NATO Security through Science Series - A:
Chemistry and Biology

Advances in Sensing with Security Applications

Edited by
Jim Byrnes

Springer

NATO OTAN

This publication is supported by:

The NATO Programme for Security through Science

Advances in Sensing with Security Applications

NATO Security through Science Series

This Series presents the results of scientific meetings supported under the NATO Programme for Security through Science (STS).

Meetings supported by the NATO STS Programme are in security-related priority areas of Defence Against Terrorism or Countering Other Threats to Security. The types of meeting supported are generally "Advanced Study Institutes" and "Advanced Research Workshops". The NATO STS Series collects together the results of these meetings. The meetings are co-organized by scientists from NATO countries and scientists from NATO's "Partner" or "Mediterranean Dialogue" countries. The observations and recommendations made at the meetings, as well as the contents of the volumes in the Series, reflect those of participants and contributors only; they should not necessarily be regarded as reflecting NATO views or policy.

Advanced Study Institutes (ASI) are high-level tutorial courses to convey the latest developments in a subject to an advanced-level audience

Advanced Research Workshops (ARW) are expert meetings where an intense but informal exchange of views at the frontiers of a subject aims at identifying directions for future action

Following a transformation of the programme in 2004 the Series has been re-named and re-organised. Recent volumes on topics not related to security, which result from meetings supported under the programme earlier, may be found in the NATO Science Series.

The Series is published by IOS Press, Amsterdam, and Springer Science and Business Media, Dordrecht, in conjunction with the NATO Public Diplomacy Division.

Sub-Series

A. Chemistry and Biology	Springer Science and Business Media
B. Physics and Biophysics	Springer Science and Business Media
C. Environmental Security	Springer Science and Business Media
D. Information and Communication Security	IOS Press
E. Human and Societal Dynamics	IOS Press

http://www.nato.int/science
http://www.springeronline.nl
http://www.iospress.nl

Advances in Sensing with Security Applications

edited by

Jim Byrnes
Prometheus Inc., Newport, RI, U.S.A.

and

Gerald Ostheimer
Prometheus Inc., Newport, RI, U.S.A.

Published in cooperation with NATO Public Diplomacy Division

A C.I.P. Catalogue record for this book is available from the Library of Congress.

ISBN-10 1-4020-4286-8 (PB)
ISBN-13 978-1-4020-4286-7 (PB)
ISBN-10 1-4020-4284-1 (HB)
ISBN-13 978-1-4020-4284-3 (HB)
ISBN-10 1-4020-4295-7 (e-book)
ISBN-13 978-1-4020-4295-9 (e-book)

Published by Springer,
P.O. Box 17, 3300 AA Dordrecht, The Netherlands.

www.springeronline.com

Printed on acid-free paper

Contents

Preface

The chapters in this volume were presented at the July 2005 NATO Advanced Study Institute on *Advances in Sensing with Security Applications.* The conference was held at the beautiful Il Ciocco resort near Lucca, in the glorious Tuscany region of northern Italy. Once again we gathered at this idyllic spot to explore and extend the reciprocity between mathematics and engineering. The dynamic interaction between world-renowned scientists from the usually disparate communities of pure mathematicians and applied scientists which occurred at our six previous ASI's continued at this meeting.

The fusion of basic ideas in mathematics, biology, and chemistry with ongoing improvements in hardware and computation offers the promise of much more sophisticated and accurate sensing capabilities than currently exist. Coupled with the dramatic rise in the need for surveillance in innumerable aspects of our daily lives, brought about by hostile acts deemed unimaginable only a few short years ago, the time was right for scientists in the diverse areas of sensing and security to join together in a concerted effort to combat the new brands of terrorism. This ASI was one important initial step. To encompass the diverse nature of the subject and the varied backgrounds of the anticipated participants, the ASI was divided into three broadly defined but interrelated areas: the increasing need for fast and accurate sensing, the scientific underpinnings of the ongoing revolution in sensing, and specific sensing algorithms and techniques.

The ASI brought together world leaders from academia, government, and industry, with extensive multidisciplinary backgrounds evidenced by their research and participation in numerous workshops and conferences. This forum provided opportunities for young scientists and engineers to learn more about these problem areas, and the crucial role played by new insights, from recognized experts in this vital and growing area of harnessing mathematics and engineering in the service of a world-wide public security interest.

The talks and the following chapters were designed to address an audience consisting of a broad spectrum of scientists, engineers, and mathematicians involved in these fields. Participants had the opportunity to interact with those individuals who have been on the forefront of the ongoing explosion of work in sensing and security, to learn firsthand the details and subtleties of this exciting area, and to hear these experts discuss in accessible terms their contributions and ideas for future research. This volume offers these insights to those who were unable to attend.

The cooperation of many individuals and organizations was required in order to make the conference the success that it was. First and foremost I wish to thank NATO, and especially Dr. F. Pedrazzini and his most able assistant, Ms. Alison Trapp, for the initial grant and subsequent help. Financial support was also received from the Defense Advanced Research Projects Agency (Drs. Joe Guerci and Ed Baranoski), USAF Rome Laboratories (Drs. Michael Wicks, Braham Himed, and Gerard Genello), US Army SMDC (Drs. Pete Kirkland and Robert McMillan), EOARD (Dr. Paul Losiewicz), Melbourne University, and Prometheus Inc. This additional support is gratefully acknowledged.

I wish to express my sincere appreciation to my assistants Marcia Byrnes and Katya Ostheimer, to my co-organizer, Gerald Ostheimer, and to the co-director, Krzysztof Kulpa, for their invaluable aid. I am also grateful to Gerald in his role as our TEXnician, for his superlative work in preparing this volume. Finally, my heartfelt thanks to the Il Ciocco staff, especially Bruno Giannasi, for offering an ideal setting, not to mention the magnificent meals, that promoted the productive interaction between the participants of the conference. All of the above, the speakers, and the remaining conferees, made it possible for our Advanced Study Institute, and this volume, to fulfill the stated NATO objectives of disseminating advanced knowledge and fostering international scientific contacts.

July 29, 2005 *Jim Byrnes*, Il Ciocco, Italy

Acknowledgments

We wish to thank the following for their contribution to the success of this conference: NATO Scientific & Environmental Affairs Division; DARPA Defense Sciences Office; European Office of Aerospace Research and Development of the U.S.A.F.; Air Force Office of Scientific Research; U.S. Air Force Research Laboratory; U.S. Air Force Rome Laboratories Sensors Directorate; U.S. Army Space and Missile Defense Command; Melbourne University, Australia; and Prometheus Inc., U.S.A.

BISTATIC AND MULTISTATIC RADAR SENSORS FOR HOMELAND SECURITY

C.J. Baker and H.D. Griffiths
University College London
U.K.

Abstract

The separation of transmitter and receiver in bistatic and multistatic radar sensors offers the system designer new and additional degrees of freedom to tailor solutions to specific applications. The receivers may be passive and hence largely immune to jamming. Passive systems that also use 'illuminators of opportunity' do not have to provide a potentially expensive transmitter. Multiple transmitters and or receivers can improve sensitivity, coverage, and importantly improve the opportunity to acquire a line of site to the target (without which detection is impossible). These advantages make this form of radar attractive for a variety of applications, many of which fit well with the needs of homeland security. Equally, however, the additional complexity of having a number of separated transmitters and receivers brings about new challenges that require careful understanding if these forms of sensors are to be routinely adopted for operational use.

In this chapter the role of active and passive techniques as a support to homeland security is explored. The essentials of bistatic and netted radar are introduced which enables the relative strengths and weaknesses of these approaches to be outlined. In this way a foundation is provided against which a variety of potential applications may be explored.

Keywords: radar; bistatic radar; multistatic radar; ambiguity function; parasitic radar; forward scatter; Babinet's principle; passive coherent location; netted radar.

1. Introduction

Radar has long been used in a variety of military and civilian applications and has become an essential component of current defensive systems. The chief reason for this is an ability to survey wide areas rapidly during the day or at night and in all weather conditions. It is the only sensor able to do this. Many countries have a network of civil

Jim Byrnes and Gerald Ostheimer (eds.),
Advances in Sensing with Security Applications, 1–22.

aviation radars that often form a part of a wider air defence capability that is able to detect aircraft out to ranges of hundreds of km. These networks are specifically designed to ensure early warning against potentially hostile threat targets. In a similar manner coastal shorelines are monitored also using a combination of civilian coastguard and military maritime systems.

However, the range and nature of potential threat targets is becoming ever more diverse. For example targets are becoming faster, more agile, stealthier and can occur in many guises. The tragic events of 9/11 are of course but one example of how the meaning of legitimate 'targets' has changed irrevocably. Others might include missile attack, uavs (including micro-lights), small high speed boats (including jetskis). In fact the potential range of threat targets is almost unlimited. This means that the source of attack can emanate in a much wider variety of new and different forms. These are not necessarily well dealt with by current conventional radar systems and alternatives merit evaluation.

Here we concern ourselves with bi and multistatic radar concepts, largely, although not exclusively, as an *appliqué* fit to existing systems. One example might be the augmentation of current air defence systems to ensure that coverage is extended into areas not well catered for by current systems (such as low level flight paths). In this way high value or physically vulnerable assets (such as nuclear power plants) may be afforded improved protection. Bistatic and multistatic radar systems have a number of advantages that make them potentially very well suited to these types of applications. They don't necessarily require expensive transmitters and are relatively immune to physical and electronic attack due to their inherent passivity and their distributed nature. They are also better able to detect stealthy targets which have been designed only to present a small cross-section to monostatic radar. However, this tends to come at the cost of increased system complexity. Nevertheless trade-offs between performance and complexity can show worthwhile benefits.

Here we introduce the principles of bistatic and multistatic radar. Firstly, we examine what is meant by these terms and then go on to develop the fundamental relationships that govern performance in terms of sensitivity, coverage, range resolution, Doppler resolution and target location accuracy. This provides the essential information necessary to understand the advantages and disadvantages of bistatic and multistatic radar operation for candidate applications. More detail can be found in the excellent text of Willis [1].

2. Definitions

A survey of the literature reveals that definitions of bistatic and multistatic radar are quite widely varying with no universal acceptance of single descriptions. The IEEE defines bistatic radar as 'a radar system that uses antennas at different locations for transmission and reception'. The distance of separation between the two is referred to as the 'baseline' range. However, there is no stipulation as to how far apart the two antennas should be. Clearly if they are near co-located, i.e. if the baseline is very small, then the system approximates monostatic radar. Very small baselines are sometimes used in cw systems where there is a need to minimise the likelihood of the transmitted signal being received directly, thus masking the presence of real targets. Suppression of the directly received signal is difficult to achieve via a single antenna typically used in monostatic radar. Here we do not consider such a system to be bistatic. If a further antenna (either transmitting or receiving or both) is added to the bistatic pair then this might be termed multistatic radar. However, other terminology often includes 'netted radar', 'multi site radar' and 'distributed radar'. The distinctions between these are, at best, somewhat blurred. Here we will use the term multistatic to mean any system comprising a bistatic pair augmented by an additional antenna (tranmit or receive). These two labels are not entirely satisfactory as there is no reason why the antennas need be 'static'!!

A particular variant on the bistatic and multistatic themes is Passive Coherent Location (PCL). This is normally taken to mean a system where transmissions are provided by a third party and only the receiver is formerly part of the design. This is sometimes also referred to as a 'hitch hiking' mode of operation. It is also usually referred to as an example of bistatic radar although, as we shall see later, many transmissions can be used from a variety of transmitter sites thus making an example of multistatic radar.

3. Bistatic essentials

In this section we review the fundamental building blocks of bistatic radar emphasising similarities and differences with the more usual monostatic counterpart. Figure 1 shows a typical bistatic geometry with clear separation of the transmitter and receiver.

It has two geometrical characteristics which differentiate it from conventional monostatic systems. These are:

- the transmitter receiver separation and ;
- the transmitter-target-receiver triangulation.

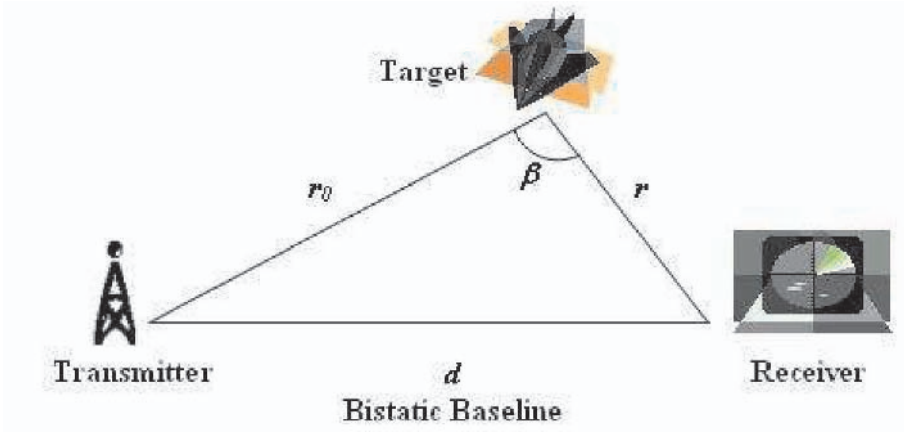


Figure 1. Bistatic geometry

The distance d represents the bistatic baseline. r_O is the transmitter to target separation and r is the target to receiver separation. β is referred to as the bistatic angle.

The baseline distance is usually fixed or slowly varying. The system performance depends to a large extent on the target position relative to the bistatic baseline and hence triangulation. Three target position areas can be described with substantially different characteristics. There is the broadside area with a bistatic angle less that 180° which provides the most common form of bistatic radar. There is also the baseline area with a bistatic angle equal to 180°. This situation corresponds to the 'forward scatter' geometry and has several implications which will be discussed later. The final area is the extended baseline behind both the transmitter and receiver. This obeys quasi-bistatic characteristics.

The bistatic radar range equation is given by

$$\frac{P_r}{P_n} = \frac{P_t G_t G_r \lambda^2 L_{pt} L_{pr} \sigma_b}{(4\pi)^3 r_0^2 r^2 k T_0 B F} \tag{1}$$

where

P_t is the transmitter power in $Watts$
G_t is the gain of the transmitter antenna
G_r is the gain of the receiving antenna
λ is the radar wavelength in $metres$
L_{pt} is the loss from transmitter to target (≤ 1)
L_{pr} is the loss from target to receiver (≤ 1)
r_o is the distance between the transmitter and target in metres

r is the distance between the receiver and target in metres
k is the Boltzmann's constant
T_0 is the noise reference temperature in kelvin
B is the receiver bandwidth in Hz
F is the receiver noise figure
σ_b is the bistatic Radar Cross Section in m^2.

The main difference between bistatic and monostatic radar is the separation of transmitter and receiver ranges. These determine the bistatic geometry. From the bistatic geometry, it can be observed that targets of constant bistatic range are described by ellipses with the transmitter and receiver as the two foci. We note that in monostatic radar these are, of course, circles. In bistatic radar these ellipses are the same as the contours of zero Doppler. The contours of maximum Doppler shift form hyperbolae as they must also cross these ellipses orthogonally. In monostatic radar the orthogonal condition again holds but this time leads to a series of lines emanating radially out from the co-located transmitter receiver pair. In bistatic radar a moving target will not present zero Doppler to two receiving sites simultaneously. This can usefully be exploited in multistatic radar systems.

Contours of constant signal to noise ratio follow the lines of ovals of Cassini as the signal to noise ratio is inversely proportional to the product of the squares of the transmitter to target and target to receiver ranges. In monostatic radar contours of constant signal to noise ratio are circles.

The bistatic radar cross section of the target is not necessarily the same as the monostatic one. For small bistatic angles of less than approximately 5°, the bistatic RCS of a complex target is equal to the monostatic RCS measured on the bisector of the bistatic angle at a frequency lower by a factor of $cos(\beta/2)$. Also, when operating in the broadside area, bistatic radar may be well suited to detecting stealthy targets. This is because a target is very unlikely to present a low bistatic cross section to two receiving sites simultaneously. This potentially makes the detection of stealthy targets easier as the reflections from it in other directions will be detected by bistatic receivers.

When a target crosses the baseline of a bistatic radar the RCS can be greatly enhanced. This is due to the forward scatter phenomenon or "Babinets" principle. Here the RCS of a target at the bistatic baseline is calculated from

$$\sigma_b = \frac{4\pi A^2}{\lambda^2} \tag{2}$$

where

A is the geometric area of the target in m^2
λ is the radar wavelength in metres.

For a sphere of radius a metres, the monostatic RCS is equal to the projected area of a sphere given by πa^2. Considering a sphere with monostatic RCS equal to $0.25m^2$ and at a wavelength equal to $0.1m$, the forward scatter RCS is:

$$\sigma_b = \frac{4\pi(0.25)^2}{(0.1)^2} = 78.5m^2 \quad (3)$$

This corresponds to an enhancement of 25dB. The forward scatter RCS will decrease as the bistatic angle decreases and ultimately reaches the monostatic RCS in the limit where the angle is equal to zero. Nevertheless, significant RCS enhancement is generally achieved at bistatic angles of 165°. An important factor is that the forward scatter RCS does not depend on material composition. As a result, bistatic radars operating in the forward scatter region may be able to detect stealthy targets and will give appreciable forward scatter RCS despite their designed low monostatic RCS. In addition it should be noted that the angular width of scattering is a function of the wavelength and hence favours low frequencies.

Bistatic clutter is a poorly understood branch of radar and few measurements have been undertaken to help develop useful models. This is a subject that requires further and quite urgent research.

There are also some important differences in the technology required to realise bistatic radar. In monostatic radar synchronisation between transmission and reception is done via a stable source, usually a local oscillator. In bistatic radar the separation of transmitter and receiver makes this much more difficult. An equivalent situation has to be achieved and this is done either via synchronised atomic clocks, a signal such as GPS or by reception of a reference signal received directly from the transmitter. The latter technique is typically used in PCL systems and we shall return to this later.

Another important difference between bistatic and monostatic radar is that a directional receive antenna must scan at a non uniform rate to follow the position of the transmitted signal through space, a process known as pulse chasing. This can be very challenging for designs based upon mechanical scanning and hence an alternative is to use one or more electronically agile beams as in phased array radar. Such phased array antennas can be expensive and in some applications will prohibit the use of the bistatic technique.

4. Passive Coherent Location (PCL)

This section looks at the intriging concept of utilising illuminators of opportunity to give a completely passive bistatic radar system. These are signals such as existing radars, communications, navigation and broadcast transmissions which happen to be in existence. Their characteristics are governed by their own missions and hence their waveforms will not necessarily have the ideal parameters for a given bistatic radar system. Nevertheless, the advantage of using such signals is their availability, fixed position and somewhat known characteristics which make building and optimising a passive receiver less complex. PCL systems can use transmissions from multiple nearby sources, hence making them an example of multistatic radar. For now, however, we can assume that they have (for any given single transmitter) the same bistatic geometry introduced in the previous section.

There has recently been an upsurge of interest in PCL radar systems that exploit illuminators of opportunity. An additional attraction is that this can dramatically reduce the costs of the system hardware. The rapid growth in number of RF emissions for TV and radio broadcasts as well as terrestrial and space based communications has resulted in a wide range of signal types available for exploitation by passive radar. Further, many such transmissions are at VHF and UHF frequencies, which allows these parts of the spectrum not normally available for radar use, and at which stealth treatment of targets may be less effective, to be used. However, the location of the transmitter and the form of the transmission to be exploited is no longer under the control of the radar designer. The multiplicity of transmissions from both terrestrial [2] and space based sources [3, 4] provide spatial and frequency diversity and can be exploited to further improve detection performance. Examples of reported operational systems include the Lockheed Martin 'Silent Sentry' system for air and space surveillance [2], the Roke Manor Research CELLDAR system for air target detection [5] and the Manastash Ridge radar for atmospheric and ionospheric studies [6]. Other reported experimental systems include those proposed by Dynetics [7] and UCL [8]–[10]. Applications include air-space surveillance [2, 5], maritime surveillance [10], atmospheric studies [6], ionospheric studies [6], oceanography [11], mapping lightning channels in thunderstorms [12] and monitoring radioactive pollution [13]. There have also been recent reports of algorithm development for interferometry [16], target tracking [17] and target classification [17, 18]. This range and diversity of systems and applications is indicative of the increasing importance of this form of sensor system.

The transmit power P_t is substantial for many passive radar sources, since broadcast and communications receivers often have inefficient antennas and poor noise figures and the transmission paths are often far from line-of-sight; thus the transmit powers have to be significantly higher to overcome the inefficiencies and losses. In the UK, the highest power FM radio transmissions are $250kW$ (ERP) per channel, with many more of lower power [19]. The highest power analogue TV transmissions are $1MW$(ERP) per channel [19]. These are omnidirectional in azimuth, and are sited on tall masts on high locations to give good coverage. The vertical-plane radiation patterns are tailored to avoid wasting too much power above the horizontal.

GSM cellphone transmissions in the UK are in the 900MHz and 1.8GHz bands. The modulation format is such that the downlink and uplink bands are each of 25MHz bandwidth, split into 125 FDMA channels each of 200kHz bandwidth, and a given basestation will only use a small number of these channels. Each channel carries 8 signals via TDMA, using GMSK modulation. Third generation (3G) transmissions are in the 2GHz band, using CDMA modulation over 5MHz bandwidth. The radiation patterns of cellphone basestation antennas are typically arranged in 120° azimuth sectors, and shaped in the vertical plane again to avoid wasting power. The patterns of frequency re-use means that there will be cells using the same frequencies within quite short ranges. Licensed ERPs are typically in the region of 400W, although in many cases the actual transmit powers are lower. The OFCOM sitefinder website [20] gives details of the location and operating parameters of each basestation throughout the UK, and an example of the information provided by this website is shown in Table 1.

In all cases it is necessary to consider the power in the portion of the signal spectrum used for passive radar purposes, which may not be the same as the power of the total signal spectrum. For example the ambiguity properties of the full signal may not be as favourable as those of a portion of the signal. This is the case for an analogue television transmission; the full signal has pronounced ambiguities associated with the 64 μs line repetition rate, but better ambiguity performance may be realised by taking just a portion of the signal spectrum at the expense of reduced signal power.

In PCL systems care must be taken to ensure that the signal received directly from the transmitter does not compete with and swamp that from the target. Typically this will be the case unless measures are taken to suppress the direct signal occurring in the indirect channel. We can formulate a simple expression for the amount of direct signal suppression required by calculating the ratio of the indirect received sig-

nal to the direct signal and requiring this to be at least the same value as that used to compute the maximum detection range. We make the simple assumption that a target can be seen above this level of direct signal breakthrough and hence that it approximates to the highest level of interference that is tolerable for single 'pulse-like' detection. There is, however, no benefit from integration as the direct leakage will also integrate up, and this may lead to a more stringent requirement needing to be set in practice. This places the direct leakage signal at the same level as the noise floor in the receiver and hence it has the attractive feature of proving equivalent performance to 'single-pulse' detection. Thus to achieve adequate suppression and hence maintenance of full system dynamic range the direct signal must be cancelled by an amount given by the magnitude of the ratio of the indirect and directly received signals, e.g.

$$\frac{P_r}{P_d} = \frac{r_b^2 \sigma_b}{4\pi r_1^2 r_2^2} \tag{4}$$

where P_r is the target echo signal, P_d is the direct signal, and r_b is the transmitter-to-receiver range (bistatic baseline). This expression is indicative only and strictly speaking the direct signal should be below that of the noise floor after integration, if integration is employed.

There are several techniques that may be used to suppress this leakage. These include: (i) physical shielding, (ii) Doppler (Fourier) processing, (iii) high gain antennas, (iv) sidelobe cancellation, (v) adaptive beamforming and (vi) adaptive filtering. Each one of these techniques will provide different suppression characteristics over the (θ, f) plane—thus physical shielding or beamforming techniques will provide suppression as a function of θ; Doppler processing or adaptive filtering will provide suppression as a function of f. The combination of high gain antennas and adaptive beam-forming also enables multiple simultaneous transmissions to be exploited. The Manastash Ridge radar [6] provides an example of the use of physical shielding; in this case suppression is achieved by siting the receiver on the other side of a large mountain which provides the screening. In other cases some simpler more localised methods may be used such as appropriate deployment of absorbing material (RAM). For the detection of moving targets Doppler or Fourier processing will automatically improve dynamic range, as the direct signal leakage will only occur at DC (with some spill over). However it should be noted that significant sidelobe leakage due to inadequate suppression of very strong directly received signals will reduce the gain from Fourier processing and hence impair dynamic range.

For the example considered here, the transmitter at Wrotham in the south-east of England is taken together with a receiver sited at the Engineering building of UCL. The transmitted power is 250 kW and broadcasts are made in the frequency range 89.1–93.5MHz. Figure 2 shows a plot of the detection range; the contour represents a signal-to-noise ratio of 15dB (and this value is used for all subsequent figures of this type). The modulation bandwidth is taken as 55kHz. A signal-to-noise ratio of 15dB or greater is maintained out to a range of nearly 30km. This performance is constrained by the effective noise figure of the receiver, and better performance would be obtained with better suppression of direct signal and noise. It should be noted that the power emitted by transmitters across the UK varies from as little as 4W to a maximum of 250kW and of course this variation has to be carefully factored in to performance predictions.

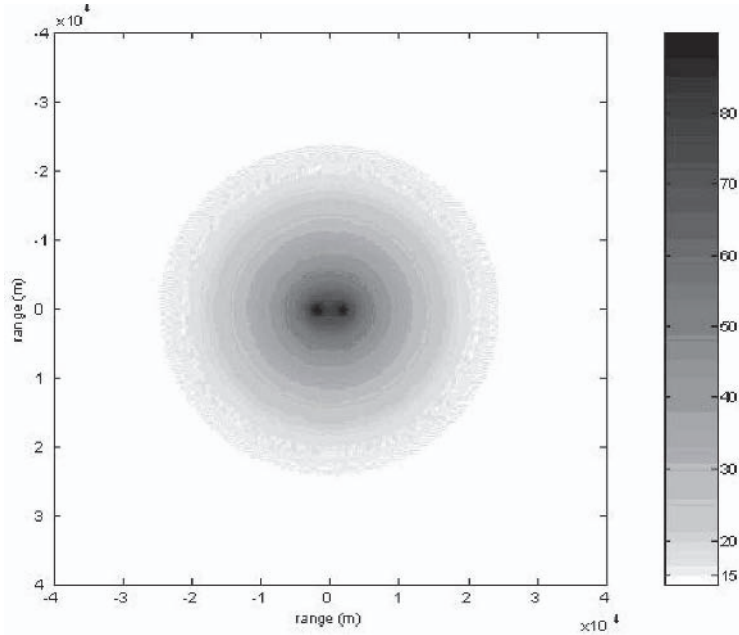


Figure 2. Detection range for an FM radio transmitter at Wrotham in south-east England and a receiver at UCL. The solid contour represents a signal-to-noise ratio of 15dB.

