks," Wireless Communications and Mobile Computing, vol. 11, pp. 277–288, 2011.
107. R. Yasmin, E. Ritter, and G. Wang, "A pairing-free ID-based one-pass authenticated key establishment protocol for wireless sensor networks," in SENSORCOMM 2011, The Fifth International Conference on Sensor Technologies and Applications, 2011, pp. 340–347.
108. U. Uludag, S. Pankanti, S. Prabhakar, and A. K. Jain, "Biometric cryptosystems: issues and challenges," Proceedings of the IEEE, vol. 92, pp. 948–960, 2004.
109. S. Cherukuri, K. K. Venkatasubramanian, and S. K. Gupta, "BioSec: A biometric based approach for securing communication in wireless networks of biosensors implanted in the human body," in Parallel Processing Workshops, 2003. Proceedings. 2003 International Conference on, 2003, pp. 432–439.
110. J. Szczepanski, E. Wajnryb, J. M. Amigó, M. V. Sanchez-Vives, and M. Slater, "Biometric random number generators," Computers & Security, vol. 23, pp. 77–84, 2004.
111. S. M. K.-u.-R. Raazi, H. Lee, S. Lee, and Y.-K. Lee, "BARI+: A biometric based distributed key management approach for wireless body area networks," Sensors, vol. 10, pp. 3911–3933, 2010.
112. C. C. Poon, Y.-T. Zhang, and S.-D. Bao, "A novel biometrics method to secure wireless body area sensor networks for telemedicine and m-health," Communications Magazine, IEEE, vol. 44, pp. 73–81, 2006.
113. S.-D. Bao, C. C. Poon, Y.-T. Zhang, and L.-F. Shen, "Using the timing information of heartbeats as an entity identifier to secure body sensor network," Information Technology in Biomedicine, IEEE Transactions on, vol. 12, pp. 772–779, 2008.
114. K. K. Venkatasubramanian and S. K. Gupta, "Physiological value-based efficient usable security solutions for body sensor networks," ACM Transactions on Sensor Networks (TOSN), vol. 6, p. 31, 2010.
115. K. Cho and D. H. Lee, "Biometric based secure communications without pre-deployed key for biosensor implanted in body sensor networks," in Information Security Applications, ed: Springer, 2012, pp. 203–218.
116. W. Wang, H. Wang, M. Hempel, D. Peng, H. Sharif, and H.-H. Chen, "Secure Stochastic ECG Signals Based on Gaussian Mixture Model for e-Healthcare Systems," 2011.
117. K. K. Venkatasubramanian, A. Banerjee, and S. Gupta, "Plethysmogram-based secure inter-sensor communication in body area networks," in Military Communications Conference, 2008. MILCOM 2008. IEEE, 2008, pp. 1–7.
118. K. K. Venkatasubramanian, A. Banerjee, and S. K. S. Gupta, "PSKA: usable and secure key agreement scheme for body area networks," Information Technology in Biomedicine, IEEE Transactions on, vol. 14, pp. 60–68, 2010.
119. G.-H. Zhang, C. C. Poon, and Y.-T. Zhang, "Analysis of Using Interpulse Intervals to Generate 128-Bit Biometric Random Binary Sequences for Securing Wireless Body Sensor Networks," Information Technology in Biomedicine, IEEE Transactions on, vol. 16, pp. 176–182, 2012.
120. A. Banerjee, K. Venkatasubramanian, and S. K. Gupta, "Challenges of implementing cyber-physical security solutions in body area networks," in Proceedings of the Fourth International Conference on Body Area Networks, 2009, p. 18.

121. J. Inman, "The antibody combining region: speculations on the hypothesis of general multispecificity," in Theoretical Immunology, G. Bell, A. Perelson, and G. Pimbley, Eds., ed: Marcel Dekker, 1978, pp. 234–278.
122. B. Rensberger, " In self-defense," in Life Itself, B. Rensberger, Ed., ed: Oxford University Press, 1996, pp. 212–228.
123. C. A. Janeway, P. Travers, M. Walport, and J. D. Capra, Immunobiology: the immune system in health and disease, 4th ed.: Garland Publishing, 1999.
124. S. A. Hofmeyr and S. Forrest, "Architecture for an artificial immune system," Evolutionary Computation, vol. 7, pp. 45–68, 2000.
125. J. O. Kephart, "Biologically inspired immune system for computers," in the Fourth International Workshop on Synthesis and Simulation of Living Systems, 1994, pp. 130–139.
126. S. Forrest, A. S. Perelson, L. Allen, and R. Cherukuri, "Self-nonself discrimination in a computer," in IEEE Symposium on Research in Security and Privacy, 1994.
127. H. Nishiyama and F. Mizoguchi, "Design of security system based on immune system," in the Tenth IEEE International Workshops on Enabling Technologies: Infrastructure for Collaborative Enterprises, ed, 2001.
128. P. K. Harmer, P. D. Williams, G. H. Gunsch, and G. B. Lamont, "An artificial immune system architecture for computer security applications," IEEE Transaction on Evolutionary Computation, vol. 6, pp. 252–280, 2002.
129. D. Ma and G. Tsudik, "Dish: Distributed self-healing," in Stabilization, Safety, and Security of Distributed Systems, ed: Springer, 2008, pp. 47–62.

Chapter 11
Wireless Sensor Microsystem Design: A Practical Perspective

Lei Wang, David R.S. Cumming, Paul A. Hammond, Jonathan M. Cooper, Erik A. Johannessen, and Kamen Ivanov

11.1 Introduction

The development of small sensory systems for use in distributed networks as well as single autonomous devices has been actively pursued for more than 50 years. The stimulant to this activity, as with most modern technologies, was the invention of the transistor that made miniaturisation possible. The earliest wireless sensor devices were developed for use in a range of applications, including wireless animal tracking and medical instrumentation [1]. Although there was considerable activity in the field during the 1950s and 1960s [2], continuing research appeared to recede in subsequent years. However, in the 1970s and 1980s, the microelectronics industry rapidly developed after the invention of the first integrated circuit (IC) microprocessor [3] for use in personal computers and work stations. The rapid growth in consumer electronic products, exemplified by mobile communications and the Internet in the 1990s, made researchers and practitioners realise the potential for personalised wireless systems incorporating location sensitive information and sensor technologies. In essence, these devices owe much to the early work of pioneers, but modern designs will depend heavily on new emerging technologies such as *System-on-Chip* (SoC) and the implementation of mobile (wireless) communication protocols.

L. Wang (✉) • K. Ivanov
Institute of Biomedical and Health Engineering, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen University Town, Shenzhen 518055, PR China
e-mail: wang.lei@siat.ac.cn

D.R.S. Cumming • J.M. Cooper
School of Engineering, University of Glasgow, Glasgow G12 8LT, UK

P.A. Hammond
Dialog Semiconductor, Edinburgh EH1 3DQ, UK

E.A. Johannessen
Førsteamanuensis, Department of Micro- and Nanosystems Technology (IMST), Vestfold University College, Postbox 2243, Tønsberg 3103, Norway

G.-Z. Yang (ed.), *Body Sensor Networks*, DOI 10.1007/978-1-4471-6374-9_11,

Table 11.1 Some key design challenges for wireless sensor micro-systems

Features	Technical challenges
Size	Devices are required to work in a constrained space
Power	Batteries are considered as too large and effective power scavenging methods still remain to be developed
Cost	Limited production numbers make for expensive products
Lifetime	Devices need to operate over extended periods from a limited power supply
Wireless	Must conform to national and international standards; size constraints of antennae, low power systems, signal absorptions (tissue)
Detection	Size constraints of sensors; performance changes over time and environment (drift)
Electronics	Partitioning into system nodes; choice of analogue vs. digital; hardware vs. software
Packaging	Assists miniaturisation; must conform to standards of applications (e.g. medical)
Stimulation	Durability of the electrode/actuator material

An ongoing theme in the execution of microsensor systems has been the use of *Application-Specific Integrated Circuits* (ASICs). Typically, the ASICs used contain less than 100,000 transistors, and are therefore small considered by the standards of SoC designs. This is partly because the ASICs used in sensors are designed by research groups with relatively modest resources. It is also due to the fact that ASICs are designed to perform one or a limited number of tasks in conjunction with a specific sensor function. This design simplicity permits extremely power conservative designs that are essential in the construction of miniaturised sensor systems.

The application of IC and ASIC technologies has also been extended beyond the boundaries of straightforward electronic design. There are considerable advantages in being able to build the sensors onto the same substrate as the electronics. This has been achieved with several chemical sensors [4], and the technology has been used to build sophisticated gas [5] and pH sensors [6]. Such technologies are relatively expensive, but from the point of view of miniaturisation, they eliminate interconnections between internal components and are therefore of particular advantage in the packaging of sensor microsystems.

Wireless sensor microsystems offer very diverse functionality which addresses a range of technical design problems. The different design skills include sensors, ASICs, wireless communication, low power, packaging, software, networking and power sources. Some of these problems and design challenges that are listed in Table 11.1 become more challenging as the devices get increasingly smaller in size.

The recent advances in miniaturised wireless sensor systems have been found to be of special benefit in the development of implantable or ingestible sensors. While some implantable biochips provide relatively simple functions such as identification, there is a growing variety of medical implants intended to monitor, support or replace human body functions. Medical implants need to incorporate many functions, which in turn may lead to quite complex systems. This is in contradiction with the rigorous constraints including size, power consumption and the required operational reliability that the same system is expected to comply with. Furthermore, scaling can also have a practical impact. A wireless system designed to operate in close proximity to, or inside, a human body must not exceed safe power

and radiation limits which may compromise the transmission range. Wireless technology is also constrained by difficulties such as antenna design [7]. However, a clear advantage of miniaturisation is that analytical tools and sensors may be able to complete assays more rapidly due to the relationship between size, diffusion and time [8]. Since wireless sensor microsystems are usually required to be small, it is often the case that functionality is stripped to a minimum, whereas in modern consumer electronics, the trend is often in the opposite direction. Applications of this principle have been found in medical implants [9] and electronic noses [10].

One group of miniature devices are those intended for diagnostics and therapy within human body. In the following section some of the latest achievements in the field of wireless microsystems will be discussed in the context of electronic body implants and ingestible laboratory in-a-pill devices as case studies for discussing many of the attributes one might expect to find in a biomedical wireless microsystem. We will also cover a range of design topics that are of relevance to medical wireless microsystem designs.

11.2 The Endoscopic Capsule

In order to diagnose a wide range of *Gastrointestinal* (GI) dysfunctions, it is useful to be able to make measurements of the physiological conditions inside a patient's gut. Conventional methods include non-invasive techniques such as radiography (X-ray, *Computed Axial Tomography* (CAT), *Positron Emission Tomography* (PET)), ultrasonography and *Magnetic Resonance Imaging* (MRI). Non-invasive methods cannot present any direct information on the biochemical environment within the GI tract, hence endoscopy, a relatively invasive technology, has gained clinical acceptance to view the GI tract and perform biopsies for subsequent analysis [11]. Sedation or a general anaesthetic is often required for animal endoscopy and the considerable inconvenience and irritation of this technique may discourage patients from undergoing the procedure [12]. The technique is also unsuitable for monitoring GI dysfunction since it cannot be used to measure conditions in the GI tract in real-time over an extended period [13]. As a consequence, there is a growing interest in the application of wireless sensor microsystems that conduct real-time measurements in the GI tract.

In 1957, the first two ingestible radiotelemetry capsules were developed independently in Stockholm [14] and New York [15]. They measured approximately 10 mm in diameter and 30 mm in length and were designed to measure temperature and pressure. The pressure sensor consisted of a diaphragm that moved an iron core inside a coil that constituted the tuning element of the single-transistor oscillator circuit that is depicted in Fig. 11.1 [14]. It produced bursts of oscillations in the 100 kHz range; the frequency was a function of pressure, and the burst repetition rate a function of temperature. The frequency and burst repetition rate were recorded externally using a standard radio receiver.

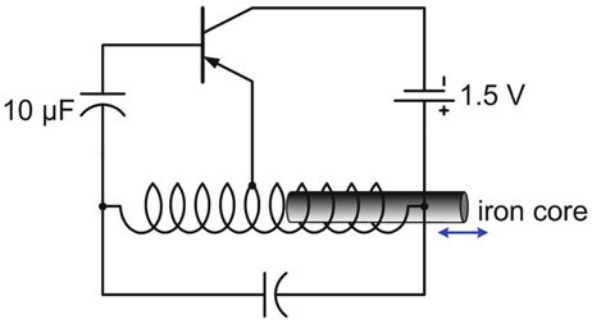


Fig. 11.1 Circuit diagram of the pressure- and temperature sensitive single-transistor oscillator developed by Mackay and Jakobson [14]

Current suppliers of endoscopic capsule systems include Given Imaging [16], Olympus, IntroMedic, Jinshan Science & Technology, RF System Lab and CapsoVision [17, 18]. Many research works still aim to advance the state-of-the-art [9, 19–31]. A temperature monitoring pill has even been developed by NASA for use by astronauts [32].

Commercial radiotelemetry capsules have been developed that have the potential to replace conventional fibre-optic endoscopy and colonoscopy. The M2A capsule (later re-branded as PillCam SB) from Given Imaging Ltd. contains a single-chip *Complementary Metal-Oxide-Semiconductor* (CMOS) image sensor, an ASIC for video transmission and white *Light-Emitting Diodes* (LEDs) for illumination [16]. The data is transmitted to an array of eight antennae worn on a belt that also allow the capsule's position to be localised, it is claimed, to within a resolution of 3.8 cm. The system is particularly well suited to detect bleeding and the software contains blood recognition algorithms to automatically highlight suspect areas. It has been widely used as a diagnostic tool [38], and to date, the PillCam capsule technology has been used by more than 1.5 million patients [33] (Fig. 11.2).

Despite significant advances in capsule technology in recent years, new technologies are emerging rapidly. The desire is to develop a complete diagnostic and therapeutic tool that is small enough to pass through the body cavities without obstruction. However, the use of diagnostic capsules is still limited to the GI tract since they are too large to enter smaller body cavities. Since the size is the main limitation for integrating more functions, the design can be simplified by designing capsules for one particular application. The main functions that are expected from a diagnostic capsule are shown in Fig. 11.3.

Almost all of the functions of the capsule system represent a technical challenge [17, 39], and current clinically and commercially approved solutions rely on relatively simple principles whereas the more sophisticated platforms are still at the experimental stage. Apart from the physical and technological limitations, another reason that prevents the fast utilisation of sophisticated functions is the need to ensure that the new systems are reliable and safe enough to operate within the human body.

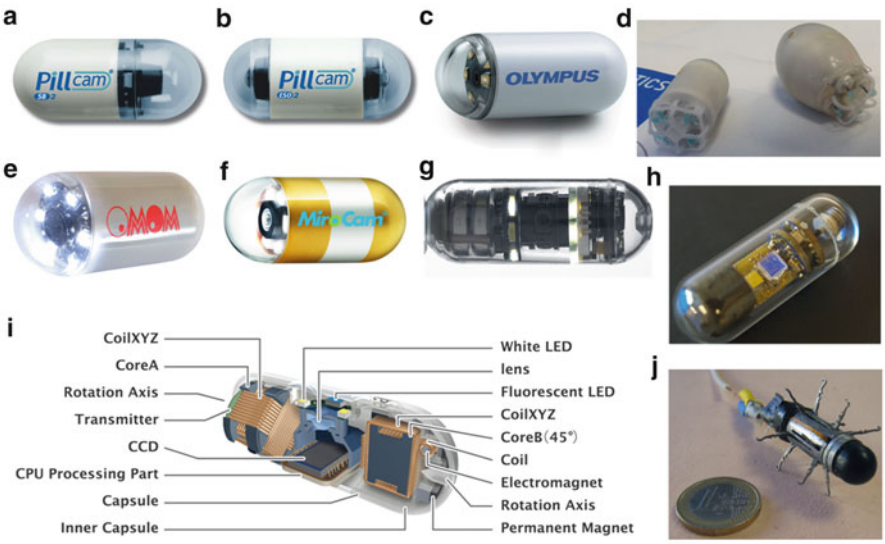


Fig. 11.2 Commercial endoscopic capsules and some research prototypes: (**a**) Small bowel capsule PillCam SB2 [33] (Reprinted with permission from Given Imaging Ltd.) and (**b**) Esophageal capsule PillCam ESO 2 with two cameras [33] (Reprinted with permission from Given Imaging Ltd.) (**c**) EndoCapsule by Olympus (Courtesy of Olympus America Inc.) (**d**) Swimming capsule developed by Tortora et al. [34, 35] (Courtesy of Dr. Gastone Ciuti, the BioRobotics Institute of Scuola Superiore Sant'Anna, Pisa) (**e**) OMOM capsule (Courtesy of Chongqing Jinshan Science and Technology Co., Ltd), (**f**) MiroCam capsule (Courtesy of IntroMedic Co., Ltd), (**g**) CapsoCam SV-1 (Courtesy of CapsoVision, Inc.) (**h**), Capsule prototype that uses autofluorescence detection from intestinal tissue [26] (Courtesy of Prof David Cumming, the University of Glasgow) (**i**) Sayaka capsule by RF SYSTEM lab., at the moment of preparation of this book still in the testing phase [36] (Courtesy of RF SYSTEM lab./RF Co., Ltd.) (**j**) Prototype of mechanism for robotic legged locomotion by Valdastri et al. [37] (Courtesy of Dr. Pietro Valdastri, STORM Lab, Vanderbilt University)

Locomotion is one of the main challenges to be resolved for successful clinical application of capsule systems [17, 30]. The PillCam SB capsule, for example, relies on passive locomotion through natural peristaltic contractions of the bowel [17]. This approach is simple and does not require additional locomotion besides natural peristalsis. However, it limits the possibility of detailed exploration of areas of interest and there is a high risk to miss pathological findings. Active locomotion allows the physician to have much more flexibility during examination. Naturally, the active motion could be initiated internally in the capsule by means of mechanical actuation. The main design problem is how to implement the actuation mechanism within the limited space of the capsule. One of the ideas is to use electro-stimulation of the bowel such that the resulting contractions can help to advance the capsule [40, 41]. A number of projects have also investigated bio-inspired locomotion mechanisms [42] by using piezoelectric deformation or the pseudo-elasticity effect of shape-memory alloys [43]. For active motion of the capsule, an important parameter is if the actuation allows for orientation and

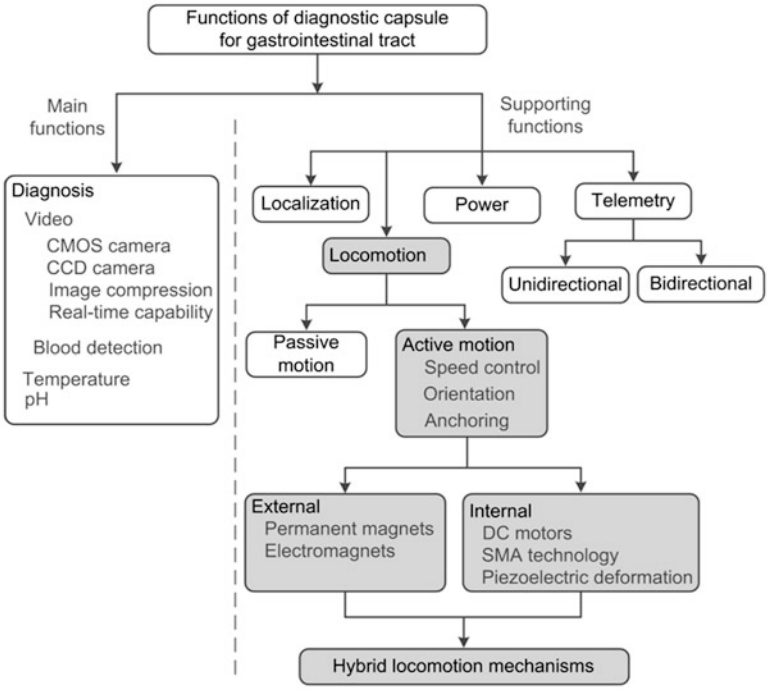


Fig. 11.3 General function of diagnostic capsules

bi-directional movement. Many of the existing ideas are limited to forward movement with limited control for orientation.

Apart from incorporating different actuation mechanisms to the capsules, it is also possible to drive the capsule externally by means of a magnetic link. This approach is much easier for practical implementation and it allows more space in the capsule to be used for other functional units. Constraints of the method include the need to observe the allowed maximum power density near the human body and the potential interferences with other devices or systems nearby. Despite these concerns, systems based on magnetic link are actively pursued both academically and commercially. The OMOM Magnetically Guided Capsule (Jinshan Science and Technology Ltd., Chongqing, China), for example, is controlled through an external permanent magnet which interacts with an on-board permanent magnet of the capsule [17, 44]. Olympus and Siemens have also pursued a joint project for magnetic navigation of capsule endoscopes [17, 45]. A study on feasibility and safety of using magnetically steered capsule in human volunteers was conducted using a specially modified capsule from Given Imaging incorporating small magnets [46]. Robotic magnetic steering was demonstrated by Carpi et al. on animal models [47]. A comprehensive review of the current capsule endoscopy projects is provided by Ciuti et al. [17] and the general requirement and design considerations of ingestible wireless capsules are provided in Fig. 11.4.

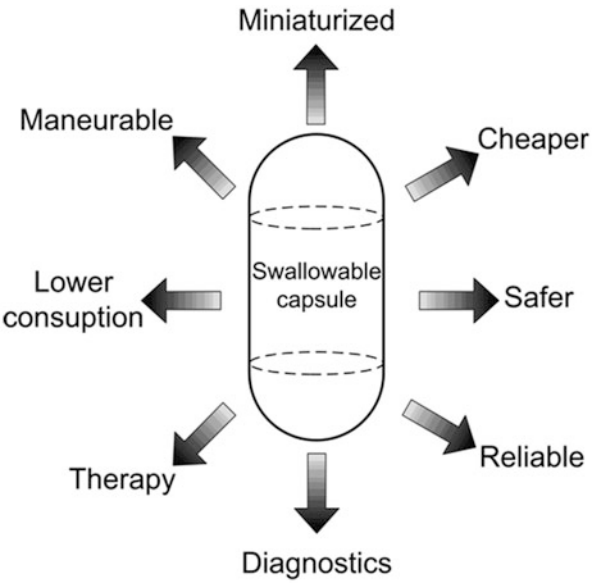


Fig. 11.4 General requirement and design perspectives for ingestible wireless capsules

11.3 Applications for Wireless Capsule Devices

Whilst this chapter is aimed at providing a practical perspective of some of the issues concerning the design of wireless microsystems, in practice, real designs are determined by the application for which they are intended. At present, the dominant measurement capabilities are image acquisition, temperature and pH, although devices capable of sample retrieval, pressure sensing and dissolved oxygen measurement, amongst others, have been explored.

pH is one of the most important physiological parameters that one can measure in any biological system. In terms of pH measurement, capsule-based systems have been employed most successfully to diagnose *Gastro-Oesophageal Reflux Disease* (GERD) [48]. These systems require a tethered capsule to monitor whether stomach acid is refluxing back into the oesophagus and causing the burning sensation commonly known as heartburn. Measurements of pH in the GI tract using 'flow-through' capsules have also been used to study *Inflammatory Bowel Diseases* (IBDs) such as Crohn's disease and ulcerative colitis [49, 50]. Knowledge of the pH profile along the GI tract has also led to the development of drug capsules that dissolve in alkaline environments, ensuring that the drug is only released beyond the acid environment of the stomach.

In 1959, Wolff [51] started the development of radiotelemetry capsules that had the potential of becoming mass produced. This would significantly reduce the production cost per unit and permit a quantity of capsules that would secure

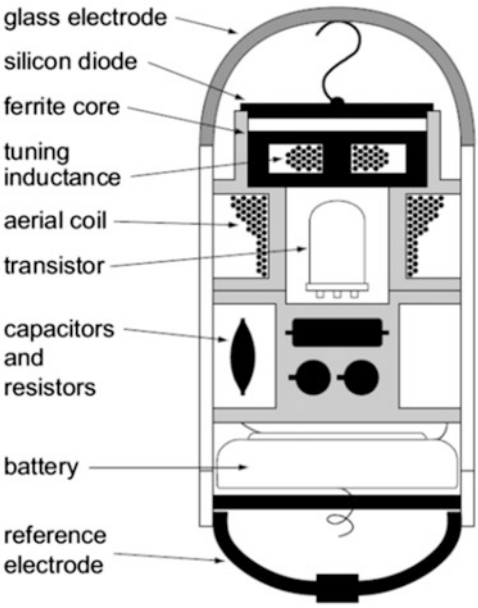


Fig. 11.5 Diagram of the cross-section through a pH-sensitive radio-telemetry capsule

statistically significant data in clinical trials. The capsules were based on single-transistor oscillators, and different versions were made which were capable of measuring pH, pressure and temperature. The temperature-measuring capsule used a coil wound on a core made from a nickel-iron alloy, which exhibited a large change in permeability with temperature. This changed the inductance of the tuning coil and shifted the frequency of transmission in a manner similar to the pressure sensor of Mackay [14]. The pH-measuring capsule used a glass electrode connected to a silicon diode, which operated as the tuning element for the oscillator. The potential of the electrode varied with pH and this voltage changed the capacitance of the diode. A reference electrode, needed to complete the pH circuit, was also included in the capsule, which is shown in Fig. 11.5. A similar pH-sensitive capsule based on a glass electrode was described by Watson and Kay [52]. It was used in the first medical study to plot the pH profile along the entire length of the GI tract [53].

These early capsules were prone to failure due to the ingress of moisture through the epoxy, which was used to seal the glass electrode in place. An improved design that housed the battery and electronics inside the glass electrode was developed by Colson [54]. The receiving antenna array was embedded in a cloth band worn around the waist. Data was collected on a portable recorder that allowed the patient to carry out their daily activities. This equipment was used in a much larger study of 66 subjects [55].

For conditions such as GERD, it is necessary to measure the pH at a fixed location, rather than measuring a flow-through profile. Early studies achieved this by ‘lowering’ the capsule into position on a piece of thread, and fixing the free end to the subject’s cheek when the capsule was in position. More recently, a capsule

has been developed that is temporarily anchored to the wall of the GI tract using an endoscopic delivery system [48] that is relatively invasive. A vacuum pump sucks the tissue into a well in the capsule and a pin is pushed through the tissue holding it in place after the delivery system has been removed. This capsule is 6 mm in diameter and 25 mm long, and it uses an antimony pH electrode. It transmits to an external pager-sized receiver, allowing patients to conduct normal activities without restriction of diet or exercise.

Data-logging telemetry capsules have been used to measure the pH inside the stomach of small animals [56] including penguins [57]. The same technology has been used in large animals such as cattle where the capsule was located in the reticulum. This is the second stomach of a ruminant animal and the measurement was used to identify effect of diet on subclinical acidosis [58]. In fact, livestock monitoring may well be the major market for capsule-based pH sensors. When combined with temperature, the data could be used by farmers to optimise feeding patterns, to detect illness, and to manage breeding. There is already a system available that combines temperature measurement with a *Radio-Frequency Identification* (RFID) chip that is unique to each animal. RFID tagging of livestock provides guaranteed information on the supply of meat from 'farm-to-fork' and standards have been developed for the manufacture of such devices and systems [59].

11.4 Technology

11.4.1 Design Constraints

The design constraints should permit the overall dimensions of a diagnostic capsule system to be small enough to permit the device to pass through all the GI sphincters with relative ease, including the lower oesophageal sphincter and the pyloric sphincter. The capsule must follow low cost methodologies since it will be a disposable device that should only be used once. Low power consumption is a requirement to minimize battery (hence overall) size and thus increase the operating time. A capsule might take a maximum of 8 h to traverse the upper alimentary tract and the small intestine, while a complete passage through the GI tract might take up to 32 h. Using readily available silver oxide battery technology, an energy storage density of 500 mWh/mL can be achieved [60], thus a suitable source, such as two SR48 cells (110 mWh each) could deliver enough energy to complete small intestinal measurements if the average power consumption is less than 20 mW.

The data sampled in the GI tract by a capsule must be retrieved accurately and securely. This usually means that the data must be wirelessly transmitted and correctly received by an external device worn by the patient. As mentioned in Chap. 5, there are a number of radio communication standards encompassing several international *Industrial*, *Scientific and Medical* (ISM) telemetry bands (pan-European medical device frequency allocations [61] and the US Federal

Communications Commission frequency allocations for biomedical telemetry and ISM devices – regulations S5.150, US209 and US350). The main bands of interest are at 418, 434, 868 and 915 MHz.

As with all measuring devices, the user must be confident that the data retrieved is accurate. The problems of accuracy can be dealt with via the normal techniques of instrument design and calibration. However, an additional constraint for wireless devices is that the data must be secure. Since capsule devices operate in the unlicensed ISM frequency bands there is a severe risk of interference from external sources operating at the same transmission frequency that could be particularly dangerous in the context of a medical device. Although the sensors and signal acquisition electronics usually require the use of analogue circuits, early devices relied on the entire system design being analogue, thus making data transfer from the devices extremely insecure. However, modern electronic techniques permit designers to convert the analogue sensor signal to the digital domain within the capsule, enabling the use of secure digital wireless techniques. These techniques ensure that data from any given capsule can be uniquely identified to avoid attributing diagnostic information to the wrong individual. Details of such designs, and others concerning wireless sensor systems, are discussed by Nikolaidis et al. and Park et al. [62, 63].

11.4.2 Microsystem Design

Although the structure and functionality of a capsule system could be diverse, many aspects of the complete solution are illustrated in Fig. 11.6. In the capsule, the sensors are the data gathering devices that are connected to the electronic instrumentation in order to acquire the external signal to be measured. In the earlier devices, all the data was managed by analogue electronics, but more recent devices convert the signals into a digital representation. In this way, a common platform can be developed in which a single controller design can be reused for successive products or different sensor modalities. This is an example of the SoC methodology, in which the majority of the components are connected together on a single chip. Common examples of products containing these SoC devices are mobile telephones, digital television and radio receivers, and computer game consoles. The design of small SoCs is particularly well suited for integration in modern capsule architectures since this device requirement does not permit assembly from ‘large’ off-the-shelf components.