Range resolution is the ability of a radar system to distinguish two closely spaced targets at differing ranges. In monostatic radar this is primarily a function of pulse length or modulation bandwidth. Azimuth

and elevation resolutions distinguish between two targets in angle space and are determined by antenna dimensions and operating wavelength. Together these specify the three-dimensional spatial resolution of any radar system. Doppler resolution is the ability to distinguish two targets by virtue of their differing velocities. In a monostatic radar system this is primarily a function of wavelength, waveform and target illumination time. It is important to understand the range and Doppler resolution of any radar system in order that overall performance may be reliably estimated. Examples might include ensuring that tracking performance is adequate for a particular application or that image quality metrics result in two-dimensional target signatures suitable for further processing tasks such as classification. In monostatic radar systems these parameters are routinely established as part of the design process and can be easily pre-determined by the radar designer to be tailored to the chosen application. However, in bistatic radar systems and more particularly in Passive Coherent Location (PCL) systems they cease to be of a routine nature and considerable care needs to be exercised in these aspects of design. Indeed in the case of PCL the lack of design control over the form, nature and origin of the transmitted waveform seems to imply severe restrictions. In practice, however, there are more freedoms than might be apparent at first sight as usually more than one transmitter may be used at any given instant of time. This is an aspect not exploited in monostatic radar systems. Nevertheless, separation of the transmitter and receiver and the time varying properties of qualifying illuminations of opportunity do result in important differences that must be thoroughly understood if PCL system design methods are to evolve to similar levels of maturity to that of monostatic radar. In this chapter we analyse these fundamental aspects of PCL radar design that determine subsequent performance. Practical measurements of transmissions of opportunity show detailed, time-dependent, behaviours that require careful consideration when developing an overall system design.

Specifically the ambiguity function has long been used to evaluate range and Doppler resolution as well as range and Doppler ambiguity. However, it was developed to capture these aspects of performance for monostatic radar systems only. Here we review the ambiguity function and in particular its bistatic formulation. This highlights the importance of system geometry with respect to target position. Results of the bistatic 'self-ambiguity' of 'on-air' signals are used to demonstrate waveform variability and its effect on range and Doppler resolution as well as detection performance. These give the best possible range and Doppler resolutions and are geometry independent. The self-ambiguity function is equivalent to the transmitter and receiver being co-located,

i.e., the monostatic case. The more general description of ambiguity in a bistatic system introduces a geometrical dependence between the transmitter, receiver and target that determines the range and Doppler resolutions.

The ambiguity function represents the output of a matched filter and may be written as

$$|\psi(R_R, f_d)^2| = \left| \int_{-\infty}^{+\infty} s_t(t) \cdot s_t^*(t + R_R) \cdot exp[j2\pi f_d t] dt \right|^2 \qquad (5)$$

where $\psi(R_R, f_d)$ is the ambiguity response at delay range and Doppler and $s_t(t)$ is the directly received transmitted signal

Computation of this function results in a three-dimensional plot for which one axis is time delay (or range), the second is Doppler frequency or radial velocity and the third is the output power of the matched filter (usually normalised to unity). The extent of the ambiguity function peak in the T_R and the f_d dimensions determines the range and Doppler resolutions respectively. As we are using the directly received signal only we term this "self-ambiguity" as there is no inclusion of any system geometry dependence on the transmitter and receiver locations.

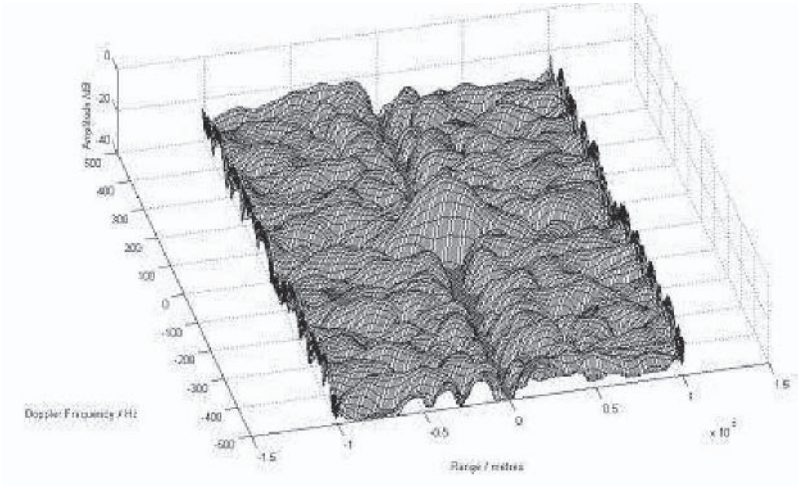


Figure 3. The ambiguity function for a BBC radio 4 transmission at 93.5MHz.

An example self ambiguity function is shown in Figure 3. The sample length taken was 80ms of stereo signal which was sampled after applying a filter of bandwidth 300kHz. Figure 3 shows an un-weighted self-

ambiguity function for a BBC Radio 4 transmission for which the signal comprises speech (in this instance an announcer reading the news). The peak of the ambiguity function is reasonably well defined but a lot of fine, semi-random structure can be seen in the regions away from the main peak which are a function of the detailed modulation present in this component of the waveform. This does not show pure noise-like behaviour but is consistent with the correlation that might be expected in a speech type signal. Cuts taken at zero range and Doppler are also shown in Figure 4 to demonstrate the range and Doppler resolutions more clearly. Side-lobe levels are very good with nearly -50dB in the frequency domain and around -25dB for range.

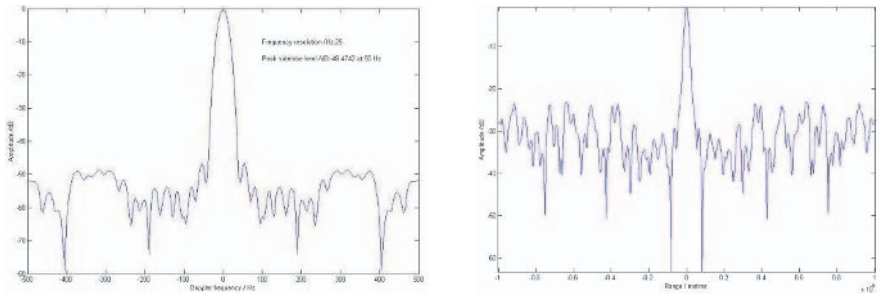


Figure 4. Range (left) and Doppler resolution (right) for the BBC radio 4 transmissions.

If a number of speech collections are analysed then we see that performance is not consistent. This is illustrated in Table 1 which shows the bandwidth in kHz for ten waveform samples.

The bandwidth is seen to vary from 500Hz to 22.2kHz. This is important in two respects. The first is that by no means all of the 150kHz modulation bandwidth is being used (in this case it is only 15% of the available bandwidth). Secondly, as the bandwidth is a function of time the performance of the radar system will also be a function of time.

So far only 'self-ambiguity' has been considered. This has been defined as the output of the matched filter response of the direct signal. Hence it may be thought of as a condition that mimics the performance of a monostatic radar system with the same waveform. In effect this yields the best possible range and Doppler resolutions with any given waveform. However, in PCL and more generally in bistatic radar the relative positions of target, transmitter and receiver govern the actual resolutions that can be achieved. Here we use the formulation presented in [21] to compute the bistatic ambiguity function for the example presented in part 1 of this chapter that uses the transmitter at Alexander

Table 1. Bandwidth variation of speech waveforms.

Collection Number	Bandwidth /kHz
1	22.2
2	0.5
3	14.8
4	9.1
5	10.1
6	0.8
7	2.2
8	4.2
9	1.0
10	5.3
Average Bandwidth /kIIz	8.0

Palace with the receiver located at University College London. From the bistatic form of the ambiguity function this can be written as:

$$|\psi(R_{RH}, R_{Ra}, V_H, V_s, \theta_R, L)| = \\ = \left| \int_{-\infty}^{+\infty} s_t(t - \tau_a(R_{Ra}, \theta_R, L)) \cdot s_t^*(t - \tau_H(R_{RH}, \theta_R, L)) \right. \\ \left. \cdot\ e^{j(2\pi f_{DH}(R_{RH}, V_H, \theta_R, L) - 2\pi f_{Da}(R_{Ra}, V_s, \theta_R, L))t} dt \right|^2 \qquad (6)$$

where R_{RH} and R_{Ra} are the hypothesised and actual ranges (delays) from the receiver to the target; V_H and V_a are the hypothesised and actual radial velocities of the target with respect to the receiver; f_{DH} and f_{Da} are the hypothesised and actual Doppler frequencies; θ_R is the angle from the receiver to the target with respect to 'North'; L is the length of the baseline formed by the transmitter and receiver.

The expression assumes the reference point of the PCL geometry to be the receiver and is essentially a straight change of variables from the equation for the self ambiguity function. The important difference is that the geometrical layout of the transmitter, receiver and target are now taken into account. This can have a significant effect on the form of the ambiguity function and the resulting range and Doppler resolutions. Indeed if a target crosses the bistatic baseline all resolution in range and Doppler is lost.

Variations in range resolution can be distilled from the ambiguity 'traffic light' plot of Figure 5. Here the range resolution only is being plotted as a function of target position. The green colour span corresponds approximately to a range resolution of up to one and a half that of the self-ambiguity case. Amber corresponds to approximately a resolution degradation of 150% of that of the self-ambiguity case and red

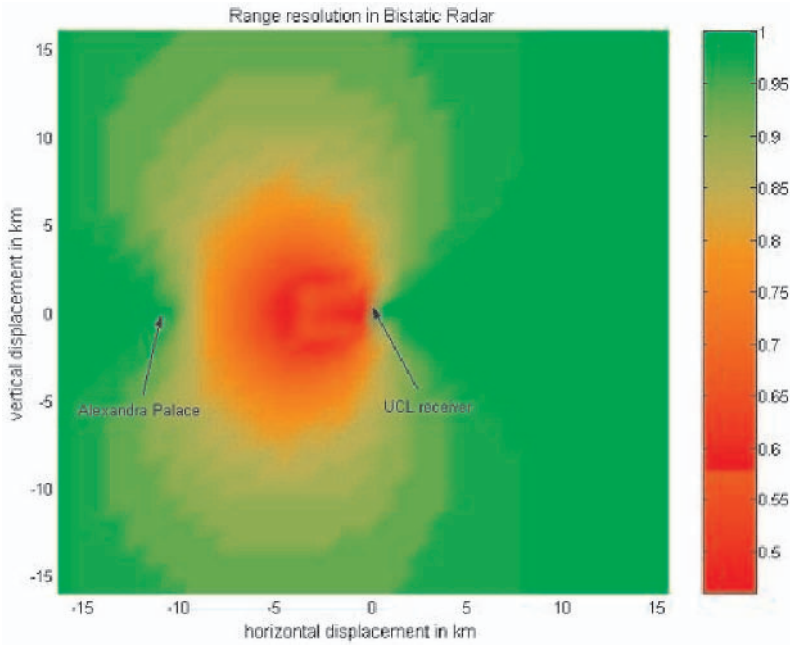


Figure 5. Traffic light plot of normalised range resolution variation for a bistatic PCL system.

approximately greater than 200%. The scale on the right hand side of the plot indicates this on a continuous colour change basis. This shows that there are significant areas where the resolution is severely degraded. There are several strategies for taking this into account in the design of the system. Two possibilities might be to either adjust the signal processing depending on the application or to simply declare a 'no-go' region where radar operation is not attempted. As the no-go area coincides with the direction of the directly received transmissions then this may be the preferred option as it aids the problem of suppression of the directly received signal.

Thus we have seen that PCL is a particular form of bistatic radar that has a number of attractive characteristics. It should not be thought of as a mature technology but more as an emerging one and extant systems are evidence of this.

5. Multistatic radar

Most current radar systems are monostatic i.e. the transmitter and receiver are co-located. The performance of these forms of radar has been greatly enhanced by the advent of high resolution imaging, low sidelobe antennas, high speed digital signal processing and other technology improvements. However, it is well known that when a target is illuminated by electromagnetic radiation scattering occurs in all directions. A single receiver remotely located will only intercept a very small portion of this energy and much of the signal and its information is lost. Netted (or multistatic) topologies can overcome this limitation and offer the potential to extend the capabilities and performance of current radar systems.

Multisatic radar has some inherent advantages. For example spatial distribution of the nodes of the network enables the area to be tailored according to the specific application of interest. Additionally, it is possible to increase sensitivity, as more of the scattered energy (in the different directions) can be collected and hence detection performance improved. Target classification and recognition can also be enhanced, as the target is observed from different perspectives. Moreover, increased survivability and reliability is achieved because of the option of having 'silent' or passive operation of the receivers. These receivers can improve the location accuracy of possible jammers by fusing the information from the network nodes. Finally, if a single node of the network is lost it can still provide a level of (reduced) performance and the network is said to exhibit graceful degradation.

Here we extend the bistatic case to netted scenarios. The topology selected is the simple case of N transmitters and one common receiver, i.e. we have in effect a series of multiple bistatic geometries with varying baselines. The reason for this choice lies in the fact that it is convenient to reference all calculations to the single receiver, thus obtaining one unified form for the ambiguity function.

The analysis is based on the matched filtering performed at the receiver. Before proceeding to the mathematical background, it is necessary to state the assumptions made when modeling the system. These are:

i. The target is considered to be a slowly fluctuating scatterer.

ii. The transmitted signal after reflection by the target is multiplied by the factor b which corresponds to the scattering characteristics of the target in the direction of the receiver. b is assumed to be a Gaussian random variable when the number of scatterers is large and none is dominant.

iii. The target's scattering properties do not change with the angle of view.

The transmitted modulation is common and has the following form:

$$s(t) = Re[u(t) \cdot exp(j\omega_c t)] \tag{7}$$

where $u(t)$ is the complex envelope of the signal and ω_c is the carrier frequency. The return signal will then be an addition of the N scattered signals from the target. An important assumption made at this point is that coherent processing of the raw data is feasible. This implies that the N echoes, which arrive at different time instances as the transmitter-target-receiver paths are different, can be processed jointly. This will involve storing a number of returns, aligning them and feeding them to the matched filter. The received signal is:

$$r(t) = \Sigma_{i=1}^{N} Re\{b_i \cdot u(t - \tau_i) \cdot e^{j\omega_c(t-\tau_i)}\} \tag{8}$$

where τ_i is the time delay and b_i is the multiplication factor. Taking into account assumption (iii), $b_i = b = constant$.

It must be noted that the echoes that arrive at the receiver do not have the same intensity, as the propagation lengths of the waves are different. Thus, a weighting must be applied, according to the signal power. The weighting factors are calculated by the following set of equations:

$$w_i = \frac{P_{Ri}}{max(P_{Ri})}, \quad i = 1, 2, ...N \tag{9}$$

and

$$P_{Ri} = \frac{P_{Ti} G_{Ti} G_r \lambda^2 \sigma_B}{(4\pi)^3 (R_R R_{Ti})^2}, \quad i = 1..., N \tag{10}$$

where the bistatic radar cross-section σ_B is considered to be a constant. In the receiver, the weighted echoes are passed through a filter matched to the original transmitted signal. Following similar analysis to [21], and excluding b, the specific instance of the ambiguity function is given by:

$$Xnetted = \left|\Sigma_{i=1}^{N} w_i X_i\right|^2 \tag{11}$$

where X_i are the bistatic ambiguity functions for the different bistatic pairs that are formed.

Examining the scenario with M receivers and one common emitter there is a significant practical limitation when attempting to implement the previous methodology. It was assumed in that analysis that the scattered signals must be processed jointly. This is easier to achieve

when there is a common receiver, whereas in this case all the receivers of the network must send their inputs to a central station for coherent processing and alignment of the signals. The real time requirement leads us to use an alternative approach. It will be assumed that each echo is processed in the corresponding receiver before being transmitted to a central unit. Thus, the processing has a distributed nature.

The outcome of this procedure will be the same as in the previous case in terms of the system's parameters (resolution and ambiguity), but not exploited via centralized processing as the ambiguity function seen by each receiver will be individually different.

The mathematical representation of the ambiguity function for this second method is now different. The matched filtering is performed for each echo and the final output will be a summation of the bistatic ambiguity functions of the various bistatic pairs. The following equation outlines this:

$$Xnetted = \Sigma_{i=1}^{N} W_i \left| X_i \right|^2 . \tag{12}$$

Examining the general case where a number of nodes are spatially distributed, a combination of the two previous methods can be used. That is each of the M receivers in the network will accept all the scattered signals, originating from the N transmitting stations, creating a series of multistatic ambiguity functions. These will be then used as inputs in the last equation to construct the radar equation type ambiguity diagram. The outcome is:

$$Xnetted = \Sigma_{j=1}^{M} \left(W_j \Sigma_{i=1}^{N} w_i X_i \right) . \tag{13}$$

Assume the transmitted signal is a coherent train comprised of three rectangular pulses. The first example refers to the case where the target is close to the baseline ($\theta_R = -80^\circ$) of one of the two bistatic pairs. For a simple bistatic radar it is well known that this scenario is detrimental to its resolution performance. The two baselines are set to $100km$ and the target is stationary with $R_R = 50km$.

The left hand plot of Figure 6 represents the contour plot of the ambiguity function for the specific scenario. The horizontal axis is the range to the target and the vertical is the velocity. The right plot shows the cuts along these axes, in a position which corresponds to the actual values of the range and the velocity of the target. The width of the main peak corresponds to the resolution of the system and any additional peaks correspond to potential ambiguities.

The simulation result outlines that there is no significant improvement in the performance of the radar network, as compared to the bistatic

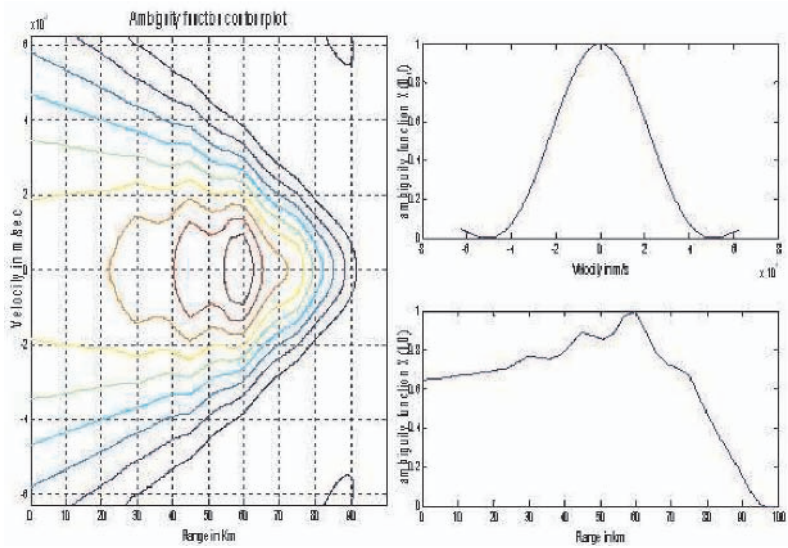


Figure 6. Contour plot and cuts along the velocity and range axis of the ambiguity function/bistatic pair dominance case.

radar. This is to be expected though, as for the specific position of the target one of the two bistatic pairs will be dominant. The echo which originates from the left transmitter will be much stronger than the one originating from the right one. Thus, the ambiguity function of the radar network will be dictated by the bistatic ambiguity function with the highest weighting factor.

There are scenarios though where the target is in such a position where none of the bistatic pairs is dominant. This is examined in the next example in a topology where the second transmitter is $100km$ above the receiver. The angle θ_R of the target is $-60°$.

The left plot represents the cuts of the bistatic ambiguity function of the first bistatic pair, in the absence of the second transmitter. Adding the emitter, and for this topology where the two weighting factors are comparable, the resolution in range and velocity has improved significantly. Moreover, the ambiguity peaks are suppressed.

Other degrees of freedom that can be varied are the baselines L_i. Returning to the geometry of the first example, where $\theta_R = -80°$, the additional transmitter is placed in the position of the receiver, thus having a combination of monostatic and bistatic radar. In this way topologies can be tailored to meet application needs. However, it is most important to note that this must be balanced against the increased complexity of

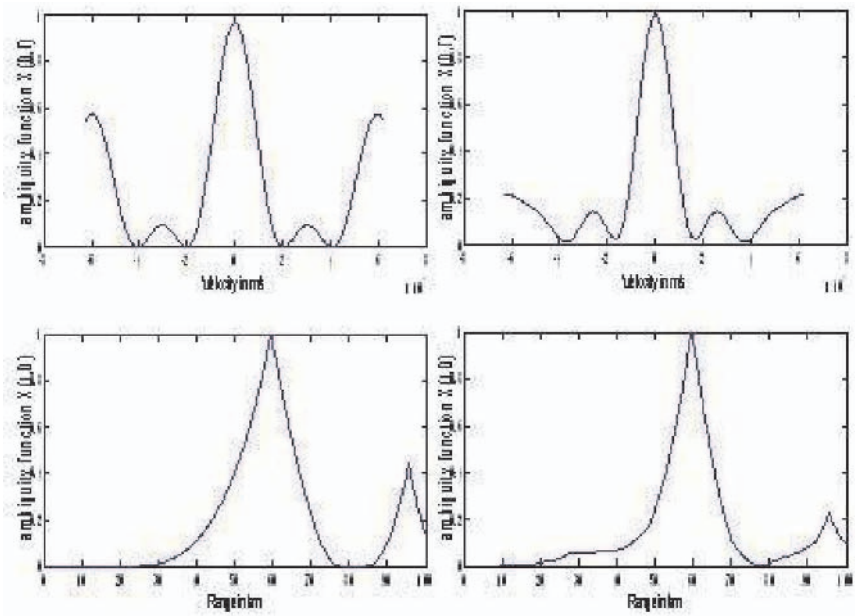


Figure 7. Bistatic and netted cuts of the ambiguity function balanced case.

multistatic radar. The exception to this is PCL where more than one transmitter is exploited. Here the optimisation of system performance will be via intelligent signal processing. It is also unlikely that it will be possible to form a fully coherent distributed sensor in the way described above. However, the ability to capture and process signals from a variety of transmitters, situated in differing locations, transmitting on different frequencies provides a great deal of diversity that can be exploited to advantage.

6. Conclusions

Bistatic and multistatic radar have significant differences to their monostatic counterparts. This offers both advantages and disadvantages, particularly in applications addressing the issues of homeland security. The particular attraction of PCL is that it may be added to augment existing capability for target detection within home air space. A multiplicity of receiver sites as well as each receiver exploiting multiple transmitters results in a system that can be tailored to the prevailing conditions. It can take advantage of the space, time and frequency diversity inherent at each receiver site, and no expensive transmitter needs to be procured. Indeed in the VHF bands one can conceive of a sys-

tem with equipment costs in the region of a few thousand dollars with a potential detection range of around 150km. For applications such as coastal surveillance for the protection of harbour areas and other points of vulnerability the lack of control over the emitted waveform may prove overly disadvantageous. Here there is a need to separate targets from (sea) clutter and generally resolutions are poor. However, as many of the targets of interest may well be moving at velocities considerably greater than that likely to be exhibited by clutter, Doppler processing might provide a route to success.

In summary, the essentials of bistatic and multistatic radar have been introduced and the resulting performance shows considerable promise for many application types. The additional attraction of the low cost of PCL makes this especially worthy of further attention.

7. Acknowlegments

The authors wish to thank Hervé Borrion for assistance in compiling this chapter into LaTeX. We also recognise the contributions made by a number of students and in particular, Ioannis Papoutsis and Daniel O'Hagan.

References

[1] Willis, N.J. *Bistatic Radar.* Artech House, 1991.

[2] Baniak, J., Baker, G. Cunningham, A.M. and Martin, L. *Silent Sentry passive surveillance* Aviation and Space Technology, 7 June 1999

[3] Cherniakov, M., Nezlin, D. and Kubin, K. *Air target detection via bistatic radar based on LEOS communication systems.* IEE Proc Radar, Sonar and Navigation Vol.149, No.1, pp 33–38, February 2002.

[4] Giffiths, H.D., Baker, C.J., Baubert, J., Kitchen. N. and Treagust, M. *Bistatic radar using spaceborne illuminator of opportunity.* IEE Conf. Publ. *Proc. RADAR 2002 Conference, Edinburgh* No.490, pp 1–5, 15–17 October 2002.

[5] http://www.roke.co.uk/sensors/stealth/celldar.asp

[6] Sahr, J.D. and Lind, F.D. *The Manastash Ridge radar: a passive bistatic radar for upper atmospheric radio science.* Radio Science Vol.32, No.6, pp 2345–2358, November-December 1997

[7] Zoeller, C.L., Budge, M.C., Jr. and Moody, M. *Passive coherent location radar demonstration* Proceedings of the Thirty-Fourth Southeastern Symposium on System Theory pp 358–362, 18–19 March 2002.

[8] Griffiths, H.D. and Long, N.R.W. *Television based bistatic radar.* IEE Proceedings Vol. 133, Part F, No.7, pp 649–657, December 1986.

[9] Griffiths, H.D., Garnett, A.J., Baker, C.J. and Keaveny, S. *Bistatic radar using satellite illuminators of opportunity* Proc. RADAR'92 Conference, Brighton, IEE Conf. Publ. No.365, pp 276–279, 12–13 October 1992.

[10] Griffiths, H.D. *From a different perspective: principles, practice and potential of bistatic radar* Proc. International Conference RADAR 2003, Adelaide, Australia, pp 1–7, 3–5 September 2003.

[11] Hawkins, J.M. *An opportunistic bistatic radar* Proc. RADAR'97 Conference, Edinburgh, IEE Conf. Publ. No.449, pp 318–322, 14–16 October 1997.

[12] Trizna, D. and Gordon, J. *Results of a bistatic HF radar surface wave sea scatter experiment* Proc. IGARSS '02' Vol.3, pp 1902–1904, 24–28 June 2002.

[13] Greneker, E.F. and Geisheimer, J.L *The use of passive radar for mapping lightning channels in a thunderstorm.* Proc. IEEE Radar Conference pp 28–33, 5–8 May 2003

[14] Yakubov, V.P., Antipov, V.B., Losev, DN. and Yuriev, I.A. *Passive radar detection of radioactive pollution* Application of the Conversion Research Results for International Cooperation SIBCONVERS '99. The Third International Symposium, Volume 2, pp 397–399, May 18–20, 1999.

[15] Howland, P. E. *Target tracking using television based bistatic radar* IEE Proc radar, Sonar and navigation, Vol 146, No 3, pp 166–174, 1999.

[16] Meyer, M. and Sahr, J.D. *Passive coherent radar scatter interferometer implementation, observations and analysis* Radio Science Vol.39, RS3008, doi: 0.1029/2003RS002985, May 2004.

[17] Herman, S. and Moulin, P.M. *A particle filtering approach to passive radar tracking and automatic target recognition* IEEE Aerospace Conference Proceedings, 2002. Vol.4, pp 1789–41808, 9–16 March 2002.

[18] Elirman, L.M. and Lanterman, A.D. *A robust algorithm for automatic target recognition using passive radar* Proceedings of the Thirty-Sixth Southeastern Symposium on System Theory, 2004 pp 102–106, March 14–16, 2004.

[19] http://www.bbc.co.uk/reception/

[20] http://www.sitefinder.radio.gov.uk/

[21] Tsao, T., Varshent, P., Weiner, D. and Schwarzlander, H. *Ambiguity function for a bistatic radar* IEEE Trans Aerospace and Electronic Systems, Vol. 33, No. 3, pp 1041–1051, 1997.

THE TERRORIST THREAT AND ITS IMPLICATIONS FOR SENSOR TECHNOLOGIES

Jennifer L. Brower
Prometheus Inc.
Newport, RI 02840, USA

Abstract Recent terrorist attacks demonstrated that even sophisticated terrorists capable of planning and executing multiple, coordinated attacks continue to rely on traditional weapons rather than risk the uncertainty of chemical, biological, radiological or nuclear (CBRN) weapons. While some terrorist organizations have the motivations and capabilities to conduct large attacks worldwide, we have not yet witnessed the use of so called weapons of mass destruction (WMD) foreshadowed by the 1995 Sarin attacks in Tokyo, the discovery of al Qaeda's crude biological weapons program in Afghanistan, and the anthrax attacks in the United States in the fall of 2001. Anti-Western extremists pose a global threat, but what do the use of traditional weapons and innovative tactics mean for the future of terrorism? This chapter describes our current understanding of the global terrorist threat including the use of CBRN weapons. A discussion of the implications for sensor research, particularly for chemical and biological agents and radioactive materials then follows.

Keywords: terrorism; al Qaeda; chemical terrorism; biological terrorism; radiological terrorism; nuclear terrorism; state sponsored terrorism; threat; sensors.

1. Introduction

The deadly terrorist bombings of July 7, 2005 in London again demonstrated that even sophisticated terrorists capable of planning and executing multiple, coordinated attacks continue to rely on traditional weapons rather than risk the technical and political uncertainty of chemical, biological, radiological or nuclear (CBRN) weapons. While terrorists have the motivations and capabilities to conduct large (and small) attacks worldwide, we have not yet witnessed the use of so called weapons of mass destruction (WMD) foreshadowed by the 1995 Sarin attacks in

Jim Byrnes and Gerald Ostheimer (eds.),
Advances in Sensing with Security Applications, 23–54.

Tokyo, the discovery of al Qaeda's crude biological weapons program in Afghanistan, and the anthrax attacks in the United States in the fall of 2001. The strike on commuter trains in Madrid, Spain; the bombing of a nightclub in Bali, Indonesia; and the attacks of September 11, 2001 in the United States demonstrate that anti-Western extremists pose a global threat, but what do the use of traditional weapons and innovative tactics mean for the future of terrorism? This chapter describes our current understanding of the global terrorist threat including the use of CBRN weapons.

The threat described in the first part of this chapter spurred increased investment in research and development technologies to prevent, detect, and respond to terrorist attacks. One specific area of research, sensors, particularly for chemical and biological agents and radioactive materials, in addition to radar and sonar, is the subject of this book. After describing the threat, this chapter goes on to discuss the use of sensors, fielded sensor capabilities, and existing gaps in sensor capabilities.

2. What is Terrorism?

2.1 Definition

There is no single definition of terrorism, and even when one can agree upon a definition, there may be disagreements about the classification of a particular incident. This chapter is written from a U.S. perspective (as the author is from the United States) and the author refers the reader to the definition of terrorism as defined by statute of the United States Government (Title 22 Chapter 28 Section 2656 f(d)):

> Terrorism is premeditated, politically motivated violence perpetrated against non-combatant targets by sub-national groups or clandestine agents, usually intended to influence an audience.

2.2 History of Modern Terrorism

Modern terrorism is generally cited as beginning in the late 1960s with the emergence of an independent Israel. In the nearly four decades since, four categories of terrorist organizations have emerged — ideological, ethno-nationalist, politico-religious, and single issue. In the 1960s through 1980s terrorism was generally practiced by members of an identifiable group with clear goals. For example, leftist terrorist organizations such as the Red Army Faction (Baader Meinhof Gang) wanted to form additional socialist states in Europe, and ethno-nationalist groups such as the Abu Nidal Organization and the Irish Republican Army (IRA) wanted separate homelands for 'their' people. These types of organizations generally chose tactics and targets to achieve their political, social,

or economic goals and claimed responsibility for their radical actions. As a result of decisions such as Roe v. Wade in the United States, and the emergence of fears of climate change and globalization, single-issue terrorism emerged in the late 1970s. It is broadly defined as "extremist militancy on the part of groups or individuals protesting a perceived grievance or wrong usually attributed to government action or inaction" [52]. In the late 1990s and early 2000s, animal rights and environmental activists were the most active domestic terrorists inside the United States in terms of number of attacks, but their belief system forbids harm to all animals (including humans) [30]. Later, groups began to emerge with less logical nationalist or ideological motivations, embracing instead more vague religious or millenarian objectives. The groups themselves were also less well defined [27]. Religious and millenarian groups such as al Qaeda and their affiliates and Aum Shinrikyo have grown to be the most dangerous terrorists based on their motivation to bring an end to modern civilization and their interest in WMD. We now turn to current thinking on the evolution of terrorism.

3. General Trends in Terrorism

Recent terrorist activity indicates that while the U.S. and its allies have enemies in many places, al Qaeda and its affiliates pose the greatest threat to Western interests. Al Qaeda — the World Islamic Front for Jihad Against the Jews and the Crusaders — is determined to bring an end to Western civilization as we know it. Note however that while the on-going attacks throughout the world are appalling in terms of the loss of human life and economic damage, the doomsday scenarios that many anticipated, involving massive casualties through the use CBRN weapons, have not materialized [22] and [37]. Lower consequence events are more likely than higher consequence events, primarily because of technical and logistical hurdles in executing large-scale attacks and because of counterterrorism efforts focused on avoiding mass attacks, but the risk of higher consequence events remains given the desire of al Qaeda (and others) to acquire or develop CBRN weapons.

A terrorism expert, Brian Jenkins, once noted, "Terrorists want a lot of people watching not a lot of people dead" [31]. At the time, only 15-20% of all terrorist incidents involved fatalities; however as early as 1987, he recognized the potential for increasing lethality. This was only one of the more recent trends he foresaw. Until the 1980s, terrorists often had specific goals for changing the behavior of a specific political body. With the evolution of religious extremism, the desensitization of terrorists and the public, growing resources and other factors, terrorism has changed

in many aspects. Terrorism experts and policy makers have examined changes in organization, motivation and capability (For instance see [22], [29], [48], and [56]. These trends were examined in the Fourth Report of the Gilmore Commission and are discussed and expanded below.

3.1 Increasing Lethality

First, selected terrorist groups are motivated and capable of killing more people in single or coordinated attacks than ever before. While the United States has been the target of terrorism for at least 35 years, more than three times as many people were killed on September 11, 2001 than in the history of modern terrorism until that day [28]. Worldwide, 14 modern terrorist operations achieved a death toll great than 100 before September 11 [32]. Since September 11, there have been several attacks with more than 100 deaths. In 2004 alone there were at least six attacks that claimed more than 100 lives each [16]. The most deadly was the Beslan school hostage crisis in Russia when 344 people were killed. Others include a TNT explosion on a ferry outside Manila which killed 118 (Abu Sayyaf Group); Ansar al-Sunnah's near simultaneous suicide attacks on two Kurdish Government targets in Arbil, Iraq, which killed 117; the slaughter by arson of 239 civilians in Northern Uganda by the Lord's Resistance Army; the multi-pronged bombing and mortar attack on the holy Shiite city of Karbala, which killed 106 (attack attributed to al Qaeda or Zarqawi loyalists, but no one claimed responsibility); and the bombing of commuter trains in Madrid Spain, which killed 191 and injured more than 600 (Abu Hafs Al Masri Brigade on behalf of al Qaeda). Other high casualty attacks include the October 2002 attack on a Bali nightclub and the Chechen attack on the Palace of Culture Theater that same month in which 162 were killed when the Russian Special Forces attempted to free the hostages using an incapacitating gas [41].

While there have been multiple attacks that have killed over 100 individuals, and despite the desire to carry out high profile mass casualty attacks, the arrests of key members of terrorist organizations have degraded al Qaeda's capability to conduct large attacks inside the United States and elsewhere. Several high casualty attacks were thwarted by law enforcement and intelligence agencies. For instance, in late 2001 intelligence from Afghanistan was used to detain 13 Jemaah Islamiyah (JI) members suspected of plotting to target U.S. Navy ships and sailors in Singapore. According to Gunaratna (2003), JI presents the largest terrorist threat in Southeast Asia with nearly 400 al Qaeda trained

members. Indonesia's tolerance has allowed a training infrastructure to thrive.

In addition, al Qaeda and its affiliates have lost several operational leaders, safe havens and sources of financing [56]. Those detained or arrested include, Khalid Sheik Muhammad, al Qaeda's operations chief and mastermind of the September 11 attacks (March 2003); Abd al-Rahim al-Nashiri, a senior operational planner in the Persian Gulf and mastermind of the USS Cole attack (November 2002); Abu Zubaydah, responsible for al Qaeda's recruitment and training and involved in the East African bombings in 1998 (March 2002); Omar al-Farouq, al Qaeda's operations chief in Southeast Asia (June 2002); Riduan Isamuddin, also know as Hambali, mastermind of the Bali bombings (August 2003); and Ibn al-Shaykh al-Libi, head of al Qaeda's training camps (December 2001). To avoid further disruptions to their attacks, terrorist groups have been forced to change tactics and targets.

3.2 Innovations in Tactics and Targets

Generally modern terrorist organizations have not been particularly innovative, often relying on a group of attack types and imitating other terrorist organizations. Even al Qaeda, an innovative group as described below, was influenced by a precursor, the Iranian-sponsored Lebanese Hezbollah, and particularly their ability to coordinate multiple attacks [24]. Al Qaeda has used coordinated, multiple attacks several times including the 1998 East Africa bombings; the September 11 attacks; and the bombings of the Interior Ministry and Recruiting Center in Riyadh Saudi Arabia in December of 2004. Affiliated terrorist groups have also imitated this tactic using multiple near-simultaneous bombings to kill scores of people. In Pakistan, the Muslim United Army simultaneously bombed 21 gas stations on May 15, 2003 using improvised explosive devices [24] and most recently on July 7, 2005 four bombs detonated nearly simultaneously across London. Also recently, in June 2005, four car bombs detonated in the early evening near Baghdad, Iraq killing at least 23 [8].

Bombings, assassinations, armed assaults, kidnappings, hijackings, and barricade and hostage incidents were the tactics used in nearly 95% of all terrorist attacks until 1987 [31]. This statistic remains in effect today. The arrest of key planners, and other disruptions, has forced al Qaeda and its affiliates to change tactics and targets and focus on smaller-scale, softer targets such as hotels, religious and holy sites, and infrastructure [20]. This is not the first time security changes have impacted tactics. For instance, terrorists commonly took control of em-

bassies in the 1970s, but as states implemented security measures, the seizures declined [31]. More recently, because of lessons learned from the Oklahoma City bombing in 1995 and the East African bombings in 1998, the U.S. secured the land access to its embassies and other government targets: al Qaeda then turned to maritime targets and attacked the USS Cole. After the U.S. decreased the vulnerability of maritime targets, al Qaeda used airliners to strike on September 11 [24]. Since September 2001, as the U.S. and others strengthened airline security, al Qaeda turned its attacks to the rail infrastructure in London and Madrid.

Because the United States has fortified its defenses, terrorists have increasingly attacked Western targets abroad. In June 2005, hotels in Indonesia were put on alert after the U.S. embassy released a statement indicating that they were the targets of an imminent terrorist attack. [1] And while no attack materialized in Indonesia, several other soft targets including a hotel in Kenya, (November 2002) and synagogues in Turkey (2003 and Tunisia (2002) have been attacked by al Qaeda. In May 2002, Abu Zubaydah reinforced the threat to soft targets, particularly places where large number of Americans gather [51].

Terrorists are also adapting technology to improve their tactics and targeting. For example, al Qaeda is adapting dual technologies such as airplanes and commercially available chemicals, agricultural fertilizers, and liquid propane and nitrogen [24]. The Tamil Tigers of Eelam (LTTE) has also tried to acquire microlite airplanes and built its own airstrip purportedly to conduct suicide missions from the air [50].

In another innovation that resulted, in part, from Bin Laden's observation of the economic damage that followed the September 11 attacks, terrorist groups are increasingly attacking economic targets. Evidence of al Qaeda's evolving strategy can be found in the tapes released periodically by Bin Laden and his associates. In a tape released on October 6, 2002, four days prior to the Bali bombing, Bin Laden and his lieutenant Ayman al-Zawahiri warned "By God, the youths of God are preparing for you things that would fill your hearts with terror and target your economic lifeline until you stop your oppression and aggression" [6]. Energy, particularly oil, has been specifically targeted. According to IntelCenter, an al Qaeda document translated in 2004 called for "hitting wells and pipelines that will scare foreign companies from working there and stealing Muslim treasures." For example, in October 2002 terrorists targeted a Malaysian oil tanker and killed a Bulgarian sailor off the cost of Yemen. Abdel Rahim Al-Nashiri, a key member of al Qaeda, reportedly financed the attack as well as the attack on the USS Cole. Nashiri was arrested in 2002. Bin Laden is also aware of the costs to defend against

potential terrorist attacks. In a tape released in late 2004, Bin Laden said, "...We are continuing this policy of bleeding America to the point of bankruptcy." [2]

The focus on economic targets is not completely new. As early as 1978, a Palestinian group called the Arab Revolutionary Army injected Israeli fruit exports with mercury to damage the Israeli economy. Officials and consumers found contaminated fruit in Holland, West Germany, Belgium and England [41]. The globalization of the world economy and the just-in-time logistics used in many Western countries has increased the visibility of economic disruption and therefore the appeal of this type of attack to terrorist groups.

3.3 Leaderless Resistance and Loose Networks

The successful disruption of al Qaeda described above has forced the group to decentralize further into a "loose collection of regional networks that operate more autonomously" [56]. A terrorist plot interrupted by U.S. and Singapore Intelligence in December 2001 demonstrates how a network of extremists from throughout Southeast Asia were willing to work in conjunction with al Qaeda leadership to plan an attack on several U.S. targets including the U.S. Embassy, a Navy ship, and Navy personnel [62]. JI was identified as the operational leader; however, eight of thirteen men arrested in January 2002 for connections to the plot had trained in al Qaeda camps in Afghanistan and Malaysia. Surveillance footage of the Singaporean targets was found in the home of an al Qaeda leader in Afghanistan, indicating close coordination between operatives in Southeast Asia and Afghanistan [62]. Although recent attacks in London bore the hallmarks of an al Qaeda attack, the quality and quantity of the explosives initially point to local militants, suggesting that the attack may have been loosely coordinated.

Cooperation in Southeast Asia is the most developed, but Islamic extremists in Central Asia, North Africa and even Europe and North America have also formed loose networks. For instance, in September 2002 the Islamic Movement of Uzbekistan (IMU) was formed, bringing together separatists from Kyrgyzstan, Tajikistan, Chechnya, and the Xingjiang Province of China [19].

Although Bin Laden's abundant resources, expertise, and agenda is attracting several groups to form loose networks, some terrorist groups have shied away from identifying themselves with Bin Laden to avoid being the focus of the war on terror or to focus on local goals. For example, when U.S. forces in Afghanistan killed several of its members, and its assets were frozen in late 2001, the Harakat ul-Mujahadeen split off the

al Almi faction (HUM-A). HUM wanted to focus on the local agenda in Jammu and Kashmir, while HUM-A wanted to continue to pursue Bin Laden's global jihad against the West. The more extreme HUM-A has gone on to attack Western targets including killing 10 French businessmen in Karachi and bombing the U.S. Consulate in Karachi in 2002 [41] and [22].

3.4 Incorporation of Technology

Terrorists are exploiting advances in information technology, such as email, the internet, encryption, and video/audio production, to coordinate internal communication and to spread their message for recruiting and fund raising purposes [25]. Terrorists are also using the internet in finance operations and to encourage hacking to deface Western websites or perpetrate denial of service attacks [46]. In late 2004, Imam Samudra, the convicted mastermind of the Bali nightclub bombing, directed compatriots to Indonesian language websites that contain instructions on online credit card fraud and money laundering [55]. This change has evolved in part as a reaction to the disruption of traditional modes of communication, financing, and safe havens. These innovations present both dangers and opportunities. While al Qaeda and others may garner support through the use of technology, the employment of information technology can be exploited for intelligence gathering [61]. Al Qaeda is aware of its vulnerabilities, and there is some evidence that al Qaeda has deliberately created noise in the system to overwhelm intelligence agencies trying to decipher terrorist communications [35]. Al Qaeda operatives have also use advanced encryption technology to prevent the Intelligence Community from gaining access to their plans. Even in the mid 1990s, Wahid El Hage used encryption to send secure e-mails while plotting the East Africa embassy bombings. Former Director of the U.S. Federal Bureau of Investigation (FBI) Louis Freeh testified as early as 1998 that "One of the most difficult challenges facing law enforcement is how rapidly criminals and terrorists — both domestic and international — adopt advanced technologies to thwart the ability of law enforcement to investigate those who wish to do harm to our Nation and its citizens. That is why encryption has become the most important issue confronting law enforcement" [23].

Terrorist exploitation of technology is further enhanced when groups share their technological advances with other groups both formally and informally. As noted above, Samudra aimed his exhortation to use the internet for fraud and money laundering at all jihadists working towards the downfall of the West.

3.5 Groups Working Together

Terrorist groups both with related and unrelated objectives are cooperating to various degrees, driven in part by the war on terrorism. In the past, organizations such as the Palestine Liberation Organization (PLO), the Provisional Irish Republican Army (PIRA), and the Basque Fatherland and Freedom (ETA) have worked together. The IRA also trained Revolutionary Armed Forces of Columbia (FARC) militants in Colombia, reportedly in exchange for $2,000,000, in the use of explosives [54].

Al Qaeda's direct training and its sharing of resources and expertise brings these interactions to the next level [22]. In response to the key arrests and the disruption of its own network, al Qaeda has reached out to foster its global impact. For instance, al Qaeda is cooperating with JI and the Moro Islamic Liberation Front (MILF) in Southeast Asia; Al Ithihad al Islami in the Horn of Africa; Al Ansar Mujahidin — Islamic International Brigade (Caucuses); the Tunisian Combatants Group in the Middle East; Jayah-e-Mohammad in South Asia; the Salafi Group for Call and Combat (GSPC) in North Africa, Europe and North America; and several other Islamist groups [24]. The interactions are beneficial to both sides. For example, al Qaeda supports a MILF-run training camp that both groups use to train themselves and others [12]. Even Sunni Muslim groups such as Hamas, al Qaeda, and Islamic Jihad are now cooperating with the Shiite Muslim group Hezbollah because of their shared hatred for the West, as evidenced by the issuance of joint press statements [22].

The interaction of multiple terrorist groups has also contributed to changes in tactics and targets. Al Ansar Mujahidin (Baryayev Gang) was clearly influenced by al Qaeda when it attacked the Moscow Theater in October of 2002. Movsar Baryayev, who was a close colleague of Ibn ul-Khattab, led the attack. Ibn ul-Khattab was a Chechen military leader, a protg of Bin Laden, and member of al Qaeda until his death in March 2002. Movsar utilized al Qaeda inspiration in the scale of the operation, the suicide potential, and the coordination [24]. The bombing of the nightclub in Bali grew from a local terrorist plot to conduct a number of small bombings on soft targets to the large-scale bombing after al Qaeda contributed bomb making expertise and resources to JI [24]. In general, al Qaeda's influence on Muslim ethno-nationalists is growing in the Philippines, Indonesia, Thailand, India-Pakistan and Russia through both imitation and the provision of direct training and resources. JI is a particularly good example of the impact al Qaeda has had. When the former Chief of JI in Singapore was arrested, he told investigators he

had been planning to hijack an Aeroflot plane from Bangkok and crash it into Singapore's main airport in 2002 to teach Russia a lesson — a clear emulation of the September 11 attacks. Further, the Bali attacks were both the first mass fatality and first suicide attack perpetrated by a Southeast Asian terrorist group — both as a result of contact with al Qaeda [24].

3.6 Threat of Individuals

Individuals acting without the support of a specific group, and who may sympathize with al Qaeda, the Palestinian cause, environmental causes or other grievances against the United States and its policies domestically or overseas also pose a threat. The threat from individual terrorists is increasing in part to the spread of propaganda and techniques and tools through the internet, and the threat is broader than that posed by al Qaeda and its affiliates [22]. For example Richard Reid, who was inspired by bin Laden, attempted to down an airliner over the United States using a shoe bomb [59]; the Egyptian, Hesham Mohamed Hadayet, killed two people at the El Al counter on July 4, 2002; and Osman Petmezci (Turkish) and Astrid Eyzaguirre (American) were stopped by German police before they could attack U.S. Army Headquarters in Heidelberg. After the arrests, police found 130 kg of bomb making components, related equipment, and a picture of Bin Laden in the couple's apartment [24]. Individual terrorists are harder to detect and stop, but they are also less likely to have sophisticated training or a wealth of resources and are therefore less likely to succeed. A particular threat in the United States is U.S. citizens who sympathize with international terrorist groups such as al Qaeda. Similar domestic threats pose challenges in other countries as well. It is to this threat that we now turn.

4. Significant Domestic Threats

While al Qaeda is considered the primary anti-Western terrorist organization, there are several other groups that have significant capabilities, and if their goals and motivations turned towards Westerners, defending against them would be a difficult endeavor. Locally these groups already present a challenge to the governments trying to protect their citizens. LTTE and FARC, two long-standing terrorist organizations, are discussed in this section because although they do not currently pose a global threat or have an anti-Western agenda, because of their organizational strength and capabilities they may pose a threat in the future.

The LTTE is one of the most deadly and persistent ethno nationalist/separatist organizations in the world. Born in 1976 out of the moderate Tamil United Liberation Front, their goal was to represent the Tamil minority in Sri Lanka and to create a separate state in the eastern and northern provinces. In addition to targeting the Sri Lankan government, LTTE has targeted civilians and other Tamil separatist groups. More than 60,000 people have been killed in the conflict since the mid 1970s with 514 fatalities attributed to the LTTE [41]. Driven in part by pressure from the global war on terrorism, the Sri Lankan government and the LTTE signed a cease-fire agreement on February 22, 2002. While the cease-fire generally held for all of 2003, violence resumed on July 8, 2004 when a female LTTE suicide bomber killed four policemen while attempting to assassinate Eelam People's Democratic Party member Douglas Devanda [41]. More recently, in February 2005, the LTTE leader Kausalyan was killed, reportedly by paramilitary Sri Lankan forces. This increase in violence has been driven by two factors: a split within LTTE over control of the Tamil vision between LTTE and another Tamil leader, Colonel Karuna and disagreement throughout Sri Lanka over the distribution of tsunami aid by the LTTE. The LTTE pose a significant local danger both in terms of the terrorist acts and in terms of recruiting children soldiers, but they have not turned their attention to the world of global terrorism or to WMD. In addition, the LTTE are well known for their innovations and adoption of technology, including the first use of women suicide bombers and their attempt to acquire microlite aircraft for terrorist attacks [49]. Continued vigilance is necessary because they are a sophisticated, organized group and splinter groups could become more radicalized. Or current interactions with Islamic terrorist groups such as Abu Sayyaf and the Moro Islamic Liberation Front could be exploited by any of the organization's splinter groups.

The FARC began as a Marxist organization determined to overthrow Columbia's government and replace it with a communist regime. The organization has wandered from its origins and increasingly focuses on the illicit drug trade and engaging in peace talks with the government. It now has more limited goals of controlling territory within the country. In addition to cocaine sales, the FARC uses kidnapping, extortion and hijacking to pad its coffers, reportedly taking in as much as $2 million per day. Its targets are private citizens and business with resources, rival communist terrorists, Columbian political and military entities, and rightist paramilitary forces. FARC does not present a direct threat to the United States, but its extensive drug trafficking network throughout the Americas, its significant resources, and its interactions with other

terrorists groups illustrate the potential for conducting significant attacks in the continental United States. This scenario will become more likely if increased cooperation among FARC and anti-Western extremists is observed, or if the U.S. war on drugs impacts their revenue stream.

4.1 Regional Assessment

The LTTE and the FARC represent two domestic threats. We now turn to an examination of some of the regions of the world with high levels of terrorist activity. According to the RAND/MIPT Terrorism Knowledge Database, the Middle East/Persian Gulf Region had both the largest number of incidents since the beginning of the new millennium and the highest number of injuries (13,663) and deaths (5,906). Almost 41% of the attacks and 75% of the deaths occurred in Iraq. Tanzim Qa'Idat Al-Jihad Fi Bilad Al-Rafidayn, a nationalist separatist group, has been one of the most prolific perpetrators since it was first mentioned in October of 2004. Israel, with 361 attacks and 681 deaths, and the Occupied Territories, with 1,607 attacks and 437 deaths, were also hit hard by terrorism. Hamas, Islamic Jihad, and Palestinian Islamic Jihad (PIJ) have perpetrated the vast majority of attacks both in Israel and in the Occupied Territories. East and Central Asia had the fewest attacks during this time period with 63 attacks and 63 deaths. Hamas presents some of the most violent opposition to Israel and has conducted numerous attacks using suicide bombers and rockets. It has caused nearly 600 deaths and 3000 injuries and is supported by Iran as well as numerous Islamic charities. PIJ is also dedicated to the destruction of Israel, but its actions have been reduced since the death of it leader Fathi Shaqai in 1995 and the start of the global war on terrorism [41].

Iraq's most infamous terrorist group is Tanzim Qa'idat al-Jihad Fi Bilad al Rafidayn (al Qaeda in Iraq), led by Abu Musab Zarqawi. This group has claimed responsibility for more than 100 attacks and 580 deaths with the stated goal of overthrowing the interim Iraqi Government, ridding the country of the American-led coalition, and forming an Islamic state. Recently, in May of 2005, multiple suicide bombers detonated blasts outside a courthouse in Baqubah killing several policemen and bystanders while attempting to kill the provincial governor of Diyala province. Tanzim Qa'idat al-Jihad Fi Bilad al Rafidayn is the successor to another deadly terrorist group, Tawhid and Jihad. This group is responsible for at least 25 incidents and nearly 200 fatalities. They used kidnappings, beheadings, assassinations and suicide bombings to move towards their goal of an Islamic State. Ansar al-Sunnah has been allied

with both of these groups, but commits deadly acts in its own name. For example, on May 11, 2005 the group perpetrated two fatal attacks - a car bombing in a Tikrit market near a police station that killed 38 and injured 84, and a suicide bombing that killed 32 recruits in Hawija. Ansar Al Islam, with similar goals, also operates within Iraq. Coalition forces decimated the organization's sanctuary in the Kurdish region of northern Iraq, and the group's leader, Mullar Krekar, was arrested in 2004 in Norway. However, Islamic jihadists are reportedly joining forces with the group, and in January of 2005, Ansar Al Islam perpetrated its first deadly attack in two years gunning down several Shiia religious leaders [41]. The number of terrorist groups allying themselves in Iraq demonstrates that Iraq is proving to be a fertile training ground and site for cross-fertilization.

Kashmir and Jammu, the disputed area between India and Pakistan is another active center of terrorist activity. Lashkar-E-Taiba (LeT), the militant arm of Markaz Dawa ul Irshad, along with Harkat ul Ansar (HuA) and Al-Badr (now known as Hizb ul Mujahidin) are responsible for much of the violence over the past 20 years. LeT is trying to establish Islamic rule over India and questions India's control of Jammu and Kashmir. As a member of Osama bin Laden's Islamic Front for Jihad against the U.S. and Israel, LeT poses a direct threat to U.S. interests, particularly in Pakistan and Afghanistan. Pakistan banned the group under pressure from the U.S., but it continues to operate. LeT relies on traditional terrorist arms and was blamed for the August 2003 bombing in downtown Bombay that killed 52 and injured more than 100 [41]. Hizb ul Mujahidin, the largest Kashmiri terrorist organization, is the militant wing of Pakistan's largest Islamic political group Jamaat-I-Islami. It primarily attacks political and military targets in Kashmir.