For capsule devices, the use of a digital architecture enables substantially more complex systems to be built that are capable of performing simultaneous measurements from several sensor channels. Figure 11.7 shows an integrated circuit designed for use in a microcapsule system that contains all the electronics required. In this implementation, a small low-power transmitter has been integrated on to the same chip with a usable operating range of only 10–20 cm. Because of the difficulty of building a suitable transmitter onto the same integrated circuit as the

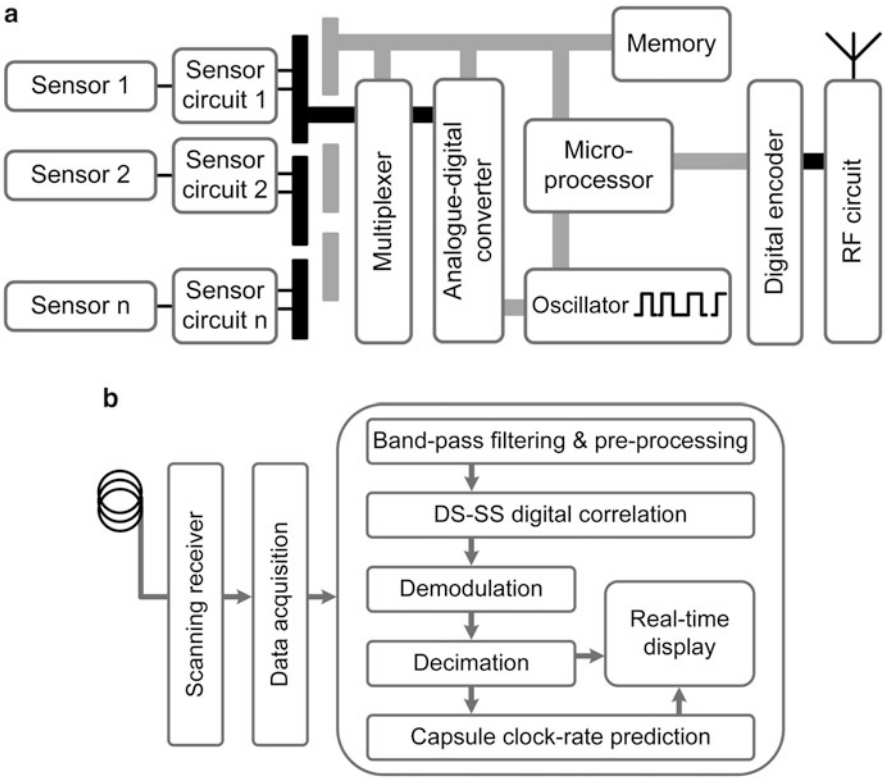


Fig. 11.6 A typical block diagram of the key features of a capsule (**a**) and a receiver (**b**). Because the receiver is not constrained by power and size it can have any reasonable level of complexity [9]

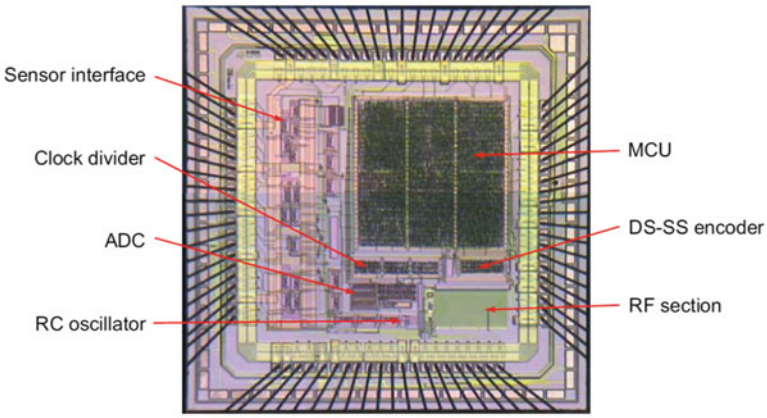


Fig. 11.7 A photomicrograph of an integrated circuit providing nearly all the electronics for a capsule device

rest of the instrument, it is usual to have a separate RF section, usually made from commercially available parts. The majority of devices have only a one-way wireless link to enable data transmission to an external device. However, a device providing a two-way link has recently been demonstrated [64]. The advantages of a two-way wireless link are improved security of the wireless connection and external control of the capsule.

Once all the measurements are combined, they are encoded and transmitted over a wireless link to a receiver outside the body. There are many possible ways of building the external system and for illustration purposes, we show a system combining an RF section, a decoder and display or data storage unit. This latter unit, usually a wearable device, could simply record data on to storage media for subsequent analysis, or could provide real-time display and analysis capabilities. Another possibility that has been investigated is to implement the external unit as a web server enabling clinicians to 'look in' from potentially any networked device with a web browser [65].

11.4.3 Integrated Sensors

As described in Chaps. 2 and 3, there are a wide variety of microsensor technologies now available, many of which can be applied to diagnostic capsule applications. A useful text describing several examples is given by Gardner [66]. Microsensors, and in particular, sensors that can be integrated into a small format and share a common platform with the electronics, are very useful for size-constrained systems such as a diagnostic capsule.

11.4.3.1 Physical

One of the most significant examples has been the use of CMOS video chips by Given Imaging Ltd [17, 30]. In addition to the ability to integrate the electronics and the sensor on the same chip, the advantages of the CMOS video approach, as opposed to using a charge coupled device, are the relatively low cost and the ability of the device to operate at relatively low voltage.

The integration of CMOS image sensors with electronics is a result of the advance of consumer electronics. Other integrated sensors, targeted at industrial and medical applications, have also been developed that are well suited to capsule systems [4]. With the advent of *Micro Electro-Mechanical Systems* (MEMS), it is now possible to pattern complex 3D structures into CMOS chips. MEMS processes can be divided into either bulk or surface micromachining. In bulk micromachining, the silicon substrate is etched away from the back of the chip or wafer, using the oxide layer as an etch stop [67, 68]. This leaves a thin membrane containing the CMOS circuits, which has excellent thermal isolation and can be used for heat-based sensors. In surface micromachining of CMOS chips, the metal layers,

or the intermetal dielectric layers, are etched away to leave freestanding structures [69]. Resonating beams for mass sensing or thin filaments for heat sensing can be made in this way. Several sensors have been fabricated using a combination of CMOS and MEMS technologies. For example, a recent capsule-type device for measuring intravascular pressure uses a MEMS capacitive pressure sensor integrated onto a CMOS chip [70, 71].

11.4.3.2 Chemical

As already discussed, pH sensors can be made by using conventional glass electrode methods, but the arrival of chip-based sensors has allowed a more integrated approach to be adopted. Using this method it has been possible to implement more than one sensor on a single chip hence increasing functionality whilst contributing to the overall aim of reducing the capsule size. Figure 11.8 shows two sensor chips that have been developed for a laboratory-in-a-pill device [9]. The chips contain a diverse range of sensor technology, not least of which is a microfabricated Ag|AgCl reference electrode.

Chemical sensors have also been realised on CMOS chips. They may be classified as follows [4]:

- *Chemomechanical sensors* typically use a polymer-coated resonating beam, diaphragm or crystal whose fundamental frequency is changed by the mass of absorbed gas molecules.
- *Catalytic sensors* have an electrically heated suspended filament that causes local oxidation reactions, and measures the heat loss as a change in temperature.
- *Thermoelectric sensors* use thermocouples to measure ambient temperature or the heat liberated or consumed by a (localised) source or chemical reaction.
- *Optical sensors* use photodiodes to measure the light output from e.g. bioluminescent bacteria when they metabolise a target compound.
- *Voltammetric sensors* are miniaturised versions of the standard two- or three-electrode cells used to measure the electron exchange currents that occur in redox reactions.
- *Potentiometric sensors* use modified field-effect transistors to measure the potential generated from the concentration of ions in a gas or liquid.
- *Conductometric sensors* use either resistors or capacitors coated with a sensitive material (polymers or metal oxides, for example) to measure changes in impedance upon exposure to the analyte solution.

Most of these CMOS-compatible sensors produce analogue outputs and need to be connected to external equipment in order to record a measurement. Recently, there have been several examples of CMOS chemical sensors that take full advantage of the ‘system-on-chip’ paradigm. The gas sensor chip described by Hagleitner [72] used a combination of chemically sensitive capacitors, resonant beams and thermocouples, as well as a temperature sensor, integrated on a single chip.

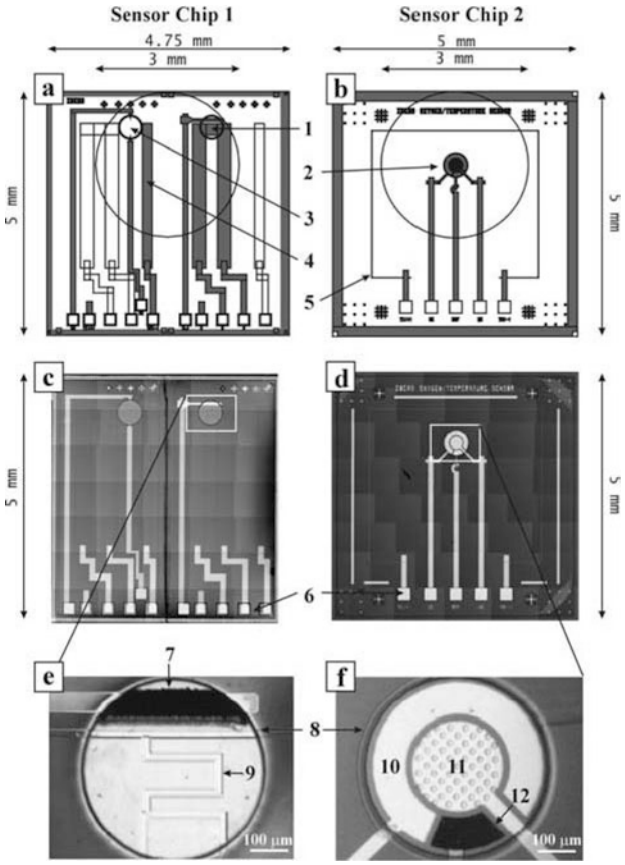


Fig. 11.8 Two sensor chips developed for a laboratory-in-a-pill device. (**a**) Schematic diagram of *Chip 1*, measuring 4.75 × 5 mm^2, comprising a pH sensor based on an ISFET (*1*), a dual electrode conductivity sensor (*3*) and a silicon diode temperature sensor (*4*); (**b**) schematic diagram of *Chip 2*, measuring 5 × 5 mm^2, comprising of an electrochemical oxygen sensor (*2*) and a Pt resistance thermometer (*5*). Once integrated in the pill, the area exposed to the external environment is illustrated by the 3 mm diameter circle; (**c**) photomicrograph of sensor *Chip 1*; and (**d**) sensor *Chip 2*. The bonding pads (*6*), which provide electrical contact to the external electronic control circuit, are shown; (**e**) close up of the pH sensor consisting of the integrated 3×10^{-2} mm^2 Ag|AgCl reference electrode (*7*), a 500 µm diameter and 50 µm deep, 10 nL, electrolyte chamber (*8*) defined in polyimide, and the 15 × 600 µm floating gate (*9*) of the ISFET sensor; (**f**) an oxygen sensor is likewise embedded in an electrolyte chamber (*8*). The three-electrode electrochemical cell comprises the 1×10^{-1} mm^2 counter electrode (*10*), a microelectrode array of 57 × 10 µm diameter (4.5×10^{-3} mm^2) working electrodes (*11*) defined in 500 nm thick PECVD Si_3N_4, and an integrated 1.5×10^{-2} mm^2 Ag|AgCl reference electrode (*12*)

All the control and sensing electronics, an *Analogue-to-Digital Converter* (ADC) and a digital interface were included on the chip. A commercial CMOS process was used, and both bulk and surface micromachining techniques were employed to define the sensing structures, after the chip had been fabricated (post processing).

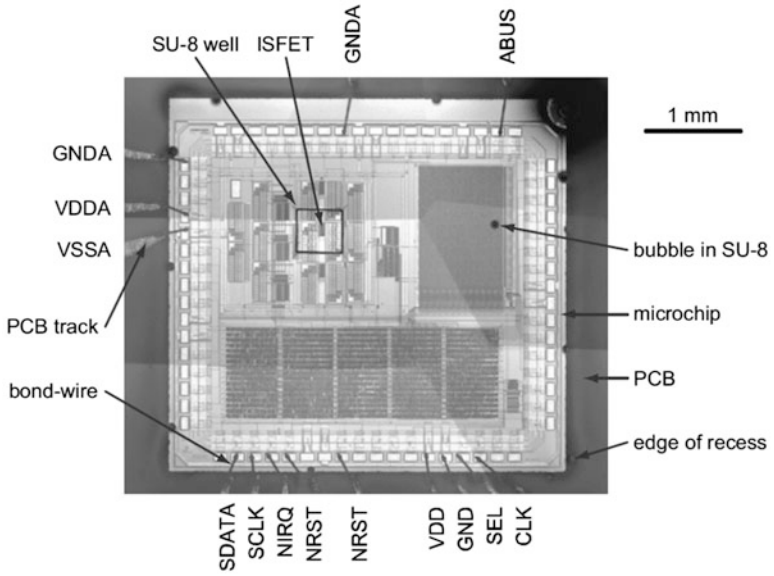


Fig. 11.9 Photomicrograph of encapsulated SoC pH meter

In another example, a fully integrated pH measuring instrument was made using a standard CMOS foundry process with no modification by micromachining [6]. Figure 11.9 shows a photomicrograph of the instrument with the individual components labelled. At the heart of this device is a floating gate *Ion-Sensitive Field Effect Transistor* (ISFET) made by taking advantage of the foundries standard process materials and design rules.

11.4.3.3 Biological

Another example, this time of a 'partial SoC' CMOS chemical sensor, uses living cells as the transducer to detect the presence of a toxin [73]. It is not a complete SoC as it requires off-chip analogue electronics and a microcontroller to make measurements. A microfluidic chamber is clamped in place above the electrodes on the chip. Heart muscle cells are injected into the chamber and cultured there. The chip allows different electrode pairs to be addressed and the system automatically selects those giving the strongest action potential signals from the cells as they beat. The system was packaged into a battery-powered handheld unit complete with pumps for the microfluidics, allowing it to be used outside the laboratory.

Living cells have also been used as bioluminescent 'bioreporters' with a CMOS SoC, to measure gas concentrations [74]. The cells used were luminescent in the presence of toluene. An integrated photodiode produced a current proportional to the light intensity, which was converted into a digital output by the on-chip processing

circuitry. Depending on the length of integration time used, concentrations as low as ten parts per billion of toluene were detected.

Clearly, biologically based sensors as described above are not of immediate application to diagnostic capsule devices, but may have an application in the future as technology moves towards highly specific discriminatory techniques.

11.5 Electronics System Design

In addition to having all the required components for the implementation of a capsule device, one must think about how the complete system is designed to achieve the desired performance. As we have seen that there is rapid progress away from bench-top instrumentation design to modern integrated circuit implementations. As a consequence, it is appropriate to use silicon design methodologies. Such methodologies have emerged from the microelectronics industry as more complex designs have been required [75, 76].

Currently, there is a trend to integrate analogue and digital functions on the same chip for both power and miniaturisation considerations. There are some specific issues regarding the choice of appropriate technology. A digital signal processor (DSP) usually involves a complex architecture and dedicated instruction sets. The use of field-programmable gate arrays (FPGA) is associated with relatively high power consumption. At the other end of the spectrum, analogue implementation of certain signal processing functions like filtering has shown to outperform its digital counterparts, but its programmability is still limited [77]. A Field-Programmable Analogue Array (FPAA) that improves the programmability of analogue processing is also not suitable due to relatively high power consumption [78]. As an alternative, an ASIC can be fully customised, providing maximal design flexibility at the lowest-possible power consumption. In an ASIC, all functional blocks can be integrated into a single piece of silicon, which means potential size reduction for the sensor nodes. This also simplifies the subsequent packaging and assembly processes. An ASIC is cost-effective when volume production is applied [79].

In Fig. 11.10, the block-diagram of a CMOS ASIC for portable medical applications is illustrated [80]. It incorporates a reconfigurable analogue front-end (AFE) module, a 12-bit ADC, a low power 32-bit RISC CPU ARM7TDMI, a scalable fast Fourier transform (FFT) module, a wireless power management unit (PMU) with a battery charger and management circuit, and a human body communication (HBC) block with integrated transmitter and receiver. The analogue signal processing chain is intended for low-frequency physiological signal (e.g. ECG/EEG/EMG) acquisition. The AFE module comprises of a low noise preamplifier and filter to detect and pre-process the signals from electrode channels. A 12-bit successive-approximation register (SAR) type ADC is used to digitise the amplified signal. The combination of fully differential AFE and ADC offers high immunity against common-mode noise and interferences. The short-time scalable FFT module serves as a digital co-processor to facilitate biomedical signal spectrum analysis. The PMU

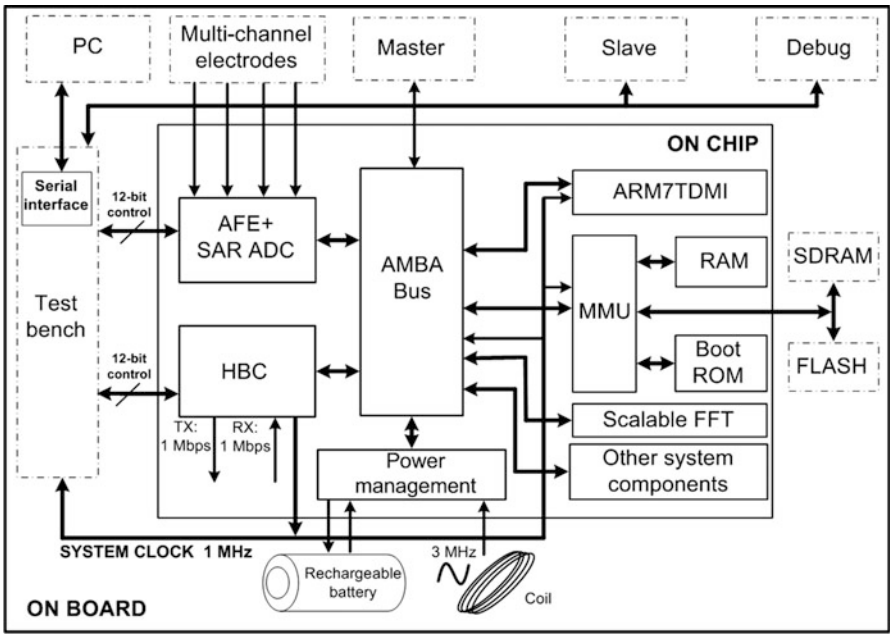


Fig. 11.10 An example of a reconfigurable ASIC used for portable medical applications [80]

with power recovery, regulation and battery management module uses the energy, induced into a coil to recharge a Li-ion battery and to power the system. The combination of ARM7TDMI and PMU ensures flexible communication and power-control modes for improved energy efficiency. Other blocks within the ASIC are the memory-management unit (MMU) and the data memory banks [80]. All blocks are fully integrated into the chip except for the power coil and the battery. The low power and low noise performance is achieved through state-of-the-art circuit design techniques. The ASIC is implemented in a standard 0.18-μm 1-poly 6-metal CMOS process and the die size of the complete chip is 5 mm $\times$ 5 mm [80]. A microphotograph of the chip is shown in Fig. 11.11.

In the following sections, we will present a methodology that is relatively simple by the standards of the industry, but encapsulates sufficient detail to enable accurate design of a sensor system.

11.5.1 Analogue Electronic Front-End Acquisition Design

The steps involved in designing an analogue circuit are illustrated in Fig. 11.12. The main difference from digital design is that both the schematic and physical designs are created by hand. An analogue circuit starts off as a high-level model or simply a list of requirements that must be met. The first attempt at a circuit design is

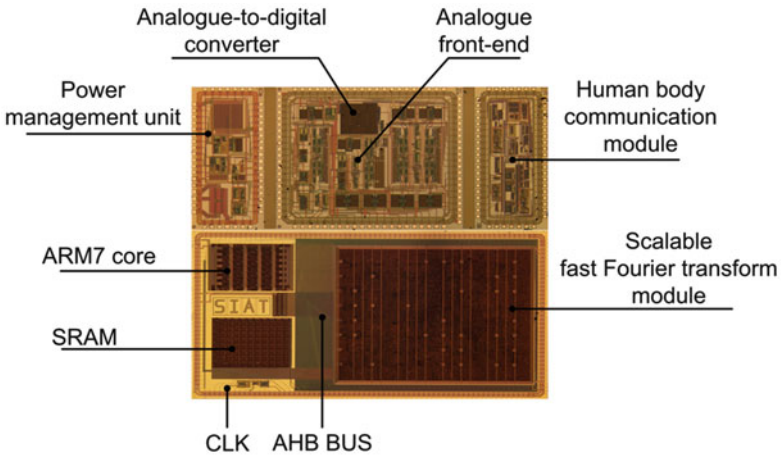


Fig. 11.11 Microphotograph of ASIC for portable biomedical applications [80]

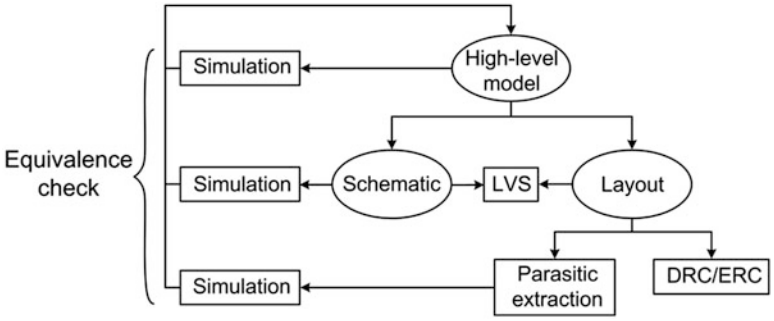


Fig. 11.12 Flowchart for the computer-aided analogue circuit design process

made using a schematic editor to draw a diagram of the components and their connections. The standard components available in a CMOS process are MOSFETs, resistors and capacitors but others such as diodes, bipolar transistors and inductors can also be created. Parameterised models for all available component types, obtained by characterizing fabricated devices, are provided by the foundry. The circuit is simulated by an analogue circuit simulator such as SPICE.

The circuit is unlikely to fulfil its requirements at the first attempt, so either the topology of the circuit or the parameters of its components are changed. Depending on the complexity of the circuit, several iterations may be required until the simulated response matches the desired response. When this is achieved, work can begin on the physical design (layout) of the circuit. A layout editor is used to draw areas of n-type and p-type silicon, polysilicon and metal that will form the components and connections. The task is usually made easier by the foundry that provides parameterised cell macros that generate the layout data for MOSFETs,

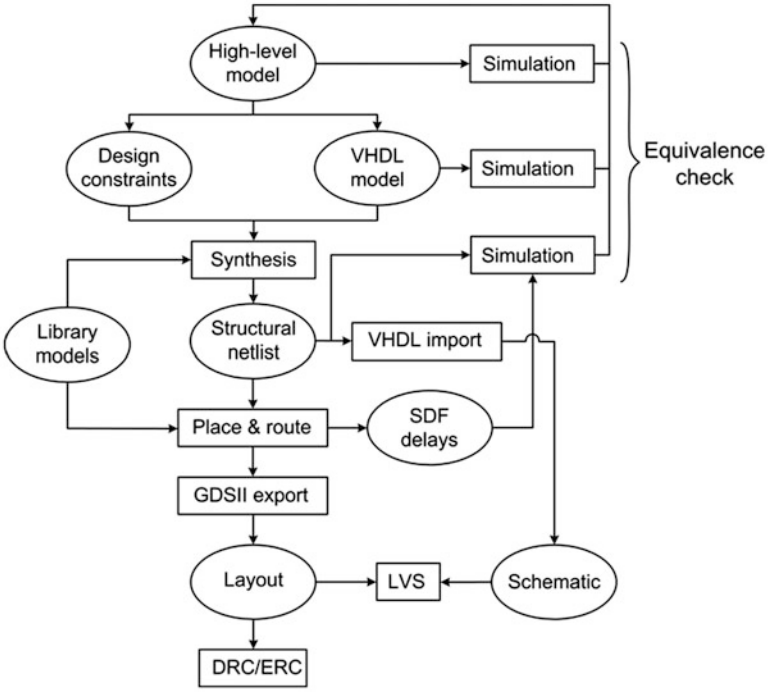


Fig. 11.13 Flowchart for the computer-aided digital circuit design process

resistors and capacitors. However, for low-noise, well-matched, or compact circuits it is often necessary to create these devices by hand. The arrangement of devices and the connections between them is also carried out by hand. Once complete, *Design Rule Checking* (DRC), *Electrical Rule Checking* (ERC) and *Layout Versus Schematic* (LVS) checks are performed to ensure that the final design is as intended. The design can be modified as required until a satisfactory conclusion is reached.

11.5.2 Digital System Design

A flow diagram of the steps involved in designing a digital circuit is shown in Fig. 11.13. In contrast to the analogue design flow, the physical design can be generated using software. The digital circuit may start off as a high-level 'behavioural' model written in a programming language such as C or Matlab. It is then re-coded into a *Hardware Description Language* (HDL) that allows the designer to describe digital circuits. The code is then compiled, simulated and debugged. Once the errors have been removed, simulation is required to ensure that the HDL code performs the functions described by the original high-level model.

If the HDL code uses only certain constructs that are allowed at a *Register Transfer Level* (RTL) (see [81] for a description of RTL), then the code may be synthesised. The synthesis process automatically generates the details of the gates and their interconnections to form a structural netlist. It attempts to optimise the netlist based on timing, area and power constraints that are set by the designer and on the information contained in the library models provided by the foundry.

After synthesis, the netlist is converted into a physical layout by the automatic 'place and route' tool. Delay information is extracted from this tool and used to annotate the structural netlist. If the simulation of this timing-accurate netlist fails to meet the specifications, then another iteration of the design loop is required. The next step in the process is to export the design, using the GDSII file format (GDSII is the standard file format for transferring/archiving 2D graphical design data) to describe the layout. The GDSII data and the HDL structural netlist are then imported into the layout editor (as used for the analogue design) to create the layout and schematic views respectively. As a final check, DRC, ERC and LVS checks are performed as for the analogue design process.

11.6 Wireless Transmission

The transmission of radio signals in and around the human body is of considerable importance to the success of a wireless capsule device. The behaviour of an electromagnetic field in the presence of a human body is influenced by the dielectric properties of human tissue. The dielectric function is frequency-dependent and the absorption of electromagnetic waves increases with frequency. Further, the reflection at boundaries, scattering, and refraction of electromagnetic fields are also frequency-dependent. The effect of capacitive loading of the surrounding tissue on the radiation source has a complex frequency and spatial dependency. However, the radiation from electrically small sources in free space increases with frequency.

Unsurprisingly, given the prevalence of mobile telephony, the majority of work that has been done to obtain a detailed understanding of radio transmission near the human body has focused on the head and neck. Early work used relatively simple body models that assumed that human tissue was homogenous [82]. As the work progressed, research moved to more sophisticated models [83]. More recently there have been a number of studies looking at the abdominal region using detailed models for both female [84] and male bodies [7]. The main results of these latter simulations are shown in Fig. 11.14. The simulations were carried out for an ingested transmitter at a number of possible locations and orientations and the data shown is therefore only representative.

Despite the fact that absorption of electromagnetic radiation increases with frequency, it is found that, up to a point, increasing the frequency can improve the far-field signal strength from an ingested source (Fig. 11.14a). The reason for this is that the size of the capsule device demands that an electrically small antenna be used, and that typically the antenna be very much smaller than $\lambda/2$

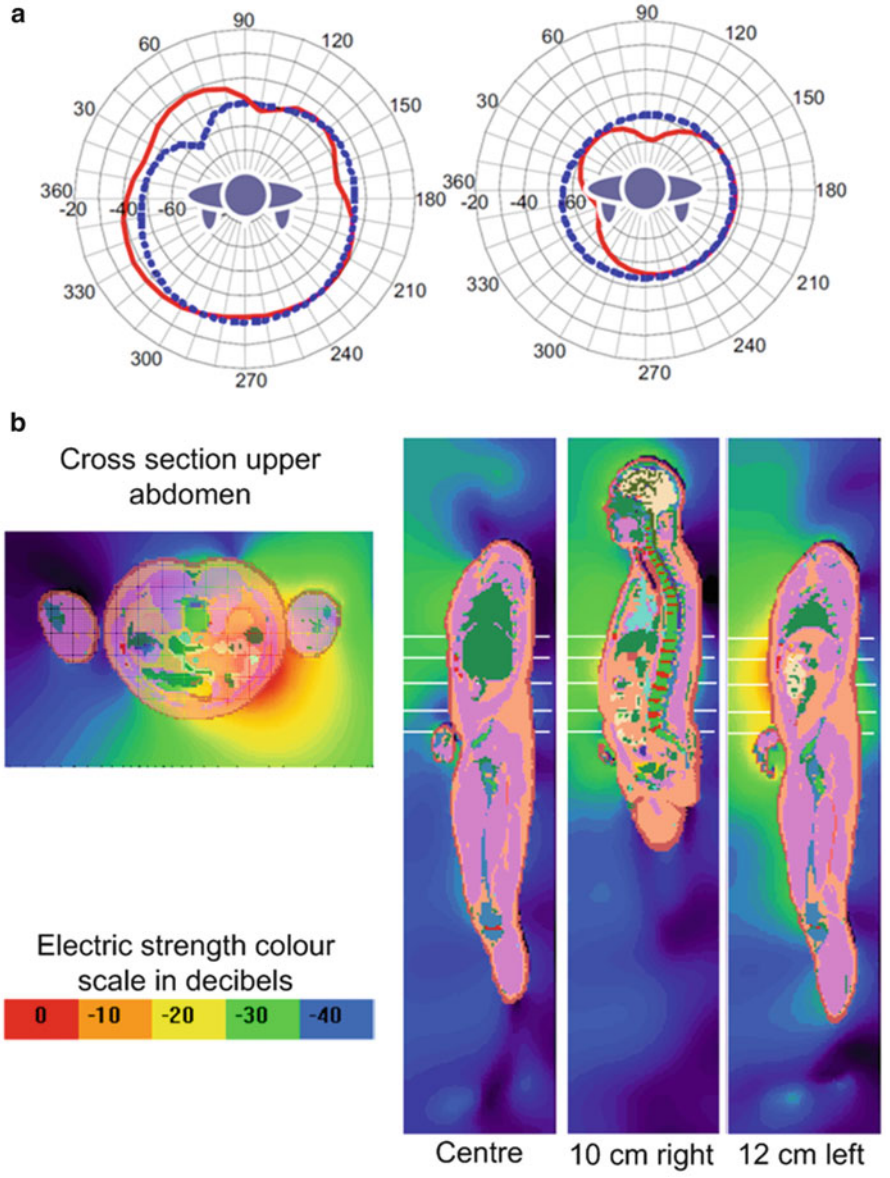


Fig. 11.14 (**a**) Far-field radiation patterns from an ingested source at 150 MHz (*left*) and 434 MHz (*right*). The *solid line* is for E-field polarization horizontal to the body, and the *dashed line* is for the vertical polarization. (**b**) The near-field pattern showing the field strength around the body

(the preferred size for the simplest radiating device), where λ is the wavelength. As a consequence, when the frequency is increased, the wavelength decreases towards the antenna dimensions, increasing the antenna's efficiency. It has been found that there is a competing effect between the increased efficiency of the antenna and the increasing absorption of the body tissue, and ultimately there is a trade-off in which an optimum frequency is found. Simulations have shown this to be in the region of 650 MHz.

The near-field pattern data is of interest since in many applications it is anticipated that the receiver antennae will be in close proximity to the body, which is for the most part located in the near-field for the wavelengths concerned. From Fig. 11.14b it can be seen that there is greater field strength to the left-hand anterior position of the abdomen due to the strongly absorbent nature of the liver on the right hand side. The results of simulations of this kind can be used directly to assist in the design of the antenna system.

An alternative to conventional propagating wireless systems is to communicate wirelessly using the inductive near-field [85]. Without dealing in detail with the electromagnetic problem here, it is possible to transmit wirelessly at low frequencies over distances that are very small compared to the propagating wavelength. In this evanescent regime, designers can take advantage of the lower absorption of RF power at lower frequencies and replace the antenna with two coils (one internally and one externally) that effectively behave as the primary and secondary coils of a transformer respectively. Detailed design and experimentation have shown that such a system can communicate effectively over the required range consuming less electrical power than a more conventional radio system [9].

The term 'Human Body Communication' (HBC) has been accepted to refer primarily two categories of solutions: galvanic coupling and capacitive coupling. The former uses a pair of electrodes in both the transmitter and receiver to transfer differential current signals, this solution is impractical for body sensor network (BSN) applications. Whilst the form factor of the complete body-worn devices is extremely miniaturised, the latter uses a single signal electrode for both transmission and reception and the ground electrodes are floating in the air. The signal forward path is established by capacitive coupling with the human body, and the signal return-path is formed by parasitic coupling through the ambient environment (Fig. 11.15). In this way high data rates and small size could be achieved [86, 87].

11.7 Power Sources

Powering ingestible capsule devices represents perhaps one of the greatest challenges in the design of these instruments. There are many possible micropower sources currently being considered and review of these has been written by Roundy [88]. Not all of the available techniques, for example solar power, are useful in the context of an ingested device. More practical techniques include: a battery; electromagnetic induction and electromechanical conversion. Other schemes, such

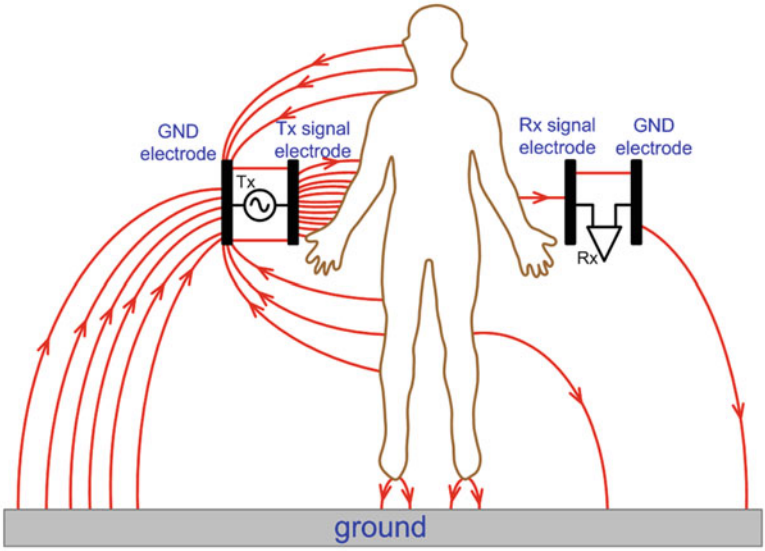


Fig. 11.15 Distribution of the main electric field in HBC [89]

as making direct use of gut mucosa as an electrolyte in an electrochemical cell arrangement, have been proposed but have not been explored in any serious way.

Although the use of batteries is by far the simplest power source it comes with certain restrictions. Not all types are favourable to use in implants (e.g. ZnO_2, which requires the exposure to oxygen for the battery to function), and achieving adequate power densities is difficult (e.g. AgO). The peak current delivery represents a potentially greater problem, since even short periods of high demand, for example during signal transmission, can very rapidly deplete a battery. As a consequence, when designing a micro-telemetry system, it is important to complete a detailed power budget and make design decisions that will ultimately compromise the device performance in order to ensure correct functionality during a complete gut transit.

Electromagnetic induction is an attractive option as it not only reduces the power constraint on the device, but also removes the need for batteries which could make the device smaller. The maximum power density permitted near the human body is in the region of 1 mW/cm^2 but this level varies from country to country [90]. This is a severe limitation given the power requirement and distance over which power must be transmitted to reach a device deeply embedded in the human abdomen. With the exception of the relatively simple RFID temperature sensing devices developed for animal use that we have already discussed, the implementations that demonstrate this technology in more sophisticated human medical devices are still few.

The Norika and Sayaka capsules by RF System Lab use inductive method for powering [17]. This makes possible the use of a CCD image sensor which provides better quality compared to the CMOS image sensors [91], which are usually utilised.

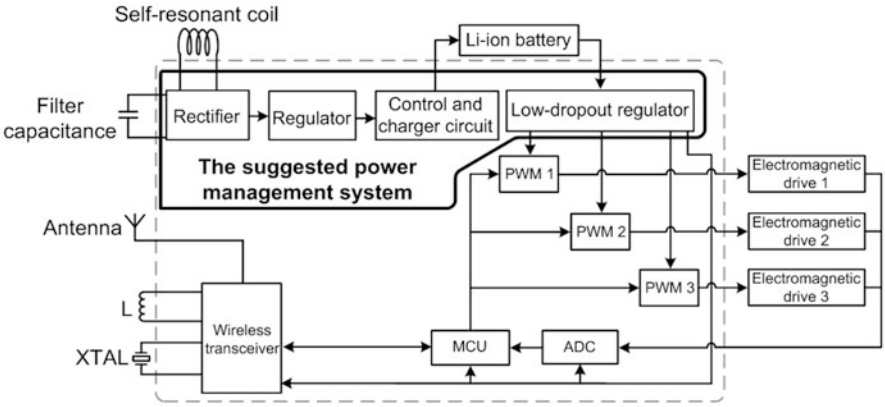


Fig. 11.16 Block-diagram of energy transmission system for powering of capsules [27]

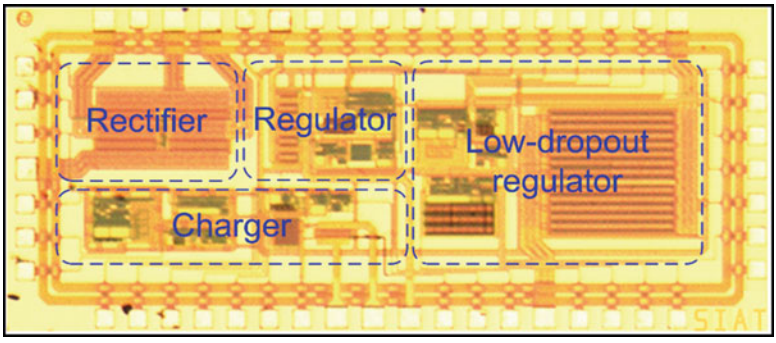


Fig. 11.17 Chip micrograph of the proposed wireless charging microsystem [27]

Figure 11.16 shows an example of a wireless power management system for endoscopic capsule robot [27].

The power is transferred to the capsule using inductive coupling at resonance frequency. The sinusoidal voltage generated in the receiving coil is rectified by a full-wave rectifier. The output voltage of the rectifier is unstable and has some ripples. A high power supply rejection ratio (PSRR) regulator is used to provide a stable and clean voltage source for the battery charger circuit that charges a Li-ion battery. The charger delivers 12 mA of constant current for charging the battery. The charger is able to control the switching between charging and discharging modes. A customised rechargeable Li-ion button battery is installed in the capsule. The battery has dimensions of 12 mm by 10 mm and capacity of 30 mAh. The maximum discharging current and the normal operating voltage range of this battery are 100 mA and 3–4.2 V, respectively. The output voltage of the Li-ion battery is regulated by the low-dropout (LDO) regulator , which provides sufficient power for the three electromagnetic drivers (Fig. 11.17).

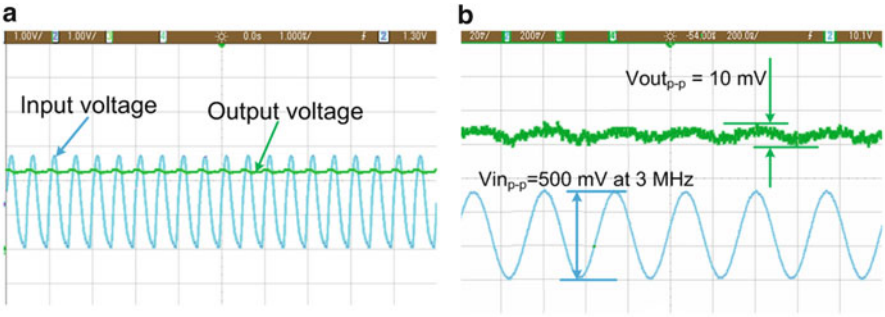


Fig. 11.18 Experimental results for the suggested charging microsystem: (**a**) input and output voltage of the rectifier (**b**) measured PSSR of the proposed regulator for signal with frequency of 3 MHz [27]

The wireless power supply microsystem for endoscopic capsule robot has been implemented using a 0.18 μm CMOS high-voltage technology. The chip size of this microsystem is 2.40 mm^2 (2.4 mm × 1 mm).

The input voltage of the regulator is in the range of 4.5–6 V. The regulator is able to operate with input voltages down to 4.48 V delivering an output voltage of 4.4 V. The variation of the output voltage is about 1.64 mV when the input voltage changes from 4.48 to 6 V. As shown in Fig. 11.18b, the regulator achieves 34 dB PSRR in the case when a 50 nF off-chip capacitor is used and the signal passed to the rectifier is with a frequency of 3 MHz. The total standby current of the proposed low-drop regulator is 85.1 μA (including the consumption of the protection circuit and the voltage reference) under an input voltage of 6 V. The control circuit turns on the charger when the battery voltage is less than 3.175 V and when the voltage of the battery reaches 4.275 V the charging stops and power is delivered to low-dropout regulator. The output voltage of the LDO is 3 V [27].

In some cases, like the bionic eye implant [92], power could be transferred wirelessly by light source, for example by laser, and received by a photovoltaic cell.

11.8 Packaging

Having decided upon the internal architecture of a laboratory-in-a-pill, it must all be packaged into a capsule. The package must be mechanically strong, chemically inert and allow access between the sensors and their environment. For optical devices there is always the prospect of an obstruction that blocks the lens, but the clear plastic dome structures that are used in current products, such as the PillCam capsule from Given Imaging, are relatively easy to manufacture and are strong. It is significantly more complicated to construct packages that will permit fluid access onto to sensor devices, especially if these devices are integrated circuits or

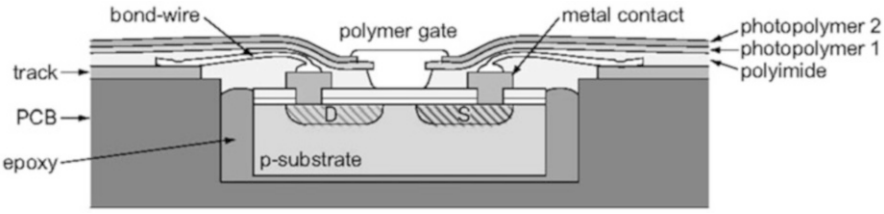


Fig. 11.19 Diagram of the cross-section through a CHEMFET device in a recessed PCB, encapsulated using a layer of polyimide and two layers of photoresist [94]

chips that will be adversely affected by current leakage due to liquids seeping into the encapsulating materials.

One of the main obstacles that prevented the commercialisation of ISFET-based devices is the repeatability and reliability of the encapsulation procedure. It is normal for the encapsulant to be applied by hand, covering the chip and bond-wires but leaving a small opening above the sensing area. Epoxy is the most extensively used material, although it is important to select a composition that is stable, a good electrical insulator and does not flow during encapsulation. Many commercially available epoxies have been assessed for their suitability by making measurements of their electrical impedance over time [93–96].

By using UV-curable polymers, it is possible to increase the automation of the packaging process by using a standard mask aligner. A lift-off technique was developed using a sacrificial layer of photosensitive polyimide to protect the ISFET gates. Alumina-filled epoxy was applied by screen printing and partially cured, before the polyimide was etched away leaving a well in the epoxy [97]. After 10 days in solution, leakage currents as high as 200 nA were observed. Better results were achieved by direct photo-polymerization of an epoxy-based encapsulant. ISFETs packaged using this method showed leakage currents of 35 nA after 3 months in solution. To avoid polarising the reference electrode, common to such devices, a leakage current of less than 1 nA is desirable [98]. This photolithographic patterning of the encapsulant was done at the wafer-level, to all the devices simultaneously. Subsequently the wafer was diced up and the individual chips were wire-bonded and coated with more encapsulant by hand.

At the chip-level, wire-bonded ISFET chips have been covered (again by hand) with a 0.5–1 mm thick photosensitive, epoxy-based film, then exposed and developed [99]. After 20 days in solution, the devices retained low leakage currents. Some degree of automation was introduced by Sibbald [94] who used a dip-coating method to apply the polymers. The chip was mounted in a recess in a PCB and the wire-bond connections were made, before coating it with a layer of polyimide. Two layers of photoresist followed, before the underlying polyimide was etched away (Fig. 11.19). The slight undercutting of the polyimide was reported to be useful in anchoring the CHEMFET membrane in place. The packaged devices showed less than 10 pA leakage current after 10 days in solution. However, the encapsulation did exhibit electrical breakdown for applied bias voltages in excess

of 1.5–2 V. This was attributed to the high electric field in the thin layer of resist covering the bond-wires. More recently, a single layer of an epoxy-based photoresist (SU-8) has been used to package a pH-sensing microchip [100]. In a separate study, photosensitive polyimide has also been used to create the wells that separate the ion-selective membranes on a multiple ISFET chip [101].

It is interesting to note that although flip-chip bonding is a well-established and robust packaging technique, it has not been applied to liquid-sensing ISFETs. In flip-chip bonding, solder bumps are patterned onto the bond-pads, allowing the chip to be directly connected to a PCB without the need for bond-wires. A gas sensor has been fabricated in this way, by bonding an ISFET to a ceramic substrate that had been coated with a suitable polymer [102]. It may be that the high cost of solder bumping, which is normally applied to a whole wafer of devices, has prevented wider use of flip-chip bonding in capsule packaging.

11.9 Conclusions

The development of microsystems for use in and around the body first came to the force before the term microsystem had even been invented. It began with the invention of the transistor, and the end is not yet in sight. However, the commercial drivers have not significantly changed: the availability of cheap technology from more mainstream research and development has enabled sensor and system designers to become more ambitious in their quest to develop new instrumentation, especially in healthcare.

In this chapter, we have outlined the broad range of wireless sensor microsystems that have been explored. Some of the most exciting applications are in human medicine and devices for use in the human gastrointestinal tract have received particular attention in recent years. In addition to reviewing these technologies, we have provided a description of many of the design challenges that must be overcome to meet the demand requirements of these applications.

References

1. Mackay RS. Bio-medical telemetry - sensing and transmitting biological information from animals and man. New York: John Wiley & Sons, 1968.
2. Mackay RS. Radio telemetering from within the body. Science, 1961; 134:1196–1202.
3. Lewin MH. Integrated microprocessors. IEEE Transactions on Circuits and Systems, 1975; 22 (7):577–585.
4. Hierlemann A, Baltes H. CMOS-based chemical microsensors. Analyst, 2003; 128:15–28.
5. Graf M, Barrettino D, Zimmermann M et al. CMOS monolithic metal-oxide sensor system comprising a microhotplate and associated circuitry. IEEE Sensors Journal, 2004; 4 (1):9–16.
6. Hammond PA, Ali D, Cumming DRS. Design of a single-chip pH sensor using a conventional 0.6 μm CMOS process. IEEE Sensors Journal, 2004; 4 (6):706–712.

7. Chirwa LC, Hammond PA, Roy S, Cumming DRS. Electromagnetic radiation from ingested sources in the human intestine between 150 MHz and 1.2 GHz. IEEE Transactions on Biomedical Engineering, 2003; 50 (4):484–492.
8. Madou MJ, Cubicciotti R. Scaling issues in chemical and biological sensors. In: Proceedings of the IEEE, 2003; 91 (6):830–838.
9. Wang L, Johannessen EA, Hammond PA et al. A programmable microsystem using system-on-chip for real-time biotelemetry. IEEE Transactions on Biomedical Engineering, 2005; 52 (7):1251–1260.
10. Beeley JM, Mills C, Hammond PA et al. All-digital interface ASIC for a QCM-based electronic nose. Sensors and Actuators B: Chemical, 2004; 103 (1–2):31–36.
11. Cotton PB, Williams CB. Practical gastrointestinal endoscopy. Oxford: Blackwell Scientific, 1980.
12. Lee R. Manual of small animal diagnostic imaging. Gloucester: British Small Animal Veterinary Association, 1995.
13. Gladman LM, Gorard DA. General practitioner and hospital specialist attitudes to functional gastrointestinal disorders. Alimentary pharmacology & therapeutics, 2003; 17 (5):651–654.
14. Mackay RS, Jacobson B. Endoradiosonde. Nature, 1957; 179:1239–1240.
15. Farrar JT, Zworykin VK, Baum J. Pressure-sensitive telemetering capsule for study of gastrointestinal motility. Science, 1957; 126:975–976.
16. Iddan G, Meron G, Glukhovsky A, Swain P. Wireless capsule endoscopy. Nature, 2000; 405 (6785):417.
17. Ciuti G, Menciassi A, Dario P. Capsule endoscopy: from current achievements to open challenges. IEEE Reviews in Biomedical Engineering, 2011; 4:59–72.
18. Fisher LR, Hasler WL. New vision in video capsule endoscopy: current status and future directions. Nature reviews Gastroenterology & hepatology, 2012; 9 (7):392–405.
19. Johannessen EA, Wang L, Cui L et al. Implementation of multichannel sensors for remote biomedical measurements in a microsystems format. IEEE Transactions on Biomedical Engineering, 2004; 51 (3):525–535.
20. Weitschies W, Karaus M, Cordini D et al. Magnetic marker monitoring of disintegrating capsules. European Journal of Pharmaceutical Sciences, 2001; 13 (4):411–416.
21. Phee L, Accoto D, Menciassi A et al. Analysis and development of locomotion devices for the gastrointestinal tract. IEEE Transactions on Biomedical Engineering, 2002; 49 (6):613–616.
22. Guo S, Sugimoto K, Fukuda T, Oguro K. A new type of capsule medical micropump. In: Proceedings of the 19th IEEE/ASME International Conference on Advanced Intelligent Mechatronics, 1999. pp 55–60.
23. Sendoh M, Ishiyama K, Arai KI. Fabrication of magnetic actuator for use in a capsule endoscope. IEEE Transactions on Magnetics, 2003; 39 (5):3232–3234.
24. Li H, Yan G, Ma G. An active endoscopic robot based on wireless power transmission and electromagnetic localization. The international journal of medical robotics + computer assisted surgery : MRCAS, 2008; 4 (4):355–367.
25. Miller KM, Mahoney AW, Schmid T, Abbott JJ. Proprioceptive magnetic-field sensing for closed-loop control of magnetic capsule endoscopes. In: Intelligent Robots and Systems (IROS), 2012 IEEE/RSJ International Conference on, 7–12 Oct. 2012. pp 1994–1999.
26. Al-Rawhani MA, Chitnis D, Beeley J, Collins S, Cumming DRS. Design and Implementation of a Wireless Capsule Suitable for Autofluorescence Intensity Detection in Biological Tissues. IEEE Transactions on Biomedical Engineering, 2013; 60 (1):55–62.
27. Peng D, Zhang J, Wang L. Experimental result on wireless power management microsystem for endoscopic capsule robot. In: 2012 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI), 5–7 Jan. 2012. pp 810–813.
28. Alonso O, Freixas L, Dieguez A. Advancing towards smart endoscopy with specific electronics to enable locomotion and focusing capabilities in a wireless endoscopic capsule robot. In: Biomedical Circuits and Systems Conference, 2009. BioCAS 2009. IEEE, 26–28 Nov. 2009. pp 213–216.

29. Sehyuk Y, Sitti M. Design and Rolling Locomotion of a Magnetically Actuated Soft Capsule Endoscope. Robotics, IEEE Transactions on, 2012; 28 (1):183–194.
30. Valdastri P, Simi M, Webster RJ, 3rd. Advanced technologies for gastrointestinal endoscopy. Annual review of biomedical engineering, 2012; 14:397–429.
31. Di Natali C, Beccani M, Valdastri P. Real-Time Pose Detection for Magnetic Medical Devices. Magnetics, IEEE Transactions on, 2013; 49 (7):3524–3527.
32. Skidmore M. Mini transmitter saves babies. NASA Aerospace Technology Innovation, 1999; 7 (1):7–8.
33. Given Imaging Ltd, http://www.givenimaging.com, 2013.
34. Tortora G, Valdastri P, Susilo E et al. Propeller-based wireless device for active capsular endoscopy in the gastric district. Minimally invasive therapy & allied technologies : MITAT : official journal of the Society for Minimally Invasive Therapy, 2009; 18 (5):280–290.
35. Carta R, Tortora G, Thone J et al. Wireless powering for a self-propelled and steerable endoscopic capsule for stomach inspection. Biosensors & bioelectronics, 2009; 25 (4):845–851.
36. Sayaka capsule system Available: http://www.rfsystemlab.com/en/sayaka, 2013.
37. Valdastri P, Webster RJ, Quaglia C et al. A new mechanism for mesoscale legged locomotion in compliant tubular environments. IEEE Transactions on Robotics, 2009; 25 (5):1047–1057.
38. Swain P. Wireless capsule endoscopy. Gut, 2003; 52 (Supplement IV):48–50.
39. Toennies JL, Tortora G, Simi M, Valdastri P, Webster RJ. Swallowable medical devices for diagnosis and surgery: The state of the art. Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science, 2010; 224 (7):1397–1414.
40. Woo SH, Kim TW, Mohy-Ud-Din Z, Park IY, Cho J-H. Small intestinal model for electrically propelled capsule endoscopy. BioMedical Engineering OnLine, 2011; 10 (1):108.
41. Mosse CA, Mills TN, Appleyard MN, Kadirkamanathan SS, Swain CP. Electrical stimulation for propelling endoscopes. Gastrointestinal endoscopy, 2001; 54 (1):79–83.
42. Weida L, Wei G, Mantian L, Yuhong Z. A novel locomotion principle for endoscopic robot. In: In: Proceedings of the 2006 I.E. International Conference on Mechatronics and Automation, 25–28 June 2006. pp 1658–1662.
43. Byungkyu K, Sukho P, Chang Yeol J, Yoon S-J. An earthworm-like locomotive mechanism for capsule endoscopes. In: 2005 IEEE/RSJ International Conference on Intelligent Robots and Systems, 2005. (IROS 2005). 2–6 Aug. 2005. pp 2997–3002.
44. Chongqing Jinshan Science and Technology Co. Ltd, Available: http://english.jinshangroup.com, 2013.
45. Keller H, Juloski A, Kawano H et al. Method for navigation and control of a magnetically guided capsule endoscope in the human stomach. In: Biomedical Robotics and Biomechatronics (BioRob), 2012 4th IEEE RAS & EMBS International Conference on, 24–27 June 2012. pp 859–865.
46. Keller J, Fibbe C, Volke F et al. Remote magnetic control of a wireless capsule endoscope in the esophagus is safe and feasible: results of a randomized, clinical trial in healthy volunteers. Gastrointestinal endoscopy, 2010; 72 (5):941–946.
47. Carpi F, Kastelein N, Talcott M, Pappone C. Magnetically Controllable Gastrointestinal Steering of Video Capsules. Biomedical Engineering, IEEE Transactions on, 2011; 58 (2):231–234.
48. Pandolfino JE, Richter JE, Ours T et al. Ambulatory esophageal pH monitoring using a wireless system. American Journal of Gastroenterology, 2003; 98 (4):740–749.
49. Sasaki Y, Hada R, Nakajima H, Fukuda S, Munakata A. Improved localizing method of radiopill in measurement of entire gastrointestinal pH profiles: colonic luminal pH in normal subjects and patients with Crohn's disease. American Journal of Gastroenterology, 1997; 92 (1):114–118.
50. Press AG, Hauptmann IA, Hauptmann L et al. Gastrointestinal pH profiles in patients with inflammatory bowel disease. Alimentary pharmacology & therapeutics, 1998; 12 (7):673–678.

51. Wolff HS. The radio pill. New Scientist, 1961; 12 (261):419–421.
52. Watson B, Kay A. Radio-telemetering with special reference to the gastrointestinal track. Biomechanics and Related Bio-Engineering Topics, 1965:111–127.
53. Watson BW, Meldrum SJ, Riddle HC, Brown RL, Sladen GE. pH profile of gut as measured by radiotelemetry capsule. British Medical Journal, 1972; 2:104–106.
54. Colson RH, Watson BW, Fairclough PD et al. An accurate, long-term, pH-sensitive radio pill for ingestion and implantation. Biotelemetry and patient monitoring, 1981; 8 (4):213–227.
55. Evans DF, Pye G, Bramley R et al. Measurement of gastrointestinal pH profiles in normal ambulant human subjects. Gut, 1988; 29 (8):1035–1041.
56. Kramer K, Kinter LB. Evaluation and applications of radiotelemetry in small laboratory animals. Physiological Genomics, 2003; 13:197–205.
57. Peters G. A new device for monitoring gastric pH in free-ranging animals. American Journal of Physiology, 1997; 273 (3 Pt 1):G748–753.
58. Enemark JM, Peters G, Jorgensen RJ. Continuous monitoring of rumen pH - a case study with cattle. Journal of Veterinary Medicine A, 2003; 50 (2):62–66.
59. ISO 11785 and ISO 3166.
60. Powers RA. Batteries for low power electronics. In: Proceedings of the IEEE, 1995; 83 (4):687–693.
61. Recommendation 70-30 relating to the use of short range devices (SRD). In: Proceedings of the European Conference of Postal and Telecommunications Administrations (CEPT), Tromso, Norway, CEPT/ERC/TR70-03, 1997.
62. Aydin N, Arslan T, Cumming DR. A direct-sequence spread-spectrum communication system for integrated sensor microsystems. IEEE Transactions on Information Technology in Biomedicine, 2005; 9 (1):4–12.
63. Nikolaidis I, Barbeau M, Kranakis E. Ad-Hoc, Mobile, and Wireless Networks: Lecture Notes in Computer Science. Heidelberg: Springer-Verlag, 2004; 3158.
64. Park HJ, Park IY, Lee JW et al. Design of miniaturized telemetry module for bi-directional wireless endoscopes. In: IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences, 2003. pp 1487–1491.
65. Wang L, Johannessen EA, Cui L et al. Networked wireless microsystem for remote gastrointestinal monitoring. In: Digest of Technical Papers for the Twelfth International Conference on Solid-State Sensors, Actuators and Microsystems, 8–12 June 2003. pp 1184–1187 vol.1182.
66. Gardner JW. Microsensors – principles and applications. Chichester: John Wiley and Sons, 1994.
67. Müller T, Brandl M, Brand O, Baltes H. An industrial CMOS process family adapted for the fabrication of smart silicon sensors. Sensors and Actuators A: Physical, 2000; 84 (1–2):126–133.
68. Tea NH, Milanovic V, Zincke CA et al. Hybrid postprocessing etching for CMOS-compatible MEMS. Journal of Microelectromechanical Systems, 1997; 6 (4):363–372.
69. Guillou DF, Santhanam S, Carley LR. Laminated, sacrificial-poly MEMS technology in standard CMOS. Sensors and Actuators A: Physical, 2000; 85(1–3):346–355.
70. Krüger C, Pfeffer JG, Mokwa W et al. Intravascular pressure monitoring system. In: Proceedings of the European Conference on Solid-State Transducers, M3C1, 2002.
71. Dudaicevs H, Kandler M, Manoli Y, Mokwa W, Spiegel E. Surface micromachined pressure sensors with integrated CMOS read-out electronics. Sensors and Actuators A: Physical, 1994; 43 (1–3):157–163.
72. Hagleitner C, Hierlemann A, Lange D et al. Smart single-chip gas sensor microsystem. Nature, 2001; 414 (6861):293–296.
73. DeBusschere BD, Kovacs GTA. Portable cell-based biosensor system using integrated CMOS cell-cartridges. Biosensors and Bioelectronics, 2001; 16 (7–8):543–556.
74. Simpson ML, Sayler GS, Applegate BM et al. Bioluminescent-bioreporter integrated circuits form novel whole-cell biosensors. Trends in Biotechnology, 1998; 16 (8):332–338.