The Russian-Chechen conflict is another hotbed of terrorist activity. The Chechen terrorists, including the Movsar Naryayev Gang (MNG), support an independent Islamic Chechnya and are heavily influenced by radical Islam. The MNG was responsible for the well-publicized hostage taking at Moscow Theater in October 2002. When Barayev died in the attack, remnants of the group reportedly formed the extremely violent Riyad us-Saliheyn Martyrs' Brigade, which has claimed responsibility for the most heinous terrorist attacks in recent Russian history including the destruction of the pro-Russian Chechen government building in 2002 which killed 72, and an attack nearly a year later on a Russian hospital which killed 52. The group, like the LTTE, has also used women suicide bombers, both to bring down an airplane in August 2004 and to bomb Russian subways [41]. The group has strong ties to al Qaeda and is

willing to perpetrate large-scale attacks, and therefore poses a threat to the West.

As shown by the activity and aims of the terrorist groups discussed in this section, regional conflicts can directly affect Western security and safety, particularly as these conflicts fuel innovation and the interaction and training of multiple terrorist groups.

5. State Sponsored Terrorism

In addition to the organized and loose networks that generally fund their own activities, there are a number of states that sponsor terrorism. This is of particular concern with regards to weapons of mass destruction, because the numerous resources that can be brought to bear in state development of CBRN weapons can in turn be transferred to terrorists. Disincentives do exist to prevent this proliferation. As of October 2004, the United States lists six countries as state sponsors of terrorism: Cuba, Iran, Libya, North Korea, Sudan, and Syria. Of these, five are pursuing WMD to one degree or another. Libya repudiated its WMD program as discussed below.

Iran is a particular focus of the United States because of the convergence of its covert nuclear weapons program and its robust support of terrorism, going so far as to provide safe haven for the West's most threatening enemy, al Qaeda. In 2003, Iran refused to identify al Qaeda members in Iran or to transfer them to their countries of origin or third countries for detention and interrogation. Iran also provides material support including money, weapons, training and refuge to Hizballah, Hamas, PLFP, and Palestinian Islamic Jihad, and allowed terrorists fighting the coalition in Iraq, including members of Ansar al-Islam, to find safe haven within its borders. One of the members of its Guardian Council, the body that determines whether laws passed by the Iranian Parliament are in line with its constitution, promoted the idea of suicide attacks on coalition forces [47].

Iran is believed to have stockpiled chemical weapons and the means to deliver them. It is also believed to have a nascent biological weapons program, although it is unlikely to have sophisticated weaponized agents at this time. There is no question that Iran has acquired dual use biotechnology equipment, but the use of that equipment in the development of biological weapons has not been confirmed. Finally Iran's clandestine nuclear program is of greatest concern globally. Iran has violated its obligations under the non-proliferation treaty and International Atomic Energy Agency commitments several times. Violations include uranium enrichment; the creation of weapons grade plutonium; the development

of uranium mines and conversion facilities; and the construction of a heavy water production plant and several other hallmarks of a nuclear weapons program. These investments, purportedly for nuclear energy, are particularly unusual for an oil-rich state [11]. Iran's weapons are primarily directed at enemies such as Israel, but because of the country's status as the most active state sponsor of terrorism through 2003 [47] and its past and present association with terrorists, the threat must be closely monitored. The on-going debate about the background of Iran's newly elected President adds to the concern.

On the other hand, the threat from Libya has been significantly reduced since Moammar Gadhafi renounced his countries WMD aspirations in December of 2003. However, Libya remains on the list of state sponsors of terrorism. Prior to Libya's commitment to repudiate terrorism, it had supported some of the most deadly terrorist attacks including the bombing of Pan Am Flight 103 over Lockerbie, Scotland on December 21, 1988. But since Gadhafi's renunciation, documented progress has been made. By January 2004, U.S. and U.K. officials had removed critical elements of Libya's nuclear and long-range ballistic missile programs and consolidated its existing chemical weapons to protect them from terrorists and ease destruction [18].

Sudan was al Qaeda's primary operation base in the early and mid 1990s, and operatives from Sudan participated in the 1998 bombings of the U.S. embassies in Kenya and Tanzania, but the country is now working to combat terrorism and protect U.S. citizens within its borders. In 2003, the country shut down a major thoroughfare in front of the U.S. embassy in Khartoum to reduce the threat and closed a training camp, expelling several Saudi citizens training at the base [47]. Because Sudan possesses only a limited chemical capability and because of its recent cooperation with the U.S., it is unlikely that Sudan will deploy WMD through its limited terrorist network.

While North Korea poses a clear strategic threat to the United States and its allies, Kim Jong Il is not known to have directed any terrorist acts since the downing of Korean Air Flight 858 in 1987 in which 115 people were killed. Kim Jong was apparently trying to derail South Korea's Olympic events. However, because North Korea has nuclear, chemical and biological weapons capabilities and has supplied ballistic missile technology to other state sponsors of terrorism, it is unclear whether North Korea might provide materiel or technology for CBRN weapons to terrorist groups [57]. North Korea's primary motivation for selling ballistic missile technology to countries such as Iran and Syria appears to be the need for hard currency as opposed to specific anti-Western aims.

Syria has not directly participated in any known terrorist acts since 1986; however, Syria provides support and refuge for Hamas, the PIL, and the PFLP, among other Palestinian liberation groups. Syria distinguishes between what it calls the legitimate fight of the Palestinians and other terrorist groups. Syria also allows Iran to supply Hizbollah in Lebanon through its border, and with Iran reportedly supported Hizbollah's deadly bombing of the Marine Barracks and U.S. Embassy in Beirut in 1983. (Target America, 2001) The secular regime in Syria has begun to cooperate with the United States in its war against the Sunni dominated al Qaeda. Although Syria has a highly developed chemical weapons program and a less developed biological weapons program, and despite its sponsorship of anti-Western terrorism in the 1970s and 1980s, it is unlikely that Syria will share its chemical and biological capabilities with terrorist groups at this time. Syrian support now focuses on terrorism with limited political aims, and it prohibits anti-Western terrorism, limiting the threat to all but its primary target, Israel [47] and [58].

Cuba, while listed as a state sponsor of terrorism by the U.S. Government, does not provide significant resources or support for the most dangerous anti-western terrorists. Cuba is listed because it provides safe harbor to members of designated terrorist groups such as the Basque Fatherland and Liberty (ETA) and the FARC as well as other fugitives from U.S. justice [47].

Taken together, the on-going threat posed by state sponsors of terrorism has been reduced since the 1970s and 1980s when ideological and ethno-nationalist separatist groups aligned their fortunes with states; however, the future threat is significant, particularly if Iran or North Korea share their technological advances with groups they support, either because they share common goals or simply for financial gain.

6. Future Threats

Because so many national security experts and policy makers have predicted the use of CBRN weapons, it continues to be surprising that terrorists have been unable to follow through on their desires on a mass scale [22]. Technical and operational challenges present significant barriers to large-scale terrorist acquisition and use of CBRN l weapons [26]. Even al Qaeda, with its significant resources and global reach has not yet demonstrated the use of a sophisticated chemical or biological weapons capability. Based on videotapes discovered in Afghanistan and statements from Ahmed Ressam, the jailed terrorist who was planning to bomb the Los Angeles airport in December 1999, as of that time, al Qaeda was only capable of poisoning a trapped dog [22].

Arrests, the loss of sanctuary and the freezing of assets have diminished some threats, but it has spawned others as terrorists adjust to being the focus of the global war on terrorism. What will be the balance? Will loss of sanctuary and financial resources prevent terrorists from developing or acquiring CBRN that can be used to mass effect? Or will the global war on terrorism embolden terrorists and states to share and use CBRN? While terrorists continue to use traditional weapons in innovative ways, a primary concern is terrorist adoption of CBRN weapons in the future. The potential for this is explored next.

As discussed above, several states that support terrorism have some CBRN capabilities, so the technical constraint alone is not limiting. Rather, the potential backlash against any state that provides a terrorist organization with CBRN has been a sufficient deterrent to this point. However groups such as al Qaeda, Aum Shinrikyo, and the Tamil Tigers have shown significant interest in one or more types of unconventional weapons.

There can be no doubt that, if given the opportunity, terrorist groups such as al Qaeda would not hesitate to use disease as a weapon against the unprotected; to spread chemical agents to inflict pain and death on the innocent; or to send suicide-bound adherents armed with radiological explosives on missions of murder [10].

With the spread of information and the desperation for hard currency of some of the state sponsors of terrorism, as well as the changing national security environment, it is possible that terrorists may build or acquire CBRN in the future.

Bolton's opinion was bolstered in June 2005 by Senator Richard Lugar's survey of 85 non-proliferation and national security analysts from the United States and other nations. It was designed in part to characterize the risks related to the terrorist use of CBRN. The survey revealed that experts believe the probability of an attack somewhere in the world with a CBRN weapon was 50% over the next five years and 70% over the next ten. An attack with a radiological weapon was seen as the most probable with the likelihood of an attack with a nuclear or biological weapon considered about half as plausible [37]. The average probability of a nuclear attack in the next ten years was nearly 30%, with experts almost evenly divided between terrorist acquisitions of a working nuclear weapon versus self-construction [37]. The average risk estimate over ten years for major chemical and biological attacks was 20%. Senator Lugar concluded "The bottom line is this: for the foreseeable future, the United States and other nations will face an existential threat from the intersection of terrorism and weapons of mass destruction."

George Tenet, the former Director of Central Intelligence went even further in his February 2004 testimony before the Select Committee on Intelligence. "I have consistently warned this committee of al-Qaeda's interest in chemical, biological, radiological, and nuclear weapons. Acquiring these remains a 'religious obligation' in Bin Ladin's eyes, and al Qaeda and more than two dozen other terrorist groups are pursuing CBRN materials."

A number of trends discussed above favor the eventual use of CBRN weapons. The willingness to commit mass murder is primary among them. Cross fertilization among terrorist groups increases the likelihood that terrorists will develop and use more sophisticated tactics and weapons as groups share information and resources on materials, methods, and tactics. Splinter groups are seen as more likely to attempt innovation; and the spread of technology will put the power to develop ever more sophisticated weapons in the hands of terrorists.

To establish themselves as significant players in the political realm, splinter groups tend to be both more violent and more experimental than their parent groups. For example, Ansar al-Islam, a splinter from the Islamic Movement of Kurdistan (IMK) that associates with al Qaeda, established a lab in northern Iraq to manufacture and test chemical and biological agents, including ricin, for use in terrorist attacks [40].

There are several specific factors that indicate terrorist groups are making progress in the pursuit of CBRN materials and technology. A few highlights include:

- The wide dissemination of information across the internet by terrorists including instructions for improvised chemical weapons [56] and the open source information in scientific journals,
- The dissemination of anthrax in the United States in the fall of 2001,
- The discovery in January 2003 of remnants of ricin, castor beans, and recipes for a half dozen other chemical and biological weapons in the London apartments of terrorists aligned with al Qaeda,
- Unearthed terrorist documents in Afghanistan indicating al Qaeda's interest in nuclear, radiological, and biological weapons [56],
- Continuing discoveries of chemical precursors in Aum hideaways in Japan.

Al Qaeda in particular continues to pursue unconventional weapons, both leveraging existing commercially available agents and technologies and creating CBRN weapons. According to Rohan Gunaratna, "The

group is also searching for new weapons such as chemical and biological agents, especially contact poisons easy to conceal and breach security" [24]. However, contact poisons and the like are unlikely to cause the mass casualties often cited by U.S. security experts. Gunaratna also notes that a fatwa issued by Sheikh Nasr bin Hamid al Fahd in May 2003 legitimizes the use of CBRN weapons. Such a fatwa is a requirement in Islam before an attack. We can learn something about past and current terrorist capabilities and motivations by examining documented cases of actual use of chemical and biological weapons.

6.1 Actual Use of CBRN

Despite significant interest in unconventional weapons, there have been few instances of widespread death or incapacitation due to CBRN use by terrorists, and the number of casualties pales in comparison to those killed by more conventional explosives, armed attacks and arson. Since 1968, more than 14,000 people have been killed by bombing, and nearly 6,000 by armed attack, but CBRN attacks have accounted for less than 20 deaths [41]. The two most notorious unconventional attacks in modern history, Aum Shinrikyo's gassing of a Tokyo subway in 1995 and the anthrax attacks in the U.S. in the fall of 2001, killed a total of 17 people. The food poisoning by the Rajneeshees in Oregon in 1984 has also received much attention. While there were no fatalities when the cult poisoned several salad bars with Salmonella, there were more than 700 injuries.

The two known deadly attacks using either chemical or biological weapons are now discussed. In each of these incidents, less than 20 people were killed, but several hundred were injured in Japan. In both cases, the resulting fear and response led to much greater disruptions and costs than the attacks themselves.

Aum's story is an illustration of how a group with significant financial resources and educated personnel may still have a hard time surmounting the technological and organizational challenges to developing a true WMD. Aum Shinrikyo, which translates as the "Supreme Truth," is a Japanese religious cult led by Shoko Asahara. Asahara drew on Christianity, Buddhism, and Hinduism to create his own religion, which, at its peak, attracted up to 40,000 followers worldwide, primarily in Russia and Japan [41]. The group first attracted the significant attention of law enforcement in 1995 after it gassed the Tokyo subway, killing 12 and injuring hundreds. Its activities and interest in unconventional weapons began long before the attack. In the early 1990s, the cult had an estimated net worth in the hundreds of millions to a billion dollars and had a

cadre of scientists including 20 university-trained microbiologists. Aum provided these members with the necessary equipment and materials, and yet the group failed in ten attempts to kill large numbers of people with either anthrax or botulinum toxin. Their failure was attributed in part to the use of a non-lethal strain of anthrax and technical difficulties in disseminating the biological agents, which proved less hardy and stickier than anticipated. Aum also tried unsuccessfully to acquire nuclear weapons and materials from Russia as well as mine uranium in Australia. Aum then turned to developing chemical agents, and while the group successfully killed individuals in Matsumoto (in 1994 Aum targeted three judges with Sarin and killed seven) and Tokyo, they did not achieve the doomsday scenarios anticipated by Asahara and dreaded by U.S. and international leaders.

Even after the arrest of its leader and other key personnel, the Japanese government did not fully outlaw the sect and a small group of followers remain. The group also retains a large network of business and influence interests. In April 2004, the Japanese Justice Ministry's Public Security Investigation Agency released a report that indicated that Aum, renamed Aleph in 2000, had set up more than 10 businesses throughout Japan. The cult purports to raise money to help victims, but the Justice Ministry claims that these businesses are designed to raise money for Aum's operating expenses [34]. Armageddon remains the cults guiding concept and the Japanese government continued to discover Sarin precursor chemicals years after the 1995 attack [39]. With its sufficient resources, followers and motivations, Aum still poses a threat. They could prove even more threatening if they could enlist help from a State or if they could illicitly purchase needed technologies to support their goal of a successful mass attack. International intelligence and law enforcement must continue to carefully watch the cult that will not go away.

In the fall of 2001, letters containing a sophisticated and lethal form of powdered anthrax were sent to news media outlets and two democratic senators (the letters to the two senators were more highly refined and therefore more deadly). Of the eleven victims of inhalational anthrax, six survived. Eleven people also came down with cutaneous anthrax. Thousands of potentially exposed individuals were prescribed the antibiotic Cipro. The perpetrator is still unknown. This attack demonstrated that an individual could create highly refined anthrax spores, which, if disseminated properly, could infect hundreds, thousands or more. What is less clear is whether the perpetrator or any other terrorist could produce larger amounts (kgs) of anthrax and efficiently disseminate the spores over a wide area. According to the nuclear threat initiative, "Producing

kilograms of dried anthrax, which would be required for a mass-casualty attack against an urban target, would entail much greater technical difficulties and hazards."

Finally in 1984, a cult contaminated the salad bars at several restaurants in the Dalles, Oregon with the non-lethal bacterium Salmonella typhimurium. The cult's leaders used the event as a drill for sickening townspeople to prevent voter participation in an upcoming election. The cult's planner was an experienced nurse and microbiologist. Ma Anand Puja ordered antibiotic test kits containing salmonella bacteria from a laboratory supply company and used the cultures as seed organisms for the mixture cult members later sprinkled at the salad bars. Though there were public health and law enforcement investigations at the time, it was not until a cult member confessed that law enforcement realized the outbreak was the result of an attack. This illustrates the difficulty of differentiating naturally occurring and man made biological events. It also reinforces the fact that known perpetrators of biological and chemical attacks typically have some scientific training. Finally, it highlights the challenge posed by dual use equipment and materials.

In addition to these three major attacks, several incidences of minor food contamination and exposure to irritating substances make up the bulk of international chemical attacks in the MIPT database. For example, in 1978 several people attending an international Assyrian Congress meeting in Sydney Australia ate food contaminated with mustine hydrochloride. No group claimed responsibility, but Iraqi delegates provided the food to delegates who had criticized the Iraqi Government. As noted above mercury-contaminated fruit was found in several European countries in 1978, and in June 2003, at least seven letters containing the irritant Adamsite (a component of rocket fuel) were distributed across Belgium by an unknown group; in October and November 2003 envelopes containing ricin were intercepted in the mail system in the United States. In January 2003 two journalists who write on terrorism were attacked at a book signing in Greece by tear gas and red paint. In addition to the anthrax attacks in the United States, anthrax was also recovered from a letter sent to the Daily Jang Newspaper and a computer company both in Karachi in October 2001. No one was injured [41] and [4]. In December 2001, police vans in the Basques region of Spain were attacked with acid and Molotov cocktails — two were injured and no one claimed responsibility. In November 2001, tear gas was used to target a man in Bishkek Kyrgyzstan. There were no deaths and few injuries in any of these incidents, further bolstering the fact that motivational and technical challenges limit the destructive power of unconventional weapons.

6.2 Development and/or Attempted Use of CBRN

The number of unsuccessful attacks or even attempts at development or acquisition of CBRN far outweighs the actual use of these weapons. The dread and fear inspired by these weapons is even more unbalanced. The case of Aum Shinrikyo described above illustrates the difficulties encountered by a sophisticated terrorist organization trying to develop and deploy chemical and biological weapons. Aum also tried unsuccessfully to buy nuclear material weapons in Russia, even though it had approximately 30,000 members in the country at the time [Bunn 2005]. There is insufficient space to discuss all of the failed attempts in detail, but suffice it to say that al Qaeda and other relatively sophisticated groups, like Aum Shinrikyo, have, according to open sources, been unable to acquire the capability to use CBRN to mass effect, even though they continue to try.

There are a number of reasons for the absence of CBRN attacks including the technical and material challenges. In addition, while al Qaeda is set to destroy the West, few other groups have the motivation to kill large numbers of people. Other factors include: terrorists prefer the certainty of conventional weapons to the uncertainty of CBRN; the weapons can be hazardous to the terrorists themselves; the response to a CBRN terrorist attack may result in further degradation of terrorist capabilities; and finally political support of the terrorist organization's base may be turned away by the use of unconventional tactics.

While there have been few successful or large-scale CBRN attacks, experts clearly believe that attacks will be more sophisticated and occur more frequently in the future. Because the threat is difficult to predict policy makers have made tremendous investments in response and recovery efforts. One small part of this effort has been an investment in the science, technology and role of sensors.

7. Preventions Efforts — The Role of Sensors

What does all of this threat information mean for the design and deployment of sensors? Because of the infinite target spectrum described above, it is not only high value, highly secured 'targets' that must be monitored, but also softer, more common targets. Sensors must be able to find the proverbial 'needle in a haystack.' Because of the wide and varied threats, sensors would ideally be multifunctional, robust, low cost, accurate, reliable, used with little training, able to remotely discern signals in a high background environment, and would provide definitive information to decision makers and require little special care such as

refrigeration or power. According to the U.S. National Science Foundation, "It is essential to be able to accurately identify and measure in real time a wide range of chemical and biological agents, at levels much lower than toxic, in vapor and on surfaces, preferably from a distant position" [43].

7.1 What Are We Trying to Detect?

The U.S. spends an estimated $3.2 billion on research and development for combating terrorism, and John Marburger, the director of the Office of Science and Technology policy, noted that "A major role for technologies in combating terrorism is the detection of chemical, biological, radiological, nuclear, or conventional weapons of mass destruction [38]. Although the first part of this chapter is devoted to understanding the threat posed by terrorists, research and development of sensors for unconventional weapons, at least in the United States, has been more focused on worst-case scenarios than on the skills and motivations of the terrorists. As a result, many of the available sensors and sensors under development are designed to detect a specific subset of weaponized CBRN agents and not the non-lethal or unknown agents that may also be encountered. That being said, a short description of the high threat agents and materials that are the focus of United States government sensor research and development follows.

Biological Agents. Several U.S. Departments including Defense, Health and Human Services, Homeland Security, and Energy have been providing funding for biological sensors. The funding is directed to high priority threat agents as defined by the Centers for Disease Control and Prevention [15]. The threats are based on the ability to cause harm rather than demonstrated terrorist potential and are divided into Class A (high threat) and Class B (medium threat) biological agents. According to the CDC, Class A agents can be easily disseminated or are highly contagious, have high mortality rates, may cause public panic, and require special training and preparation. Class B agents are moderately easy to disseminate, have moderate or low morbidity, and require enhanced attention by CDC. Class A agents including Bacillus anthracis, the causative agent for anthrax; Clostridum botulinum toxin, which causes botulism; Yersinia pestis, the agent that causes plague; Variola major, which causes smallpox; Francisella tularensis, which causes tularemia; and Filo and arena viruses, which cause hemorrhagic fevers. Category B agents cause less serious disease and include food and water safety threats and Brucella species that cause brucellosis among others.

Chemical Agents. Sensors for chemical agents have focused mostly on known military chemical agents, which fall under six broad categories: blister agents, such as mustard, phosgene and lewisite; blood agents, such as cyanide; choking agents, such as phosgene and chlorine; incapacitating agents; nerve agents, such as Sarin and Soman; and riot control agents.

The U.S. government is also focused on the risk posed by attacks on industrial chemical facilities [53]. According to Massachusetts's representative, Edward Markey, "Chemical facilities are at the top of the terrorists' target list" [14]. However, because attacks on these facilities are more likely to result in a known release of a defined chemical entity, sensors are less important than situations where either the chemical release goes undetected or where an unknown substance is released.

Radiological Isotopes. Radioisotopes are in widespread daily use. Sources include the military, medical, industrial and academic communities. Until recently, radioisotopes were not strictly controlled. In the United States alone, there are approximately 22,000 licenses maintained by the Nuclear Regulatory Commission (NRC) and individual states through a special agreement with the NRC [60]. While a so-called dirty bomb is unlikely to cause significantly more casualties than a large bomb alone, policy makers are concerned with the public reaction following such an event. The primary contaminants are alpha and gamma emitters. As discussed above, national security experts deem a 'dirty bomb' as the most likely unconventional weapon over the next decade. As a result, effective detectors for the isotopes discussed below may be critical in alerting officials before an attack occurs or reducing health effects after an attack. For several isotopes, removing clothing after exposure can reduce the hazard by 90 percent.

Common radioactive material in use today includes: the alpha emitters Americium-241 and Plutonim-238; the beta emitters Phosporus-32 and Strontium-90; and the gamma emitters Cesium-137, Cobalt-60, and Iridium-192 [44]. These materials are commonly used in smoke detectors, oil exploration, industrial gauges, food and mail irradiation, cancer therapy, industrial radiography, and in research laboratories.

Nuclear Materials. The United States has deployed sensors both nationwide and overseas for the detection of nuclear materials. Although the presence of highly enriched uranium (an indication of a functional or potential nuclear weapon) would present the greatest threat, currently deployed sensors are unable to detect this material because of its low radioactivity. The Department of Homeland Security alone spent more

than $100 million in fiscal year 2004 to develop improved sensors for nuclear and radiological materials [17].

While nuclear materials are easier to track than CBR agents because of the complex facilities required to produce them, danger exists because several countries are reportedly diverting enriched materials from nuclear power plants, and because large stockpiles of fissile materials are not always sufficiently guarded. Terrorists have been unable to harness nuclear materials; however, between 1993 and 2004, according to the International Atomic Energy Agency, there were 650 documented instances of illegal transfers of nuclear and radiological materials [3]. In 2005, Russia's defense ministry reportedly prevented two terrorist attempts to infiltrate nuclear weapons sites [5]. The United States is working with Russia to increase the security of all nuclear stockpiles, yet much remains to be done [42]. Improved sensors are needed for fissile materials such as plutonium-239 and uranium-235, fussionable materials such as deuterium and tritium, and source materials such as tritium, polonium, beryllium, lithium-6 and helium-3.

7.2 Fielded Sensor Capabilities

Operational sensors are most effective at detecting substances in order to improve response — after the fact. Most detectors are not yet capable of providing warning. Once the presence of a potential hazard is detected by deployed sensors, more sophisticated instruments may be used — often off site — to further characterize the threat. Current sensors for chemical and biological agents use techniques such as ion mobility spectrometry, gas chromatography/mass spectrometry, blackbody infrared spectrometry, antibody kits, UV-induced fluorescence, and surface acoustic wave sensors [43]. Light detection and ranging systems (LIDAR) are being developed for remote detection. Some of these strategies are based on older technology due to the time it takes to fully field research. As a result they are often logistically difficult because they are large, expensive and require consumables as well as electricity and training [33].

Recognizing that nuclear materials are widely available and the terrorists' interests in radiological and nuclear devices, the United States Congress appropriated $300 million to the Department of Homeland Security to install radiation detectors at U.S. borders. Through 2005, DHS had installed 470 radiation portal monitors throughout the country including mail facilities and land and sea entries into the United States. The U.S. has also supported the installation of detectors at the borders of the states of the former Soviet Union through its Departments of State,

Energy and Defense. The General Accountability Office of the United States reported that currently deployed radiation detection equipment cannot sufficiently detect nuclear materials when they are shielded by lead or other metals, and that the equipment is least capable of detecting highly enriched uranium (HEU) because of the low relative radioactivity noted above. The detectors were also limited by the manner in which they were used [9]. For instance, to limit the number of false alarms from materials such as kitty litter and ceramics, border agents reportedly lowered the threshold sensitivity. The inspectors also allowed trucks to pass through monitors at rates of speed too high to efficiently detect radiation. The detectors were also adversely affected by environmental conditions such as wind, moisture and cold [3]. These deficiencies point to the need to improve sensor design.

The New York Times was even more unforgiving, "The federal government's efforts to prevent terrorist from smuggling a nuclear weapon into the United States are so poorly managed and reliant on ineffective equipment that the nation remains extremely vulnerable to a catastrophic attack" [36]. The newspaper reported that detectors at the Port Authority of New York and New Jersey suffer as many as 150 false alarms per day from 22 monitors, more than an order of magnitude greater than the predicted rate. Newly developed sensors must work quickly enough to facilitate the flow of goods and services across borders, and they must be both selective for and sensitive to low levels of radiation from materials of concern without a high rate of false alarms.

Biological. Biological detection is complicated by the fact that there are thousands of pathogens that might be used as biological weapons, and the means to detect microorganisms are often species specific. Current detection systems can be divided into three categories: environmental, hand-held mobile, and surveillance. All of the commercially available sensors are 'detect to respond' rather than 'detect to prevent' or warn. Environmental monitoring is generally defined as continuous or semi-continuous sampling of the environment in a fixed place. The U.S. Biowatch system is an environmental monitoring system dispersed nationwide in urban centers. It is designed to detect a biological event in 36 hours by filtering air at known time intervals, storing the samples, and amplifying the samples with polymerase chain reaction (PCR) twice if something is detected. Fluorescent-labeled probes for specific agents are introduced during PCR to allow detection of known threat agents. Some of the biggest challenges include understanding background concentrations of the agents being analyzed and sampling in a variety of different environmental backgrounds. Once a biohazard is detected, it

is sent to an approved laboratory for confirmatory testing. This type of system focuses on detecting known threats. Most mobile detectors are also pathogen specific. They are reduced in size and weight from the environmental samplers. Syndromic surveillance is also being used to detect attacks with biological pathogens. This involves the large-scale collection of health-related data that precede diagnosis but indicate the presence of an outbreak. (CDC, 2005)

Chemical. Exquisitely sensitive chemical agent sensors are available, but work best under laboratory conditions. Environmental chemical sensors suffer many of the same issues as biological detectors. They lack sensitivity, are not sufficiently mobile or flexible, and require trained users. Several types of chemical detectors are in use and are mentioned above.