75. Zhang Y, Ma KK, Yao Q. A software/hardware co-design methodology for embedded microprocessor core design. In: IEEE Transactions on Consumer Electronics, 1999; 45 (4):1241–1246.
76. Kundert K, Chang H, Jefferies D et al. Design of mixed-signal systems-on-a-chip. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, 2000; 19 (12):1561–1571.
77. Wang L, Thiemjarus S, Lo B, Yang GZ. Toward a mixed-signal reconfigurable ASIC for real-time activity recognition. In: 5th International Summer School and Symposium on Medical Devices and Biosensors, 2008. ISSS-MDBS 2008., 1–3 June 2008. pp 227–230.
78. Wang L, Ren S, Liu B, Wang G, Zhang J. A CMOS discrete-time reconfigurable analogue ASIC for low power biomedical signal filtering. In: 3rd International Conference on Bioinformatics and Biomedical Engineering, 2009. ICBBE 2009., 11–13 June 2009. pp 1–4.
79. Nie ZD, Wang L, Chen WG, Zhang T, Zhang YT. A low power biomedical signal processor ASIC based on hardware software codesign. Conference proceedings: Annual International Conference of the IEEE Engineering in Medicine and Biology Society IEEE Engineering in Medicine and Biology Society Conference, 2009:2559–2562.
80. Zhang J, Nie Z, Huang J, Yu L, Wang L. Towards low frequency low noise low power body sensor network-on-chip. In: International Conference on Green Circuits and Systems (ICGCS), 21–23 June 2010. pp 115–117.
81. Sjöholm S, Lindh L. VHDL for designers. Prentice Hall, 1997.
82. Toftgard J, Hornsleth SN, Andersen JB. Effects on portable antennas of the presence of a person. IEEE Transactions on Antennas and Propagation, 1993; 41 (6):739–746.
83. Okoniewski M, Stuchly MA. A study of the handset antenna and human body interaction. IEEE Transactions on Microwave Theory and Techniques, 1996; 44 (10):1855–1864.
84. Scanlon WG, Burns B, Evans NE. Radiowave propagation from a tissue-implanted source at 418 MHz and 916.5 MHz. IEEE Transactions on Biomedical Engineering, 2000; 47 (4):527–534.
85. Ahmadian M, Flynn BW, Murray AF, Cumming DRS. Data transmission for implantable microsystems using magnetic coupling. IEE Proceedings on Communications, 2005; 152 (2):247–250.
86. Nie Z, Ma J, Li Z, Chen H, Wang L. Dynamic propagation channel characterization and modeling for human body communication. Sensors, 2012; 12 (12):17569–17587.
87. Ruoyu X, Wai Chiu N, Hongjie Z, Hengying S, Jie Y. Equation environment coupling and interference on the electric-field intrabody communication channel. IEEE Transactions on Biomedical Engineering, 2012; 59 (7):2051–2059.
88. Roundy S, Steingart D, Frechette L, Wright P, Rabaey J. Power sources for wireless sensor networks. LNCS, 2004; 2920:1–17.
89. Wen-cheng W, Ze-dong N, Feng G, Teng-fei L, Lei W. Experimental Studies on Human Body Communication Characteristics Based Upon Capacitive Coupling. In: Body Sensor Networks (BSN), 2011 International Conference on, 23–25 May 2011 2011. pp 180–185.
90. United Kingdom regulations for adults exposed to radiation in the band from 10 MHz to 60 MHz. http://www.who.int/docstore/peh-emf/EMFStandards/who-0102/Worldmap5.htm.
91. McCaffrey C, Chevalerias O, O'Mathuna C, Twomey K. Swallowable-capsule technology. IEEE Pervasive Computing, 2008; 7 (1):23–29.
92. Revolutionary sight restoration. Brochure by Nano Retina Inc, available at http://www.nano-retina.com.
93. Chovelon JM, Jaffrezic-Renault N, Cros Y, Fombon JJ, Pedone D. Monitoring of ISFET encapsulation aging by impedance measurements. Sensors and Actuators B: Chemical, 1991; 3 (1):43–50.
94. Sibbald A, Whalley PD, Covington AK. A miniature flow-through cell with a four-function CHEMFET integrated circuit for simultaneous measurements of potassium, hydrogen, calcium and sodium ions. Analytica Chimica Acta, 1984; 159:47–62.

95. Grisel A, Francis C, Verney E, Mondin G. Packaging technologies for integrated electrochemical sensors. Sensors and Actuators, 1989; 17 (1–2):285–295.
96. Gràcia I, Cané C, Lora-Tamayo E. Electrical characterization of the aging of sealing materials for ISFET chemical sensors. Sensors and Actuators B: Chemical, 1995; 24 (1–3):206–210.
97. Muñoz J, Bratov A, Mas R et al. Planar compatible polymer technology for packaging of chemical microsensors. Journal of the Electrochemical Society, 1996; 143 (6):2020–2025.
98. Matsuo T, Esashi M. Methods of ISFET fabrication. Sensors and Actuators, 1981; 1:77–96.
99. Bratov A, Muñoz J, Dominguez C, Bartrolí J. Photocurable polymers applied as encapsulating materials for ISFET production. Sensors and Actuators B: Chemical, 1995; 25 (1–3):823–825.
100. Hammond PA, Cumming DRS. Encapsulation of a liquid-sensing microchip using SU-8 photoresist. Microelectronic Engineering, 2004:73–74:893–897.
101. Tsukada K, Sebata M, Miyahara Y, Miyagi H. Long-life multiple-ISFETS with polymeric gates. Sensors and Actuators, 1989; 18 (3–4):329–336.
102. Fleischer M, Ostrick B, Pohle R et al. Low-power gas sensors based on work-function measurement in low-cost hybrid flip–chip technology. Sensors and Actuators B: Chemical, 2001; 80 (3):169–173.

Chapter 12
Wearable Sensor Integration and Bio-motion Capture: A Practical Perspective

Zhiqiang Zhang, Athanasia Panousopoulou, and Guang-Zhong Yang

12.1 Introduction

In the previous chapters, we have discussed the fundamentals of BSN hardware and processing techniques including multi-sensor fusion, context aware and autonomic sensing. In this chapter, we will use bio-motion analysis as an exemplar to demonstrate how some of these methods are used for practical applications involving multiple wearable sensors.

Motion Capture (Mocap) and reconstruction is the process of recording the general body movement of a human subject or living being and translating the movement onto a 3D model such that the model performs the same actions as the subject [1]. The Mocap technology has been used for a variety of applications, from delivering realistic animation in filming and entertainment to assessing the performance of professional athletes. Clinically, motion reconstruction systems are increasingly used to analyse the biomechanics of patients. The analysis provides an objective measure of physical function to aid interventional planning, evaluate the outcomes of surgical procedures and assess the efficacy of treatment and rehabilitation [2, 3]. Thus far, a number of motion-tracking technologies have been developed and they can be mainly classified as optical tracking, mechanical tracking and inertial-sensor based tracking systems [4].

12.1.1 Optical Tracking Systems

Optical motion tracking systems utilise data captured from image sensors to triangulate the 3D position of a subject between one or more calibrated cameras.

Z. Zhang • A. Panousopoulou • G.-Z. Yang, PhD (✉)
The Hamlyn Centre, Imperial College London, London SW7 2BY, UK
e-mail: g.z.yang@imperial.ac.uk

G.-Z. Yang (ed.), *Body Sensor Networks*, DOI 10.1007/978-1-4471-6374-9_12,

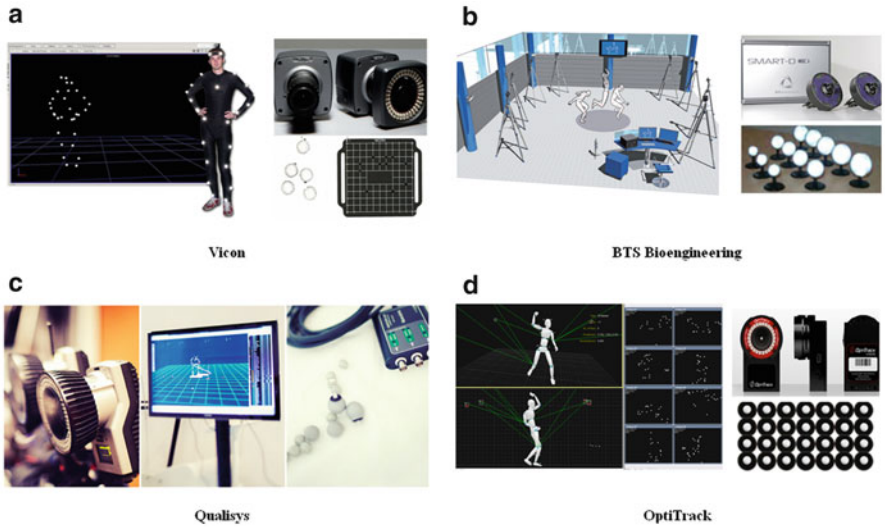


Fig. 12.1 Examples of the marker based optical tracking systems (**a**) Vicon offers a number of different systems: Vicon MX, Bonita and Vicon Motus, (**b**) BTS BioEngineering also provides several systems for kinematic analysis: POSEIDON, OEP system and SMART-DX (**c**) Qualisys motion capture system and (**d**) OptiTrackMotive platform

These cameras typically work in the infrared range to simplify the tracking process. Data acquisition is traditionally implemented using special optical markers attached to a subject, as shown in Fig. 12.1. By extracting locations of the markers in 3D from the images recorded by the infrared cameras, the systems can produce data with 3 Degrees-of-Freedom (DoF) for each marker, and rotational information can be inferred from the positions of three or more markers. Commercial systems such as Vicon [5], BTS [6], Qualisys [7] and OptiTrack [8] are widely used. Because of their high accuracy and fast update rates, these systems are commonly used as laboratory based bio-motion analysis platforms, as well as in the film industry. The disadvantages of these systems are associated with their high cost, complexity of the system setup and the inconvenience of these obtrusive markers, thus prohibiting its routine use in the free-living environment [9, 10].

Emerging techniques in computer vision have led to the development of the marker-less approach for generating human motion. This is achieved by tracking surface features identified dynamically for each part of the body [11, 12]. The marker-less systems, such as OpenStage 2 and BioStage developed by Organic Motion [13], as shown in Fig. 12.2, do not need subjects to wear special markers for location tracking. Instead, the underlying vision algorithms are able to analyse multiple streams of optical input, breaking them down into constituent body components for tracking through multi-view geometrical and kinematic modelling.

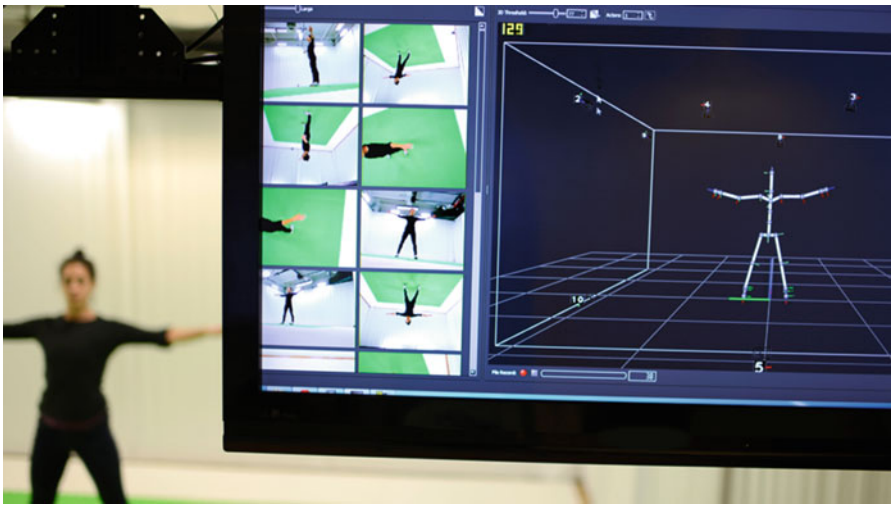


Fig. 12.2 Organic Motion's OpenStage 2: the only professional system on the market that delivers accurate tracking without the use of special markers

The main advantage of this type of system is that there is no need for special markers but its accuracy is inferior to that of marker-based platforms. Future research will likely lead to the development of special garment with fully integrated markers using infrared sensitive textiles, and therefore combines the advantages of both marker-based and marker-less techniques.

12.1.2 Mechanical-Based Tracking Systems

Mechanical motion capture systems directly track body joint movement using an exoskeleton or skeletal-like structure attached to the human body. The exoskeleton is a rigid structure of jointed, straight metal or plastic rods linked together with potentiometers or goniometers that articulate at the joints of the body for motion measurement. When the subject moves, the articulated mechanical parts follow the same movement by measuring the subject's relative motion. Mechanical motion capture systems not only can provide real-time motion estimation, but also are able to provide intelligent limb movement supports [14, 15]. For this reason, most platforms tend to be integrated as robotic platforms so as to provide mechanical support, feedback and control for limb rehabilitation applications. Some of the well-known commercial platforms are shown in Fig. 12.3. These include, for example, Gypsy 7 [16], Ekso Bionics [17], Armeo [18] and Lokomat [19].

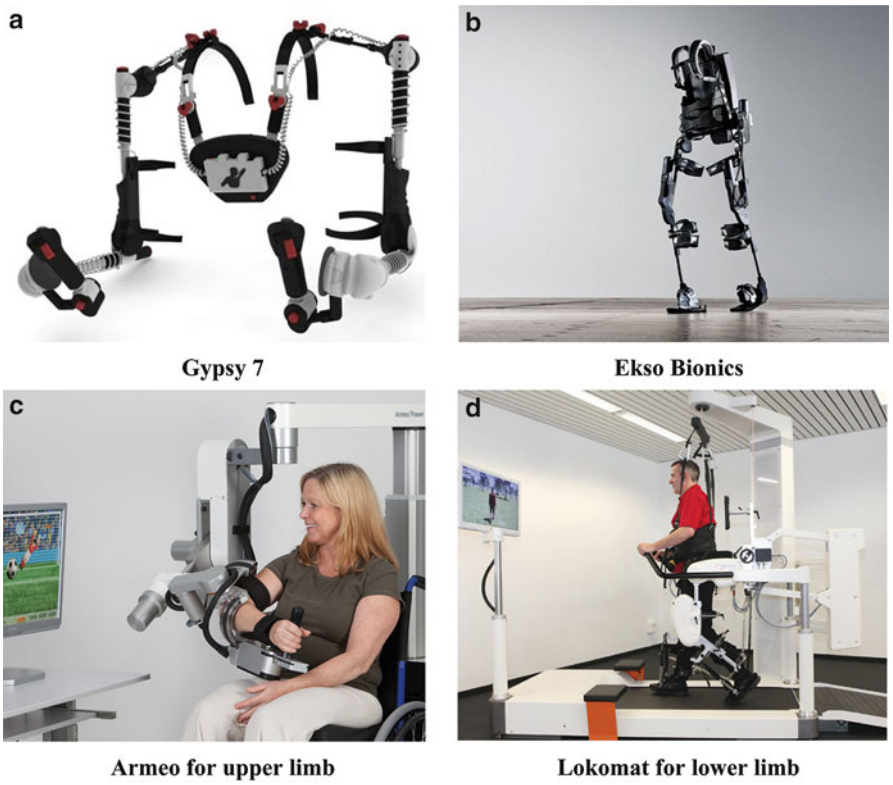


Fig. 12.3 Mechanical-based tracking systems using exoskeletons/skeletal structures, (**a**) Gypsy 7: the latest incarnation of Animazoo's state of the art exo-skeletal gyroscopic hybrid technology; (**b**) Ekso Bionics: enables individuals with lower extremity paralysis to stand up and walk over ground with a weight bearing, four point reciprocal gait; (**c**) ArmeoPower: the first commercially available robotic arm exoskeleton for neurorehabilitation and (**d**) Hocoma's Lokomat: a driven gait orthosis that automates locomotion therapy on a treadmill and improves the efficiency of treadmill training

12.1.3 Wearable Inertial-Sensor Based Tracking Systems

In contrast to vision based or mechanical-based tracking systems, inertial motion capture is based on miniature inertial sensors combined with biomechanical models and sensor fusion algorithms. It offers much greater flexibility without spatial constraints as no external cameras, mechanical supports, emitters or markers are required. The inertial sensors are usually placed at the surface of human body with data that is wirelessly transmitted to a computer for real-time capture and 3D reconstruction. Due to the inherent measurement drift, other micro sensors, such as magnetometers, ultrasonic sensors or ambient cameras are used in parallel [20, 21]. The main benefits of wearable inertial sensor based systems are that they provide a cost-effective means of real-time capture in free-living environment. In recent years, there has been extensive development in this area and many

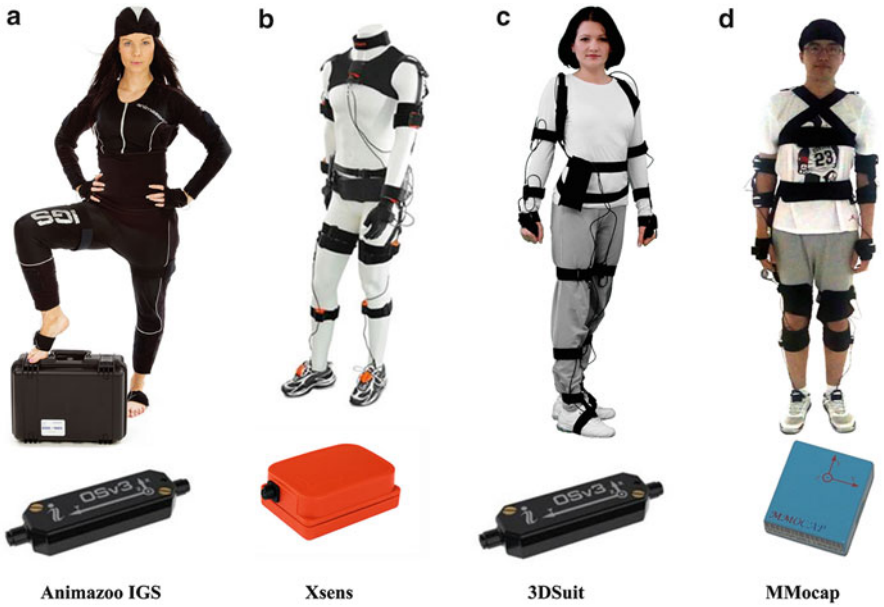


Fig. 12.4 Wearable inertial sensor based motion tracking systems, (**a**) Animazoo's IGS Motion capture systems; (**b**) Xense: the leading innovator in 3D wearable motion tracking technology and products; (**c**) 3DSuit: the affordable solution in inertial sensor based motion capture developed by Inertial Labs [28], and (**d**) MMocap: solution provider for digital 3D animation and digital neuro-rehabilitation

research platforms have been proposed. Established commercial systems include, for example, Animazoo [22], Xsens [23], 3D Suit [24] and MMocap [25] as illustrated in Fig. 12.4.

With recent advances in wireless sensing technologies, many of the hardware related issues including sensor miniaturisation and embodiment have been addressed. However, for practical use there are still many technical hurdles related to the handling of sensor drift, consistency in sensor placement, and accurate orientation estimation and motion reconstruction. Technically, these are closely related to the sensor fusion principles described in the previous chapters of this book. The main tasks for wearable motion estimations include orientation estimation of each body segment where sensors are placed and motion/posture estimation based on the estimated orientation and biomechanical modelling.

In the following sections of this chapter, we will use biomotion capture based on inertial and magnetic sensor units as the exemplar to demonstrate the key processing steps involved. We will start with the orientation representation, followed by the sensor fusion for orientation estimation, and biomechanical analysis for human motion estimation. In linking to Chap. 5, issues related to wireless communication and quality of service (QoS) will also be discussed.

12.2 Orientation Representation: Quaternion

12.2.1 Quaternion Definition

According to Euler's rotation theorem, the orientation of a rigid body is expressed as a single rotation from a reference placement with one point fixed. Such a rotation can be split into three axial rotations using Euler angles, which gives rise to the singularity problem due to the loss of one degree-of-freedom whenever the two axes of the rotations are in parallel (also called gimbal lock). Alternative representation is to use rotation matrix with a total of nine parameters. It is a very useful and popular way to represent rotations, but is less concise than other representations. In this chapter, unit quaternion is selected to represent the rotation, which is expressed as a single rotation about a given axis to avoid the singularity problem. Moreover, the quaternion is computationally a relatively efficient parameterisation scheme since it only has one redundant parameter, as opposed to the six redundant elements of the rotation matrix [26].

A quaternion q is defined by the extension of the complex numbers as [27]:

$$q = q_1 i + q_2 j + q_3 k + q_4 \tag{12.1}$$

where q_1, q_2, q_3, q_4 are real numbers and i, j, k are hyper imaginary numbers having the properties:

$$\begin{gathered} i^2 = -1, j^2 = -1, k^2 = -1 \\ ij = -k, ji = k \\ jk = -i, kj = i \\ ki = -j, ik = j. \end{gathered} \tag{12.2}$$

With ideas from both vector and matrix algebra, the quaternion q can be viewed as the real or scalar part q_4 and the imaginary or vector part $q_1 i + q_2 j + q_3 k$; therefore, the quaternion can be written as a four-dimensional column vector as:

$$q = \begin{bmatrix} v_q & q_4 \end{bmatrix}^T \tag{12.3}$$

where

$$v_q = \begin{bmatrix} q_1 & q_2 & q_3 \end{bmatrix}^T. \tag{12.4}$$

If q_1, q_2, q_3, q_4 satisfy

$$|q| = \sqrt{q_1^2 + q_2^2 + q_3^2 + q_4^2} = 1 \tag{12.5}$$

the quaternion is called a unit quaternion or quaternion of rotation. For any unit quaternion q, a rotation angle θ can be found which satisfies:

$$v_q = \frac{v_q}{|v_q|} \sin(\theta/2),\ q_4 = \cos(\theta/2) \tag{12.6}$$

where $v_q/|v_q|$ is the unit vector which represents the rotation axis, which is unchanged during the rotation, and θ is the rotation angle. According to the Euler's rotation theorem, the rotation will never change if rotating around $v_q/|v_q|$ for another 2π, thus we can get a new quaternion with vector part

$$\frac{v_q}{|v_q|} \sin\left(\frac{\theta + 2\pi}{2}\right) = -\frac{v_q}{|v_q|} \sin(\theta/2) \tag{12.7}$$

and scalar part

$$\cos\left(\frac{\theta + 2\pi}{2}\right) = -\cos(\theta/2). \tag{12.8}$$

Therefore, the quaternion q and the quaternion $-q$ describe the same rotation, so only the quaternion with positive scale part is considered in this chapter.

12.2.2 Quaternion Algebra

Since the scalars and vectors are in the subspace of quaternion, the rules in scalar and vector algebra also apply to quaternion. Let us consider the following two quaternions:

$$q = \begin{bmatrix} v_q \\ q_4 \end{bmatrix} = [q_1\ \ q_2\ \ q_3\ \ q_4]^T \tag{12.9}$$

and

$$p = \begin{bmatrix} v_p \\ p_4 \end{bmatrix} = [p_1\ \ p_2\ \ p_3\ \ p_4]^T \tag{12.10}$$

the addition and subtraction of them can be defined as [29]:

$$q \pm p = \begin{bmatrix} v_q \pm v_p \\ q_4 \pm p_4 \end{bmatrix} = [q_1 \pm p_1\ \ q_2 \pm p_2\ \ q_3 \pm p_3\ \ q_4 \pm p_4]^T. \tag{12.11}$$

The quaternion addition and subtraction obey associative and commutative laws.

Quaternion multiplication, designated by $\bigotimes$, is defined as [30]:

$$\begin{aligned} q \bigotimes p &= (q_1 i + q_2 j + q_3 k + q_4)(p_1 i + p_2 j + p_3 k + p_4) \\ &= (q_4 p_1 + q_1 p_4 - q_2 p_3 + q_3 p_2) i + (q_4 p_2 + q_2 p_4 - q_3 p_1 + q_1 p_3) j \\ &\quad + (q_4 p_3 + q_3 p_4 - q_1 p_2 + q_2 p_1) k + q_4 p_4 - q_1 p_1 - q_2 p_2 - q_3 p_3 \\ &= \begin{bmatrix} q_4 p_1 + q_1 p_4 - q_2 p_3 + q_3 p_2 \\ q_4 p_2 + q_2 p_4 - q_3 p_1 + q_1 p_3 \\ q_4 p_3 + q_3 p_4 - q_1 p_2 + q_2 p_1 \\ q_4 p_4 - q_1 p_1 - q_2 p_2 - q_3 p_3 \end{bmatrix} \\ &= \mathcal{L}(q) p \\ &= \mathcal{R}(p) q \end{aligned} \tag{12.12}$$

where

$$\mathcal{L}(q) = \begin{bmatrix} q_4 & q_3 & -q_2 & q_1 \\ -q_3 & q_4 & q_1 & q_2 \\ q_2 & -q_1 & q_4 & q_3 \\ -q_1 & -q_2 & -q_3 & q_4 \end{bmatrix} \tag{12.13}$$

and

$$\mathcal{R}(p) = \begin{bmatrix} p_4 & -p_3 & p_2 & p_1 \\ p_3 & p_4 & -p_1 & p_2 \\ -p_2 & p_1 & p_4 & p_3 \\ -p_1 & -p_2 & -p_3 & p_4 \end{bmatrix}. \tag{12.14}$$

The stew-symmetric matrix/cross-product operator $\lfloor v_q \times \rfloor$ for any vector v_q can be defined as:

$$\lfloor v_q \times \rfloor = \begin{bmatrix} 0 & -q_3 & q_2 \\ q_3 & 0 & -q_1 \\ -q_2 & q_1 & 0 \end{bmatrix} \tag{12.15}$$

so the cross-product can then be written as:

$$v_q \times v_p = \lfloor v_q \times \rfloor v_p. \tag{12.16}$$

Therefore, the $\mathcal{L}(q)$ and $\mathcal{R}(p)$ can be further simplified as:

$$\mathcal{L}(q) = \begin{bmatrix} q_4 I_{3\times 3} - \lfloor v_q \times \rfloor & v_q \\ -v_q^T & q_4 \end{bmatrix} \tag{12.17}$$

and

$$\mathcal{R}(p) = \begin{bmatrix} p_4 I_{3\times 3} + \lfloor v_p \times \rfloor & v_p \\ -v_p^T & p_4 \end{bmatrix}. \tag{12.18}$$

The conjugate of any quaternion q is the quaternion just with negated imaginary parts, and can be written as:

$$q^{'} = -q_1 i - q_2 j - q_3 k + q_4 \tag{12.19}$$

or

$$q^{'} = \begin{bmatrix} -v_q \\ q_4 \end{bmatrix} = [\,-q_1 \quad -q_2 \quad -q_3 \quad q_4\,]^T. \tag{12.20}$$

The quaternion inverse, i.e., the multiplicative inverse can be defined as:

$$q^{-1} = \frac{1}{q} = \frac{q^{'}}{\|q\|}. \tag{12.21}$$

In the special case of a unit quaternion, its conjugate is its inverse.

$$\|q\| = 1 \Rightarrow q^{-1} = q^{'}. \tag{12.22}$$

12.2.3 Quaternion and Rotation Matrix

A quaternion can be used to rotate any 3D point in space around an arbitrary axis. Given a point/vector v_η, we can build a quaternion:

$$\eta = \begin{bmatrix} v_\eta \\ 0 \end{bmatrix} \tag{12.23}$$

then given the unit quaternion q which provide the desired axis and angle of rotation, the rotated point/vector can be written as [31]:

$$\begin{aligned}
\eta^{rotated} &= \begin{bmatrix} v_\eta^{rotated} \\ 0 \end{bmatrix} = q \bigotimes \eta \bigotimes q^{'} \\
&= \begin{bmatrix} q_4 I_{3\times 3} - \lfloor v_q \times \rfloor & v_q \\ -v_q^T & q_4 \end{bmatrix} \begin{bmatrix} v_\eta \\ 0 \end{bmatrix} \bigotimes q^{'} \\
&= \begin{bmatrix} q_4 v_\eta - v_q \times v_\eta \\ -v_q^T v_\eta \end{bmatrix} \bigotimes q^{'} \\
&= \begin{bmatrix} q_4 I_{3\times 3} + \lfloor v_q \times \rfloor & v_q \\ -v_q^T & q_4 \end{bmatrix} \begin{bmatrix} q_4 v_\eta - v_q \times v_\eta \\ -v_q^T v_\eta \end{bmatrix} \\
&= \begin{bmatrix} \left(2q_4^2 - 1\right) I_{3\times 3} - 2q_4 \lfloor v_q \times \rfloor + 2 v_q v_q^T & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} v_\eta \\ 0 \end{bmatrix}.
\end{aligned} \tag{12.24}$$

However, the rotated point/vector can also rotated by the corresponding 3×3 rotational matrix R_q as:

$$v_\eta^{rotated} = R_q v_\eta \tag{12.25}$$

thus the relationship between the quaternion q and its corresponding rotational matrix R_q can be written as:

$$\begin{aligned} R_q &= \left(2q_4^2 - 1\right) I_{3\times 3} - 2q_4 v_q \times + 2v_q v_q^T \\ &= \begin{bmatrix} q_1^2 - q_2^2 - q_3^2 + q_4^2 & 2(q_1q_2 + q_3q_4) & 2(q_1q_3 - q_2q_4) \\ 2(q_1q_2 - q_3q_4) & -q_1^2 + q_2^2 - q_3^2 + q_4^2 & 2(q_2q_3 + q_1q_4) \\ 2(q_1q_3 + q_2q_4) & 2(q_2q_3 - q_1q_4) & -q_1^2 - q_2^2 + q_3^2 + q_4^2 \end{bmatrix} \\ &= \begin{bmatrix} 2q_1^2 + 2q_4^2 - 1 & 2(q_1q_2 + q_3q_4) & 2(q_1q_3 - q_2q_4) \\ 2(q_1q_2 - q_3q_4) & 2q_2^2 + 2q_4^2 - 1 & 2(q_2q_3 + q_1q_4) \\ 2(q_1q_3 + q_2q_4) & 2(q_2q_3 - q_1q_4) & 2\left(q_3^2 + 2q_4^2 - 1\right) \end{bmatrix} \\ &= \begin{bmatrix} 1 - 2q_2^2 - 2q_3^2 & 2(q_1q_2 + q_3q_4) & 2(q_1q_3 - q_2q_4) \\ 2(q_1q_2 - q_3q_4) & 1 - 2q_1^2 - 2q_3^2 & 2(q_2q_3 + q_1q_4) \\ 2(q_1q_3 + q_2q_4) & 2(q_2q_3 - q_1q_4) & 1 - 2q_1^2 - 2q_2^2 \end{bmatrix}. \end{aligned} \tag{12.26}$$

The inverse problem is to determine the quaternion q when the rotational matrix is known. Given the rotational matrix R_q

$$R_q = \begin{bmatrix} c_{11} & c_{21} & c_{31} \\ c_{12} & c_{22} & c_{32} \\ c_{13} & c_{23} & c_{33} \end{bmatrix} \tag{12.27}$$

it can be derived as the follows by defining T:

$$\begin{aligned} T &= c_{11} + c_{22} + c_{33} + 1 \\ &= -\left(q_1^2 + q_2^2 + q_3^2 - 3q_4^2\right) + 1 \\ &= 4q_4^2 \end{aligned} \tag{12.28}$$

if $T > 0$ then

$$q = \begin{bmatrix} \dfrac{c_{23} - c_{32}}{2\sqrt{T}} & \dfrac{c_{13} - c_{31}}{2\sqrt{T}} & \dfrac{c_{12} - c_{21}}{2\sqrt{T}} & \dfrac{\sqrt{T}}{2} \end{bmatrix} \tag{12.29}$$

if $T < 0$ then identify which major diagonal element (c_{11}, c_{22} or c_{33}) has the greatest value, i.e.,

(a) if c_{11} is the greatest

$$q = \begin{bmatrix} \dfrac{\sqrt{2 + 2c_{11} - T}}{2} & \dfrac{c_{12} + c_{21}}{4q_1} & \dfrac{c_{13} + c_{31}}{4q_1} & \dfrac{c_{33} - c_{32}}{4q_1} \end{bmatrix} \tag{12.30}$$

(b) if c_{22} is the greatest

$$q = \begin{bmatrix} \dfrac{c_{12}+c_{21}}{4q_2} & \dfrac{\sqrt{2+2c_{22}-T}}{2} & \dfrac{c_{23}+c_{32}}{4q_2} & \dfrac{c_{31}-c_{13}}{4q_2} \end{bmatrix} \tag{12.31}$$

(c) if c_{33} is the greatest

$$q = \begin{bmatrix} \dfrac{c_{13}+c_{31}}{4q_3} & \dfrac{c_{23}+c_{32}}{4q_3} & \dfrac{\sqrt{2+2c_{33}-T}}{2} & \dfrac{c_{12}-c_{21}}{4q_3} \end{bmatrix}. \tag{12.32}$$

12.2.4 Quaternion Integration

The quaternion integration has the general form of [32]:

$$q_t = \Theta(t, t-1)q_{t-1} \tag{12.33}$$

where $\Theta(t,\ t-1)$ only relates to the rotational velocity $\omega(t)$. Under certain assumptions, we can obtain closed form solutions for this equation. The simplest assumption is that $\omega(t)$ is constant over the integration period. If $\omega(t) = \omega$ and remains a constant, $\Theta(t,\ t-1)$ can be simplified as:

$$\Theta(t, t-1) = \Theta(\Delta t) = \exp\left\{\frac{1}{2}\mathcal{R}(\omega)\Delta t\right\}. \tag{12.34}$$

We can rewrite matrix exponential $\Theta(t,t-1)$ using its Taylor series expansion as:

$$\Theta(\Delta \mathrm{t}) = \mathrm{I}_{4\times4} + \frac{1}{2}\mathcal{R}(\omega)\Delta \mathrm{t} + \frac{1}{2!}\left(\frac{1}{2}\mathcal{R}(\omega)\Delta \mathrm{t}\right)^2 + \frac{1}{3!}\left(\frac{1}{2}\mathcal{R}(\omega)\Delta \mathrm{t}\right)^3 + \cdots \tag{12.35}$$

The matrix $\mathcal{R}(\omega)$ has the following properties,

$$\begin{aligned} \mathcal{R}(\omega)^2 &= -|\omega|^2 \mathrm{I}_{4\times4} \\ \mathcal{R}(\omega)^3 &= -|\omega|^2 \mathcal{R}(\omega) \\ \mathcal{R}(\omega)^4 &= |\omega|^4 \mathrm{I}_{4\times4} \end{aligned} \tag{12.36}$$

and so on. By substituting these properties into Eq. 12.34, we can get

$$\Theta(\Delta \mathrm{t}) = \cos\left(\frac{|\omega|\Delta \mathrm{t}}{2}\right)\mathrm{I}_{4\times4} + \frac{1}{|\omega|}\sin\left(\frac{|\omega|\Delta \mathrm{t}}{2}\right)\mathcal{R}(\omega). \tag{12.37}$$

However, $\Theta(\Delta t)$ is the multiplication matrix associated with a specific quaternion Δq, so we can rewrite Eq. 12.33:

$$q_t = \Theta(t, t-1)q_{t-1} = q_{t-1} \otimes \Delta q \tag{12.38}$$

where

$$\Delta q = \begin{bmatrix} \dfrac{\omega}{|\omega|} \sin\left(\dfrac{|\omega|\Delta t}{2}\right) \\ \cos\left(\dfrac{|\omega|\Delta t}{2}\right) \end{bmatrix}. \tag{12.39}$$

To measure rotation, the simplest way is to use a gyroscope. However, a gyroscope only measures the angular velocity and thus an integration step is required to derive the rotational angles. In this case, sensor drift presents a major problem to accurate integration. The use of quaternion does not solve the problem itself but presents a convenient mathematical formulation for error correction. To demonstrate the effect of the inertial drift during gyroscope integration, Fig. 12.5 shows a typical gyroscope signal when the sensor node was placed stationary on a table for 2 min. The quaternion integration results are also shown. It is evident that the quaternion integration starts to drift from the ground-truth after only 10 s, and the drift increases exponentially over time. To overcome this issue, the gyroscope signal should be combined with accelerometer and magnetometer measurements using sensor fusion algorithms, which will be discussed in the subsequent section.

12.3 Bayesian Fusion for Orientation Estimation

Now with the quaternion representation of rotation explained, we can move on to orientation estimation by using an inertial/magnetic sensor unit, which typically contains a three-axis gyroscope, accelerometer and magnetometer. The gyroscope senses the angular velocity, which can be numerically integrated to obtain the orientation, but the integration is subject to boundless orientation drift error. Usually, the accelerometer and the magnetometer mainly measure the gravity and the local magnetic field, with respect to the global coordinate system resolved in the sensor local coordinate system, so the accelerometer and magnetometer measurements can provide two constant reference vectors (the gravity for the vertical reference and the local magnetic field for the horizontal reference) to compensate for the gyroscope integration error. The purpose of sensor fusion is to combine these three types of sensor measurements to ensure robust orientation estimation.

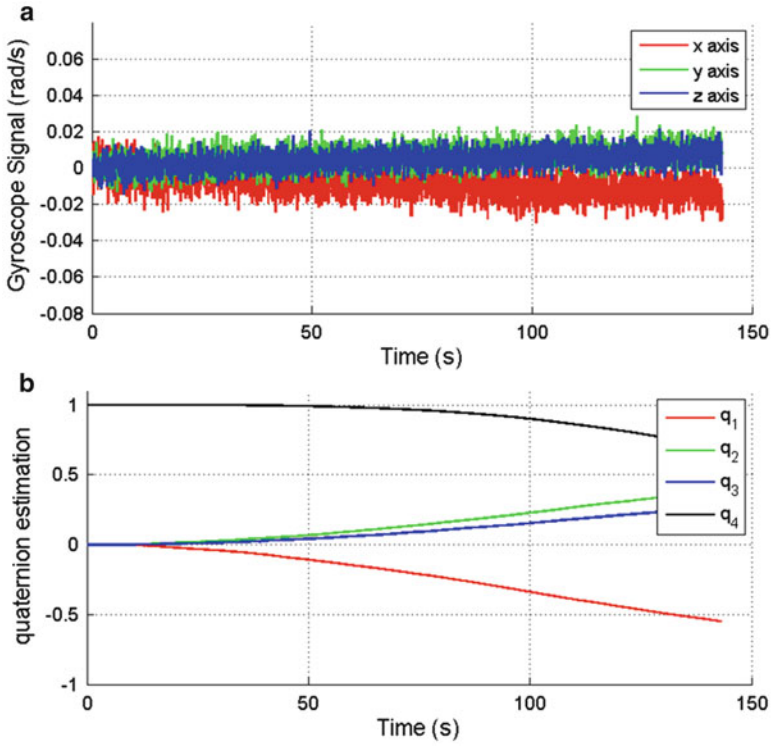


Fig. 12.5 Example of gyroscope signal integration in the presence of sensor drift. (**a**) The raw gyroscope signal when the node is stationary and (**b**) the corresponding quaternion integration showing the effect of sensor drift

Most orientation estimation and tracking problems are formulated using a dynamic system and a state space approach. Under the formulation of a dynamic system, the state at time t is usually denoted as x_t, which may include its orientation and angular velocity. The sensor measurement at time t is y_t, e.g., the measured acceleration, angular velocity and magnetic field direction. The objective is to find the best possible estimate of the filtering probability density function (pdf) $P(x_t|y_{1:t-1})$, where $y_{1:t} \triangleq \{y_1, y_2, \cdots, y_t\}$. The Bayesian fusion theory provides a solid framework for computing conditional pdf recursively [33].

12.3.1 Bayesian Fusion Theory

The dynamic system is usually modelled as a first-order Markov process, represented as a dynamic equation:

$$x_t = f(x_{t-1}, u_t, r_t) \tag{12.40}$$

where $f : \mathbb{R}^{n_x} \times \mathbb{R}^{n_u} \times \mathbb{R}^{n_r} \rightarrow \mathbb{R}^{n_x}$ is possibly a nonlinear function of the state, u_t is the input vector, r_t is the stochastic independent identical distribution process noise with zero mean and covariance matrix Q_t, and n_x, n_u, n_r are the dimensions of the state, input and the process noise vectors, respectively. The measurement equation is normally written as:

$$y_t = h(x_t, e_x) \tag{12.41}$$

where $h : \mathbb{R}^{n_x} \times \mathbb{R}^{n_e} \rightarrow \mathbb{R}^{n_y}$ is possibly a nonlinear function, e_t is the stochastically independent identical distribution measurement noise with zero mean and covariance matrix R_t. For these equations, n_y, n_e are the dimensions of the measurement and the measurement noise vectors, respectively.

In the case of Bayesian fusion, the orientation estimation problem is to recursively calculate the conditional pdf $P(x_t|y_{1:t-1})$ at time t given the measurement $y_{1:t}$ up to time t. It is assumed that the initial pdf $p(\mathrm{x}_0|\mathrm{y}_0) \equiv p(x_0)$ is known as the prior, thus the pdf $P(x_t|y_{1:t-1})$ can be recursively calculated in two stages of Bayesian filtering: prediction and update.

Suppose that the pdf at time $t-1$ is available, the prediction step uses the dynamic model to obtain the prior probability density function of the state at time t via the Chapman-Kolmogorov equation:

$$p(x_t|y_{1:t-1}) = \int p(x_t|x_{t-1})p(x_{t-1}|y_{1:t-1})dx_{t-1}. \tag{12.42}$$

At time t, the measurement y_t becomes available and it can be used to update the predicted pdf and obtain the required posterior pdf of the current state:

$$p(x_t|y_{1:t}) = \frac{p(y_t|x_t)p(x_t|y_{1:t-1})}{p\left(y_t|y_{1:t-1}\right)} \tag{12.43}$$

where $p(y_t|y_{1:t-1})$ is constant and calculated as:

$$p(y_t|y_{1:t-1}) = \int p(y_t|x_t)p(x_t|y_{1:t-1})dx_t. \tag{12.44}$$

Equations 12.42 and 12.43 comprise the recursive Bayesian filtering; however, the above method is only a conceptual solution and in most cases, a closed form solution does not exist, thus forcing the use of approximations such as Kalman filter and its variations [34, 35].

12.3.2 Dynamic and Measurement Model

The probability density functions $p(x_t|x_{t-1})$ and $p(y_t|x_t)$ are the key components in the recursive Bayesian filtering. They are usually implicitly given by the dynamic

equation and measurement equation [36]. However, Eqs. 12.40 and 12.41 are too general, and it is often sufficient to assume that the process and measurement noises are both additive and there is no control vector, so:

$$\begin{aligned} x_t &= f(x_{t-1}) + r_t \\ y_t &= h(x_t) + e_t \end{aligned} \tag{12.45}$$

$p(x_t|x_{t-1})$ and $p(y_t|x_t)$ can be explicitly written as Gaussian distributions:

$$\begin{aligned} p(x_t|x_{t-1}) &\propto N(x_t - f(x_{t-1}), Q_t) \\ p(y_t|x_t) &\propto N(y_t - h(x_t), R_t) \end{aligned} \tag{12.46}$$

To find the dynamic and measurement models, the first step is to determine the state vector x_t. Since the attitude of sensor unit is of our main interest here, we define the x_t to include the orientation and angular rate as:

$$x_t = \begin{pmatrix} q_t^{se} \\ \omega_t^s \end{pmatrix} \tag{12.47}$$

where q_t^{se} is a four dimensional unit quaternion, which describes the orientation of the sensor unit with respect to the reference system, and ω_t^s is the angular rate vector given in the sensor coordinate system. The dynamic equation can be expanded as:

$$\begin{aligned} q_t^{se} &= \Theta(\omega_{t-1}^s, \Delta t) q_{t-1}^{se} + r_{t-1}^q \\ \omega_t^s &= \omega_{t-1}^s + r_{t-1}^{\omega} \end{aligned} \tag{12.48}$$

where

$$\Theta(\omega_t^s, \Delta t) = exp\left\{\frac{1}{2}\mathcal{R}(\omega_{t-1}^s)\Delta \mathrm{t}\right\}. \tag{12.49}$$

The inertial and magnetic sensor units are usually subjected to a calibration procedure to define the exact physical alignment of each component, as well as the gains, offsets and their temperature behaviour. There is extensive literature describing this procedure and details about sensor calibration can be found, for example in Jurman et al. [37], Shen et al. [38], and Zhang and Yang [39]. Here we assume that all the three sensors are well calibrated and can provide accurate 3D acceleration, angular rate and magnetic field in metric units of the sensor coordinate system.

The calibrated gyroscope signal $y_{\omega,t}$ contains measurements of the angular velocity ω_t^s from body to reference expressed in the sensor coordinate system, so the measured angular velocity can be written as:

$$y_{\omega,t} = \omega_t^s + e_{\omega,t} \tag{12.50}$$

where $e_{\omega,t}$ is assumed to be zero mean white noise.

The calibrated accelerometer signal $y_{a,t}$ mainly contains the gravity vector with respect to the reference coordinate system resolved in the sensor coordinate system. By defining g as the vector of the gravitational field resolved in the reference coordinate system, the expected measurements of this field can be given by the transformation of g to the local sensor coordinate system. Thus the measured acceleration $y_{a,t}$ can be written as:

$$y_{a,t} = \left(q_t^{se}\right)^{-1} \otimes g \otimes q_t^{se} + e_{a,t} \tag{12.51}$$

where $e_{a,t}$ is assumed to be zero mean white noise.

Similarly, the calibrated magnetometer signal $y_{m,t}$ mainly contains the magnetic filed vector with respect to the reference coordinate system resolved in the sensor coordinate system. By defining m as the vector of the magnetic field resolved in the reference coordinate system, the expected measurements of this field are given by the transformation of m to the local sensor coordinate system. Thus the measured magnetic filed $y_{m,t}$ can be written as:

$$y_{m,t} = \left(q_t^{se}\right)^{-1} \otimes m \otimes q_t^{se} + e_{m,t} \tag{12.52}$$

where $e_{m,t}$ is also assumed to be zero mean white noise.

12.3.3 Kalman Filtering

The Bayesian estimation framework discussed in Sect. 12.3.1 can be used to fuse the measurements from the inertial and magnetic sensors, and Eqs. 12.48, 12.49, 12.50, 12.51, and 12.52 can be applied to perform the prediction and update steps recursively as given in Eqs. 12.42 and 12.43. However, it is too generic in the sense that we have to determine which state estimation algorithm to use. In practice, the Kalman filter has been widely used as a two-step fusion scheme for linear dynamical system. Because of the nonlinearity of the Eqs. 12.48, 12.49, 12.50, 12.51, and 12.52, an extended Kalman filter can be adopted; however, better performance tends to be achieved by using an unscented Kalman filter [40]. The extended Kalman filter only uses the first order terms of the Taylor series expansion of the nonlinear function, and it often gives poor performance when the models are highly nonlinear. Unlike the extended Kalman filter, the unscented Kalman filter uses a deterministic sampling technique known as the scaled unscented transformation to pick a minimal set of sample points (called sigma points) around the mean. These sigma points are then propagated through the nonlinear functions, from which the mean and covariance of the estimate are then recovered. In contrast to particle filters, the small number of points used by the unscented Kalman filter makes this estimator particularly appealing for real-time applications with limited computational power [41]. Moreover, the advantage of the unscented Kalman filter, with respect to the extended Kalman filter, is that the absence of linearisation

improves the estimate on performance and avoids the calculation of Jacobian matrices, so we have selected unscented Kalman filter here.

The quaternion-based unscented Kalman filter mainly consists of the following four steps:

1. ***Initialisation***: At time $t-1$, we will get the maximum a posteriori estimation of the state vector x by a Gaussian distribution $N(\mu_{t-1}, \Sigma_{t-1})$. We can construct another Gaussian distribution for unscented Kalman filter recursion, where the mean is $x_{t-1}^{\alpha} = [\mu_{t-1}^{T}, 0, 0]^{T}$, while the covariance is

$$P_{t-1}^{\alpha} = \begin{bmatrix} \Sigma_t & 0 & 0 \\ 0 & Q_t & 0 \\ 0 & 0 & R_t \end{bmatrix}. \tag{12.53}$$

The constructed Gaussian direction can be represented by a set of $2L+1$ sample point χ_{t-1}^{i} and weight W_{t-1}^{i}, denoted as sigma points $(\chi_{t-1,}^{i} W_{t-1}^{i})$ as

$$\begin{aligned}
\chi_{t-1}^{0} &= x_{t-1}^{\alpha} \\
\chi_{t-1}^{i} &= x_{t-1}^{\alpha} + \left(\sqrt{(L+\lambda)P_{t-1}^{\alpha}}\right)_i \qquad i = 1, 2\cdots, L \\
\chi_{t-1}^{i} &= x_{t-1}^{\alpha} - \left(\sqrt{(L+\lambda)P_{t-1}^{\alpha}}\right)_{i-L} \quad i = L+1, L+2\cdots, 2L \\
W_{t-1}^{0,m} &= \frac{\lambda}{L+\lambda} \\
W_{t-1}^{0,c} &= \frac{\lambda}{L+\lambda} + \left(1-\alpha^2+\beta\right) \\
W_{t-1}^{j,c} &= W_{t-1}^{j,m} = \frac{1}{2(L+\lambda)} \qquad j = 1, 2, \cdots, 2L
\end{aligned} \tag{12.54}$$

where $\lambda = \alpha^2(L+\kappa) - L$, κ, α and β are positive scaling parameters. The term $\left(\sqrt{(L+\lambda)P_{t-1}^{\alpha}}\right)_i$ is the ith column or row of the matrix square root of $(L+\lambda)P_{t-1}^{\alpha}$, with $L = n_x + n_r + n_e$.

2. ***Prediction***: using the prediction model $f(x_{t-1})$ defined in Eq. 12.48, the predicted estimation $\mu_{t|t-1}$ and covariance $\Sigma_{t|t-1}$ can be calculated as:

$$\begin{aligned}
\chi_{i,t|t-1}^{x} &= f\left(\chi_{i,t-1}^{x}\right) + \chi_{i,t-1}^{v} \\
\mu_{t|t-1} &= \sum_{i=0}^{2L} W_{t-1}^{i,m} \chi_{i,t|t-1}^{x} \\
\widetilde{\mu}_{i,t|t-1} &= \chi_{i,t|t-1}^{x} - \mu_{t|t-1} \\
\Sigma_{t|t-1} &= \sum_{i=0}^{2L} W_{t-1}^{i,c} \ \widetilde{\mu}_{i,t|t-1} \left(\widetilde{\mu}_{i,t|t-1}\right)^{T}
\end{aligned} \tag{12.55}$$

where $\chi_{t-1}^{i} = \left[\left(\chi_{i,t-1}^{x}\right)^{T} \quad \left(\chi_{i,t-1}^{v}\right)^{T} \quad \left(\chi_{i,t-1}^{xe}\right)^{T} \right]$ and the measurement innovation and covariance can be calculated as:

$$\begin{aligned} \gamma_{i,t|t-1} &= h\left(\chi_{i,t|t-1}^{x}\right) + \chi_{i,t-1}^{e} \\ y_{t|t-1} &= \sum_{i=0}^{2L} W_{t-1}^{i,m} \ \gamma_{i,t|t-1} \\ \widetilde{y}_{i,t|t-1} &= \gamma_{i,t|t-1} - y_{t|t-1} \\ \mathrm{S}_{t} &= \sum_{i=0}^{2L} W_{t-1}^{i,c} \ \widetilde{y}_{i,t|t-1} \left(\widetilde{y}_{i,t|t-1}\right)^{T}. \end{aligned} \tag{12.56}$$

3. ***Update***: The update step is as follows:

$$\begin{aligned} \mu_{t} &= \mu_{t|t-1} + K_{t}\left(y_{t} - y_{t|t-1}\right) \\ \Sigma_{t} &= \Sigma_{t|t-1} - K_{t} S_{t} (K_{t})^{T} \end{aligned} \tag{12.57}$$

where

$$\begin{aligned} K_{t} &= Cov_{t|t-1} (S_{t})^{-1} \\ Cov_{t|t-1} &= \sum_{i=0}^{2L} W_{t-1}^{i,c} \widetilde{\mu}_{i,t|t-1} \left(\widetilde{y}_{i,t|t-1}\right)^{T}. \end{aligned} \tag{12.58}$$

4. Renormalisation: In contrast to the normal unscented Kalman filter where the estimate μ_t can be acquired directly after the update stage, the quaternion Kalman filter needs a unit quaternion in order to represent an orientation. Thus, the quaternion part q_t^{se} in estimate μ_t needs to be normalised after each update stage as follows:

$$q_{t}^{se} = \frac{q_{t}^{se}}{\left\| q_{t}^{se} \right\|}. \tag{12.59}$$

To illustrate how the quaternion-based unscented Kalman filter can be used to overcome the gyroscope integration drift, Fig. 12.6 shows the accelerometer and magnetometer measurements when the sensor is stationary. The estimated quaternion is also shown in the figure, clearly illustrating the gyroscope integration drift that can be compensated by fusing all the sensor measurements together.

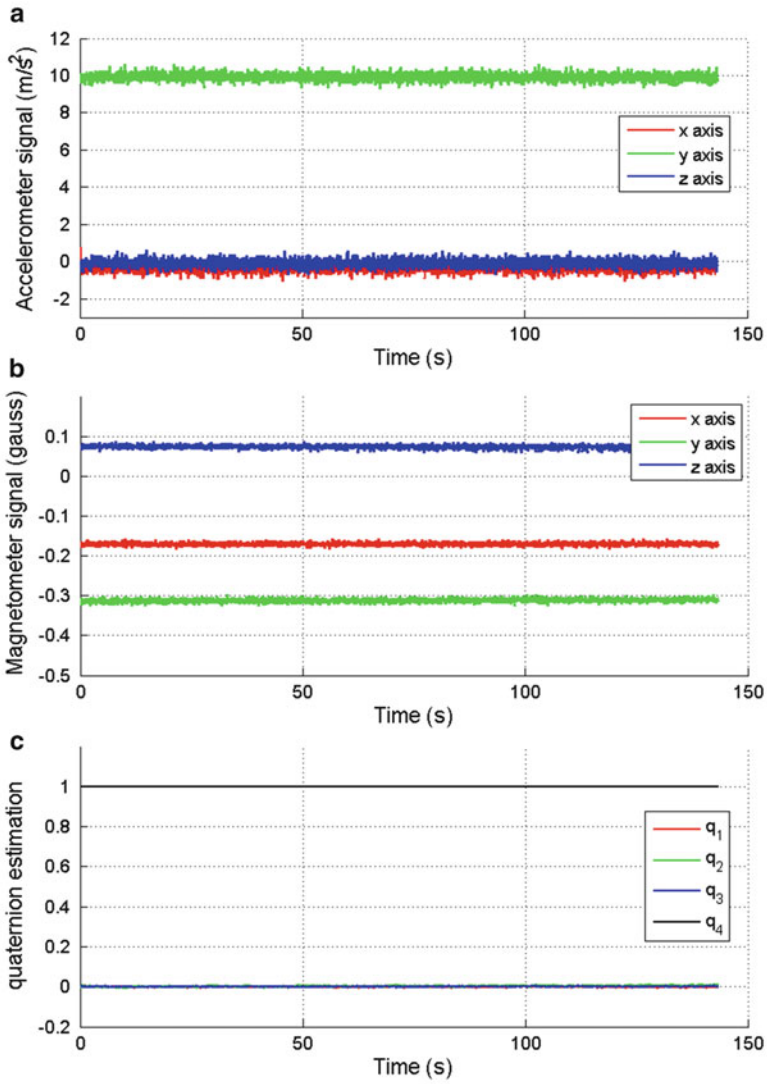


Fig. 12.6 Example of quaternion-based unscented Kalman filter for the same gyroscopic signal as shown in Fig. 12.5. (**a**) The accelerometer measurements; (**b**) the magnetic sensor measurements and (**c**) the estimated quaternion using the quaternion-based unscented Kalman filter

12.3.4 Temporary Interference and Processing

As mentioned earlier, the gyroscope senses the angular velocity, while the accelerometer and the magnetometer mainly measure the gravity and local magnetic field with respect to the reference coordinate system resolved in the local sensor

coordinate system. In fact, the accelerometer is generally used to measure the sum of gravity and the linear acceleration of the rigid body, while the magnetometer measures the sum of local magnetic field and magnetic disturbance. In the previous section, we assume that the linear acceleration and magnetic disturbance are small and can be well modelled by the measurement noise. However, this assumption is not applicable to situations where relatively large linear acceleration exists due to dynamic motion or magnetic disturbance due to ferromagnetic material. To make the orientation-estimation method more robust and resilient, two types of solutions have been proposed:

- *Temporary interference and disturbance estimation*: the basic idea is to expand the state vector and add the linear acceleration and magnetic disturbance as the part of the state vector. The interference or disturbance is then estimated by using Bayesian fusion. For instance, Veltink and Luinge [42] proposed a model to separate the gravitational acceleration and the linear acceleration and then derive the inclination from 3-D accelerometers. Young [43] attempted to estimate the linear accelerations of inertial sensors attached to rigid bodies based on the assumption that the length between the sensor and rotation centre was known. The main disadvantage of this method is that it can only deal with relatively small interference or disturbance, and it is problematic where there is large interference.
- *Measurement noise adjustment*: the basic idea behind this method is to detect whether there is acceleration interference or magnetic disturbance. If the inference/disturbance is detected, the Bayesian will make less use of accelerometer/magnetometer information by increasing the covariance matrix of accelerometer/magnetometer measurement noise. For instance, Sabatini [44] developed a quaternion-based extended Kalman filter, where the perturbed situations are handled by switching observation variances. Similarly, Kang and Park [45] introduced an attitude-estimation method by using an extended Kalman filter with a fuzzy-logic-based tuning algorithm. Four fuzzy rules were designed to change the characteristics of the filter to ensure confidence of the outputs of the gyros. The main disadvantage of this solution is the responding speed, and the covariance matrix increment is not fast enough to handle the outburst due to extremely large inference/disturbance [46].