Radiation. Radiation portal monitors have been in use for 20 years at U.S. nuclear facilities and are being used as part of the Second Line of Defense (SLD) program at Russian borders. At Los Angeles and Oakland ports, every container that is unloaded from a ship is screened before it leaves for its terrestrial destination and at other U.S. ports a portion of cargo is screened [7]. However, like the GAO, the National Institute of Standards and Technology in the United States recently evaluated 31 commercially available radiation detectors and found that most detectors could accurately measure gamma rays but not low energy x-rays [45]. Most current detectors were originally designed to be used under controlled conditions and not to detect terrorist events, where the instruments must be more flexible and detect a wider array of particles.

8. Improving Sensors

The research described in the remainder of this volume may advance the sensing of several of the materials listed above. To make radiation detectors useful for the detection of radiological materials and weapons, the instruments must be able to detect unknown types of radiation quickly, over a wide range of energies without delicate calibrations, and in many environments. In general, to improve sensor technology, sensors should address a wide variety of agents, be inexpensive, require little training to use and understand, be both accurate and reliable, be capable of withstanding extreme environments, require little or no power or reagents, be capable of remote detection and identification, and be able to discern signals in a high background environment regardless of environmental media.

9. Conclusions

While much of this chapter is focused on extreme Islamist terrorism, it should be emphasized that only a tiny fraction of the world's 1.44 billion Muslims support terrorism. Terrorism is a mindset and a tactic of extremes, either right or left, ethno nationalist, or religious. There has been much progress in the war on terrorism, but as demonstrated by recent attacks, we must remain vigilant for many reasons including:

1 Some of the most skilled and resolute terrorists remain at large including Osama Bin Laden, his deputy Ayman al-Zawahiri and Abu Musab al-Zarqawi,

2 Al Qaeda is resilient and has morphed from a more hierarchical group into a distributed organization, which will be even more difficult to defend against,

3 The war in Iraq has energized al Qaeda affiliates and other Islamic fundamentalist groups to fight the United States and other members of the coalition,

4 Regional organizations have also been impacted by the war on terrorism, but remain serious threats,

5 Cross-fertilization is increasing,

6 The spread of technology progresses onward, and it can be adapted for terrorist purposes.

Given the successes of the war on terrorism and the caveats listed above, research on sensors should address near term threats such as metals in weapons, explosives and improvised explosive devices (IEDs), and suicide packs, while continuing to address the longer-term threats of CBRN.

References

[1] Adnki.com, Indonesia: Jakarta Hotels on Full Alert, June 3, 2005, available at www.adnki.com/index_2level.php?cat+Terrorism&loid=8.0.173629318&par=0

[2] Aljazeera.net, "Full Transcript of bin Laden's Speech, " November 1, 2004, available at English.aljazeera.net.

[3] Aloise, Gene, "Combating Nuclear Smuggling: Efforts to Deploy Radiation Detection Equipment in the United States and Other Countries, June 21, 2005, GAO-05-840T.

[4] "Anthrax cases hit Pakistan," BBC News, November 2, 2001, available at news.Bbc.co.uk.

[5] "Army prevents terrorist attacks on nuclear sites," RBC, June 22, 2005, available at www.rbcnews.com.

[6] Associated Press, "Bin Laden tape: 'Youths of God' plan more attacks," October 7, 2002 accessed at http://www.smh.com.au/articles/2002/10/07/1033538881353.html, as cited in Gilmore 2002).

[7] Associated Press, "L.A. Port Getting Radiation Detectors," June 4, 2005, available at www.msnbc.msn.com/id/8092280.

[8] Associated Press, "Multiple Car Bombs Kill 23 in Baghdad," June 22, 2005, available at www.foxnews.com.

[9] Barrett, Devlin, "False Alarms Plague Port Anti-Nuke System," June 21, 2005, Associated Press, available at www.sfgate.com.

[10] Bolton, John R., Undersecretary of State For Arms Control and International Security, "The International Aspects of Terrorism And Weapons of Mass Destruction, "Second Global Conference On Nuclear, Bio/Chem Terrorism: Mitigation And Response, The Hudson Institute, Washington, DC Friday, November 1, 2002 as Released By The State Department and cited in Gilmore 2002.

[11] Bolton, John, R. "Iran's Continuing Pursuit of Weapons of Mass Destruction," Testimony before the House International Relations Committee Subcommittee on the Middle East and Central Asia, June 24, 2004, available at www.state.gov/t/us/rm/33909.htm.

[12] Bonner, Raymond, "Philippine Camps are training al Qaeda's Allies, Officials Say," New York Times, May 31, 2002 as cited in Gilmore 2002.

[13] Bunn, Matthew with Anthony Weir and Josh Freidman, "The Demand for Black Market Fissile Material," NTI, June 16, 2005, available at www.nti.org/e_Research/cnwm/threat/demand.asp.

[14] CBS News, "Chemical Threats Close to Cities," July 6, 2005, available at www.cbsnews.com/stories/2005/07/06/national/main706788.shtml.

[15] Centers for Disease Control And Prevention, "Syndromic Surveillance: an Applied Approach to Outbreak Detection," June 6, 2005 available at www.cdc.gov/epo/dphsi/syndromic.htm.

[16] A Chronology of Significant Terrorism for 2004, National Counterterrorism Center, United States, available at www. Fas.org/irp/threat/nctc2004.pdf.

[17] Department of Homeland Security, FY 2004 Budget Fact Sheet," October 1, 2003, available at www.dhs.gov/dhspublic/display?content=1817.

[18] DeSutter, Paula A. "U.S. Government Assistance to Libya in the Elimination of Its Weapons of Mass Destruction," Testimony before the Senate Foreign Relations Committee, February 2, 2004, available at www.state.gov/t/vc/rls/rm/2004/29945.htm.

[19] FBIS, "Russian Newspaper on Union of Islamic Movements in Central Asia," Moscow Pravda, September 16, 2002 as cited in Gilmore 2002.

[20] Finn, Peter and Dana Priest, "Weaker al Qaeda Shifts To Smaller-Scale Attacks," The Washington Post, October 15, 2002, http://www.washingtonpost.com/wp-dyn/articles/A25832-2002Oct14.html, as cited in Gilmore 2002.

[21] First Report to the President and Congress, 1999, The Advisory Panel to Assess Domestic Response Capabilities for Terrorism Involving Weapons of Mass Destruction, December 15, 1999, RAND.

[22] Fourth Annual Report to the President and Congress of the Advisory Panel to Assess Domestic Response Capabilities for Terrorism Involving Weapons of Mass Destruction: Implementing the National Strategy, December 15, 2002, RAND.

[23] Freeh, Louis, "1999 Budget Request," testimony before the Senate Appropriations Subcommittee for the Departments of Commerce, Justice, and State, the Judiciary and Related Agencies, March 3, 1998, available at www.fas.org/irp/congress/1998_hr/s980303f.htm.

[24] Gunaratna, Rohan, "The Future of Al Qaeda and the Islamist Terrorist Threat to Southeast Asia and Australia," Australian Security in the 21st Century, delivered at the Parliament House Canberra, May 27, 2003 available at www.mrcltd.org.au/uploaded_documents/thefutureofalqaeda.pdf.

[25] Higgins, Andrew, Karby Leggett, and Alan Cullison, "How al Qaeda put the Internet to use," The Wall Street Journal, November 11, 2002, http://www.msnbc.com/news/833533.asp?0si as cited in Gilmore 2002.

[26] Hinton, Henry L. testimony before the U.S. Senate Committee on Governmental Affairs, 17 October 2001, GAO-02-162T, p. 4 as cited in Gilmore 2002.

[27] Hoffman, Bruce, "Terrorism Trends and Prospects" Chapter Two in Countering the New Terrorism, Ian Lesser et al., 1999, RAND, MR-989-AF.

[28] Hoffman, Bruce, "RE-Thinking Terrorism in Light of a War on Terrorism," testimony before the subcommittee on Terrorism and Homeland Security, House Permanent Select Committee on Intelligence, U.S. House of Representatives, September 26, 2001, available at www.rand.org/publications/CT/CT182?CT182.pdf.

[29] Hoffman, Bruce, "Al Qaeda, Trends in Terrorism And Future Potentialities: An Assessment," 2003, RAND P-8078.

[30] Jarboe, James F., FBI, "The Threat of Eco-Terrorism," testimony before the House Resources Committee, Subcommittee on Forests and Forest Health, February 12, 2002, available at www.fbi.gov/congress/congress02/jarboe021202.htm.

[31] Jenkins, Brian M., "The Future Course of International Terrorism," The Futurist, July–August, 1987, available at www.wfs.org/jenkins.htm.

[32] Jenkins, Brian M. "The Organization Men: Anatomy of a Terrorist Attack," in James F. Hoge, Jr. and Gideon Rose, How Did This Happen? Terrorism and the New War (NY: Public Affairs, 2001).

[33] Kosal, Margaret, "The Basics of Chemical and Biological Weapons Detectors", November 24, 2004, Monterey Institute of internationals Studies, available at cns.miis.edu/pubs/week/031124.htm.

[34] Kyodo News Service, Japan, "Aum Shinrikyo Sets Up More Than 10 Business Entities" April 16, 2004, available at www.religion newsblog.com/6796.

[35] Lake, Eli, "Al Qaeda's Disinformation War," October 30, 2002, The New Republic Online.

[36] . Lipton, Eric, "U.S. Borders Vulnerable, Witnesses Say," June 22, 2005, New York Times.

[37] Lugar, Richard, "The Lugar Survey on Proliferation Threats and Responses," June 2005, United States Senator For Indiana, Chairman Senate Foreign Relations Committee.

[38] Marburger, John, Statement before the subcommittee on emerging threats and capabilities committee on armed services, United States Senate, April 10, 2002, available at armed-services.senate.gov/statement/2002/April/Marburger.pdf.

[39] Marshall, Andrew, "It Gassed the Tokyo Subway, Microwaved Its Enemies and Tortured Its Members. So Why is the Aum Cult Thriving?" The Guardian, July 15, 1999.

[40] The MIPT Terrorism Annual 2002, with contributions from D. Brannan, P. Chalk, K. Cragin, and S. Daly, National Memorial Institute for the Prevention of Terrorism, available at www.mipt.org.

[41] MIPT Terrorism Knowledge Database, available at www.tkb.org.

[42] National Academy of Sciences, Making the Nation Safer: the Role of Science and Technology in Countering Terrorism, 2002.

[43] "The New Challenges of Chemical and Biological Sensing: National Science Foundation Workshop," January 9-10, 2002, Arlington Virginia, available at www.chemistry.gatech.edy/sensing forum-02/welcome.html.

[44] News & Terrorism: Communicating in a Crisis, Fact Sheet from the National Academies and the Department of Homeland Security, Radiological attack: Dirty Bomb and Other Devices, available at nae.edu/NAR/pubundcom.nsf/weblinks/CGOZ-646NVG/$file/radiological%20attack.pdf.

[45] Pappalardo, Joe, "Security Beat," National Defense, July 2005 available at www.nationaldefensemagazine.org/issues/2005/jul/security_Beat.htm.

[46] Patterns of Global Terrorism 2001, Office of the Coordinator for Counterterrorism, U.S. State Department, May, 2002 available at www.state.gov/s/ct/rls/pgtpt/2001.

[47] Patterns of Global Terrorism 2003, Office of the Coordinator for Counterterrorism, U.S. State Department, April 29, 2004 available at www.state.gov/s/ct/rls/pgtpt/2003/31644.htm.

[48] Perl, Raphael, "Terrorism and National Security Trends," CRS Issue Brief for Congress, December 21, 2004, available at www.fas.org/irp/crs/IB10119.pdf.

[49] Raman, B., "The LTTE: The Metamorphosis," Paper no. 448, South Asia Analysis Group, April 29, 2002 available at 222.saag.org/papers5/paper448.html.

[50] Raman B., "The World's First Terrorist Air Force," Observer Research Foundation, available at www.observerindia.com/analsysis/A445.htm.

[51] Shannon, Elaine, "Another warning from Zubaydah," Time, May 11, 2002, http://www.time.com/time/nation/article/0,8599,236992,00.html, as cited in Gilmore, 2002.

[52] Smith, G. Davidson, Combating Terrorism, London, Routledge, 1990, p. 7 as cited in G. Davidson Smith, "Single Issue Terrorism," Commentary No. 74, Canadian Security Intelligence Service, Winter 1998, available at www.csis-scrs.gc.ca/eng/comment/com74_e.html

[53] Stephenson, John B., "Homeland Security: Federal and Industry Efforts are Addressing Security Issues at Chemical Facilities, but Additional Action is Needed," April 27, 2005, GAO-05-631T.

[54] "Summary of Investigation of IRA Links to FAC Narco Terrorists in Colombia," Majority Staff of the U.S. House International Relations Committee, April 24, 2002, available at wwwa.house.gov/international_Relations/107/findings.htm.

[55] Swartz, Jon, "Terrorists' Use of Internet Spreads," USAToday, February 20, 2005, available at www.usatoday.com/money/industries/technology/2005-02-20-cyber-terror-usat_x.htm.

[56] Tenet, George, Director of Central Intelligence, "The Worldwide Threat 2004: Challenges in a Changing Global Context," testimony before the Senate Select Committee on Intelligence, February 24, 2004.

[57] Terrorism: Questions and Answers: North Korea, Council on Foreign Relations, 2004, available at cfrterrorism.org/sponsors/northkorea.html.

[58] Terrorism: Questions and Answers: Syria, Council on Foreign Relations, 2004, available at cfrterrorism.org/sponsors/northkorea.html.

[59] United States District Court, District Court of Massachusetts, United States of America v. Richard Colvin Reid, http://news.findlaw.com/hdocs/docs/terrorism/usreid011602ind.html, as cited in Gilmore 2002.

[60] U.S. Nuclear Regulatory Commission, The Regulation and Use of Isotopes in Today's World, available at www.nrc.gov/reading-rm/doc-collections/nuregs/brochures/br0217/r1/br0217r1.pdf.

[61] Williams, Mike "Analysis: What next for al Qaeda?" November 22, 2001, http://news.bbc.co.uk/1/hi/world/south_asia/1678467.stm, as cited in Gilmore 2002.

[62] Zakis, Jeremy and Steve Macko, "Major Terrorist Plot in Singapore Discovered: al Qaeda Believed well Established in the Asian Region", January 12, 2002 available at www.emergency.com/2002/jamaah_islamiyah.htm as cited in Gilmore 2002.

ADVANCES IN SENSORS; THE LESSONS FROM NEUROSCIENCES

M. Costa
Department of Physiology and Centre of Neuroscience
School of Medicine
Flinders University
Adelaide 5006
Australia

Abstract This is a short review of how neuronal sensors fit in the broader biological context of animal survival. This may help those involved in the development of engineered sensors to put in perspective their task with what the evolutionary process has achieved. Most of the information reported here is available in the educational field of neuroscience, with mention of some recent relevant findings. I have attempted to place these findings in an evolutionary perspective as it clarifies better the intrinsic role of some of the extraordinary particularities of the biological solutions of neuronal sensors.

Keywords: action potential; axon; ion channel; mechanoreceptor; membrane potential; neurons; neurotransmitter; olfaction; proteins; synaptic gap.

1. Energies that affect earth living organisms survival

Organisms are exposed on earth to an environment with different physical energies, including gravity, and in general kinetic energy, chemicals, temperature variations, sound and noise, and electromagnetic waves. Evolutionary processes, not too surprisingly, have endowed living multicellular organisms with sensors specific for these physical energies within some selective ranges, mainly those best represented on the earth's surface. Such sensors are intrinsically linked with the ability of organisms to behave in an appropriate way, 'taking in account' the sensed environment. This implies the ability to comply with essential functions of all living systems. These include escape from harm, feeding, breathing and drinking, reproducing and exploring. Escape from harm is the fundamental capacity that ensures the survival of each individual organism

Jim Byrnes and Gerald Ostheimer (eds.),
Advances in Sensing with Security Applications, 55–67.

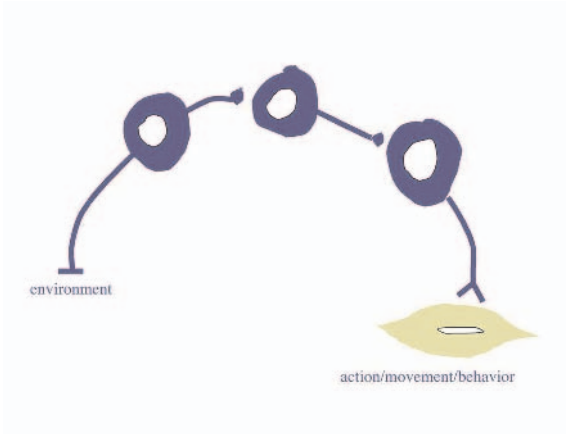


Figure 1. Rudimentary nervous system.

and leads, in humans and most vertebrates to the simple withdrawal from painful stimuli. Feeding is required as food is the fuel needed to be burned to provide the energy, which is stored as chemical energy by living beings. Breathing guaranties the intake of oxygen for the oxidation (burning) of the food fuel. Most living chemical reactions occur in water, an essential component of life on earth. Individuals disappear (die) and are replaced by other derived individuals similar, but not identical, and this enables 'species' to change adaptively by the evolutionary process of natural (Darwinian) selection. A great deal of animal behaviour is dedicated to the reproductive function. Finally living organisms show spontaneous motion (exploratory locomotion). This gives the advantage of most animals versus plants, to find suitable new environments rather that adapting to an adverse local one. Human migration in prehistory from Africa testifies to the importance of this function for the success of human kind on earth.

2. The emergence of a nervous system

The most significant occurrence in evolution of multicellular organisms is the differentiation of cells that can initiate and carry signals at large distances within the organism. Such cells are called nerve cells or neurons. The main property of such cells is the ability to sense and amplify very small signals to initiate macroscopic behaviour of the entire organism. The diagram in Figure 1 shows the minimal organization of a nervous system with a sensory neuron that responds to some physical

energy, a motor neuron that acts on a movement cell (muscle cell) and in between an 'inter-neuron'.

As multicellular organisms became more complex in evolution, including humans, the relative number of interneurons increased. Most of the human brain and spinal cord is made of interneurons. Although the diagram is highly simplified it implies that each of the three classes of neurons will have quite different functions. The sensory neurons, also called primary afferent neurons, are the link between the organisms as a whole and the environment (including its own internal environment). The sensory neurons detect stimuli in a very localized fashion and at a particular, albeit ongoing, time. Yet the organism needs to detect changes in the environment over larger areas surrounding the organism and over longer time periods. This involves the detection of what, where and when something is 'out there'. The 'what' requires neural processes of construction of some 'percept' involving many neurons. The 'where' involves construction of neuronal maps of some sensed information (visual, acoustic, body etc. maps). The 'when' involves the inaccurate process of memory and prediction that enable past events to affect behavior and future events to be predicted to some degree. Different time range memory systems are known in higher vertebrates and these include working, short and long-term memory. Since the essential elements of the nervous system are the nerve cells, a short description of the unique properties of neurons follows.

3. Neurons as excitable cells

Neurons are specialised cells with a typical shape characterised by a cell body, in the range of several tens of micrometres, and long cytoplasmic processes, nerve fibres either called axons or dendrites. The identity of nerve cells as being formed by cell body and fibres has been recognised since the first half of the nineteenth century, when a young student, Deiter, dissected a large nerve cell from one of the collections of nerve cells (nuclei) in the brainstem. Towards the end of the nineteenth century Camillo Golgi, an Italian neurologist, developed a stain that highlighted extremely well only a few nerve cells from a brain packed with myriad nerve cells. This staining method enabled Ramon y Cajal, a Spanish histologist, to confirm his theory that the nervous system is made of individual nerve cells (neuron theory). Neurons, as most other cells in multicellular organisms, share the machinery for life, including burning fuel, synthesis of its own components (e.g. proteins) and excretion of residues. In particular neurons have very little store of fuel to burn and thus need a continuous supply of fuel and oxygen. In addi-

tion, neurons and muscle cells share the characteristic of 'excitability'. Excitability is a term that implies responsiveness of a cell to a small signal leading to a disproportionately large response. Neurons are in an electrical state distant from equilibrium. In a typical neuron the charges on the internal cytoplasmic fluid are more numerous than on the external liquid, producing a difference in potential of about -65mv. This means that they have an electrical polarity. The transient sudden loss of such polarity is what has been named action potential, nerve impulse or spike. In order to understand how this loss of polarity occurs we must first understand how the electrical polarity is generated.

Neurons, like all other cells, are surrounded by a cell membrane made of a phospholipid bilayer, which is self-assembled in an aqueous environment. Within this compartmentalised 'bag', the cytoplasm with organelles, performs the specific cell functions. The electrical polarity of the cell is due to the presence in the membrane of a large macromolecule that actively separate the ions Na+ and K+, the sodium-potassium pump. This molecule is a protein. All proteins are formed by single strands of sequences of the basic elements, aminoacids. These are small molecules made in the cells or imported from the outside, characterised by an amino group at one end and by a carboxylic group at the other. These endings make aminoacids capable of binding together in strings (linked by peptidic bonds). Such strings form peptides and proteins. The single strand is usually bent to form spirals, which then form rods and globules. The final shape of a protein thus depends on the 'tertiary' structure. The sodium-potassium pump is a protein with a pore that can open and close inside or outside the neuron membrane, transporting sodium ions out and potassium ions in the neuron. This pump ensures that there is little sodium inside the cell and more potassium than outside. Since there are fixed negative charges anchored in the molecules inside the neuron, this pump maintains the electrical polarity of the nerve cell. Thus this polarity keeps the cell membrane at 'rest' far from its equilibrium. If the ions were able to move freely across pores in the membrane they would rush across along what is called the concentration gradient until the rush is counteracted and balanced by the charges reaching their 'equilibrium potential'. Indeed there are special proteins in the membrane, which form pores, normally closed or almost closed and that are specific for different ions. The main determinants of the neuronal activity are channels for sodium and for potassium, which have a remarkable property of opening in response to a change in voltage off the membrane. When the membrane becomes less negative (it is said then to depolarise), these channels open with a positive feedback process. This non-linearity ensures that past a 'threshold' value of

the membrane potential, the sodium channel opens in an explosive-like manner, enabling sodium ions to rush inside the cell along its concentration gradient until it reaches its equilibrium potential (around 40 mv positive). This process lasts only about one millisecond before stopping spontaneously, and is followed by the opening of the potassium channel, which is also voltage dependent. The potassium ions rush out, removing the excess of positive charges from the inside, thus re-establishing the electrical negative polarity. This transient change in polarity of the neuronal membrane (depolarisation followed by repolarisation) is the event called action potential. Other ions such as chloride and calcium also play minor roles in this process. This explosive, self amplifying, all-or-none event, is the way in which nerve cells carry signals. The entire language of the nervous system is coded by action potentials.

Understanding of the structure and function of these channel molecules has developed enormously in recent years thanks to molecular biology and the technique of patch-clamping, which enables one to detect not just the electrical events in the entire neuron, with a microelectrode inserted in the nerve cell, but the electrical events in small patches of neuronal membranes, sucked by a glass pipette, containing a single channel molecule. The action potential sets currents that affect the neighbouring patch of membrane, which then becomes locally depolarised. This 'passive' depolarisation of adjacent patches of membrane triggers there the generation of a new action potential. The sequential activation of local action potentials results in the 'conduction' of action potential along the entire length of the neuronal membrane. As the neuron has very long processes (axon), this signal travels from its origin to the end. This is the propagation of nerve impulses within the nervous system. The speed of conduction depends on the diameter of the axon, the larger the greater the speed. As greater speeds are advantageous for fast responses, in evolution most axons have become covered by fatty sheets secreted by satellite cells (glial cells), to insulate long stretches of the axons compelling action potentials to jump along the axon, thus increasing its speed of conduction. The speed ranges from a few cm/s up to 120m/s. These electrical signals can also jump from one neuron to a next neuron in the circuit. This is called electrical transmission. This mode of transmitting neuronal impulses is the exception rather than the rule in the nervous system. The most common mode of transmission is via a chemical, which is released at the end of a neuron, acting then on the next neuron. This 'chemical transmission' is mediated by several types of small molecules, thus named neurotransmitters. Amongst these are amines such as acetylcholine, noradrenaline, adrenaline, dopamine and serotonin. Also aminoacids such as glutamate, aspartate, glycine and

GABA are widely used transmitters. Also ATP, the molecule usually associated with chemical energy in cells, and nitric oxide act as transmitters. Several small neuropeptides, short sequences of amino acids, act as transmitters.

The process of chemical transmission involves the arrival of an action potential at the nerve ending. This triggers the entry of calcium ions in the cells via specific protein channels. Calcium triggers an amplifying cascade of chemical reactions between macromolecules and small vescicles (bags) containing several thousands of neurotransmitter molecules. This results in the opening of these vescicles (synaptic vescicles) to the outside of the neuron. When the vescicles open and release the transmitter, the distance to the next neuron is very small (on the order of tens of nanometers). This gap is called synaptic gap and the 'synapse' is the ensemble of the 'pre-synaptic' membrane, the gap and the 'post-synaptic' membrane. Transmission due to the diffusion of the neurotransmitter is very fast, being measured in microseconds. The postsynaptic membrane (i.e. the membrane of the next neuron) contains special proteins that selectively bind the neurotransmitter molecules. These proteins are called 'transmitter receptors' and form an enormous variety of molecular families adapted to the tens of different transmitters as keys to their locks. The binding of transmitters with their receptor triggers an amplifying cascade of events inside the neuron that leads to opening or closing of ion channels, thus changing the electrical state of the membrane. If the transmission leads to a depolarisation, and consequently to the triggering of an action potential, it is said to be excitatory. Transmission can also be inhibitory if the result of transmitter action is the increase in polarity of the membrane (hyperpolarisation) taking the neuron further away from its threshold of firing action potentials. Indeed much of the control of neuronal operation in the nervous system involves inhibitory transmission, without which the nervous system would run wildly uncontrolled. The entire armamentarium of pharmacological drugs used for pleasure, such as morphine and heroin, cannabis, ectasy, psychedelic drugs, most poisons (e.g. sarin), some pesticides and most therapeutic drugs that affect organs (e.g. high blood pressure), human behavior and mental diseases (e.g. schizophrenia, depression, Parkinson), all act on some of the processes involving chemical transmission. The function of nerve cells and the neural circuits they form is thus determined by a number of highly non-linear amplification processes that ensure that very small signals are amplified to become macroscopic events.

4. Sensory neurons

The first neuron in the simplified neural circuit in Figure 1 is the sensory neuron. The sensory neurons share all the properties of other neurons such as excitability and transmission to other neurons via chemical transmitters. They are activated from their resting state by events in the environment. Different classes of sensory neurons then exist to detect the physical energies mentioned above. These are the true 'biological sensors'. Of course all living cells are capable of responding to changes in the environment to some degree and thus are also 'sensors'. The activation of sensory neurons can be direct or may involve the activation of an intermediate, non-neuronal cell (sensory cell). The following list describes the major classes of sensory neuronal receptors according to the physical energy they sense.

Classification of neuronal sensory receptors

- Mechanoreceptors
  - kinetics muscle, joints, pressure on body position and movements of body parts (sense of self; proprioception)
  - internal sensors for tensions and pressures (heart, blood vessels, gut, bladder, lungs etc.)
  - hearing and balance
- Chemoreceptors
  - feeding
    - olfaction (olfactory lobe)
    - taste
    - internal sensors for nutrients
  - body homeostasis (control of sugars, blood O2 and CO2 etc.)
  - defense (tissue damage)
  - reproduction
    - olfaction (pheromones and neuroethology)
- Photoreceptors
  - vision; photons receptors

5. Sensory transduction

The process by which a small change in the environment is sensed and amplified to become neural activity is called sensory transduction. Since different sensory neurons respond to different stimuli (small changes in the environment), neurophysiologists predicted the presence of special molecules on the surface of the nerve endings of the sensory neurons or of the sensory cells, specifically adapted to respond to different stimuli. Each of these molecules should be able to translate the stimulus in a depolarisation of the sensory neuron that would trigger action potentials, thus initiating sensory neural activity. This is achieved by the opening of ion channels by the stimulus, which leads to the graded depolarisation of the sensory ending.

The prototype of sensory neuron studied earlier because of its accessibility, is a mechanosensitive sensory neuron with nerve endings in the skin, surrounded by a small capsule forming an onion-like passive structure, named after its discoverer 'Pacinian corpuscle'. Controlled graded deformation of this corpuscle results in a graded depolarisation of the nerve ending (receptor potential). When the depolarisation reaches a threshold this will trigger action potentials, which will be conducted towards the central nervous system. There is a very good range of stimuli within which the receptor potential and the resulting frequency of firing of action potentials are linearly correlated. The stimulus-action potential frequency response curve is in fact sigmoid, as are most biological transfer functions.

The excellent relation between amount of membrane depolarisation and frequency of firing of action potentials has been demonstrated in most neuron types. Thus frequency of firing of action potentials is probably the most important determinant in coding intensity of sensory signals (stimuli). In addition to frequency of firing different neurons, including sensory neurons, possess additional ion channels that either lead to spontaneous depolarisations or to oscillations that lead to specific windows in which action potentials can occur, resulting in patterns of firing in bursts. The pattern of firing of action potentials is the other major, less well understood, way to encode neural information.

6. Molecules of sensory transduction

The molecular nature of the neuronal receptors is now becoming understood with the advent of molecular biological techniques. The molecular structure of the mechanosensitive channels has been established only recently. In principle mechanosensitive channels must be opened by mechanical deformation of the neural membrane in which they are em-

bedded, either by the stretched membrane pulling on the molecule or via connections with the intracellular 'skeleton' of the neuron itself. One of the simplest models studied is the mechanosensitive channels in bacteria [11]. There is a plethora of families of molecules with mechanosensitivity including: DEG/ENaC family, TRP families, MSC families, some K+channels etc., and these are present in the entire spectrum of living species. One of the recent findings on mechanosenstive channels relates to the hearing system.

7. Hearing system and mechanosensation

The hearing apparatus in terrestrial animals transforms sound waves into nerve impulses. This is performed by a special set of sensory receptor cells (hair cells) located in the organ of Corti in the inner ear. These cells in turn transmit the message to the primary sensory neurons for hearing (an example of indirect activation of sensory neurons). The hair cells are so named because they have a number of small protrusions-like cilia or hair that are bent by the motion imparted by the sound waves that vibrate the eardrum membrane, with amplification of such vibration by an ossicular chain of levers to impart a similar vibration, via an internal liquid, to a thin membrane on which the hair cells are sitting. The bending of the hairs results in depolarisation of the hair cells, which then transmit a chemical signal to the sensory neurons, modifying their neural activity (frequency of firing). The molecular mechanisms by which the hairs, or cilia, open the mechanosensitive channels and their location have been elucidated [3]. The mechanosensitive molecule belongs to the class of the TRP (Transient Receptor Potential) channels previously identified as the gene product defective in a blind Drosophila mutant (see [8]). Molecules with similar structure (homologs) have now been found in many animals, both invertebrate and vertebrate. The TRP superfamily consists of 28 mammalian members as well as 13 Drosophila proteins [4]. These channel protein molecules have a molecular architecture similar to that of voltage-gated ion channels, with four subunits arranged to form a channel. TRPs are generally non-selective channels for positive ions. TRP channels are conspicuously involved in the mechanosensory sensory function. However they are also involved in many other sensory functions including vision, taste, olfaction, pheromone sensitivity, osmosensitivity, nociception and thermosensation. This is surprising and points to a fundamental similarity between molecular mechanisms unrelated to the classic distinction of the sensory 'modalities'.

Some of the TRPs are activated directly by sensory stimuli, but others are activated by a variety of intracellular chemical messengers [12] Initial

studies demonstrate that these sensor molecules can be integrated with microchips for potentially enormous artificial uses [9].

8. Temperature receptors

In mammals, four TRPVs (members of the vanilloid subfamily of TRP channels) are activated at distinct heat thresholds (33–52 ^{o}C), whereas TRPM8 (of the melastatin subfamily) and ANKTM1 are activated at cold (17–25 ^{o}C) temperatures (see [12]). A big surprise in this field is that these molecules were found to also respond also to a variety of well-known natural chemicals from the external environment. These include hot chilli (capsicum sativum, capsaicin) receptor channels, found by Caterina et al [2] to be also sensitive to heath. Other natural substances that activate TRP receptors include cold mint (menthol, eucaliptol, anisette), piperine (black pepper), resiniferatoxin (euphorbia resinifera-cactus), camphor (cinnamonum camphora-laurel), isothiacyanates (mustard, wasabi, horseradish), cinnamaldehyde (cinnamon oil), THC (marijuana; cannabis sativa), allicin (garlic; allium sativum) [12]. Naturally the burning pain caused by capsaicin directly links temperature receptors with pain receptors.

9. Pain receptors

Pain is one of the most mysterious of the senses and certainly one, which although essential for survival, receives little sympathy by the 'users'. Noxious (harmful) stimuli to the skin are known to elicit two kind of painful sensation (see [7]). One occurs earlier (fast pain) and is due to direct activation of pain fibres (depolarisation of axons) by excessive mechanical activation and involves channels of the ENaC family. The second pain is felt with some delay (slow pain) and is initiated by acid (hydrogen ions) acting on the receptor channel TRP-VR1 and the ASICs. The pungent pain produced by capsaicin is also due to its action on the receptor channel TRP-VR1. Noxious heath acts via the TRP-VR1 and the VR1-L. When tissue surrounding the pain nerve ending is injured, a soup of chemicals is produced and released locally. Many of these released substances are able to initiate or positively modulate the activation of pain fibres. Examples of such substances are histamine, serotonin (5HT), ATP, bradykinin, prostaglandins, PARs, neuropeptide such as tachykinins, bradykinin, CGRP etc., and different nerve growth factors. It is still uncertain whether there is also a formation of endogenous capsaicin-like substances (endogenous vanilloids). Conversely there are also substances capable of reducing the excitability of the pain nerve ending. Amongst these are the cannabinoids (endogenous molecules such

as anandamide or exogenous ones like THC, the active principle of marijuana), opioids (endogenous molecules like enkephalins, endorphins and dynorphins, or exogenous ones like heroin and its derivative morphine). Thus peripheral receptors for pain abound and modulating mechanisms render the field most complex for suitable control of this defense function. Chemical sensors par excellence are the olfactory and taste receptors.

10. Olfaction

Olfaction, once thought to be a primitive sense, is now recognized as an elaborate sensory system that deploys a large family of odorant receptors to analyse the chemical environment. Interactions between these receptors and their diverse natural binding molecules (ligands) translate the world of odors into a neural code. Humans have about 350 odorant receptors. Rodents have more than a thousand.

In vertebrates the neurons for olfaction are located in the nose mucosa and consist of short neurons with a peripheral ending endowed with odorant receptors for a large number of molecules in the environment. Each receptor neuron only contains one odorant receptor and is connected directly with the olfactory lobe of the brain. The vertebrate olfactory system must cope with a staggering developmental problem: how to connect millions of olfactory neurons expressing different odorant receptors to appropriate targets in the brain.

The story of search for odorant receptors has come of age in recent years with the awarding of the Nobel price in Medicine in 2004 to two investigators — Dr. Richard Axel and Dr. Linda S. Buck — for their discovery in the early 1990s of the genes that code for odorant receptors in the rat [1]. The discovery of odorant-receptor genes in the rat provides a missing link between the molecular biology odorant receptors and the physiological properties of sensory neurons. Indeed the odorant receptor also determines the connectivity of the sensory neurons, as shown in experiments in which the receptor has been removed from single olfactory neurons and replaced by different ones [5, 6].

All animals exhibit innate behaviors in response to specific sensory stimuli that are likely to result from the activation of developmentally programmed neural circuits. Even the activation of single classes of olfactory neurons can trigger complex behaviors [10]. The authors observed that Drosophila exhibit robust avoidance to odors released by stressed flies. When stressed, the flies emit an odorant mixture that elicits avoidance in other flies. CO2 is the active component of this mixture. Specific blockade of the activation of a particular odorant receptor

also blocked the response in the other flies. Thus CO2 is the signal molecule for this avoidance behavior.

11. Vision

Vision is undoubtedly the most important of the human senses. The process of transduction of photons into neural activity occurs in the retina by a process with significant similarities to chemical receptors including odorant reception. Photons reach the retina via the optical system of the eye and activate special protein receptors (the visual pigments). These are located on the surface of the receptor cells (non-neuronal) called cones (separate classes containing pigments for red, green and blue wavelengths) or rods (with a broad spectrum pigment of wavelength, thus suitable for black and white night vision). Photons acting on the receptor induce a conformation change of the protein and this initiates a cascade of chemical events inside the receptor cell that leads to a change in its electrical property with chemical signals, in turn acting on the primary sensory neurons in the retina.

12. General view of the sensory systems

Despite the large variety of stimuli the sensory neurons respond to, the mechanisms of activation (transduction) utilize remarkably similar molecular processes. However, their activation is only the first step in the process of constructing and utilizing the sensory information. Neurons work in large groups and, as signals go through subsequent neural synaptic stations, undergo modifications all resulting in a simplification and amplification of relevant features. One of the most common mechanisms is lateral inhibition. Activation of parallel sensory pathways also activates cross-talking neurons, which are inhibitory to their neighbor. The result of such an arrangement is that signals with a small advantage in amplitude become larger at the expense of the surrounding weaker signals. By the time the signal reaches the cortex, the original weak but larger signal becomes amplified. This enables a better discrimination of what happens in the outside world. A similar process of lateral inhibition in the retina ensures a better detection of edges for shape recognition. Sensory neural systems may well be a good lesson for those involved in the important task of developing better sensors for a variety of aims. In parallel to developing better amplification and discrimination of signals, it is equally important to improve the quality of analysis of the information for decision-making that should be based on wise knowledge, avoiding the proverbial badly adaptive 'knee jerk' reactions.

References

[1] L. Buck and R. Axel A novel multigene family may encode odorant receptors: a molecular basis for odor recognition. *Cell* **65**:175–187, 1991.

[2] Caterina et al. The capsaicin receptor: a heat-activated ion channel in the pain pathway. *Nature* **389**, 1997.

[3] D.P. Corey et al. Vertebrate hearing mechanosensory channel. *Nature* **432**:723–730, 2004.

[4] A. Fleig and R. Penner The TRPM ion channel subfamily: molecular, biophysical and functional features. *TIPS* **25** (12), 2004.

[5] R.V. Friedrich Odorant receptors make scents. *Nature* **430**, 2004.

[6] E.A. Hallem et al. *Cell* **117**:965–979, 2004.

[7] D. Julius and A.I. Basbaum. Molecular mechanisms of nociception. *Nature* **413**, 2001.

[8] S-Y. Lin and D P Corey. TRP channels in mechanosensation. *Current Opinion in Neurobiology* **15**:350–357, 2005.

[9] E. Neher Molecular biology meets microelectronics. *Nature Biotechnology* **19**, 2001.

[10] G.S.B Suh et al. A single population of olfactory sensory neurons mediates an innate avoidance behaviour in Drosophila. *Nature Neurosci* **431**, 2002.

[11] S. Sukharev and A. Anishkin. Mechanosensitive channels: what can we learn from 'simple' model systems? *TINS* **27**(6), 2004.

[12] Voels et al. Sensing with TRP channels. *Nature Chemical Biology* **1**, 2005.

CHEMICAL SENSORS AND CHEMICAL SENSOR SYSTEMS: FUNDAMENTALS LIMITATIONS AND NEW TRENDS

Andrea Orsini, Arnaldo D'Amico
University of Roma "Tor Vergata"
Dept. of Electronic Engineer,Via del Politecnico, 1 00133 Roma
andrea.orsini@psm.rm.cnr.it, damico@eln.uniroma2.it

Abstract

Chemical sensors are becoming more and more important in any area where the measurement of concentrations of volatile compounds is relevant for both control and analytical purposes. They have also found many applications in sensor systems called electronic noses and tongues.

This chapter will first consider fundamentals of sensor science including a brief discussion on the main terms encountered in practical applications, such as: sensor, transducer, response curve, differential sensitivity, noise, resolution and drift.

Basic electronic circuits employed in the sensor area will be discussed with a particular emphasis on the noise aspects, which are important for achieving high resolution values in those contexts where measurement of the lowest concentration values of chemicals is the main objective.

All the most relevant transducers such as: MOSFET, CMOS, Surface Plasmon Resonance device, Optical Fibre, ISFET, will be covered in some detail including their intrinsic operating mechanisms and showing their limitation and performance. Shrinking effects of these transducers will also be commented on.

The electronic nose and electronic tongue will be described as systems able to give olfactory and chemical images, respectively, in a variety of applications fields, including medicine, environment, food and agriculture.

Finally some future trends will be outlined in order to predict possible applications derived from today's micro and nanotechnology developments.

Jim Byrnes and Gerald Ostheimer (eds.),
Advances in Sensing with Security Applications, 69–94.

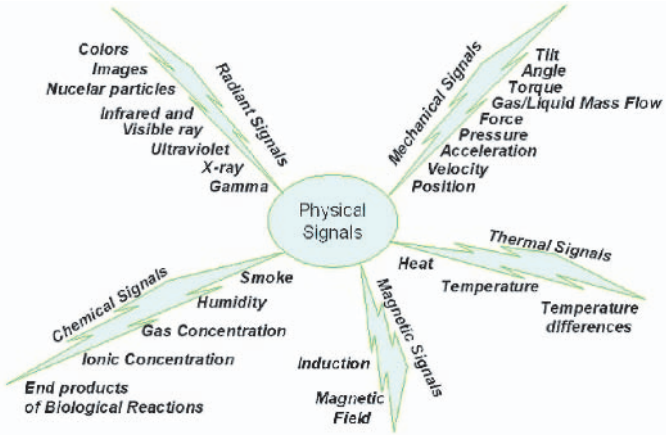


Figure 1. Signal Domains and Parameters.

1. Introduction–Parameters

In the course of the last twenty years, using techniques borrowed from standard silicon technology, silicon sensors became fundamental for the measurement of most physical and chemical parameters. Figure 1 shows the physical domains and the parameters for which silicon sensors have been introduced.

In the case of chemically sensitive devices, the interaction of a given volatile compound or ions in solution can produce one of the following changes: mass, charge, temperature, refractive index, magnetic field, work function. For each of these changes suitable transducers are now available.

Generally speaking sensors are devices able to interface the chemical, physical and biological world with that of electronics and /or electro - optics for processing, storing, communications and data presentation. In the following we introduce the most important sensor parameters and review the most successful chemical sensors able to reveal mass, charge and refractive index variations due to absorption-desorption processes involving volatile compounds.

Response Curve

The response curve (RC) represents the calibrated output response of a sensor as a function of the measurand/s applied to its input. For instance, in the case of a chemical sensor based on conductivity (G), it is recommended to use one of the following notations [1] for the output response:

- G (*conductance*);
- G/G_0 (*relative conductance*);
- $(G - G_0)$ (*conductance change*);
- $(G - G_0)/G_0$ (*relative conductance change*).

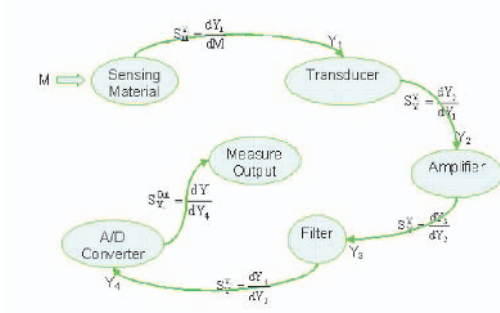


Figure 2. A complex Sensor

In the case of a sensor whose output is a frequency, the representations of the output responses may be as follows:

- f (*frequency*);
- f/f_0 (*relative frequency*);
- $(f - f_0)$ (*frequency change*);
- $(f - f_0)/f_0$ (*relative frequency change*).

It is worth mentioning that for the communication of useful information the operating point of the sensors must always be specified in terms of sensor temperature, electrical or magnetic polarization, and number of fundamental blocks of the sensor model (figure 2).

Sensitivity

Sensitivity (S) is defined as the derivative of the response (as a function of the operating point) with respect to the measurand (M) and, with reference to the four cases above, we have:

- $S = dG / dM$;
- $S = d(G/G_0) / dM$;
- $S = d(G - G_0) / dM$;

- $S = d(G - G_0)/G_0 \ / \ dM$.

In the case of a linear response without offset, the previous sensitivity relationships simplify to:

LINEAR RESPONSE	
Conductivity Based Sensor	Frequency Output Sensors
■ $S = G \ / \ M$; ■ $S = (G/G_0) \ / \ M$; ■ $S = (G - G_0) \ / \ M$; ■ $S = (G - G_0)/G_0 \ / \ M$.	■ $S = f \ / \ M$; ■ $S = (f/f_0) \ / \ M$; ■ $S = (f - f_0) \ / \ M$; ■ $S = (f - f_0)/f_0 \ / \ M$.

In the case of a piecewise linear response, for each of the segments the sensitivities can be simplified as follows:

PIECEWISE LINEAR RESPONSE	
Conductivity Based Sensor	Frequency Output Sensors
■ $S = \Delta G \ / \ \Delta M$; ■ $S = \Delta(G/G_0) \ / \ \Delta M$; ■ $S = \Delta(G - G_0) \ / \ \Delta M$; ■ $S = \Delta(G - G_0)/G_0 \ / \ \Delta M$.	■ $S = \Delta f/ \ \Delta M$; ■ $S = \Delta(f/f_0) \ / \ \Delta M$; ■ $S = \Delta(f - f_0) \ / \ \Delta M$; ■ $S = \Delta(f - f_0)/f_0 \ / \ \Delta M$.

With reference to fig. 2 we define the different sensitivities as follows:

- ${}^{i}S = \frac{dY_1}{dM}$ Internal S
- ${}^{T}S = \frac{dY_2}{dY_1}$ Transduction S
- ${}^{A}S = \frac{dY_3}{dY_2}$ I, V, G, R Amplifier S
- ${}^{F}S = \frac{dY_4}{dY_3}$ Filter S
- ${}^{A/D}S = \frac{dY_{OUT}}{dY_4}$ Analog/Digital Conversion S
- ${}^{T}S_0 = \frac{dY_2}{dY_1} * \frac{dY_1}{dM} = \frac{dY_2}{dM}$ Overall Transduction S
- ${}^{A}S_0 = \frac{dY_3}{dY_2} * \frac{dY_2}{dY_1} * \frac{dY_1}{dM} = \frac{dY_3}{dM}$ Overall Amplifier S
- ${}^{F}S_0 = \frac{dY_4}{dY_3} * \frac{dY_3}{dY_2} * \frac{dY_2}{dY_1} * \frac{dY_1}{dM} = \frac{dY_4}{dM}$ Overall Filter S

- ${}^{A/D}S_0 = \frac{dY_{OUT}}{dY_4} * \frac{dY_4}{dY_3} * \frac{dY_3}{dY_2} * \frac{dY_2}{dY_1} * \frac{dY_1}{dM} = \frac{dY_{OUT}}{dM}$ Overall A/D S

We apply some of the above definitions to practical examples related to temperature and chemical sensors.

Noise (N)

Resolution can only be determined after noise evaluation of the sensor, keeping in mind that noise is related to the operating point. In practical situations different kinds of noises can be encountered: Thermal, Flicker, Generation–Recombination, Shot, and others that are seen in special cases but are not so frequent. The most important parameter used for the characterization of noise devices is the Noise Spectral Density by which, through integration, it is possible to estimate the mean square value of the output voltage:

$$V^2 = \int_{f_{low}}^{f_{high}} S(f) \cdot df \tag{1}$$

where:

- S(f) = 4 k T R for thermal noise,
- S(f) = 2 q I for shot noise,
- S(f) = k V^2/f^{α} with α close to 1 for flicker noise and
- S(f) = k_1 k_2 / $(1+w^2\tau^2)$ for g-r noise.

Resolution

At the theoretical level resolution (R) is defined as the amount of the measurand which gives a signal to noise ratio equal to one at the output and can be expressed, in a simplified form, as:

$$R = (NoiseVoltage)/S. \tag{2}$$

In practice it is defined as the amount of the measurand which gives a signal to noise ratio equal to 3 or 6 or 9, according to the kind of application, at the output and can be expressed in a simplified form as:

$$R = (3 \quad or \quad 6 \quad or \quad 9) * (Noise \quad Voltage)/S, \tag{3}$$

which means, in all cases, that sensors showing the same output noise present the better solution when *Sensitivity* is higher. Since sensitivity is a function of the operating point, so is resolution.

Resolution can be defined in two other ways: minimum detectable signal level applicable when the response is in the vicinity of the output noise level, and minimum detectable signal change level applicable at any operating point along the domain of the response curve when the measurand change is close to the noise level.

Drift

Drift (D) represents a slow, unpredictable fluctuation of the output signal. It has no statistical meaning. Its presence can sometimes be reduced by a careful design of all the individual sensor parts, but cannot be eliminated. It is a rather complex phenomenon, probably due to the aging effects of the microscopic constituents of the sensing material.

EXAMPLES
Metallic and Semiconductor Thermistor

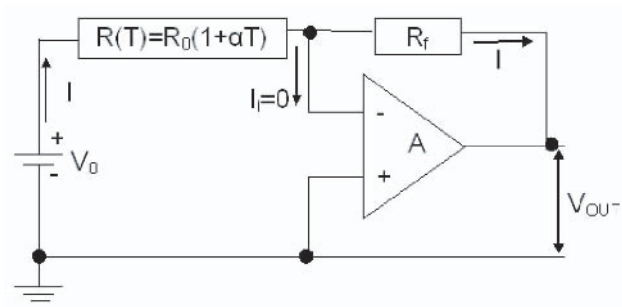


Figure 3. Schematic design of thermistor signal read-out circuit.

As a first example, let us consider a metallic thermistor inserted in fig. 3, whose resistance is, in a first approximation, expressed as: $R(T)=R_0(1+\alpha T)$. R(T) is the resistance of a PTC thermistor at a given temperature T, R_0 is the resistance at T_0, and I represents a suitable DC (or AC current), while A is the constant gain of a low noise amplifier, operating in a suitable bandwidth. Let us suppose that the injected current I does not induce, through the heating process, a detectable change of the resistance value.

In order to allow the resistance measurement, the current injection is obtained in the circuit represented in fig. 3, by applying a voltage V_0. This current, due to the virtual ground condition determined by the circuit configuration (very high input impedance), will cross the feedback resistor R_f and determine an output voltage. In this example M and Y_j (j=1,2,3) are the quantities:

$$M = T; Y_1 = R, Y_2 = I, Y_3 = V_{OUT}. \tag{4}$$

The Sensitivities can then be written as:

$$^{i}S = \frac{dY_1}{dM} = \alpha R_0 \tag{5}$$

$$^{T}S = \frac{dY_2}{dY_1} = -V_0 \frac{1}{R(T)^2} \tag{6}$$

$$^{A}S = \frac{dY_3}{dY_2} = -R_f. \tag{7}$$

The Overall Amplifier Sensitivity can be expressed as:

$$^{A}S_0 = {}^{A}S * {}^{T}S * {}^{i}S = \frac{\alpha * V_0 * R_f}{R_0 * (1 + \alpha T)^2} \tag{8}$$

and finally the Resolution is given by:

$$\Delta T = \frac{V_{noise}}{^{A}S_0}. \tag{9}$$

This means that in order to determine the temperature resolution, the sensitivity should be estimated and noise measurement performed at the output of the circuit.

It is worth mentioning that the overall sensitivity can be modified by changing both the value of the polarization current I and the amplification value A. As a particular case, when $I = 1\mu$A and A$= 10^6$, the product of I and A is equal to 1, so that the Overall Sensitivity coincides with the Internal Sensitivity divided by R(T).

All the changes in I and A should be done in order to have negligible self-heating of the thermistor, and an amplifier must be selected having a noise as low as possible in order to obtain an optimal resolution value for the determination of small temperature changes.

A second example is the case of a negative temperature coefficient (NTC) thermistor, such as a semiconductor, characterized by:

$$R(T) = R_0 \cdot e^{-\beta T} \qquad ^{i}S = \frac{dY_1}{dM} = -\beta \cdot R(T). \tag{10}$$

The Overall Amplifier Sensitivity ^{A}S$_0$ can be expressed as:

$$^{A}S_0 = {}^{A}S * {}^{T}S * {}^{i}S = \frac{-\beta * e^{\beta T} * V_0 * R_f}{R_0} \tag{11}$$

and the Resolution is given by:

$$\Delta T = \frac{V_{noise}}{\prod^{i} S_i} = \frac{V_{noise}}{^{A}S_0}. \tag{12}$$

2. Fundamentals Devices

The MOSFET, the most important microelectronic device, can be reduced in dimension to reach a minimum feature size of 0.1 micron but even lower dimensions (0.05 microns) are foreseen, as demonstrated by recent advanced experiments.

Relevant for our discussion is the genesis of the sensitivity behaviour in a class of devices all generated from the well known MOSFET structure (ISFET and GASFET). In particular, the influence of charges into the gate oxide on the threshold voltage and MOSFET behaviour under shrinking conditions will be discussed.

The MOSFET operation

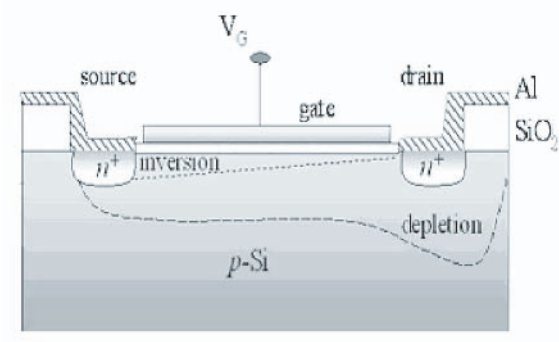


Figure 4. Cross-sectional view of a n-channel MOSFET.

The charge control equation related to this device in the quasi-linear region is approximately given by:

$$I_{DS} = \mu_n C_{ox} \frac{W}{L} [(V_{GS} - V_T) V_{DS} - \frac{V_{DS}^2}{2}] \tag{13}$$

where

- C_{ox} = oxide capacitance
- V_{DS} = dc drain-source voltage
- V_T = threshold voltage
- V_{GS} = gate voltage
- μ_n = inverted channel electrons mobility.