In this section, we will introduce a vector selection scheme to deal with the temporary linear acceleration interference and magnetic disturbance. It can overcome the responding speed problem and protect the orientation estimation algorithm in conditions such as intensive movement and magnetic disturbance. The basic idea for the vector selection scheme is to detect whether the sensor measurements $y_{a,t}$ and $y_{m,t}$ are perturbed and then replace the degraded measurements with more reliable ones. For this purpose, since gyroscope measurements are related to the angular velocity regardless of the kinematic conditions and the magnetic environment, they can be considered as more reliable information than the degraded measurements of accelerometer and magnetometer. It should be noted

Table 12.1 Four cases of vector selection

	Case I	Case II	Case III	Case IV
$s_{a,t}$	$y_{a,t}$	$y_{a,t}$	$y_{a,t\|t-1}$	$y_{a,t\|t-1}$
$s_{m,t}$	$y_{m,t}$	$y_{m,t\|t-1}$	$y_{m,t}$	$y_{m,t\|t-1}$

Case I: slow motion with no magnetic disturbance; Case II: slow motion but with magnetic disturbance; Case III: fast motion with no magnetic disturbance and Case IV: fast motion but with magnetic disturbance

that the unbounded orientation error involved in the integration step in Eq. 12.48 mainly comes from the accumulation of the gyro measurement errors, but such errors in a short period are much smaller than those caused by $y_{a,t}$ and $y_{m,t}$ [47].

Based on the reasons mentioned earlier, the new measurement $s_{a,t}$ and $s_{m,t}$ are obtained before Kalman filtering through the predicted measurement. According to Eq. 12.48, the predicted quaternion $\hat{q}^{se}_{t|t-1}$ can be calculated as:

$$\hat{q}^{se}_{t|t-1} = \Theta\left(\omega^{s}_{t-1}, \Delta t\right)\hat{q}^{se}_{t-1}. \tag{12.60}$$

Hence the new vectors $s_{a,t}$ and $s_{m,t}$ can be replaced by the following equations:

$$s_{a,t} = \begin{cases} y_{a,t|t-1}, & if \left| \|y_{a,t}\| - \|g\| \right| > \varepsilon_{a_1} \ \ or \ \ \|y_{a,t} - y_{a,t|t-1}\| > \varepsilon_{a_2} \\ y_{a,t}, & otherwise \end{cases} \tag{12.61}$$

$$s_{m,t} = \begin{cases} y_{m,t|t-1}, & if \left| \|y_{m,t}\| - \|m\| \right| > \varepsilon_{m_1} \ \ or \ \ \|y_{m,t} - y_{m,t|t-1}\| > \varepsilon_{m_2} \\ y_{m,t}, & otherwise \end{cases} \tag{12.62}$$

where $y_{a,t|t-1}$ and $y_{m,t|t-1}$ are the predicted measurements of accelerometer and magnetometer, i.e., $y_{a,t|t-1} = \left(\hat{q}^{se}_{t|t-1}\right)^{-1} \otimes g \otimes \hat{q}^{se}_{t|t-1}$ and $y_{m,t|t-1} = \left(\hat{q}^{se}_{t|t-1}\right)^{-1} \otimes m \otimes \hat{q}^{se}_{t|t-1}$. ε_{a_ζ} and ε_{m_ζ} $(\zeta = 1, 2)$ are the criterion thresholds. Table 12.1 summarises the four cases of the measurement vector selection.

Figure 12.7 gives an example of the vector selection scheme after deliberately putting iron materials around the stationary sensor node to simulate the magnetic field interference. It is evident that the estimated quaternion is hardly affected by the interference with the vector selection scheme, while quaternion is severely affected without such scheme.

12.4 Human Body Motion Reconstruction

For human body analysis, it is common to model it as an articulated structure with 15–19 links/segments. With this model, the overall posture of the body can be represented by determining the orientation of each segment from the sensors positioned on these segments.

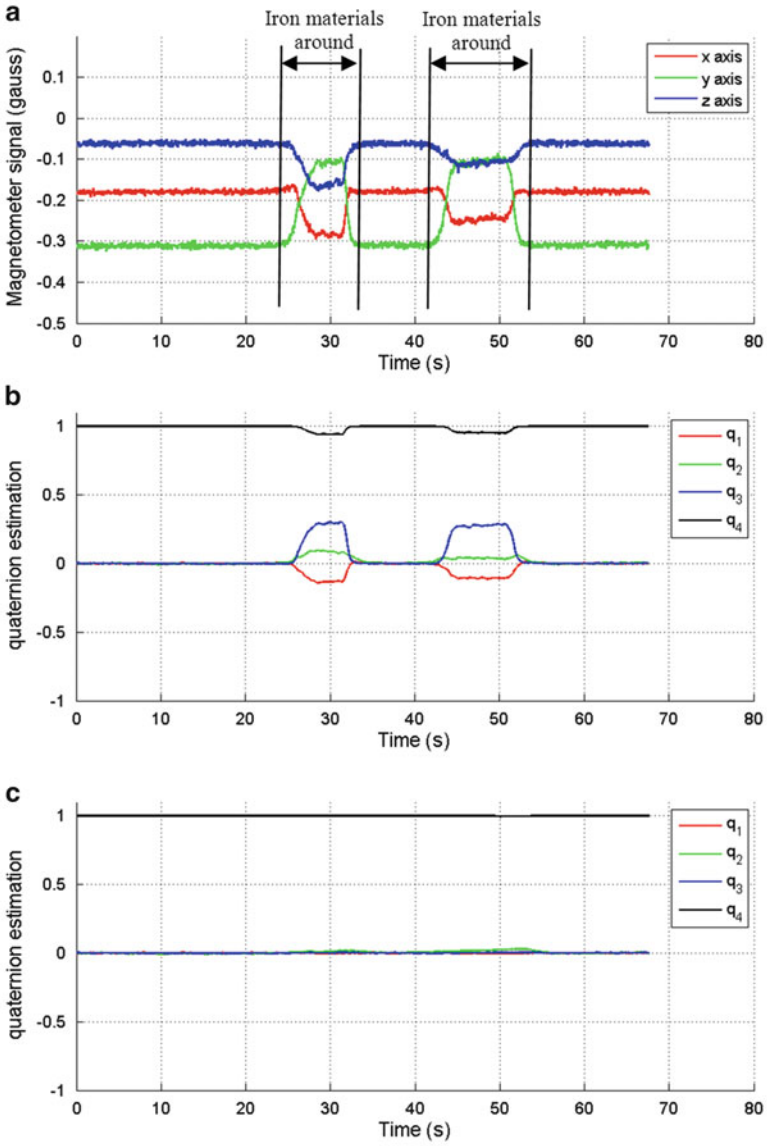


Fig. 12.7 An example of vector selection scheme for dealing with magnetic disturbance. (**a**) The magnetic sensor measurements after putting iron materials around the stationary sensor node; (**b**) the estimated quaternion without vector selection scheme and (**c**) the estimated quaternion with vector selection scheme

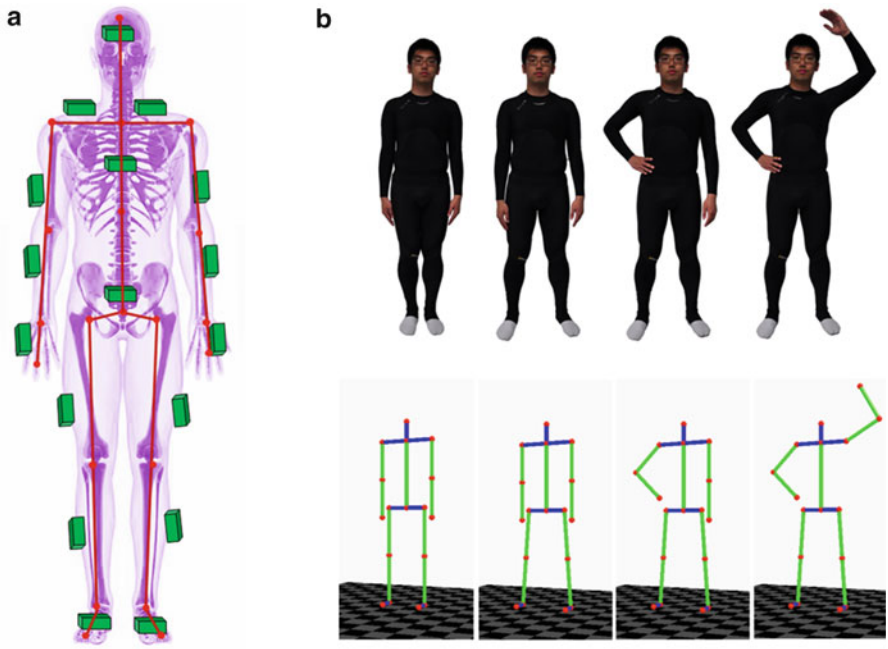


Fig. 12.8 The biomechanical model based motion estimation. (**a**) Model definition: The *red dots* are the joints; the lines indicate the body segments in the biomechanical model and the cubes represent the inertial/magnetic sensor units placed on the segments; (**b**) examples of real time posture reconstruction from inertial/magnetic sensor units

12.4.1 Human Biomechanical Model

In general, the human skeleton can be abstracted as a link structure articulating body segments as shown in Fig. 12.8. In the model, the red dots indicate the joints whiles the lines represent the body segments. The root point is selected as the specific point at the bottom of the sacrum of human body, and the trajectory of this point is then taken as the global displacement for the subject. In order to make the biomechanical model and the real human subject as close as possible, the articulated model for an individual subject needs to be built with individual specific parameters, such as height and segment lengths, which can be manually measured based on the anatomical landmarks on the body segments.

The skeleton model of the upper body is composed of spine, shoulders, upper arms, forearms, hands, neck and the head. However, there are several different simplifications. For example, the spine is usually divided into five regions: cervical (neck bones); thoracic (in the chest); lumbar (low back); sacral (attached to the pelvis) and coccygeal (the tail bone). Each region has a number of vertebral bones and several joints. However, the spine can normally be simplified as 1–4 segments

and the neck bone is always taken as the part of the head. Another example is the hand which is normally ignored since it will not affect the general movement of upper limb. In this section, we will simplify the spine and head into three different segments, and will also regard the hand as one rigid segment. The lower body skeleton model is defined to consist of seven body segments and eight joints. The seven body segments include left and right pelvis, femur, tibia and feet, while the eight joints represent left and right hips, knees, ankles and toes, as shown in Fig. 12.8. Similar to the hand, the foot is also simplified as one rigid segment.

In this model, all the segments are assumed to have the simple rigid geometry, and deformation is neglected during movements. From the root, the model forms a topological tree structure with the joints obeying a parent–child relationship, and every joint has a parent joint except the root joint. To retain human posture and location information, the root joint is analysed in DoFs with 3D position and 3D orientation. All other joints are in three DoFs with orientation only.

The inertial/magnetic sensors are placed on each body segment to capture the orientation of the corresponding body segment. The placement of the inertial/magnetic sensor units on the human body is shown in Fig. 12.8. In this setup, three coordinate systems are involved:

- *Earth* (e): The motion including pose and displacement is estimated with respect to this coordinate system. It is fixed to the earth and used as the reference, which can be aligned in any way. Here it is specifically defined as follows: the x-axis points forwards, z-axis points down and y-axis is perpendicular to x-axis and z-axis as a right-handed coordinate system.
- *Sensor* (s): It corresponds to the axes of three orthogonally mounted inertial sensors and magnetometers in the sensor unit, which is time varying, and all the sensor measurements are expressed in this coordinate system.
- *Body* (b): It is attached to a body segment whose rotation is to be measured, which is also time varying. Although the sensor and the body segment are rigidly attached to each other, the body coordinate system does not coincide with the sensor coordinate system. They are separated by a constant translation and rotation. Here only the rotation is considered if we are mainly interested in the orientation of the segments.

12.4.2 Posture Estimation

By mapping the orientation of inertial/magnetic sensors to segments of the biomechanical model, the posture of the subject can be reconstructed, as illustrated in Fig. 12.8. This means that the key issue for posture estimation is to convert the estimated orientation from the sensor coordinate to the body segment coordinate [48].

As mentioned earlier, the orientation $\hat{q}_t^{se}$ of each inertial/magnetic sensor unit can be estimated by using the Bayesian fusion. In order to express the body segment kinematics in the global coordinate system, we need to determine the orientations of each segment by converting the sensor orientation to segment orientation via sensor-to-segment calibration. During calibration, the subjects are required to remain at given postures, where the orientation of each body segment q_{cal}^{be} is known in advance while the sensor orientation q_{cal}^{se} can be also obtained from the sensor measurements. By denoting the rotation from the sensor to body segment q^{sb} we have:

$$q_{cal}^{se} = q_{cal}^{be} \otimes q^{sb} \tag{12.63}$$

so

$$q^{sb} = \left(q_{cal}^{be}\right)^{-1} \otimes q_{cal}^{se}. \tag{12.64}$$

The orientation of corresponding body segment can then be calculated as:

$$\hat{q}_t^{be} = \hat{q}_t^{se} \otimes \left(q^{sb}\right)^{-1}. \tag{12.65}$$

Once the orientations of all the body segments are known, the posture is reconstructed by mapping the real body segments orientations to the biomechanical model segments, and some exemplary results are shown in Fig. 12.8b.

12.5 Applications of Bio-motion Analysis

Hitherto, the bio-motion analysis has been widely used for sport performance research, medical rehabilitation and entertainment. In medical and sport performance research, the bio-motion analysis systems are often integrated with force plates and electromyography. When an individual walks across the force plates embedded into the floor for gait analysis, or performs functional tasks using both hands for upper limb analysis, the kinematics, kinetics (forces responsible for joint movement) and electromyography are collected to ensure an integrated analysis of the human bio-motion.

Another example of bio-motion analysis application is physical rehabilitation. By tracking the movements and trajectories of motor-disorder patients, the bio-motion systems can evaluate their performance of performing a set sequence of motion tasks and generate feedbacks. They can also be linked to the clinical Active Range of Motion scale and motor feature indices to evaluate the quality of the movement and help the patients to speed up the recovery progress. Take the stroke rehabilitation for example; the main objective is to help the patients to relearn the physical motion skills. Existing rehabilitation methods heavily rely on the therapists in the lab environments, which is labour-intensive and depends on the experiences of the therapists. However, the bio-motion systems can be used as an

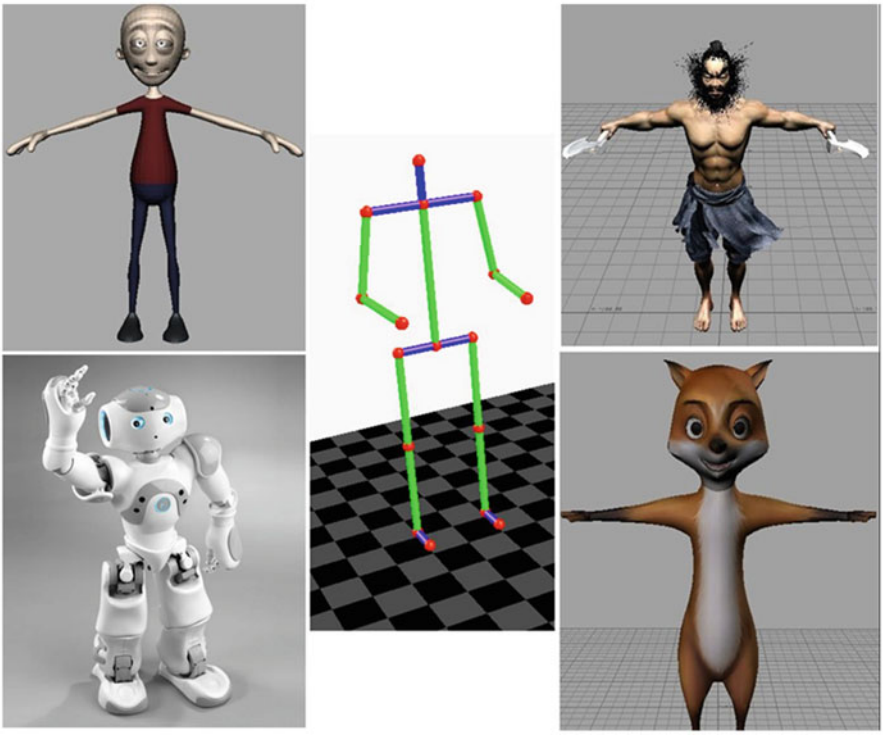


Fig. 12.9 Example of bio-motion analysis for entertainment. The captured human movement can drive different characters to move

adjunction to the therapists in daily living environments. It can help to guide the patients' motor movement training by giving instructions and demonstrations. Via capturing a set of touch-and-go movements, motor feature indices such as speed and smoothness can then be derived to evaluate the recovery progress objectively. According the evaluation results, the system can also generate training plan to automatically guide the rehabilitation process.

In addition to biomedical applications, the bio-motion analysis systems have also been used for entertainment, such as animated computer games, as shown in Fig. 12.9. Instead of using the stick-man to visualise the movement, we can match the real human movements to cartoon characters or avatars, which provide more realistic appearance of the characters.

12.6 Network and Quality-of-Service for Bio-motion Analysis

The bio-motion systems mentioned above require the integration of a set of miniaturised inertial sensors over a reliable network backbone. A fundamental aspect of BSNs is their wireless inter-connectivity, for which extensive research as described in Chap. 5 has been carried out. Issues related to the reliability and quality-of-service for bio-motion analysis haven't yet been fully addressed. This is due to a number of factors, resulting from a combination of: (a) the bio-motion-specific application requirements and (b) the limitations imposed by the operational characteristics. In particular, a wearable inertial sensing platform requires a relatively high sampling and transmission rate. The resulting network is further characterised by highly varying mobility, as a result of the body motion and the wireless links are affected by varying (posture-related) RF-attenuation and interference effects.

In Chap. 10, we described the network aspects of autonomic sensing, which could address some of these questions. In order to highlight how such a design could handle the highly dynamic network conditions, we describe below a self-reconfigurable BSN for bio-motion applications.

By taking an IEEE 802.15.4-compliant BSN for example, we consider a network of nine sensor nodes, encapsulating the inertial sensing, as well as a sink node. The communication model adopts a client-server approach; the sink node (client) requests periodically data samples (REQ) from the sensor nodes and each sensor node (server) responds to the request with a data packet (REP).

In order to minimise congestion resulting from the high data rates, a TDMA-based approach is adopted for accessing the common medium access. With this approach, each node has a dedicated slot for transmitting its own data. The slots allocation is made in a localised fashion and is based on the timestamp of a REQ from the sink node. In particular, by using the method described in [49], a linear regression model is implemented on each sensor, allowing on-node calculation of the actual transmission delay between the sink node and the sensor node. As shown in Fig. 12.10, nodes can thereby self-allocate their transmissions in non-overlapping time periods.

This scheme was implemented in TinyOS [50] and evaluated on a subject in an office environment. Despite the moderate speed of movement, as shown in Fig. 12.11, the results can be highly variable in terms of the volume of information

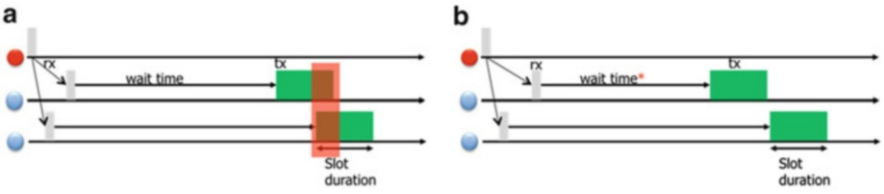


Fig. 12.10 An example of the slot allocation with respect to the actual transmission delay, based on Maroti et al. [49]. If the linear regression model is not adopted (**b**), Nodes may attempt to transmit on overlapping slots (**a**)

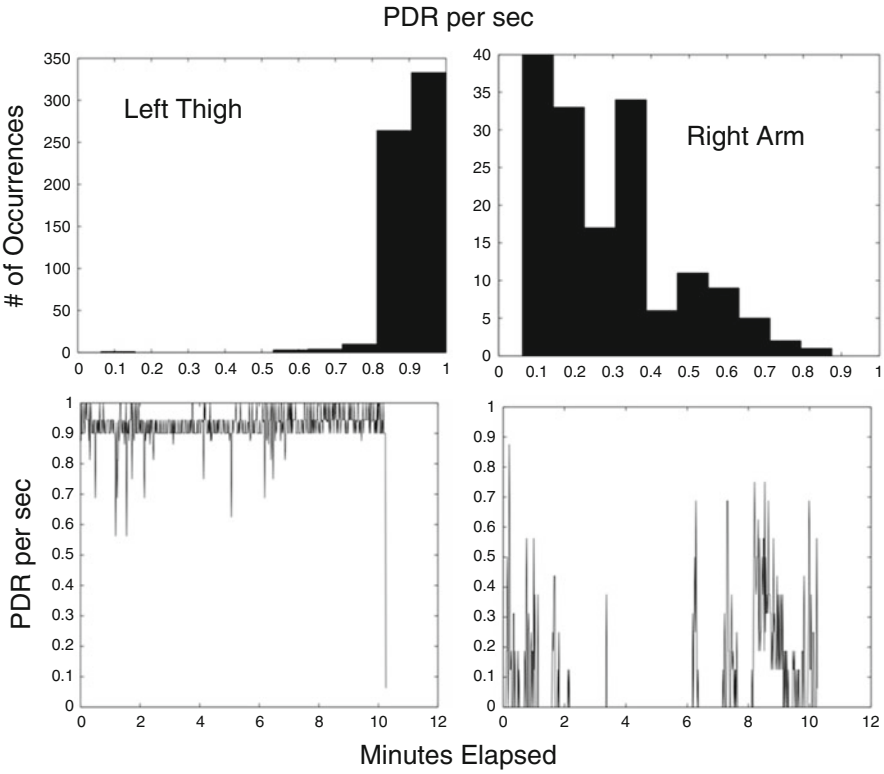


Fig. 12.11 The Packet Delivery Ratio (*PDR*) of the successfully received packets from the sink node for the sensors placed on the left thigh (*left*) and the right arm (*right*). The histogram of PDR (*top*) and the PDR versus the duration of the experiment (*bottom*)

that is available at the sink node for motion reconstruction. More specifically, these results indicate that there is a constant and consistent stream of information arriving from the bio-motion sensor placed on the left thigh throughout the entire duration of the experiment, since the percentage of the successfully received packets per second remains constantly above 90 %. By contrast, a severe lack of information is recorded on the information arriving from the right arm, since the percentage of successfully received packets per second is below 60 %, while for a significant amount of time remains as low as 10 %. In bio-motion analysis terms, this can have a major impact on the human motion estimation, due to the unbalanced volume of the available raw information.

These results show that the transmission range is extremely uneven due to motion for different sensor nodes. Consequently, in conjunction with the self-allocation on transmission slots, the sensors need to adaptively configure the topology of the network, whenever they fail to reach to the sink node with a direct transmission hop. This is achieved by a simple, yet robust, technique that relies on the passive feedback that each sensor node can capture regarding the network

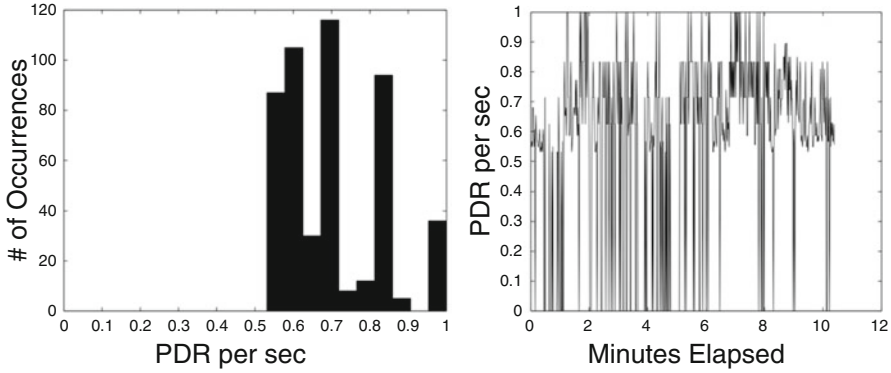


Fig. 12.12 The PDR of the successfully received packets from the sink node for the sensor placed on the right arm with the on-demand topology reconfiguration. The histogram of PDR (*right*) and the PDR versus the duration of the experiment (*left*)

status. In particular if a sensor node cannot receive a distant REQ from the sink, but can capture data activity from its neighbours, it automatically changes mode and self-inhibits the transmission links towards the sink node. This self-inhibition is accompanied by a self-activation on transmission requests within the local network neighbourhood for piggybacking its data, along with the normal single-hop traffic, towards the sink. The result extends the star topology as mentioned in Chap. 5, generated in a reconfigurable and on-demand fashion. The results are presented in Fig. 12.12 and illustrate the improvement in QoS-terms. It can be seen that there is a significant improvement in the volume of the information that captured at the right arm and arrives the sink, since the percentage of the successfully received packets remains above 55 %. In addition, as shown in the right diagram of Fig. 12.12, the flow of information from the sensor placed at the right arm towards the sink is continuous. This improvement is due to the on-demand reformation of the network topology; when the node on the right arm cannot detect the presence of the sink node, the network self-adapts its mode to build an extended-start link, between the specific node and the sink node.

The discussion thus far indicates how a set of preliminary experiments, along with the corresponding observations, can facilitate the design of a BSN that responds more efficiently to the demands of bio-motion-related applications. Driven by the necessity of seamless adaptation in different types of environments while operating continuously on one or more individuals, a more systematic approach should be adopted for addressing the related network re-configurability issues. This implies a two-fold research direction, related to how the captured biometrical information can be exploited for delivering a more balanced BSN network: (a) the design of cross-layer network protocols and (b) the on-node and in-network biomotion data processing for compressing the volume of information that travels towards a central decision point. Equivalently, increasing the intelligence of BSN nodes and signifying their role in deriving context from raw information, is expected to allow a fairer and

balanced flow of information towards the sink node, by promoting the self-adaptation to the varying spatio-temporal conditions and thereby improve the quality of information available for human motion estimation.

12.7 Conclusions

In this chapter, we have presented the basic concept of human motion capture using wearable inertial and magnetic sensor units. We started from the basics of quaternion based orientation representation, to the relationship between quaternion and rotation matrix for motion analysis. The Bayesian fusion framework is provided, followed by a detailed implementation of quaternion-based unscented Kalman filter for orientation estimation. One key challenge for accurate orientation estimation is interference and magnetic disturbance. A vector selection scheme has been proposed to handle such a challenge. This is to protect the orientation estimation algorithm against undesirable situations such as intensive movements and magnetic disturbance. Based on the sensor fusion results, detailed algorithm design in terms of how to estimate human motion based on biomechanical model has been presented. With reference to Chap. 5, issues related to network QoS have been highlighted.

It should be noted that in this chapter, we assumed that the sensor node can provide accurate sensor measurement and didn't pay any attention to the sensor calibration. In general, the main sources of sensor error include temperature related drift, bias, scale factor and misalignment; therefore, appropriate sensor calibration is critical to the accuracy and overall performance of the sensor fusion and motion estimation techniques. Therefore, further work is required for continuous self-calibration with consideration of different temporal characteristics of the sensors combined with the use of temperature controlled casing designs to minimise these errors. It is also possible to model and incorporate temperature related drift characteristics as the prior combined with real-time temperature monitoring to cater for these changes.

References

1. De Aguiar E, Theobalt C, Stoll C, and Seidel HP, Marker-less deformable mesh tracking for human shape and motion capture, In: Proceeding of IEEE Conference on Computer Vision and Pattern Recognition, 2007; 1–9.
2. King B and Paulson L, Motion capture moves into new realms. Computer, 2007, 40(9), 13–16.
3. Wong C, Zhang Z, Kwasnicki R, Liu J, and Yang GZ, Motion Reconstruction from Sparse Accelerometer Data Using PLSR, In: Proceedings of ninth International Conference on Wearable and Implantable Body Sensor Networks (BSN), 2012; 178–183.

4. Yun X and Bachmann E, Design, Implementation, and Experimental Results of a Quaternion-Based Kalman Filter for Human Body Motion Tracking, IEEE Transactions on Robotics, 2006, 22(6): 1216–1227.
5. Vicon Motion Capture System, http://www.vicon.com/
6. BTS SMART-D Motion Capture System, http://www.btsbioengineering.com/
7. Qualisys Motion Capture Systems, http://www.qualisys.com
8. OptiTrack Motion Capture Systems, http://www.naturalpoint.com/optitrack/
9. Moeslund TB, Hilton A, Krüger V, A survey of advances in vision-based human motion capture and analysis, Computer vision and image understanding, 2006; 104(2): 90–126.
10. Vlasic D, Adelsberger R, Vannucci G, Barnwell J, Gross M, Matusik W, Popović J, Practical motion capture in everyday surroundings, ACM Transactions on Graphics (TOG), 007; 26(3): 35.
11. Hasler N, Rosenhahn B, Thormahlen T, Wand M, Gall J, Seidel HP, Markerless motion capture with unsynchronized moving cameras, In proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2009; 224–231.
12. Liu Y, Stoll C, Gall J, Seidel HP, Theobalt C, Markerless motion capture of interacting characters using multi-view image segmentation, In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2011; 1249–1256.
13. Organic Motion, http://www.organicmotion.com/
14. Yoon J, Novandy B, Yoon CH, Park KJ, A 6-DOF gait rehabilitation robot with upper and lower limb connections that allows walking velocity updates on various terrains, IEEE/ASME Transactions on Mechatronics, 2010; 15(2): 201–215.
15. Schabowsky CN, Godfrey SB, Holley RJ, Lum PS, Development and pilot testing of HEXORR: hand EXOskeleton rehabilitation robot, Journal of Neuroengineering and Rehabilitation, 2010; 7(1):36.
16. Gypsy 7 Motion Capture System, http://www.metamotion.com/gypsy/gypsy-motion-capture-system.htm
17. Ekso Bionic Suit, http://www.eksobionics.com/ekso
18. ArmeoPower Arm Exoskeleton, http://www.hocoma.com/products/armeo/
19. Lokomat Gait Orthosis, http://www.hocoma.com/products/lokomat/
20. Tao Y, Hu H, A novel sensing and data fusion system for 3D arm motion tracking in tele-rehabilitation, IEEE Transactions on Instrumentation and Measurements, 2008; 57(5):1029–1040.
21. Huang S, Sun S, Huang Z, Wu J, Ambulatory real-time micro-sensor motion capture, In: Proceedings of the 11th ACM international conference on Information Processing in Sensor Networks, 2012; 107–108
22. Animazoo Mocap Suits, http://www.animazoo.com/
23. Roetenberg D, Luinge H and Slycke P, Xsens MVN: full 6DOF human motion tracking using miniature inertial sensors, Xsens Motion Technologies BV Tech. Rep., 2009.
24. 3DSuit 3DTracker, http://www.3dsuit.com/
25. Micro-Sense Mocap, http://www.microsenstech.com and http://snarc.ia.ac.cn
26. Zhang Z, Wu J, A novel hierarchical information fusion method for three-dimensional upper limb motion estimation, IEEE Transactions on Instrumentation and Measurement, 2011; 60(11):3709–3719.
27. McAllister LB, A quick introduction to quaternions, Pi Mu Epsilon Journal, 1989; 9(1):23–25.
28. Inertial Labs Inc, http://www.inertiallabs.com
29. Chou JCK, Quaternion kinematic and dynamic differential equations, IEEE Transactions on Robotics and Automation, 1992; 8(1):53–64.
30. Trawny N, and Stergios IR, Indirect Kalman filter for 3D attitude estimation, University of Minnesota, Dept. of Comp. Sci. & Eng., Tech. Rep, 2005.
31. Kuipers JB, Quaternions and rotation sequences, Princeton university press, 1999, 127–143.
32. Sabatini AM, Quaternion-based strap-down integration method for applications of inertial sensing to gait analysis, Medical and Biological Engineering and Computing, 2005, 43(1): 94–101.

33. Chen Z, Bayesian filtering: From Kalman filters to particle filters and beyond, Statistics, 2003; 182(1):1–69.
34. Arulampalam MS, Maskell S, Gordon N, Clapp T, A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking, IEEE Transactions on Signal Processing, 2002; 50(2):174–188.
35. Koch W, On Bayesian tracking and data fusion: a tutorial introduction with examples, IEEE Aerospace and Electronic Systems Magazine, 2010; 25(7):29–52
36. Hol JD, Schön TB, Luinge H, Slycke PJ and Gustafsson F, Robust real-time tracking by fusing measurements from inertial and vision sensors, Journal of Real-Time Image Processing, 2007; 2(2):149–160.
37. Jurman D, Jankovec M, Kamnik R, & Topič M, Calibration and data fusion solution for the miniature attitude and heading reference system, Sensors and Actuators A: Physical, 2007; 138(2):411–420.
38. Shen SC, Chen CJ and Huang HJ, A new calibration method for MEMS inertial sensor module, in: proceeding 11th IEEE International Workshop on of Advanced Motion Control, 2010.
39. Zhang Z, Yang GZ, Calibration of Miniature Inertial and Magnetic Sensor Units for Robust Attitude Estimation, IEE Transactions on Instrumentation and Measurements, 2014; 63(3):711–718.
40. Kraft E, A quaternion-based unscented Kalman filter for orientation tracking, In: Proceedings of the sixth International Conference of Information Fusion, 2003.
41. Zhang ZQ, Ji LY, Huang ZP and Wu JK, Adaptive Information Fusion for Human Upper Limb Movement Estimation, IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans, 2012; 42(5):1100–1108.
42. Luinge H and Veltink P, Inclination measurement of human movement using a 3-D accelerometer with autocalibration, IEEE Trans. Neural Syst. Rehabil. Eng., 2004; 12(1): 112–121.
43. Young A, Use of body model constraints to improve accuracy of inertial motion capture, in: proceeding of International Conference on Wearable and Implantable Body Sensor Networks (BSN), 2010; 180–186.
44. Sabatini A, Quaternion-based extended Kalman filter for determining orientation by inertial and magnetic sensing, IEEE Transactions on Biomedical Engineering, 2006; 53(7): 1346–1356.
45. Kang C and Park C, Attitude estimation with accelerometers and gyros using fuzzy tuned Kalman filter, in: proceeding of European Control Conference, 2009; 3713–3718.
46. Sun S, Meng X, Ji L, Wu J, Wong WC, Adaptive sensor data fusion in motion capture, In: Proceedings of 13th Conference on Information Fusion, 2010; 1–8.
47. Zhang Z, Meng X, Wu J, Quaternion-Based Kalman Filter With Vector Selection for Accurate Orientation Tracking, IEEE Transactions on Instrumentation and Measurement, 2012; 61(10):2817–2824.
48. Meng XL, Zhang ZQ, Sun SY, Wu JK, Wong WC, Biomechanical model-based displacement estimation in micro-sensor motion capture, Measurement Science and Technology, 2012: 23(5): 055101.
49. Maróti M, Kusy B, Simon G and Lédeczi, The flooding time synchronization protocol, In: Proceedings of the 2nd international conference on Embedded networked sensor systems, 2004; 39–49.
50. TinyOS operating system, http://www.tinyos.net/

Appendix A: Wireless Sensor Development Platforms

Benny Lo and Guang-Zhong Yang

A.1 Introduction

The development of BSN has been facilitated by the rapid advances in *Wireless Sensor Networks* (WSNs) in recent years. BSN harnesses several allied technologies that underpin the development of pervasive sensing for healthcare, wellbeing, sports and other applications that require "ubiquitous" and "pervasive" monitoring of physical, physiological and biochemical parameters in any environment and without activity restriction and behaviour modification. Key to the development of BSN are technologies that address miniaturised biosensor design, suitable for both wearable and implantable devices, biocompatibility and materials to ensure long-term deployment, low-power wireless communication, integrated circuits and systems, power scavenging techniques from the body, autonomic sensing and standards and integration. Major technical hurdles of BSNs are related to continuous sensing and monitoring, requiring long-term stability of the sensors and low-power operation, also necessitating bio-inspired design (e.g., bio-inspired mix-signal ASIC) and power scavenging techniques ultimately for battery-less operation. For device level inter-connectivity, BSNs can be wired (e.g., interconnecting with smart fabric) or wireless (making use of common wireless sensor networks and standards.

Since the introduction of the concept of WSN, a large number of development platforms have been introduced [1, 2]. As shown in Fig. A.1, there are nearly 120 new WSN platforms introduced since 1998. The number of new platforms introduced peaked in 2004–2007 and remained relatively static 2008 onwards. The reason for the decline is largely due to the maturing of the sensing platforms

B. Lo • G.-Z. Yang, PhD (✉)
The Hamlyn Centre, Imperial College London, London SW7 2BY, UK
e-mail: g.z.yang@imperial.ac.uk

G.-Z. Yang (ed.), *Body Sensor Networks*, DOI 10.1007/978-1-4471-6374-9,

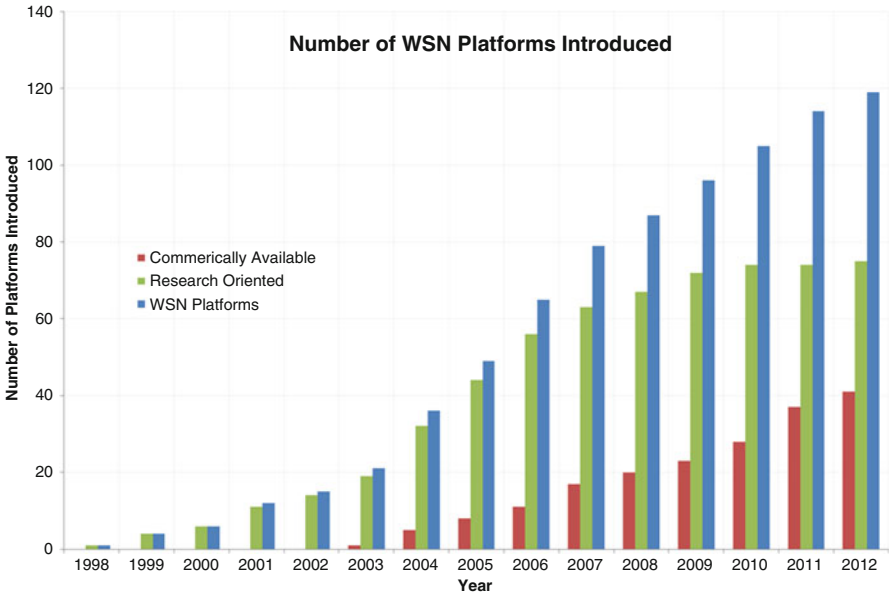


Fig. A.1 Number of new WSN hardware platforms introduced in recent years

and the increasing number of commercially available platforms. This is apparent by comparing the trend of the platforms designed for research purposes and those commercial platforms as shown in Fig. A.1.

A more detailed list of WSN platforms is provided in Table A.1. It must be pointed out that this table is by no means exhaustive and unintentional omission is likely. The general design and requirements for BSNs can be different from typical WSN applications. However, many of the WSN development platforms have been adapted or directly adopted for wearable BSN applications. For implantable applications, bespoke electronics and new ASIC designed are used due to stringent power and resource constraints. After the introduction of dedicated BSN platforms, such as the BSN node [3, 4], Pluto [5], Shimmer [6], purposely built platforms have been used in BSN research. This appendix outlines the system architecture of common WSN platforms and provides an overview of some of the main hardware components involved.

A.2 System Architecture

The system architectures of all WSN development platforms are relatively similar, and they mainly consist of the following components:

- Processor – the computer of the sensor node
- Wireless Communication – the wireless link between sensor nodes
- Memory – external storage for sensor readings or program images

Table A.1 Wireless sensor network development platforms

Platforms	CPU	Clock (MHz)	RAM/Flash/ EEPROM	Radio transceiver	BW (bps)	Freq. (MHz)	OS	Year	Organisation
WeC	Atmel AT90LS8535	4	512/8 K/32 K	RFM TR1000	10 k	916.5	TinyOS	1998	UC Berkeley
AWAIRS 1	Intel StrongArm SA1100	59–206	1 M/4 M	Conexant RDSSS9M	100 k	900	MicroC/OS	1999	Rockwell
Micromote	Atmel AT90S8535 Atmel AT90LS2343	4 10	512/8 K/32 K 128/2 k	RFM TR1000	10 k	916.5	TinyOS	1999	UC Berkeley
Rene 1	Atmel AT90LS8535	4	512/8 K/32 K	RFM TR1000	10 k	916.5	TinyOS	1999	UC Berkeley
Rene 2	Atmel Atmega 163	8	1 K/16 K/32 K	RFM TR1000	10 k	916.5	TinyOS	2000	UC Berkeley
uAMPS	StrongARM SA1100	206	16 M/512 K (ROM)	National LMX3162	1 M	2,450	uOS (eCOS microkernel)	2000	MIT
BT node[ab]	Atmel Atmega 128 L	8	4 K/128 K/4 K	ZV4002 BT/TI (Chipcon) CC1000	1 M 38.4 k	2,400 868	TinyOS	2001	ETH Zurich
Dot	Atmel Atmega 163	8	1 K/16 K/32 K	RFM TR1000	10 k	916.5	TinyOS	2001	UC Berkeley
RSC WINS	DEC SA1100	206	1 M/4 M	RDSSS9M	100 k	900		2001	Rockwell
SpotON	Dragonball EZ	16	2 M/2 M	RFM TR1000	10 k	916.5		2001	University of Washington/ Intel
Smart-its[b] (Lancaster)	Microchip PIC18F252	8	3 K/48 K/64 K	Radiometrix	64 k	433	Smart-its	2001	Lancaster
Smart-its (Teco)	Atmel ATMega 103 L	4	4 K/128 K	Ericsson BT	1 M	2,400	Smart-its	2001	University of Karlsruh
CENS Medusa MK2	Atmel ATMega128L Atmel AT91FR4081	40	136 K/1 M	RFM	10 k	916	Palos	2002	UCLA
iBadge	Atmel Atmega103L	6	4 K/128 K	Ericsson BT	1 M	2,400	Palos	2002	UCLA
Mica	Atmel Atmega 128 L	4	4 K/128 K/512 K	RFM TR1000	40 k	916.5	TinyOS	2002	UC Berkeley
iMote1	Zeevo ZV4002 (ARM)	12–48	64 K/512 K	Zeevo BT	720 k	2,400	TinyOS	2003	Intel
Mica2[ab] [10, 11]	Atmel Atmega 128 L	8	4 K/128 K/512 K	TI (Chipcon) CC1000	38.4 k	900	TinyOS	2003	UC Berkeley/ Crossbow
Mica2Dot[ab]	Atmel Atmega 128 L	4	4 K/128 K/512 K	TI (Chipcon) CC1000	38.4 k	900	TinyOS	2003	UC Berkeley/ Crossbow
Mantis Nymph	Atmel Atmega128L	4	4 K/128 K/512 K	TI (Chipcon) CC1000	38.4 k	900	Mantis	2003	U Colorado
RFRAIN	TI (Chipcon) CC1010	3–24	2 K/32 K	TI (Chipcon) CC1010	76.8 k	0.3–1,000	RFRAIN Libraries	2003	MIT

(continued)

Table A.1 (continued)

Platforms	CPU	Clock (MHz)	RAM/Flash/ EEPROM	Radio transceiver	BW (bps)	Freq. (MHz)	OS	Year	Organisation
U3	Microchip PIC18F452	0.031–8	1 K/32 K/256	CDC-TR-02B	100 k	315	Pavenet	2003	U Tokyo
AquisGrain	Atmel Atmega128L	4	4 K/128 K/512 K	TI (Chipcon) CC2420	250 k	2,400	–	2004	Philips Research
BEAN	TI MSP430F149	8	2 K/60 K/512 K	TI (Chipcon) CC1000	76.8 k	0.3–1,000	YATOS	2004	Universidade Federal de Minas Gerais
BSN node v2[b] [3, 4]	TI MSP430F149	8	2 K/60 K/512 K	TI (Chipcon) CC2420	250 k	2,400	TinyOS	2004	Imperial College London/ Sensixa
CIT sensor node	Microchip PIC16F877	20	368/8 K	Nordic nRF903	76.8 k	868	TinyOS	2004	Cork Institute of Technology
DSYS25[ab]	Atmel Atmega 128	4	4 K/128 K	Nordic nRF2401	1 M	2,400	TinyOS	2004	UCC
eXtreme Scale Mote (XSM)	Atmel ATmega128L	8	4 K/128 K	TI (Chipcon) CC1000	76.8 k	433	TinyOS	2004	Ohio State U and Crossbow
Fleck 1 Fleck 2	Atmel Atmega 128	4	4 K/128 K/512 K	Nordic nRF903	76.8 k	902–928	TinyOS	2004	CSIRO
MicaZ[ab]	Atmel Atmega 128 L	8	4 K/128 K	TI (Chipcon) CC2420	250 k	2,400	TinyOS	2004	Crossbow
MITes[b]	Nordic nRF24E1	16	512/4 K	Nordic nRF24E1	1 M	2,400	–	2004	MIT
Parasitic	Silicon C8051F311	25	1.2 k/16 k	BlueRadios BR-C11A	1 M	2,400		2004	MIT
Particle2/29[a]	Microchip PIC 18 F6720	20	4K128K/512 K	RFM TR1001	125 k	868.35	Smart-its	2004	Lancaster/University of Karlsruhe
Pluto[b] [5]	TI MSP430F149	8	4 K/60 K/512 K	TI (Chipcon) CC2420	250 k	2,400	TinyOS	2004	Harvard
ProSpeckz[b]	Cypress CY8C2764	12	256/16 K	TI (Chipcon) CC2420	250 k	2,400	Speckle net	2004	University of Edinburgh
Spec	8-bit AVR-like RISC core	4–8	3 K	FSK Transmitter	100 k		TinyOS	2004	UC Berkeley
Telos[ab]	TI MSP430F149	8	2 K/60 K/512 K	TI (Chipcon) CC2420	250 k	2,400	TinyOS	2004	UC Berkeley/ Sentilla (Moteiv)

Ember RF module[a]	Atmel Atmega 128 L	8	4 K/128 K	Ember 2420	250 k	2,400	EmberNet	2005	Ember
EnOcean TCM120	Microchip PIC18F452	10	1.5 K/32 K/256	Infineon TDA 5200	120 k	868	TinyOS	2005	Helmut Schmidt University
eyesIFXv2[b]	TI MSP430F1611	8	10 K/48 K	Infineon TDA5250	64 k	868	TinyOS	2005	TU Berlin
Hogthrob	Atmel Atmega 128 L Xilinx Spartan 3 XC3S400	8 48	4 K/128 K 56 K	Nordic nRF2401	1 M	2,400		2005	Technical University of Denmark
iMote2	Intel PXA 271	13–104	256 K/32 M	TI (Chipcon) CC2420	250 k	2,400	TinyOS	2005	Intel
MediMesh[b]	TI MSP430F1611	8	10 K/48 K	TI (Chipcon) CC2420	250 k	2,400	TinyOS	2005	Chinese University of Hong Kong
Mulle[a]	Renesas M16C/62P	10	31 K/384 K/2 M	Mitsumi WML C46AHR	2.178 M	2,400	Mulle	2005	Eistec
RISE	TI (Chipcon) CC1010 EM	3–24	2 K/32 K	TI (Chipcon) CC1010 EM	76.8 k	0.3–1,000	TinyOS	2005	UC Riverside
ScatterWeb ESB	TI MSP430F149	8	2 K/60 K	RFM TR1001	115.2 k	868	TinyOS/Contiki	2005	Freie Universität Berlin
Stack[b]	Silicon C8051F206	25	1 K/8 K	RFM TR1000	115.2 k	916		2005	MIT
Solar biscuit	Microchip PIC 18LF452	7.3728	1 K/32 K	TI (Chipcon) CC1000	76.8 k	315		2005	The University of Tokyo
Tmote sky/TelosB[ab] [12]	TI MSP430F1611	8	10 K/48 K/1 M	TI (Chipcon) CC2420	250 k	2,400	TinyOS	2005	UC Berkeley/ Sentilla (Moteiv)
XYZ sensor node	OKI ML67Q500x (ARM/THUMB)	1.8–57.6	4 K/256 K/512 K	TI (Chipcon) CC2420	250 k	2,400	SOS	2005	Yale
Ant[ab]	TI MSP430F1232	8	256/8 K	Nordic nRF24AP1	1 M	2,400	Ant	2006	Dynastream Innovation Inc.
BT node rev3[ab] [13]	Ateml ATmega128L	8	4 K/128 K/4 K	ZV4002 BT/TI (Chipcon) CC1000	1 M 38.4 k	2,400 868	BTnut TinyOS	2006	ETH Zurich/Art of Technology
Cookies	Analog Device ADuC841 Xilinx XC3S200 Spartan3	16 48	4 K/62 K 234 K	ConnectBlue OEMSPA13i	1 M	2,400		2006	Universidad Politécnica de Madrid
DSRPN	TI OMAP5912 TMS320C55x	192 200	250 K 128 K			340		2006	Chinese Academy of Science

(continued)

Table A.1 (continued)

Platforms	CPU	Clock (MHz)	RAM/Flash/ EEPROM	Radio transceiver	BW (bps)	Freq. (MHz)	OS	Year	Organisation
e-Watch[b] [14]	Philips LPC2106 ARM7TDMI	60	64 K/128 K	SMARTM Bluetooth	1 M	2,400		2006	Carnegie Mellon University
IECAS	Silicon C8051F121	100	84 K/128 K	TI (Chipcon) CC2420	250 k	2,400		2006	Institute of Electronics Chinese Academy of Science
LEAP	Intel PXA255	400	128 M/64 M/ 192 M	TI (Chipcon) CC2420 Atheros 5006XS	250 k 20 M	2,400 2,437	LEAP software framework	2006	UCLA
Nano-Qplus	Atmel ATmegal28	8	4 K/128 K	TI (Chipcon) CC2420	250 k	2,400	Nano-Qplus OS	2006	ETRI Korea
SENTIO	Atmel Atmega 128L	8	4 K/128 K	TI (Chipcon) CC2420	250 k	2,400		2006	Mid Sweden University
Shimmer[b] [6]	TI MSP430F1611	8	10 K/48 K/2G	TI (Chipcon) CC2420 Mitsumi WML-C46N	250 k 2.178 M	2,400 2,400	TinyOS	2006	Intel
Tinynode 584[a]	TI MSP430F1611	8	10 K/48 K/512 K	Semtech XE1205	152 k	868/902	TinyOS	2006	EPFL/Tinynode
TUTWSN[b]	Microchip PIC18LF4620	10	4 K/64 K/1 K	Nordic nRF905	50 k	433/868/ 915		2006	Tampere University of Technology
Tyndall Mote[b]	Atmel Atmega 128L	8	4 K/128 K/32 K	Nordic nRF2401	1 M	2,400	TinyOS	2006	Tyndall
ubER-Badge[b] [15]	TI MSP430F149	8	4 K/60 K/256 M	TI (Chipcon) CC1010	76.8 k	0.3–1,000		2006	MIT
uPart0140ilmt[a]	Microchip rfPIC16F675	4	64/1 K	Microchip rfPIC16F675	19.2 k	868	Smart-it	2006	University of Karlsruh
ZN1	Renesas H8S/2218	4–24	128 K/128 K (ROM)	TI (Chipcon) CC2420	250 k	2,400		2006	Hitachi
AVRraven[a]	Atmel AtMega1284p Atmel ATmega3290p	20	128 K/16 K/256 K	Atmel AT86RF230	250 k	2,400	Atmel Studio	2008	Atmel
FireFly	Atmel ATmega1281	16	8 K/128 K/2G	TI (Chipcon) CC2420	250 k	2,400	Nano-RK	2007	Carnegie Mellon University
Fleck 3	Atmel Atmega128L	8	4 K/128 K/1 M	Nordic nRF905	50 k	433/868/ 915	TinyOS	2007	CSIRO

GWNode [16]	Microchip PIC18LF8722	40	128 K/64 K/1 K	BiM 1	10 k	173		2007	Southampton
iSense[ab]	Jennic JN5148	4–32	128 K/512 K	802.15.4	250 k	2,400	iSense	2007	coalesenses
JN5121[ab]	OpenRSIC1000	16	96 K/64 K	IEEE802.15.4	250 k	2,400	JenNet/Jenie	2007	Jennic
mPlatform[b] [17]	TI MSP430F1611 OKI ML67Q5003 XC2C512 CoolRunner-II CPLD	8 60 32–200	10 K/48 K 32 K/512 K	TI (Chipcon) CC2420	250 k	2,400	mPlatform	2007	Microsoft
NeoMote[a] (EcoWizard)	Atmel ATMega 128L	8	4 K/128 K	TI (Chipcon) CC2420	250 k	2,400		2007	MEMSIC (Crossbow)/ Sumitomo Precision
NWSP[b]	Altera EP2C20F256C7N	500	26 K	National LMX9830	1 M	2,400	eCos	2007	Nokia
Power meter WSN	TI (Chipcon) CC1010	3–24	2 K/32 K	TI (Chipcon) CC1010	76.8 k	0.3–1,000		2007	University of New South Wales Sydney
S-Mote	TI (Chipcon) CC2430	32	8 K/128 K	TI (Chipcon) CC2430	250 k	2,400	RETOS	2007	Yonsei University Korea
Sun SPOT[ab] [18]	ARM920T	180	512 K/4 M	TI (Chipcon) CC2420	250 k	2,400	Java J2ME	2007	Orcale (Sun)
Tmote mini[ab]	TI MSP430F1611	8	10 K/48 K/1024 K	TI (Chipcon) CC2420	250 k	2,400	TinyOS	2007	Sentilla (Moteiv)
WeBee3[b]	TI (Chipcon) CC2431	16	8 K/128 K	TI (Chipcon) CC2430/ CC2431	250 k	2,400		2007	Lucerne University of Applied Sciences
ZigBit[a]	Atmel ATmega1281V	16	8 K/128 K	Atmel AT86RF230RF	250 k	2,400	ZigBit Development Kit/TinyOS	2007	MeshNetics
Arduino BT[ab]	Atmel ATmega328	16	2 K/32 K/1 K	Blue giga WT11i	3 M	2,400	Arduino IDE/Java	2008	Arduino
BSN node v3[ab] [19]	TI MSP430F1611	8	10 K/48 K/4 M	TI (Chipcon) CC2420	250 k	2,400	TinyOS/BSNOS	2008	Imperial College London/ Sensixa

(continued)

Table A.1 (continued)

Platforms	CPU	Clock (MHz)	RAM/Flash/ EEPROM	Radio transceiver	BW (bps)	Freq. (MHz)	OS	Year	Organisation
Dalian WSN	LPC2138	30	32 K/512 K	TI (Chipcon) CC2420	250 k	2,400		2008	Dalian University of Technology
EPIC mote	TI MSP430F1611	8	10 K/48 K/512 K	TI (Chipcon) CC2420	250 k	2,400	TinyOS	2008	UC Berkeley
MASN	Ember EM250 (XAP2b)	12	5 K/128 K	Ember EM250	250 k	2,400	TinyOS	2008	Rochester Institute of Technology
Shimmer 1.3[ab]	TI MSP430F1611	8	10 K/48 K/2G	TI (Chipcon) CC2420 Roving Networks RN-42 (Bluetooth)	250 k 3 M	2,400 2,400	TinyOS	2008	Shimmer
WBSN[b]	TI MSP430F1611	8	10 K/48 K/124 K	TI (Chipcon) CC2420	250 k	2,400	TinyOS	2008	Queen Mary's University of London
FemtoNode	FemtoJava	56	0.5 K	TI (Chipcon) CC2420	250 k	2,400	API-Wireless	2009	Universidade Federal do Rio Grande do Sul
iCubes	Silicon C8051F320	25	2 K/16 K	TI (Chipcon) CC2500	500 k	2,400	Over The Air Programmed (OTAP)	2009	Technical University of Crete
Kmote-B[a]	TI MSP430F1611	8	10 K/48 K/512 K	TI (Chipcon) CC2420	250 k	2,400	TinyOS	2009	INETCH Co./ TinyOS Mall
PowWow – CAIRN	TI MSP430F1612	8	5 K/55 K	TI (Chipcon) CC2420	250 k	2,400	PowWow	2009	Inria
SenseNode[a]	TI MSP430F1611	8	10 K/48 K/1024 K	TI (Chipcon) CC2420	250 k	2,400	TinyOS	2009	Genet
Sensium[ab]	Toumaz Sensium TZ1030	1	64 k	Toumaz Sensium TZ1030	50 k	928			Toumaz
Shimmer 2[ab]	TI MSP430F1611	8	10 K/48 K/2G	TI (Chipcon) CC2420 Roving networks RN-42 (Bluetooth)	250 k 3 M	2,400 2,400	TinyOS	2009	Shimmer
T-node[a]	Atmel ATMega 128L	8	4 K/128 K	TI (Chipcon) CC1000	38.4 k	868	SOWNet	2009	SOWNet Technologies
Tempo 3.1[b] [20]	TI MSP430F1611	8	10 K/48 K	Roving Networks RN-41 (Bluetooth)	1 M	2,400	TEMPOS	2009	University of Virginia

WeBee3G	TI (Chipcon) CC2430	16	8 K/128 K	TI (Chipcon) CC2430	250 k	2,400		2009	Lucerne University of Applied Sciences
Egs	Atmel SAM3U	96	52 K/256 K	TI (Chipcon) CC2520 WML-C46	250 k 2.178 M	2,400 2,400		2010	Harbin Institute of Technology
ENS[b]	TI MSP430F2410	16	4 K/56 K	TI (Chipcon) CC2420	250 k	2,400	ENS Network software	2010	University of Edinburgh
G-node[a]	TI MSP430F2418	16	8 K/116 K/1 M	TI (Chipcon) CC1101	1.2 k	868/915	SOWNet	2010	SOWNet Technologies
Indriya CS-03A14[a]	Atmel ATMega 128L	8	4 K/128 K	XBee	250 k	2,400	TinyOS	2010	Indrion
RecoNode	Xilinx Virtex-4 FX20	450	128 M/	TI (Chipcon) CC2520	250 k	2,400	RTOS/TinyOS	2010	University of Denver
Shimmer 2R[ab]	TI MSP430F1611	8	10 K/48 K/2G	TI (Chipcon) CC2420 Roving Networks RN-42 (Bluetooth)	250 k 3 M	2,400 2,400	TinyOS	2010	Shimmer
VE209-ST VEmesh[a]	TI MSP430	16	512B/8 K	Semtech SX1211/SX1231	50 k	902		2010	Virtual Extension
XBee-PRO ZB[ab]	Freescale HCS08	50.33	2 K/32 K	ZigBee	250 k	2,400	XBee SDK	2010	Digi International
Z1[a]	TI MSP430F2617	16	8 K/92 K	TI (Chipcon) CC2420	250 k	2,400	TinyOS/ContikiOS	2010	Zolertia
AS-XM1000[a]	TI MSP430F2618	16	8 K/116 K/1024 K	TI (Chipcon) CC2420	250 k	2,400	TinyOS/ContikiOS	2011	Advanticsys
Indriya_ DP_01A11[a]	TI MSP430F2618	16	8 K/116 K	TI (Chipcon) CC2520	250 k	2,400	TinyOS	2011	Indrion
Indriya_ DP_03A20[a]	Atmel ATMega 128L	8	4 K/128 K	GainSpan GS1011 M	1 M	2,400	TinyOS	2011	Indrion
IRIS[a]	Atmel ATmega1281	16	8 K/128 K/512 K	Atmel AT86RF230	250 k	2,400	Mote Works/Mote Runner	2011	MEMSIC (Crossbow)
Libelium Wasp mote[ab]	Atmel ATMega1281	8	8 K/128 K/2GB	XBee module	250/230/ 230 k	2,400/900/ 848	Libelium (Over the Air) OTA	2011	Libelium
LOTUS[a]	NXP LPC1758	100	64 K/512 k/64 M	Atmel AT86RF231	250 k	2,400	Free RTOS/ TinyOS/Mote Runner	2011	MEMSIC (Crossbow)
MTM-CM5000-MSP[a] MTM-CM3000-MSP[a] MTM-CM3300-MSP[a] MTM-CM4000-MSP[a]	TI MSP430F1611	8	10 K/48 K/1024 K	TI (Chipcon) CC2420	250 k	2,400	TinyOS/ContikiOS	2011	Advanticsys
Preon32[a]	ARM Cortex-M3	72	64 K/256 K	Atmel AT86RF231	2 M	2,400	Java	2011	Virtenio
WiSMote Dev[a]	TI MSP430F5437	8	16 K/256 K/1 M	TI (Chipcon) CC2520	250 k	2,400	Contiki	2011	Arago Systems

(continued)

Table A.1 (continued)

Platforms	CPU	Clock (MHz)	RAM/Flash/ EEPROM	Radio transceiver	BW (bps)	Freq. (MHz)	OS	Year	Organisation
CoSeN/MantaroBlocks[a]	Atmel ATmega32A4	32	4 K/32 K/1 K	Atmel AT86RF231	250 k	2,400	Atmel Studio	2012	UMBC/Mantaro
FireFly3	Atmel ATmega128RFA1	16	16 K/128 K/4 K	Atmel ATmega128RFA1	250 k	2,400	Nano-RK	2012	Carnegie Mellon University
M12 module[a]	Freescale MC13224v (ARM7)	26	96 K/128 K	Freescale MC13224v (802.15.4)	250 k	2,400	Contiki OS	2012	RedWire
panStamp[a]	Atmel Atmega328p	8	2 K/32 K/1 K	TI (Chipcon) CC1101	1.2 k	868/915	panStamp	2012	panStamp
WiSMote mini[a]	Atmel ATmega128RFA2	16	16 K/128 K/4 K	Atmel ATmega128RFA2	2 M	2,400	Contiki	2012	Arago Systems

[a]Commercially available platforms
[b]Designed or used for BSN application

- Sensor Interface – interface with sensors and other devices
- Power Supply – the power source of the sensor node

A.2.1 Processor

Most WSN platforms are based on COTS (*Commercial Off-The-Shelf*) components, and the development of WSN depends extensively on the rapid advances of microprocessors. The majority of the WSN platforms use 8-bit or 16-bit RISC processors, such as the Atmel ATmega and *Texas Instruments*' (TI) MSP430 microcontrollers, due to their low power design, integrated multi-sensor interfaces and widely available developing tools. More recently, some platforms have started to use the ARM Cortex processors due to an increase in computational power of the sensor nodes.