In the saturation region, at and above the pinch-off point, neglecting the effective channel length change due to the V_{DS} value, due to the condition $\partial I_{DS}/\partial V_{DS} = 0 (thus V_{DS} = V_{GS} - V_T)$, this equation becomes:

$$I_{DS} = \mu_n C_{ox} \frac{W}{L}[(V_{GS} - V_T)^2]. \tag{14}$$

Considering the above two equations, it is possible to derive the two transconductance expressions, $g_{m,LIN}$ and $g_{m,SAT}$ as follows:

$$g_{m,LIN} = \frac{\partial I_{DS}}{\partial V_G} = \mu_n C_{ox} \frac{W}{L} V_{DS} \tag{15}$$

$$g_{m,SAT} = \frac{\partial I_{DS}}{\partial V_G} = \mu_n C_{ox} \frac{W}{L}(V_{GS} - V_T). \tag{16}$$

It is worth pointing out that g_m has the same meaning as the output current-input voltage gate sensitivity (S).

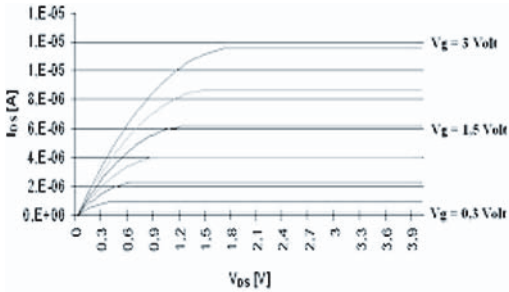


Figure 5. $I_{DS} - V_{DS}$ characteristic with V_{Gn} as input parameter($V_{Gn} > V_{Gn-1}$).

From the transconductance expressions we see that the first (eq. 15) is linear with V_{DS} and the second (eq. 16) is related to both MOSFET gate voltage V_{GS} and its threshold voltage V_T. In both cases C_{ox} plays an important role. In fact in order to increase the sensitivity, the gate oxide thickness should be as thin as possible. $g_{m,SAT}$, according to equation 16, depends on V_T and it is known that V_T depends on V_{FB}, the flat band voltage, according to:

$$V_T = V_{FB} + V_C + 2|\Phi_P| + \frac{1}{C_{ox}}\sqrt{2\varepsilon_S q N_a(2|\Phi_P| + V_C - V_B)} \tag{17}$$

where

- V_C= voltage applied at drain
- Φ_P= potential in a doped region
- ε_S= relative dielectric constant
- N_a= acceptor atomic density
- V_B= voltage applied at substrate.

Influence of charges into the oxide layer in a MOS system

It is important to briefly recall the influence of a given charge distribution present in the oxide on V_{FB}, and, as a consequence, on V_T [2]. If

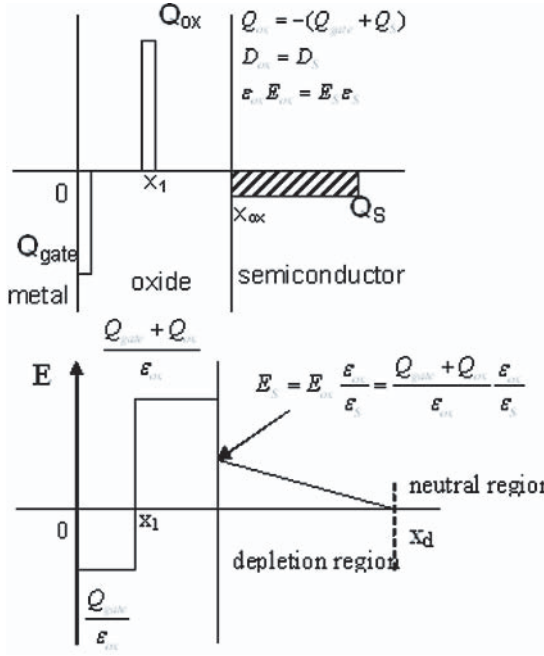


Figure 6. example of sheet of charge and approximate plot of the electric field in the Q_{ox} in the oxide layer MOS system.

a charge distribution $\rho(x)$ is present in the oxide, V_{FB} is given by:

$$V_{FB} = \Phi_{MS} - \frac{Q_f}{C_{ox}} - \frac{1}{C_{ox}} \int_0^{x_{ox}} \frac{x\rho(x)}{x_{ox}} dx \tag{18}$$

where

- Φ_{MS}= metal semiconductor work functions difference
- Q_f= fixed charge at the $SiO_2 - Si$ interface.

In fact the effect of any oxide charge is to shift the flat band voltage from its value related to the ideal case (absence of charges into the oxide). If this charge is stable, the shift induces a stable change of the threshold voltage V_T. If this charge is *not stable*, the transconductance value will not be stable and, as a consequence, the I_{DS} will experience

an unwanted change that may affect the useful signal in real operative conditions.

On the other hand, *fixed (stable) charges* into the oxide may be tolerated, and their influence overcome, if a calibration procedure is considered before any estimation of the MOSFET output.

In order to give an example of oxide charge induced MOS behaviour, let us consider the case of fig. 6 where a sheet of charge Q_{ox} is inside the oxide at x_1, and where the corresponding electric field is drawn according to Gauss' law:

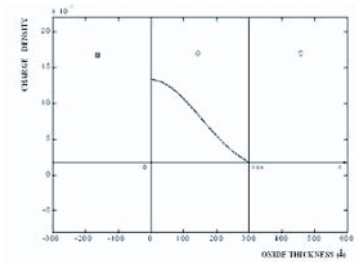


Figure 7. Example of charge half Gaussian distribution inside the oxide layer of a MOSFET structure.

Due to the electric field presence in the semiconductor, Q_{ox} will influence its flat band voltage. It is possible to derive, from simple considerations related to electrostatics, that the V_{FB} variation ΔV_{FB}, when a charge Q_{ox} is in the oxide at position x_1, is given by [2]:

$$\Delta V_{FB} = -\frac{Q_{ox} x_1}{C_{ox} x_{ox}}. \tag{19}$$

This means that if $x_1 = x_{ox}$, ΔV_{FB} will be equal to $-Q_{ox}/C_{ox}$, as expected.

On the other hand if $x_1 = 0^+$ (in other words, if Q_{ox} is just inside the oxide at $x = 0^+$), $-Q_{ox}$ will fully compensate the positive charge, and ΔV_{FB} will be zero. In fact the electric field into the silicon will be about zero and Q_{ox} will not influence the quiescent point.

It is worth noting that if Q_{ox} is moved just outside the oxide layer (for instance at $x = 0^-$), then ΔV_{FB} will be different from zero because the situation will correspond to a positive voltage directly applied to the gate by a virtual metal gate represented by the sheet of charge of intensity $+Q_{ox}$. As an example, if the oxide charge can be represented by half of the Gaussian distribution, as follows

$$\rho(x) = \frac{1}{\sigma\sqrt{2\pi}} exp[-\frac{x^2}{2\sigma^2}] \quad , \tag{20}$$

the flat band voltage contribution will be given by:

$$\Delta V_{FB} = \frac{\sigma}{C_{ox} x_{ox} \sqrt{2\pi}} \{exp[-\frac{x^2}{2\sigma^2}] - 1\} \quad . \tag{21}$$

The ISFET operation

In principle the ISFET is derived from a MOSFET, where the metal is replaced by the couple solution-reference electrode and where a CIM (Chemically Interactive Material) is deposited on the SiO_2, the gate oxide.

The purpose of the CIM is to attract ions present in the solution; the incorporated ions represent an equivalent charge near the SiO_2 and it is equivalent to a voltage applied to the gate. In particular cases the CIM can be the SiO_2 itself or a thin film of Si_3N_4 or other insulators. As an example, in order to obtain an ISFET sensitive to K^+ ions in solution, valynomicine may be utilized as a CIM.

An approximate expression for the V_{FB} of an ISFET is:

$$V_{FB} = E_{ref} - \Delta\phi_i - (\phi_0 + \frac{RT}{F} \ln a_i) - \phi_{Si} - \frac{Q_{SS}}{C_{ox}} - \frac{1}{C_{ox}} \int_0^{x_{ox}} \frac{x\rho(x)}{x_{ox}} dx \tag{22}$$

$$E_{ref} = \mu_+^{El} - \mu_+^{met} \tag{23}$$

where

- $\Delta\phi_i$= correction factor
- ϕ_0= standard potential of the oxide-electrolyte interface
- a_i= ion activity of the electrolyte
- ϕ_{Si}= silicon work function
- Q_{SS}= charge of the surface states
- $\int_0^{x_{ox}} \frac{x\rho(x)}{x_{ox}} dx$= oxide charge.

At this point it is worth mentioning that the space charge regions present from the reference electrode up to the semiconductor show different electric fields, and as a consequence, different potential drops. The contribution of all of them determines the operating point (QP) of the system.

Around this point any change of the electrolyte concentration would change the I_{DS} current by a certain sensitivity value. Once the noise at QP is measured, then the resolution at QP may be evaluated by

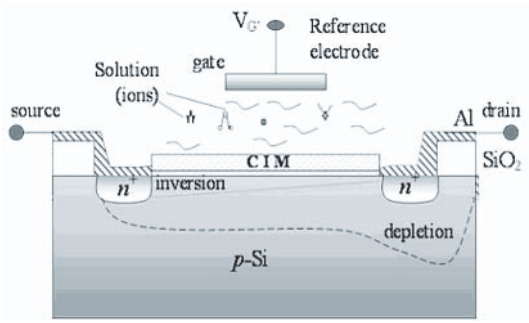


Figure 8. cross sectional view of an ISFET in a solution with the reference electrode.

$R_{QP} = V_{N,QP}/S_{QP}$, where $S_{QP} = \partial I_{DS}/\partial C_{onc}$ and all quantities are dependent on the quiescent point (QP).

Moreover it is important to stress that any change of one of the above mentioned space charge regions, either in solids or in liquids, will yield, as final effect, a change of the output current of the system.

Finally we observe that changes of the I_{DS} can only come from changes related to what is in between the gate and the semiconductor. Charges out of the gate have no influence on the system. This observation will prove useful in understanding the GASFET operation.

Essential is the presence of the reference electrode, which means an electrode whose potential is solution-independent and should also be temperature independent, at least in a specific range around the QP.

Any charge change occurring only "between" the reference electrode and the semiconductor is a candidate for a change of I_{DS}. In particular one of the most important points is the surface potential at the oxide-solution interface (φ_0) if no CIM is present, or the surface potentials between the CIM and the solution and the potential between the SiO_2 and the CIM, in the presence of a given CIM. The ISFET operation may be represented by the following "changes-flow" which may be considered as superimposed on the quiescent point determined by the reference electrode potential:

$$\Delta\varphi_{0SiO_2-El.} \rightarrow \Delta V_{FB} \rightarrow \Delta V_T \rightarrow \Delta I_{DS} \rightarrow \Delta V_{out} \tag{24}$$

$$\Delta\varphi_{0CIM-El.} \rightarrow \Delta\varphi_{SiO_2} \rightarrow CIM \rightarrow \Delta V_{FB} \rightarrow \cdots \tag{25}$$

In the absence of a given CIM the device shown in fig. 8 may become the well known pH sensor.

It is worth mentioning that a variety of insulators may be utilized, as well, for pH measurements, such as oxides of Ti, Ta, Al, Ir. All of

them, due to their different permittivity values, give rise to different sensitivities for the ISFET and also for the GASFET.

The GasFET operation

GasFETs are devices similar to MOSFET, and are able to show a certain sensitivity to volatile compounds. The intrinsic sensitivity mechanism is based on the possibility of generating either a positive or negative charge to be deposited in the vicinity of the SiO_2-CIM interface or even on the top of the CIM. A conductive gate must be present in order to get the correct operating point of the FET structure and an air gap must be present too between the CIM and the gate in order to permit the gas to flow.

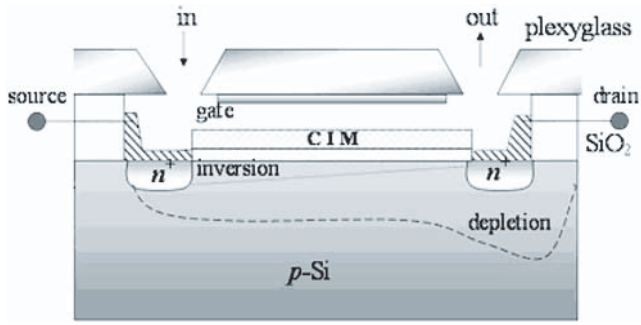


Figure 9. cross sectional view of a GASFET where the CIM is deposited on the SiO_2 layer.

For example the Pd gate MOSFET is sensitive to hydrogen due to the catalytic behaviour of palladium (very few metals are transparent to a given gas at standard pressure and temperature). The diffusion of hydrogen atoms at the Pd-SiO_2 interface and, in a first approximation, the consequent local change of the work function of the Pd-H structure, induces a change in V_{FB} and, as a consequence, in the I_{DS}, with a certain degree of sensitivity.

The necessity to have an air gap with the aim of allowing the volatile compounds to enter and leave the system, has an influence on the overall sensitivity. In fact it becomes reduced due to the presence of additional capacitors in series with that due to the SiO_2 layer. The total capacitance is made up by the presence of C_{ox}, due to the SiO_2 layer, C_{CIM},

due to the CIM layer and C_{air} due to the air gap:

$$C_{TOT} = \frac{1}{\frac{1}{C_{ox}} + \frac{1}{C_{CIM}} + \frac{1}{C_{air}}} \tag{26}$$

$$g_{m,SAT} = \frac{\partial I_{DS}}{\partial V_G} = \mu_n C_{TOT} \frac{W}{L}(V_G - V_T). \tag{27}$$

Due to the fact that C_{air} is small enough, we get in a first approximation:

$$C_{TOT} = \frac{C_{ox}C_{air}C_{CIM}}{C_{air}C_{ox} + C_{ox}C_{CIM} + C_{CIM}C_{air}} \approxeq C_{air}. \tag{28}$$

In order to have a large $g_{m,SAT}$ a high C_{air} value should be used or, in other words, the air gap should be as small as possible.

Two different situations are possible as far as the GasFET architecture is concerned.

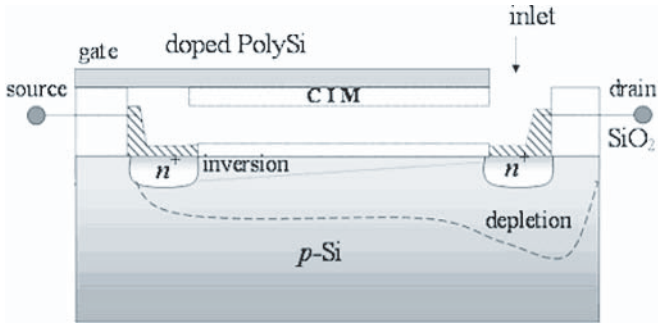


Figure 10. cross sectional view of a GASFET where the CIM is deposited underneath the gate.

Considering the two architectures of figs. 9 and 10, we have two kinds of sensitivity: one is related to the ratio of the output current variation ΔI_{DS} with respect to ΔV_G; the other is related to the ratio of the output current variation ΔI_{DS} with respect to the equivalent voltage change due to a charge variation occurring on behalf of the adsorbing CIM. The configuration of fig. 9 is certainly preferred to that of fig. 10, because in fig. 9 the charge variation at the CIM level is closer to the SiO_2 layer, when compared to the situation occurring in fig. 10. As a consequence, a greater effect will be induced on the depletion-inversion regions of the silicon underneath the SiO_2.

MOSFET shrinking

It is not obvious that the sensor technology takes advantage of the integrated transistor shrink (ITRS) trend. In fact actual MOSFET dimen-

sions offered by microelectronic technology are already even too small for most of the sensor applications.

First, the decreasing supply voltages tend to reduce the dynamic range of analog circuitry, unless this is compensated for by a substantial increase of power dissipation. Next, device miniaturization enhances, rather than reduces, 1/f noise. Finally, the very high cost of silicon real estate after a deca-nanometer process could become incompatible with the requested size of on-chip sensors.

According to the scaling rules, the MOSFET transconductance is expected to remain constant if both the lateral and vertical device dimensions are reduced by the same scaling factor λ. For a short-channel MOSFET, the saturation current I_{DSAT} may indeed be expressed as

$$I_{DSAT} = WC_{ox}(V_{GS} - V_T - V_{DSAT})v_{sat} \tag{29}$$

where W is the device width, C_{ox} is the oxide capacitance per unit area, V_{GS} is the gate-source voltage, V_T is the threshold voltage, V_{DSAT} the saturation voltage and v_{sat} is the carrier saturation velocity. The device transconductance turns out to be

$$g_m = WC_{ox}[1 - (dV_{DSAT}/dV_{GS})]v_{sat} \tag{30}$$

and, under the assumption that (dV_{DSAT}/dV_{GS}) is roughly constant for a given ratio V_{DD}/V_T , it turns out that W scales with λ and $C_{ox} = (\kappa \ \ \varepsilon \ /t_{ox})$ with $1/\ \lambda$, regardless of the voltage scaling factor. Thus, the MOSFET transconductance is expected to basically remain constant through several technology generations, provided high-κ dielectrics under development compensate for the non-scalability of the oxide thickness t_{ox}. On the other hand, the transconductance per unit width (g_m/W) is expected to increase by the scaling factor λ.

Connected with the transconductance scaling rules is the behaviour of the thermal noise, and the mean square value of the current $<i_d^2>$ in saturation reads: $< i_d^2 > = 4KT\ \gamma g_m$. Here γ is a factor which, within a simplified MOSFET model, equals 2/3, but it may become larger under strong non equilibrium conditions, where the average energy of the carriers increases well above (3/2)KT. Due to the insensitivity of the device transconductance to the scaling factor, the thermal noise is expected to stay nearly constant as the device size is scaled down. The major problem comes instead from the flicker noise, otherwise referred to as the 1/f noise, which increases inversely with the gate area. The flicker noise is modeled as a voltage source of value: $< v_g^2 >= K/(WLC_{ox}f)$, in series with the gate. Here K is a process-dependent constant.

Therefore the flicker noise is expected to grow with γ as the device size is scaled down. In deep submicron MOSFETs the corner frequency at which thermal noise equals flicker noise may be as large as 100 MHz, indicating that, at low frequency, 1/f noise is the most severe noise source which affects sensor performance.

When the available gate area is further reduced and the number of devices may increase, another crucial point is the complexity of the deposition of a suitable and possibly different chemically interactive material (CIM) on their gates. This problem may have a solution in those cases where different kinds of CIMs are mixable without altering their individual properties. In fact with only one deposit, if the gate areas are sufficiently small most of the gates will be covered by different compositions of CIMs. This process would allow the easy fabrication of single chip electronic noses with the possibility of a high degree of redundancy.

3. Thermopiles

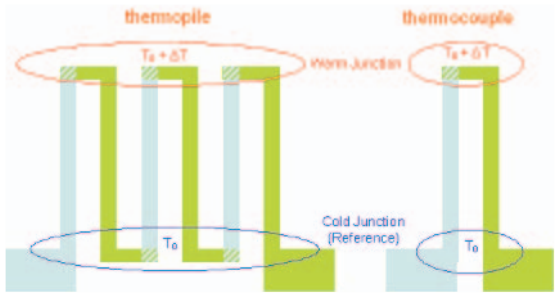


Figure 11. Schematic design of Thermocouple and Thermopile.

Thermopiles are considered temperature sensors and are fabricated incorporating a number of thermocouples. Each thermocouple is formed by a couple of different materials (Metal1-Metal2, Metal-Semiconductor, Semiconductor-Semiconductor) and responds to a temperature difference localized between the two junctions (cold junction and warm junction), see fig. 11. One of the two junctions can be considered the reference one.

During operation the voltage developed at the thermopile output is proportional to the thermoelectric power of each of the two different materials and to the temperature difference between the warm and cold junction (Seebeck effect).

When constituted of metals, thermopiles exhibit a very low noise, in particular only thermal noise if the voltage amplifier used for signal amplification has a very high input impedance.

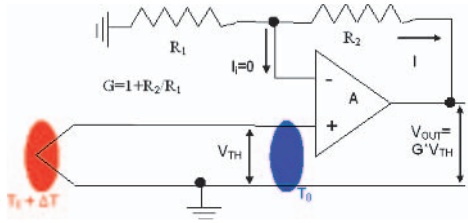


Figure 12. Schematic design of thermocouple or thermopile signal read-out.

Thermopiles can be deposited through thermal evaporation or even sputtering on either hard or soft substrates.

A thermopile can also be used as a chemical sensor if one of the two materials is a catalytic metal for a given volatile compound. In this case it is necessary to keep the warm and cold junctions at constant temperature. During absorption of the volatile compound on behalf of the catalytic material the thermoelectric power may change, giving rise to an output voltage which can be related to the concentration of the volatile compound. A typical example is the thermopile as hydrogen sensor, where one of the two materials is palladium, a standard hydrogen catalyzer.

4. Kelvin Probe

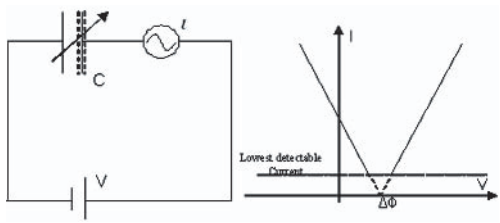


Figure 13. Schematic design of Kelvin Probe circuit and its signal output.

The importance of surfaces has grown along with the development of chemical sensors in recent years, due to the interaction between a given volatile compound and the surface of a chemically interactive material.

The Kelvin Probe technique allows measurement of the Work Function of a given surface, not only in stationary conditions but also during absorption – desorption processes.

In its simple form the Kelvin Probe is shown in figure 13, where the test plate is left fixed while the other plate of the capacitor can

be mechanically moved in different ways, in particular, as one possible example, with a piezoelectric system.

According to electromagnetic theory, any time a charge capacitor changes its value a displacement current is generated, expressed as $I = dQ/dt = CdV/dt + VdC/dt$.

The experiment is conducted measuring the current corresponding to different voltages (positive and negative) applied to the capacitor. Since the overall voltage applied to the capacitance is $V - \Delta\Phi$ the displacement current is given by: $I = (V - \Delta\Phi)dC/dt$.

A plot of current vs. voltage allows the $\Delta\Phi$ to be determined as the intersection of the current amplitudes on the X-axis.

5. Bulk Acoustic Waves

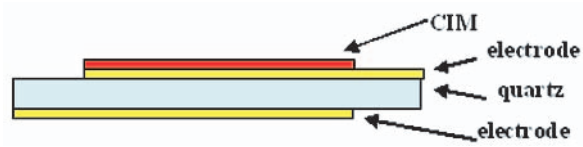


Figure 14. Quartz resonator as microbalance for chemical sensing.

A schematic diagram of a bulk acoustic wave (BAW) chemical sensor is composed of a BAW piezoelectric resonator with one or both surfaces covered by a membrane (CIM) (fig. 14).

The BAW structure is usually connected to a suitable amplifier to form an oscillator whose resonant frequency is related to both the physical and geometrical characteristics of the device.

Any change in the physical properties of the membrane due to adsorption or absorption of chemical species from either the gas or liquid phase affects the resonant frequency of the structure.

The resonator is usually made of quartz and both longitudinal and shear modes can be used. As to the quartz, crystallographic cuts showing a highly stable temperature operation dependence are carefully selected in order to improve the possibility of obtaining satisfactory resolution values.

6. Surface Acoustic Waves

Surface acoustic wave (SAW)-type chemical sensors exploit the propagation loss of the acoustic waves along layered structures consisting of at least a substrate covered by the CIM.

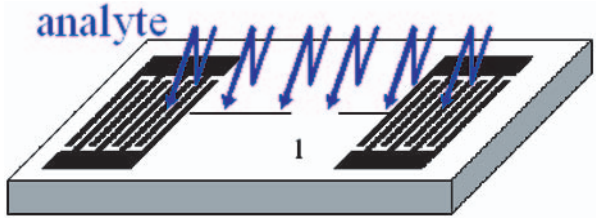


Figure 15. Basic structure of a SAW chemical sensor.

Changes produced by the measurand on the properties of the CIM can affect both the phase velocity and the propagation loss of the acoustic wave. There are examples of SAW sensors based on the measurements of the changes in the phase velocity.

A SAW device is configured as a delay line and fed by a radio frequency signal. Any change in the velocity Δv is detected as a change $\Delta\Phi$ in the phase delay of the wave, thanks to a phase detector that gives a voltage proportional to the difference of phase between signal input and output.

$$\varphi = 2\pi\frac{1}{\lambda} = 2\pi\frac{lf}{v} \tag{31}$$

$$\Delta\varphi = 2\pi lf \cdot \Delta(\frac{1}{v}) = -\varphi_0\frac{\Delta v}{v}. \tag{32}$$

7. Natural and Artificial Olfaction

In the last decade much effort has been oriented to the fabrication of artificial olfaction machines able to determine chemical images (also odor images) of complex volatile compounds. Today many different electronic noses and tongues are available for odor detection and classification and for the creation of chemical images of liquids.

Characteristics comparison of natural and artificial olfaction

Natural olfaction	Artificial Olfaction
■ Receptors: -Non selective -Ultrahigh Redundancy(10^8) -Biochemical transduction signal: pattern of spikes. ■ Sample Delivery: -Actuation of sniffing -Two sources of odor (outside and inside) ■ Signal processing: Data synthesis ■ Data analysis: -Ultra Wide Database -Drift compensation -High integration with other senses	■ Sensors: -Non Selective -Low Redundancy (10) -Chemical transduction signal: steady signal ■ Sample Delivery: -Continuous sniffing -A source of odor (outside) ■ Signal processing: One sensor–one signal ■ Data analysis: -Limited database -Poor drift compensation -Integration with other instruments

These systems are formed by a number of cooperating individual non-selective sensors, whose outputs are processed to form chemical images or, in the presence of odors, olfactory images.

Natural olfaction does not give analytical information about the inhaled air, but rather it provides signals to the brain in order to get, at the perception level, a qualitative description of the sniffed air. Also natural olfaction utilizes a huge number (millions) of non selective receptors which show sensitivity to thousands of different odors. The artificial olfaction system has a smaller number (from 5 to 50) of sensors and, after a suitable data analysis technique, it is possible to obtain images of the volatile compound clusters present in the environment. The sensors used most for artificial nose applications are those based on quartz micro-balances, operating at room temperature, or those employing metal oxide semiconductor materials such as SnO_2 (operating in the temperature range (200-500) $^\circ$C, doped with different catalysts, in order to give a higher sensitivity toward gasses in the detection processes.