A.2.2 Wireless Communication

Wireless communication is the most power-demanding component of WSNs. This often accounts for more than 50 % of the overall power budget of a sensor node [7]. Parallel to the development of micro-power radio transceivers, such as the Pico radio [8], existing research in WSNs has been focused on developing energy-efficient protocols and routing strategies. The three main components of wireless communication include the radio transceiver, antenna, and communication protocols.

A.2.2.1 Radio Transceiver

Among different radio transceivers, the TI (Chipcon) CC2420 is the most popular of the IEEE 802.15.4 chipsets used in the WSN platforms. Since the introduction of *SoC* (System on a Chip) by integrating both microcontroller and radio transceiver onto a single chip (e.g., Nordic nRF24E1 and ATmega128FEA1), new WSN platforms are increasingly using bespoke SoC chipsets to minimise the footprint and reduce the overall size of the sensor node.

A.2.2.2 Antenna

For WSN and BSN, antenna design is important, although the majority of the current WSN platforms use COTS antennas. Among these, the ceramic antennas are most commonly used in WSN platforms due to their small footprints and consistent signal quality for mass production. For BSNs, particularly for implantable applications, miniaturised antenna design remains a key research topic, as mentioned in earlier chapters of this book.

A.2.2.3 Communication Protocol

Depending on the hardware and the operating systems used, early WSN platforms mainly relied on proprietary communication protocols. The introduction of the IEEE 802.15.4 standard enables the standardisation of communication between WSN platforms. Many recent WSN platforms have adopted this standard as the basis for their wireless communication protocol. Examples of this include Telos, MicaZ, Pluto, iMote2, Tmote sky, XYZ node, ProSpeckz and the BSN node.

Due to the broad range of WSN applications, the 802.15.4 standard defines only the MAC and physical layers of the communication protocol. This enables the design of application-specific protocols. For example, Zigbee is built based on the 802.15.4 standard and is designed to ease the inter-operation between different devices. Zigbee specifies all the protocol layers required for forming a wireless network and it also provides an interface for application development.

As mentioned in Chap. 5, although Zigbee is designed for low power sensor communication and enables interoperability, the power consumption of Zigbee is still relatively high for BSN applications. Following the release of the Bluetooth Low Energy (Bluetooth SMART) and the standardisation of IEEE 802.15.6, it is envisaged that new BSN platforms will increasingly rely on these new communication protocols, whereas WSN applications will remain to use 802.15.4 and Zigbee protocols.

A.2.3 Memory

Since limited *Random Access Memory* (RAM) is provided by MCUs, most WSN platforms are designed with an external flash memory or *Electrically Erasable Programmable Read-Only Memory* (EEPROM). Due to the non-volatile nature of the EEPROM, it is used in most embedded systems for storing configuration information because it does not require power to retain the stored data. It is also used as an immediate storage for sensor readings. For instance, in order to perform feature extraction or filtering of the sampled data, the EEPROM can be used as a processing buffer for these algorithms. Another use of the EEPROM is for storing program images. In addition to EEPROM, some WSN platforms have designed with a micro-SD card interface for data logging. Although micro-SD card consumes more power than EEPROM, the large storage capacity is ideal for long-term data logging.

A.2.4 Sensor Interface

To enable practical application development, most WSN platforms offer analogue and digital sensor interfaces.

A.2.4.1 Analogue Interface

Sensors such as simple photo resistors and thermistors, or more complex gyroscope and condenser microphones, generally provide analogue readings. Most WSN or wearable BSN platforms are equipped with ADC interfaces for data sampling and acquisition. For instance, the Atmel Atmega128L MCU has an eight-channel 10-bit ADC that can sample at a rate up to 15.4 ksps (*kilo-samples per second*), whereas the TI MSP430 microcontroller has a 12-bit ADC, which provides a higher precision than that of the Atmel processor. In addition to ADCs, some platforms are equipped with *Digital-to-Analogue Converter* (DAC) for controlling sensors or actuators.

A.2.4.2 Digital Interface

Since analogue readings are prone to voltage drift caused by the depletion of the battery power, sensors such as the 9-axis inertial motion unit MPU-9150 [9] provide direct digital readings. As sensor data is relatively small in size, serial communication is mainly used for interfacing with digital sensors, and the three most commonly used serial communication protocols are I^2C, SPI and UART.

A.2.4.3 Integrated Sensors

To ease application development, many WSN or wearable BSN platforms have built-in sensors such as humidity, temperature, inertial, magnetic and photo sensors. With integrated sensor board design, the hardware platform can be made more compact and immune to noise induced by cables and connectors. However, integrating sensors on the hardware platforms can limit the general use of the platforms as different applications may have varying sensor requirements.

A.2.5 Power Supply

Currently, power supply is the main determining factor for the size and lifetime of the WSN or BSN hardware. Similar to mobile phones, the battery or alternative power source is often the largest single component of these sensor nodes. To miniaturise the sensor nodes, a number of alternatives have been proposed. With recent advances in power harvesting technologies, it has been demonstrated that WSN sensor nodes can be powered by motion, temperature gradient, inductive coupling and other alternatives. However, batteries still remain the main source of power for current WSN or wearable BSN platforms. Among different battery technologies, Lithium-ion or Lithium Polymer batteries are the most popular choice

for sensor hardware because of their high power density. Although zinc-air batteries have a higher energy capacity than that of Lithium batteries, the high rate of power drains from the current radio transceivers limits the direct use of zinc-air battery for BSN applications. To simplify sensor deployment, most WSN platforms have integrated batteries. As discussed in earlier chapters of this book, the development of new power scavenging techniques coupled with ultra-low power BSN designs could provide significant improvements in BSN design in future years.

A.3 Conclusions

In this chapter, we have outlined the common WSN and wearable BSN development platforms that have emerged in recent years. Although many of the WSN hardware platforms can be adapted for BSN applications, due to the specific requirements and constraints imposed by BSNs, a number of dedicated platforms have been proposed for BSN research and development. As the BSN platform technologies mature, recent research has been increasingly focused on the development of optimised BSN platforms for both wearable and implantable applications. Many of them are entering into routine clinical use, thus realising the original goal of the BSN in supporting pervasive health monitoring, early detection and personalised treatment of diseases.

References

1. (2004). *WsLAN summary: WP1000 – state of the art*. Available: http://www.sintef.no/units/informatics/projects/wslan/WP1000-summary.pdf
2. M. Weiser, “Hot topics: ubiquitous computing,” *Computer,* vol. 26, pp. 71–72, Oct 1993 1993.
3. J. W. Ng, B. P. Lo, O. Wells, M. Sloman, N. Peters, A. Darzi, C. Toumazou, and G.-Z. Yang, “Ubiquitous monitoring environment for wearable and implantable sensors (UbiMon),” in *International Conference on Ubiquitous Computing (Ubicomp)*, 2004.
4. B. Lo, S. Thiemjarus, R. King, and G.-Z. Yang, “Body sensor network-a wireless sensor platform for pervasive healthcare monitoring,” in *The 3rd International Conference on Pervasive Computing*, 2005, pp. 77–80.
5. V. Shnayder, B.-r. Chen, K. Lorincz, T. R. F. Jones, and M. Welsh, “Sensor networks for medical care,” in *Conference On Embedded Networked Sensor Systems: Proceedings of the 3 rd international conference on Embedded networked sensor systems*, 2005, pp. 314–314.
6. B. Kuris and T. Dishongh. (2006, SHIMMER - Sensing Health with Intelligence, Modularity, Mobility, and Experimental Reusability Available: http://www.snm.ethz.ch/snmwiki/pub/uploads/Projects/SHIMMER_HWGuide_REV1P3.pdf
7. V. Shnayder, M. Hempstead, B. Chen, G. W. Allen, and M. Welsh, “Simulating the power consumption of large-scale sensor network applications,” in *the Second ACM Conference on Embedded Networked Sensor Systems*, Baltimore, MD, 2004, pp. 188–200.
8. M. J. Ammer, M. Sheets, T. Karalar, M. Kuulusa, and J. Rabaey, “A low-energy chip-set for wireless intercom,” in *the Annual ACM IEEE Design Automation Conference*, Anaheim, CA, 2003, pp. 916–919.

9. (2012). *Nine-Axis (Gyro + Accelerometer + Compass) MEMS MotionTracking™ Devices*. Available: http://www.invensense.com/mems/gyro/nineaxis.html
10. K. Mechitov, W. Kim, G. Agha, and T. Nagayama, "High-frequency distributed sensing for structure monitoring," in *Proc. First Intl. Workshop on Networked Sensing Systems (INSS 04)*, 2004.
11. J. Jeong, S. Kim, and A. Broad, "Network reprogramming," *University of California at Berkeley, Berkeley, CA, USA,* 2003.
12. G. Werner-Allen, K. Lorincz, M. Ruiz, O. Marcillo, J. Johnson, J. Lees, and M. Welsh, "Deploying a wireless sensor network on an active volcano," *Internet Computing, IEEE,* vol. 10, pp. 18–25, 2006.
13. J. Beutel, "Fast-prototyping Using the BTnode Platform," in *Design, Automation and Test in Europe, 2006. DATE '06. Proceedings*, 2006, pp. 1–6.
14. U. Maurer, A. Rowe, A. Smailagic, and D. P. Siewiorek, "eWatch: a wearable sensor and notification platform," in *Wearable and Implantable Body Sensor Networks, 2006. BSN 2006. International Workshop on*, 2006, pp. 4 pp.-145.
15. M. Laibowitz, J. Gips, R. AyIward, A. Pentland, and J. A. Paradiso, "A sensor network for social dynamics," in *Information Processing in Sensor Networks, 2006. IPSN 2006. The Fifth International Conference on*, 2006, pp. 483–491.
16. A. Elsaify, P. Padhy, K. Martinez, and G. Zou, "GWMAC- A TDMA Based MAC Protocol for a Glacial Sensor Network," presented at the 4th ACM PE-WASUN 2007, 2007.
17. D. Lymberopoulos, N. B. Priyantha, and F. Zhao, "mPlatform: a reconfigurable architecture and efficient data sharing mechanism for modular sensor nodes," in *Information Processing in Sensor Networks, 2007. IPSN 2007. 6th International Symposium on*, 2007, pp. 128–137.
18. R. B. Smith, "SPOTWorld and the Sun SPOT," in *Information Processing in Sensor Networks, 2007. IPSN 2007. 6th International Symposium on*, 2007, pp. 565–566.
19. J. Andersen, B. Lo, and Y. Guang-Zhong, "Experimental platform for usability testing of secure medical sensor network protocols," in *Medical Devices and Biosensors, 2008. ISSS-MDBS 2008. 5th International Summer School and Symposium on*, 2008, pp. 179–182.
20. A. T. Barth, M. A. Hanson, H. C. Powell, and J. Lach, "TEMPO 3.1: A Body Area Sensor Network Platform for Continuous Movement Assessment," in *Wearable and Implantable Body Sensor Networks, 2009. BSN 2009. Sixth International Workshop on*, 2009, pp. 71–76.

Appendix B: BSN Software and Development Tools

Joshua Ellul, Benny Lo, and Guang-Zhong Yang

B.1 Introduction

Practical applications of Body Sensor Networks (BSN) give rise to programming challenges that are not present in traditional computing systems. The typical memory resources available on sensor nodes are extremely limited, typically equipped with tens of kilobytes of program space and 10 KB of volatile memory. Therefore, algorithm developers must ensure that memory is not used imprudently. More challengingly, typical applications impose prolonged lifetimes, whilst sensor nodes are deployed with extremely limited power resources. Therefore, software developers must ensure that algorithms are highly efficient and that the processor and any peripheral hardware are put into sleep modes as much as possible. Sensor nodes are required to communicate with other nodes in order to relay sensed information or to send updates or configuration messages into the network. Therefore, nodes usually cannot sleep indefinitely and are required to follow wakeup schedules so that they can communicate in predetermined communication windows. Further programming complexity is increased due to internal clock drift which sensor nodes are prone to, and therefore clocks must be corrected to accommodate for such drift typically done using time synchronisation techniques. The complexity does not end there; the wireless medium over which sensor nodes communicate only allows one node to communicate at a time (within the same transmission range). Therefore, Medium Access Control (MAC) protocols must be implemented to avoid wireless collisions. Also, sensor networks aim to maximise lifetime by optimising the route taken from sensor nodes to base stations to consume the least amount of energy. A plethora of MAC and routing protocols

J. Ellul • B. Lo • G.-Z. Yang, PhD (✉)
The Hamlyn Centre, Imperial College London, London SW7 2BY, UK
e-mail: g.z.yang@imperial.ac.uk

G.-Z. Yang (ed.), *Body Sensor Networks*, DOI 10.1007/978-1-4471-6374-9,

have been proposed that focus on different aspects including scalability, dynamicity, responsiveness and density amongst other attributes.

The challenges that face sensor node application programming also stem from the low level embedded programming expertise required. Wireless sensor nodes typically comprise of a microcontroller, a wireless transceiver, a number of sensors and other peripherals such as additional flash storage. Drivers must be implemented for each hardware peripheral which typically communicate over SPI or I^2C using a strict communication protocol often involving low level registers. Such hardware devices often utilise General Purpose Input/Output (GPIO) pins for status updates, which are in turn wired up to microcontroller interrupts. Moreover, substantial internal microcontroller peripherals are wired to interrupts. Thus, knowledge and programming experience of interrupt based systems is required to program such low level embedded systems.

Sensor node applications are predominantly developed in C or flavours of C such as nesC [1]. Therefore, development challenges also include those faced by traditional low-level systems developed in C, including allocation and deallocation of memory by the programmer, as well as no type safety mechanisms. The most popular sensor network operating system, TinyOS, allows developers to code applications in nesC which exposes an event based programming paradigm. The abstraction layers provided are very low, and, as noted by [2], "*it is often difficult to implement even simple programs*."

As described above, the development learning curve for sensor networks is steep. Higher-level languages providing higher abstractions can be used to lower the learning curve substantially. Higher-level abstractions can help by hiding the lower-level embedded systems and sensor networks specific requirements. Nodes can be put to sleep automatically by underlying drivers; default routing and MAC protocols can be used and swapped without the higher level developer having to change any code in relation to this; drivers for different hardware peripherals can be used and then exposed to the higher level language by abstracting the communication protocols, interrupts and register access. However, in using a higher-level language such as Java, the developer is also relieved of memory management since this will be taken care of by a garbage collector. Implicit type safety also ensures that the programmer will not incorrectly cast types. More so, the majority of the available workforce is already familiar with high level languages, such as Java. Therefore, the gradient of the learning curve can be drastically decreased be removing the requirement to learn a new language. It has been shown that higher level languages such as Java provide a more efficient development and maintenance environment compared to lower level languages such as C [3]. Therefore, by using a higher-level language, the development and maintenance costs of BSNs can be reduced, and thus increase the successful adoption of such technology.

B.2 BSN Requirements and Issues

BSNs consist of a number of sensor nodes that can sense the environment, process the sensed data and transmit (and receive) the processed data for an extended period of time. The requirements and challenges inherent in a BSN can directly influence

programmability. Therefore, the main requirements and issues of BSNs will be described here with a focus on how they affect ease of programming.

Body sensor nodes are most often equipped with a limited energy source and are usually expected to operate for an extended period of time. A primary requirement of BSNs lies in the fact that they are deployed in environments monitored in an unobtrusive manner. This relies on miniaturisation of nodes and therefore battery size, which in turn requires an energy efficient system. In order to meet the expected lifetime, and given the limited battery source, the system must be as energy-efficient as possible. This involves ensuring that the wireless transceiver, sensors and any other devices are only used and turned on when they are required. This becomes more complex when considering other aspects including MAC protocols, routing protocols and time synchronisation amongst other factors. Programmers are commonly required to explicitly turn different hardware components on and off, and even in to different sleep modes. Such fine-grained control of hardware is often intimidating to programmers since they are not familiar with such low level fine-grained control. Computational efficiency on the other hand is often considered to be of minor importance and therefore cheaper, more energy-efficient (and slower) processors than those used in larger platforms can be used. However, computational efficiency is often discarded without considering the impact it may have on quality of service and energy expenditure. Computational inefficiency can impact energy expenditure both directly and indirectly [4]. Therefore, programming environments should be both as energy and computationally efficient as possible without requiring extensive effort from the system developer.

Body sensor nodes communicate with each other over the same physical, wireless medium. Therefore, protocols and overall system implementation must be able to scale with the network size and limited bandwidth. Developers typically have to cater for such low-level intrinsic properties when really, they should concentrate their effort on application requirements. In addition to the work involved in deploying a BSN, one must also keep in mind that like other computing platforms, sensor nodes may be required to be updated from time to time due to various reasons including bug fixes, new application requirements and even complete retasking of a network. Therefore, software reconfiguration and reprogrammability is essential, however this is often left up to the developer to implement.

The underlying theme from the above is that the low level internals of BSNs is more often than not left up to application developers to implement, when really they should be focusing on the application specific requirements. It was for this reason that a new operating system was designed to facilitate ease of programming for BSNs.

B.3 Operating Systems for BSNs

The first task required to implement a BSN application is to choose an operating system. The choice of operating systems will also greatly affect many aspects of the development life cycle. Table B.1 provides a development-centric comparison of

Table B.1 Operating systems for BSNs

Operating system	Programming paradigm	Pros	Cons
BSNOS	Java Single-threaded	Easy to use High-level programming language Integrated development environment	Not suitable for complex networking protocol research or implementation Currently only supports the BSN platform
TinyOS	nesC Event driven, TOS thread support can be added	Supports many hardware platforms Extensive code available	Requires learning a new programming language Even simple applications can be hard to implement Does not support the BSN platform out-of-the-box
Contiki	C Event driven and protothread support	Supports many hardware platforms Extensive code available	Requires C programming experience Does not support the BSN platform out-of-the-box
Mantis OS	C Multi-threaded	Supports several platforms	Requires C Programming experience Does not support the BSN platform out-of-the-box
LiteOS	LiteC++ Events and multi-threaded	Supports several platforms Easy to use	Does not support the BSN platform out-of-the-box

popular operating systems typically used traditional for wireless sensor networks and a new operating system specifically proposed for ease of development of BSNs.

The programming language and Application Programming Interface (API) are prime factors that determine ease of developing applications. TinyOS [5] is a popular operating system for wireless sensor networks. Applications are written in nesC [1], an event-driven, component-based language. Although, the event-based paradigm does couple well with microcontroller based interrupts, event handling is non-trivial for those beginning programming. Also, the requirement to learn a new programming language for only BSN development is unfavourable. Contiki [6] and MANTIS OS [7] require applications to be programmed in C. Although C is the primary language of choice for embedded systems developers, it does not provide high level abstractions that would be sufficient for novice programmers. LiteOS [8] provides a UNIX-like operating system abstractions and requires applications to be programmed in LiteC++. The operating systems provide extensive support for different platforms, networking protocols and other low-level features which are useful to those conducting research in the low-level wireless sensor networks field, however the abstraction provided is still

too low for non-electronics and computer scientists. BSNOS [9] is an operating system targeted specifically to facilitate ease of programming for BSNs and is further described in the next section.

B.4 BSNOS – An Operating System for BSN

Operating systems developed for wireless sensor networks and bare metal coding have been popular choices for developing body sensor node applications. Expert knowledge of C and embedded programming has thus been required to create BSN applications, which even highly experienced programmers still find as a challenging task. Application deployment, and therefore data collection is usually therefore delayed due to few such expert programmers being available.

BSNs face different challenges and are prone to different requirements than that of traditional wireless sensor networks. WSN environments are typically much larger than that of the body. Also, the human body is geometrically the same to other body deployments (unlike environments where wireless sensor networks are used). Thus, BSNs require networks of smaller sizes with known locations for node placement. A BSN does not tend to change in size. A nodes placement on a body is unlikely to move, although the body is. Body sensor nodes do not require long-range communication and more often than not can communicate directly with all other nodes. Therefore, body sensor networks do not require the complex MAC and routing protocols, which are a primary requirement for most wireless sensor networks.

A smaller sized network implies that BSNs are relieved from scalability issues usually prone to larger networks. However, smaller sized networks also means that nodes cannot rely on neighbouring nodes for redundancy. Therefore, BSN applications are required to be more robust and resilient to noise and errors. More so, long sleep periods are a luxury that typical body sensor networks do not have since if an event is missed it could be vital. Short sleep times in turn affect the lifetime of the sensor node, which is already a challenge due to constrained battery sizes. That said, most body sensor network deployments do not require extended lifetimes since they are usually deployed temporarily to analyse or detect specific conditions or applications.

BSNOS [9] is an operating system specifically targeted to the challenges and requirements of BSNs with focus on ease of development for early stage programmers as well as scientists from other fields of research. Most wireless sensor networks programming tools require C programming as a means of development. BSNOS on the other hand, exposes a Java programming environment to lower the learning curve required to develop BSN applications. To facilitate an efficient execution platform, a run-time compiler similar to [10] is used in BSNOS. Figure B.1 presents an overview of the BSNOS main components and development tools.

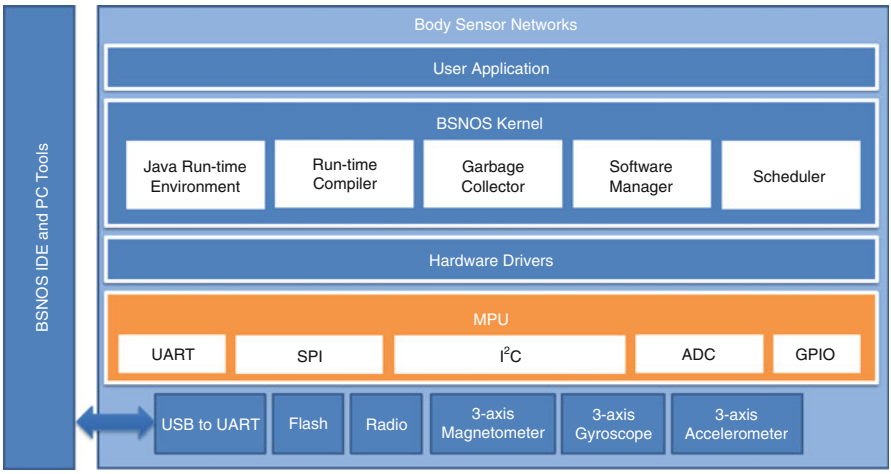


Fig. B.1 An overview of BSNOS and its main components

User applications sit at the top of the software stack. The applications are loaded on top of the BSNOS kernel, which is the main operating system component. The BSNOS kernel is comprised of five main subcomponents. At the core of the kernel lies the Scheduler. The Scheduler is responsible for scheduling execution of timers, threads and events. Body sensor networks may require updates and bug fixes occasionally. The Software Manager meets the requirement of software updates by exposing routines to dynamically install and remove classes and functions.

The run-time environment exposes a subset of Java functionality including object and array manipulation amongst other features. In order to provide an efficient Java execution platform, a run-time compiler has been implemented. One of Java's major benefits is that it relieves developers from memory management. When memory previously used by an application is no longer used, a check is required to release such memory so that it can be used by other parts of the application. This is the job of the Garbage Collector.

The kernel interacts with hardware drivers which communicate directly with the underlying hardware including the microcontroller (Texas Instruments MSP430F1611) and its associated peripherals, 32 Mb Flash storage (Adesto AT45DB321D), a 2.4 GHz RF transceiver (Texas Instruments CC2420), a 3-Axis digital magnetometer/compass (Honeywell HMC5843), a 3-axis digital gyroscope (InvenSense ITG-3200) and a 3-axis accelerometer (Analog Devices ADXL330).

Since ease of use is one of the main driving factors for BSNOS an application programming interface (API) that provides a very high level abstraction of the underlying hardware is provided. A sample of the API made available is presented in Table B.2.

Table B.2 BSNOS sample API

Module	Function	Description
Accelerometer	void performAccelSample()	Instructs the Accelerometer driver to acquire an accelerometer reading
	short getAccelX()	Returns the X-axis reading for a previously acquired accelerometer reading
Radio	void appendShortToRadio (short s)	Appends a 16-bit value to the radio message buffer
	byte sendRadioMsg(short dest)	Send a radio message comprised of the content in the radio message buffer to the destination address specified in the parameter
LEDs	void toggleLed(byte ledNr)	Toggles the state of the LED indicated by the parameter
Timer	void waitMS(short ms)	Waits for the specified number of milliseconds

```
public static void main() {
    while ( true ) {
        BSN.performAccelSample();
        BSN.appendShortToRadio( (short) BSN.getAccelX() );
        BSN.appendShortToRadio( (short) BSN.getAccelY() );
        BSN.appendShortToRadio( (short) BSN.getAccelZ() );
        BSN.sendRadioMsg( (short) BSN.BROADCAST_ADDR );
        BSN.waitMS( (short) 20 );
    }
}
```

Fig. B.2 A simple sample and send program

The high-level programming interface enables for BSN programs to be written in not more than a few lines of code. A simple sample and send program is demonstrated in Fig. B.2.

To further provide a platform allowing for ease of development of BSN applications an Integrated Development Environment (IDE), BSNOS IDE, is provided along with BSNOS. The IDE is based on the Eclipse platform and bundled as a single executable. The IDE, besides providing Code Completion, Import/Export functionality and all other Eclipse based features, also provides an easy way to create BSN specific applications including Projects Wizards for both Body Sensor Nodes, base stations and PC based software; downloading software to sensor nodes by the touch of a button and classes to easily browse functionality available via BSNOS.

**More details on BSNOS and downloads are available from:* http://www.bsn-web.org/

References

1. Gay D, Levis P, von Behren R, Welsh M, Brewer E, Culler D, editors. The nesC language: A holistic approach to network embedded systems. The ACM SIGPLAN 2003 conference on Programming language design and implementation; 2003.
2. Sugihara R, Gupta RK. Programming models for sensor networks: A survey. ACM Transactions on Sensor Networks. 2008;4(2):8:1–8:29.
3. Butters AM. Total cost of ownership: A comparison of C/C++ and Java. 2007.
4. Ellul J. Run-time compilation techniques for wireless sensor networks: University of Southampton; 2012.
5. Levis P, Madden S, Polastre J, Szewczyk R, Woo A, Gay D, et al., editors. TinyOS: An operating system for sensor networks. Ambient Intelligence; 2004.
6. Dunkels A, Grönvall B, Voigt T. Contiki - a lightweight and flexible operating system for tiny networked sensors. Proceedings of the 29th Annual IEEE International Conference on Local Computer Networks 2004. p. 455–62.
7. Bhatti S, Carlson J, Dai H, Deng J, Rose J, Sheth A, et al. MANTIS OS: an embedded multithreaded operating system for wireless micro sensor platforms. Mobile Networks and Applications. 2005;10(4):563–79.
8. Cao Q, Abdelzaher T, editors. liteOS: a lightweight operating system for C++ software development in sensor networks. Proceedings of the 4th international conference on Embedded networked sensor systems; 2006.
9. Ellul J, Lo B, Yang G-Z. The BSNOS platform: a body sensor networks targeted operating system and toolset. The 5th International Conference on Sensor Technologies and Applications 2011. p. 381–6.
10. Ellul J, Martinez K. Run-time compilation of bytecode in sensor networks. The 4th International Conference on Sensor Technologies and Applications 2010. p. 133–8.

Index

G.-Z. Yang (ed.), *Body Sensor Networks*, DOI 10.1007/978-1-4471-6374-9,