Varieties of polymers are also employed as sensitive material for electronic nose applications, and the operating temperature may reach about 100 $^\circ$C. In the case of quartz microbalance-based sensors a large role is played by the chemically interactive material (CIM) on which it is deposited. A rather efficient room temperature operating CIM is the metal-porphirin, by which it is possible to construct varieties of nostrils, just changing the type of coordinated metal. Interesting metals success-

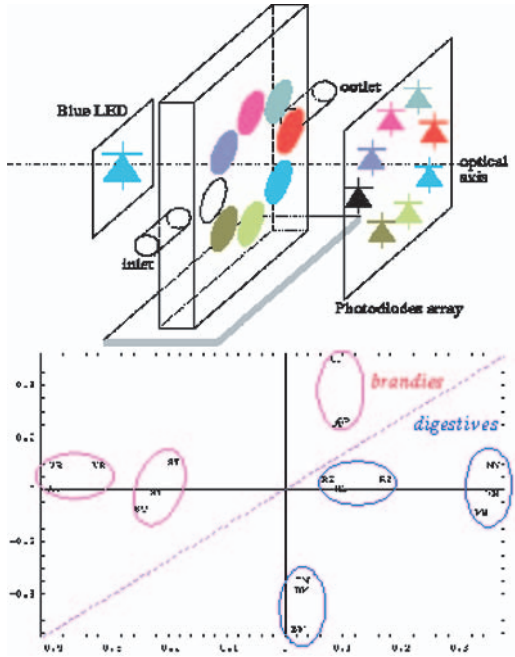


Figure 16. Opto-nose and multi-component analysis of nose output.

fully employed are: cobalt, zinc, copper, magnesium, iron, etc. Fig. 16 shows an opto-nose and the principal components analysis result related to the discrimination skin to distinguish brandies and digestives, while fig. 17 shows the typical multi-component analysis data of a nose output [3].

Future perspectives for the electronic nose research field are listed below. They concern both expected sensing and technical sensing and performance. Improvement of sensing performance of the instrument:

- Increase of both selectivity and sensitivity of the instrument towards different samples;
- Introduction of enrichment techniques of headspace sampling;
- Optimization of molecular structures towards more specific applications;
- Generation of more reproducible and long-life sensors.

Improvement of technical performances of the instrument:

- Creation of a portable nose powered by batteries, able to detect increasing concentration of gases (or odors);

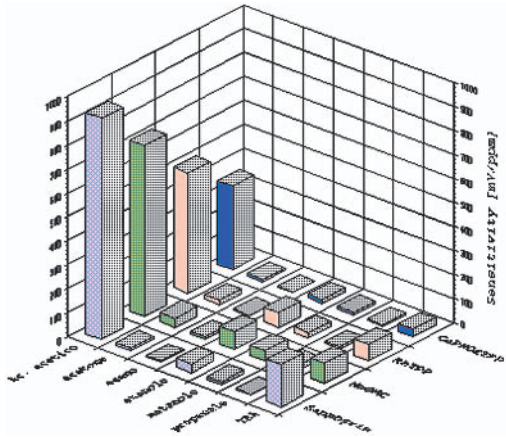


Figure 17. Typical responses of the different sensors.

- Miniaturization of the instrument.

8. Optical Fibre Sensor

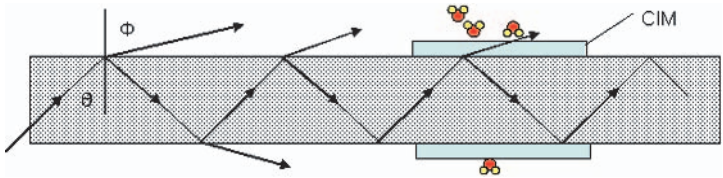


Figure 18. Propagation of light in a typical fibre sensor.

Optical fibres behave as wave guides for light. The original ones were made of glass but now polymeric materials are mostly used as they are cheaper than glass and metal wires.

The light can propagate along the fibre usually by totally internal reflection (TIR) as indicated in fig. 18. TIR is controlled by the angle of incidence and also by the refractive indices of the media involved in the propagation.

From Snell's Law, $\sin(\Theta_i) * n_i = \sin(\Theta_i) * n_r$. We have *TIR* when $\sin(\Theta_i) > n_r/n_i$, while we will have *refraction and reflection* when $\sin(\Theta_i) < n_r/n_i$. In practical cases properties of light, such as phase, polarization and intensity, can be modulated inside the wave guide by a given measurand, which is interacting, for instance, with a CIM lying within the penetration depth for the evanescent field of the light localized near the external guide surface.

Integrated guides can also be used as a chemical sensor. Fig. 19 shows a Mach-Zehnder interferometer where one of the two branches has been covered by palladium, a catalytic metal for H_2 [4]. The output phase change measurement determination of parts per million of H_2 has proven to be possible.

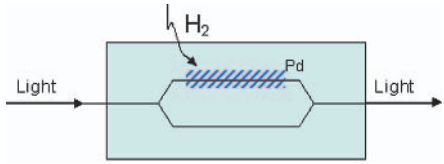


Figure 19. Mach-Zehnder Interferometer.

9. Surface Plasmon Resonance

The term surface plasmon resonance (SPR) can refer to the phenomenon itself or to the use of this phenomenon to measure biomolecules binding to surfaces. This method is now widely used in the biosciences and provides a generic approach to measurement of bio molecule interactions on surfaces.

The phenomenon of SPR is directly related to Snell's Law (see fig. 18). In fact when radiation passes to a medium with lower dielectric constant there is a critical angle beyond which the refracted beam cannot propagate in the other medium.

The decrease in reflectivity at the SPR angle (2_{SP}) is due to absorption of the incident light at this particular angle of incidence. At this angle the incident light is absorbed and excites electron oscillations on the metal surface.

It is important to understand why reflectivity is sensitive to the refractive index of the aqueous medium if the light is reflected by the gold film. This sensitivity is due to an evanescent field which penetrates approximately 200 nm into the solution [5].

The evanescent field appears whenever there is resonance between the incident beam and the gold surface and is not present when there is no plasmon resonance, that is, where the reflectivity is high.

$$\text{Total Internal Reflectance:} \qquad \sin(\Theta_i) = \frac{n_r}{n_i}. \tag{33}$$

Incident light can excite a surface plasmon when its x axis component equals the propagation constant for the surface plasmon. SPR occurs when this projected distance matches the wavelength of the surface plasmon (fig. 20).

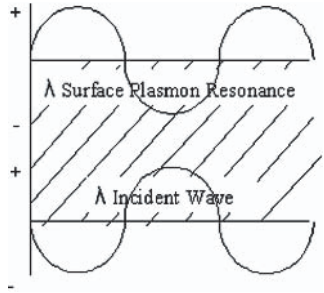


Figure 20. Condition of Surface Resonance.

Resonance cannot be excited with light incident from the air or from a medium with lower dielectric constant ($k_{SP} > k$). The refractive index of a prism reduces the wavelength to $\lambda = \lambda_0/n_P$.

A condition of resonance can be obtained by:

- $k_{SP} = k_0(\frac{\varepsilon_r \cdot \varepsilon_i}{\varepsilon_r + \varepsilon_i})$
- $k_x = k_0 \cdot n_i \cdot \sin(\Theta_i)$
- Condition of resonance: $k_{SP} = k_x$

10. Conclusions

In view of the future development of sensors driven by increasing demand for accuracy and precision, and by the opening of new fields close to the biological area (which is oriented toward nano-biosensor fabrication), it appears even more important to properly use the most relevant sensor keywords, such as: response curve, sensitivity, noise, drift, resolution, and selectivity.

The correct understanding of these words and their implications is of fundamental importance for the scientific and industrial community interested in sensor science development, since it allows the correct dissemination of both experimental and theoretic results, even if other important terms in the sensor field have not been discussed in this chapter, such as speed of response, reversibility, repeatability, reproducibility, and stability, to which some attention was paid during the presentation of this work at the ASI.

Some fundamental transducers have also been considered to explain intrinsic sensing and sensitivity mechanisms, without disregarding comments on noise which are fundamental to the determination of resolution.

References

[1] A.D'Amico, C.DiNatale. A contribution on some basic definitions of sensors properties. *Sensors Journal, IEEE*, vol. 1, Issue 3, Oct 2001 Page(s):183 - 190

[2] R.Muller, T.Kamins, M.Chan. *Device electronic for integrated circuits* John Wiley & Sons, 3rd edition, 2003.

[3] C.DiNatale, D.Salimbeni, R.Paolesse, A.Macagnano, A.D'Amico. Porphyrins-based opto-electronic nose for volatile compounds detection, *Sensors and Actuators B* 65 (2000) 220-6.

[4] A.D'Amico et al. Integrated optic sensor for the detection of H2 concentrations *Sensors and Actuators B* 7 (1992) 685-8.

[5] N.J. Walker. A technique whose time has come, Science 296 (2002) pp. 557–559.

WIRELESS SENSOR NETWORKS FOR SECURITY: ISSUES AND CHALLENGES

Tolga Onel, Ertan Onur, Cem Ersoy
Department of Computer Engineering
Boğaziçi University
Bogazici University
Bebek 34342 Istanbul, Turkey

Hakan Delic
Department of Electrical and Electronics Engineering
Boğaziçi University
Bogazici University
Bebek 34342 Istanbul, Turkey

Abstract In this chapter, the sensing coverage area of surveillance wireless sensor networks is considered. The sensing coverage is determined by applying Neyman-Pearson detection and defining the breach probability on a grid-modeled field. Using a graph model for the perimeter, Dijkstra's shortest path algorithm is used to find the weakest breach path. The breach probability is linked to parameters such as the false alarm rate, size of the data record and the signal-to-noise ratio. Consequently, the required number of sensor nodes and the surveillance performance of the network are determined. For target tracking applications, small wireless sensors provide accurate information since they can be deployed and operated near the phenomenon. These sensing devices have the opportunity of collaboration amongst themselves to improve the target localization and tracking accuracies. Distributed data fusion architecture provides a collaborative tracking framework. Due to the present energy constraints of these small sensing and wireless communicating devices, a common trend is to put some of them into a dormant state. We adopt a mutual information based metric to select the most informative subset of the sensors to achieve reduction in the energy consumption, while preserving the desired accuracies of the target position estimation.

Keywords: wireless sensor network; detection theory; Kalman filtering; target intrusion detection; false alarm.

Jim Byrnes and Gerald Ostheimer (eds.),
Advances in Sensing with Security Applications, 95–119.

1. Introduction

Wireless sensor devices that are employed for security applications have several functionalities. The first one is the distributed detection of the presence of a target or an and the estimation of parameters of interest. The target may be tracked for various purposes. The detection, estimation and tracking efforts may or may not be collaborative. The second task involves wireless networking to organize and carry information. Issues related to distributed detection and estimation have long been studied. Moreover, wireless sensor networking is addressed in the literature to a certain extent in the context of ad hoc networking. However, there is not much work done on how the wireless networking constraints affect the distributed detection and estimation duty of the wireless smart sensor networking devices.

The sensing and communication ranges of some propriety devices are listed in [43]. For example, the sensing range of the Berkeley motes acoustic sensor, HMC1002 magnometer sensor and the thrubeam type photoelectric sensor are nearly one meter, 5 meters and 10 meters respectively. The communication range of the Berkeley motes MPR300, MPR400CB and MPR520A are 30, 150 and 300 meters, respectively. The ratio of the communication and sensing ranges shows that the network must be densely deployed. The high redundancy level of the network necessitates energy conservation.

For surveillance wireless sensor networks (SWSN), depending on the sensing ranges and the coverage schemes of the sensors, as well as the deployment density of the network, the sensing coverage area may contain breach paths. The probability that a target traverses the region through the breach path gives insight about the level of security provided by the SWSN. Some of the design issues related with security applications are [29]:

1 How many sensor must to be deployed to provide a required security level [30]?

2 How could the sensor detection process be modeled and how is the sensing coverage determined?

3 What are the effects of geographic properties of the field on target detection?

4 How should the sensors be deployed in the region [37]?

5 What is the weakest part of the coverage and how can the breach paths be discovered [11, 45]?

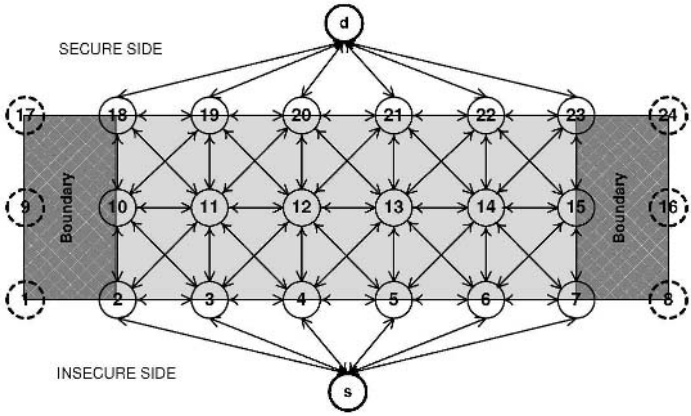


Figure 1. A sample field model constructed to find the breach path for length is 5 m., width 2 m., boundary 1 m., and grid size 1 m. ($N = 8, M = 3$).

6 How could the false alarms be minimized and the decisions be improved with collaboration?

7 What are the effects of the signal properties on the sensing coverage?

8 What is the impact of sensor scheduling on the sensing coverage [35] [39] [42]?

9 Non-communicating What should the effective communication and sensing ranges of the sensors be [8, 39]?

10 Should incremental deployment be considered?

Intrusion Detection

The security level of a WSN can be described with the breach probability that can be defined as the miss probability of an unauthorized target passing through the field. We define the weakest breach path problem as finding the breach probability of the weakest path in a SWSN. To calculate the breach probability, one needs to determine the sensing coverage of the field in terms of the detection probabilities.

In order to simplify the formulations, we model the field as a cross-connected grid as in Fig. 1 The field model consists of the grid points, the starting point and the destination point.

The target aims to breach through the field from the starting point that represents the insecure side to the destination point that represents the secure side from the SWN viewpoint. The horizontal and vertical

axes are divided into $N-1$ and $M-1$ equal parts, respectively. In this grid-based field model along the y-axis, we add boundary regions to the two sides of the field. Thus, there are NM grid points plus the starting and destination points.

Sensor deployment has a direct impact on the performance of target detection. Chvatal's art gallery problem [10] is to determine the minimum number of guards required to cover all points in a gallery. The similarity between the art gallery and sensor placement problems is established in [12], where algorithms are proposed to find effective locations for the sensor nodes. One algorithm tries to maximize the average coverage of the grids and the other tries to maximize the coverage of the least effectively covered grid. The goal is to determine the required number of sensor nodes and their places to provide a coverage threshold that defines the confidence level of the deployment.

Another approach to the breach path problem is finding the path which is as far as possible from the sensor nodes as suggested in [26], where the maximum breach path and maximum support path problems are formulated. In the maximum breach path formulation the objective is to find a path from the initial point to the destination point where the smallest distance from the set of sensor nodes is maximized. In the former problem, the longest distance between any point and the set of sensor nodes is minimized. To solve these problems, Kruskal's algorithm is modified to find the maximal spanning tree, and the definition of a breach number tree is introduced as a binary tree whose leaves are the vertices of the Voronoi graph.

The weakest breach path is also referred to as the best coverage problem in [23]. The energy considerations are modeled, a graph is created and the distributed Bellman-Ford algorithm is used to find the shortest path. Several extensions to the solutions are provided such as finding the best path with the minimum energy consumption and finding the path where the length is bounded.

In [25], Megerian *et al.* introduce the exposure concept as the ability to observe a target moving in a sensor field. By expressing the sensibility of a sensor in a generic form, the field intensity is defined as the sum of the active sensor sensibilities. The exposure is then defined as the integral of the intensities (involving all sensors or just the closest one) on the points in a path in the sensor field.

The field to be monitored is usually narrow and long in perimeter security applications. Thus, non-uniform deployment may be necessary. He *et al.* conclude that the sensor nodes generate false alarms at a non-negligible rate [18], and an exponentially weighted moving average on the sensor node is sufficient to eliminate transient alarms.

Due to the scarcity of energy resources of sensor nodes, energy conservation at all layers of the sensor network models is a widely studied topic. One method of energy conservation is applying a well-designed sleep schedule of sensor nodes [35] [39] [42]. However, for surveillance applications sleep scheduling of sensor nodes may produce insecure regions in the field. Thus, the primary concern in designing a sleep scheduling for surveillance wireless sensor networks is maintaining the coverage area. In [39], a coverage configuration protocol is presented that provides varying degrees of coverage depending on the application. Defining the coverage as the monitoring quality of a region, an analysis of the sensing coverage and communication connectivity is provided in a unified framework rather than an isolated one.

Target Tracking

Target tracking, in other words the processing of the measurements obtained from a target in order to maintain an estimate of its current state, has major importance in Command, Control, Communications, Computer, Intelligence, Surveillance and Reconnaissance (C4ISR) applications. Emerging wireless sensor technologies facilitate the tracking of targets just from within the phenomenon. Due to environmental perturbations, observations obtained close to the phenomenon are more reliable than observations obtained far from it. Wireless communication characteristics of the emerging wireless sensor nodes provide an excellent distributed coordination mechanism to improve global target localization accuracies. However, again, there is an inherent energy constraint for wireless sensor devices. In order to conserve the valuable battery energy of wireless sensor devices, some of the sensors should go into the dormant state controlled by the sleep schedule [41]. Only a subset of the sensors are active at any instant of time. Otherwise, a bulk of redundant data would be wandering in the network.

Collaborative target tracking has inherent questions such as how to dynamically determine who should sense, what needs to be sensed, and whom the information must be passed on to. Sensor collaboration improves detection quality, track quality, scalability, survivability, and resource usage [44].

There is a trade-off between energy expenditure and tracking quality in sensor networks [31]. Sensor activation strategies are *naive activation* in which all the sensors are active, *randomized activation* in which a random subset of the sensors are active, *selective activation* in which a subset of the sensors are chosen according to some performance criterion,

and *duty cycled activation* in which the sensors are active for some duty cycle and in dormant state thereafter.

In information driven sensor querying (IDSQ) [9, 44], the so-called cluster heads decide on the sensors to participate actively in the tracking task. In [24], a dual-space paradigm is presented in which the subset of sensors towards whom the target is approaching are selected to be active. In the location-centric approach to collaborative sensing and tracking, addressing and communication is performed among geographic regions within the network rather than individual nodes [34, 5]. This makes localized selective-activation strategies simpler to implement. Prediction based target tracking techniques based on Pheromones, Bayesian, and Extended Kalman Filter are presented in [6, 7], and a real implementation can be found in [27]. Multiple target tracking is examined in [4] [15] [22].

Censoring sensors [1] [17] [32] [33] is one approach to control the network traffic load. Sensors that are deemed as noninformative do not send their decisions or observations if their local likelihood ratio falls in a certain single interval. A special case of this phenomenon occurs when the lower bound of the no-send region interval used is zero. In this particular case, the problem reduces to sending the local decision/observation if the local likelihood ratio is above some threshold. A deficiency with this approach occurs for tracking applications if all the sensor local likelihood ratios fall in the no-send region, and no belief about the target state is shared among the sensors.

Research has focused on how to provide full or partial sensing coverage in the context of energy conservation [36, 42]. Nodes are put into a dormant state as long as their neighbors can provide sensing coverage for them. Such solutions regard the sensing coverage to a certain geographic area as binary, i.e., either it provides coverage or it does not [41]. These approaches consider the sensor selection problem only in terms of coverage and energy savings aspects, without paying attention to the detection quality. In tracking applications, when selecting a subset of sensors to contribute to the global decision, we have to take into account how informative the sensors are about the state of the target. In [9, 14], the sensor which will result in the smallest expected posterior uncertainty of the target state is chosen as the next node to contribute to the decision. It is shown in [14] that minimizing the expected posterior uncertainty is equivalent to maximizing the mutual information between the sensor output and the target state. In [38], an entropy-based sensor selection heuristic is proposed for target localization. The heuristic in [38] selects one sensor in each step and the observation of

the selected sensor is incorporated into the target location distribution using sequential Bayesian filtering.

2. Neyman-Pearson Detection

Using the field model described in section 1, detection probabilities are to be computed for each grid point to find the breach probability. The optimal decision rule that maximizes the detection probability subject to a maximum allowable false alarm rate α is given by the Neyman-Pearson formulation [20]. Two hypotheses that represent the presence and absence of a target are set up. The Neyman-Pearson (NP) detector computes the likelihood ratio of the respective probability density functions, and compares it against a threshold which is designed such that a specified false alarm constraint is satisfied.

Suppose that passive signal reception takes place in the presence of additive white Gaussian noise (AWGN) with zero mean and variance σ_n^2, as well as path-loss with propagation exponent η. The symbol power at the target is ψ, and the signal-to-noise power ratio (SNR) is defined as $\gamma = \psi/\sigma_n^2$. Each breach decision is based on the processing of L data samples. We assume that the data are collected fast enough so that the Euclidean distance d_{vi} between the grid point v and sensor node i remains about constant throughout the observation epoch. Then, given a false alarm rate α, the detection probability of a target at grid point v by sensor i is [20, 30]

$$p_{vi} = 1 - \Phi\Big(\Phi^{-1}(1-\alpha) - \sqrt{L\gamma_{vi}}\Big)$$

where $\Phi(x)$ is the cumulative distribution function of the zero-mean, unit-variance Gaussian random variable at point x, and

$$\gamma_{vi} = \gamma A d_{vi}^{-\eta}$$

represents the signal-to-noise ratio at the sensor node i, with A accounting for factors such as antenna gains and transmission frequency. Active sensing can be accommodated by properly adjusting the constant A.

Because the NP detector ensures that

$$\lim_{d_{vi}\to\infty} p_{vi} = \alpha,$$

instead of p_{vi} we use [30]

$$p_{vi}^* = \begin{cases} p_{vi} & \text{if } p_{vi} \geq p_t, \\ 0 & \text{otherwise,} \end{cases}$$

where $p_t \in (0.5, 1)$ is the threshold probability that represents the confidence level of the sensor. That is, the sensor decisions are deemed sufficiently reliable only at those d_{vi} distances where $p_{vi} > p_t$. Depending on the application and the false alarm requirement, typically $p_t \geq 0.9$. Note that p^*_{vi} is not a probability measure, but we shall nevertheless treat it as one in the ensuing calculations.

For those sensor types where the detection probability can not be explicitly tied to signal, noise and propagation parameters (e.g. infrared), the sensing model proposed by Elfes can be used [13]. The detection probability is defined such that different sensor types are represented by generic parameters. When the sensor-to-target distance is smaller (larger) than a threshold, the target is absolutely (not) detected. Elfes's model is employed in [29], where the required number of sensors is determined for a target breach probability level under random sensor placement.

The detection probability p_v at any grid point v is defined as

$$p_v = 1 - \prod_{i=1}^{R} (1 - p^*_{vi}) \tag{1}$$

where R is the number of sensor nodes deployed in the field. The miss probabilities of the starting and destination points are one, that is $p_0 = 0$ and $p_{NM+1} = 0$. More clearly, these points are not monitored because they are not in the sensing coverage area. The boundary regions are not taken into consideration.

The weakest breach path problem can now be defined as finding the permutation of a subset of grid points $V = [v_0, v_1, \ldots, v_k]$ with which a target traverses from the starting point to the destination point with the least probability of being detected, where $v_0 = 0$ is the starting point and $v_k = NM+1$ is the destination point. Here we can define the breach probability P of the weakest breach path V as

$$P = \prod_{v_j \in V} (1 - p_{v_j}) \tag{2}$$

where p_{v_j} is the detection probability associated with the grid point $v_j \in V$, defined as in (1). A sample sensing coverage and breach path is shown in Fig. 2. Using the two-dimensional field model and adding the detection probability as the third axis, we obtain hills and valleys of detection probabilities. The weakest breach path problem can be informally defined as finding the path which follows the valleys and through which the target does not have to climb hills so much. For a number of quality of deployment measures that can be utilized to evaluate a sensor network's intrusion detection capability, see [28].

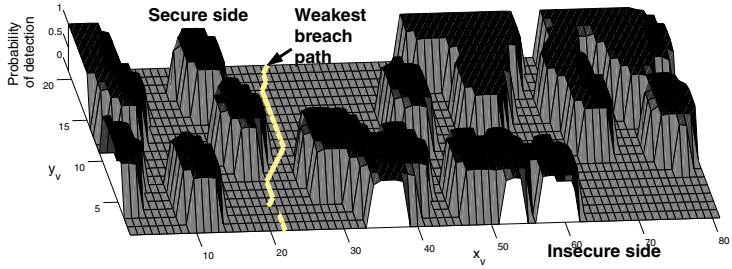


Figure 2. A sample sensing coverage and breach path where the field is 70×20 m., the boundary is 5 m. wide and the grid size is 1 m. ($N = 81, M = 21, L = 100, R = 30, \alpha = 0.1, \eta = 5, \gamma = 30$ dB.) [30].

In order to solve the weakest breach path problem, where we construct a graph to model the field, Dijkstra's shortest path algorithm can be employed [40]. The detection probabilities associated with the grid points cannot be directly used as weights of the grid points, and consequently they must be transformed to a new measure d_v. Specifically, let

$$d_v = -\log(1 - p_v)$$

be the weights of the grid points.

Using Dijkstra's algorithm, the breach probability can be defined as the inverse transformation of the weight d_{NM+1} of the destination point which is

$$P = 10^{-d_{NM+1}}. \tag{3}$$

The resulting path V is used to calculate the breach probability in (2), which is equal to the value computed in (3) [29].

3. Breach Probability Analysis [30]

The system parameter values depend on the particular application. When a house or a factory is to be monitored for intrusion detection, the cost of false alarms is relatively low. On the other hand, the financial and personnel cost of a false alarm is significantly higher when the perimeter security of a nuclear reactor is to be provided by deploying a SWSN to monitor unauthorized access. The cost of a false alarm might involve the transportation of special forces and/or personnel of related government agencies to the site, as well as the evacuation of residents in the surrounding area.

In simulations, an area with dimensions 100 m. $\times$ 10 m. is secured by a WSN. The grid size is taken as one meter so that the detection probabilities of targets on adjacent grid points are independent. The

boundary width is 10 m. The false alarm rate is set to 0.01, which is rather demanding on the network. Other nominal values are $\eta = 3, L = 100, \gamma = 30$ dB, $p_t = 0.9$ and $R = 31$. The results are the averages of 50 runs.

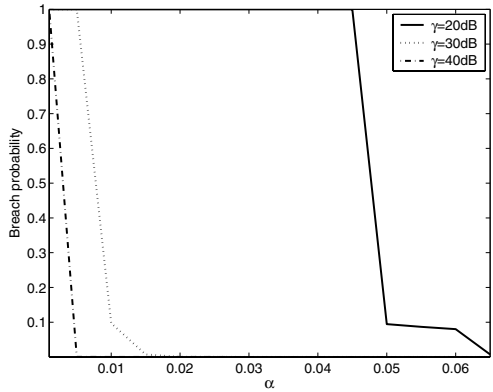


Figure 3. The effect of α on the breach probability.

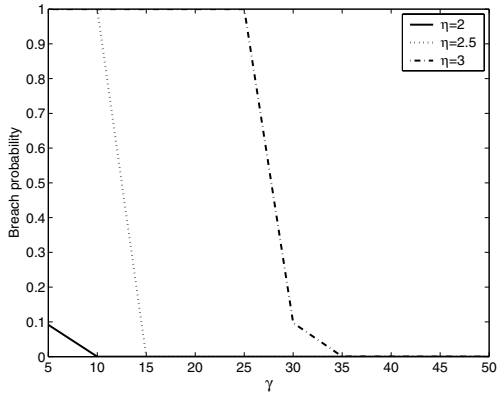


Figure 4. The effect of γ on the breach probability.

The breach probability P is quite sensitive to the false alarm rate α. As shown in Fig. 3b, as α increases, the SWSN allows more false alarms, and consequently, the NP detection probability and the detection probability p_v of the targets at grid point v both increase in α. Consequently, the breach probability decreases.

As the signal-to-noise ratio γ increases, the detection performance improves (see Fig. 4), and the breach probability decreases. In an obstacle-face environment, which corresponds to $\eta = 2$, and SNR level of $\gamma = 10$

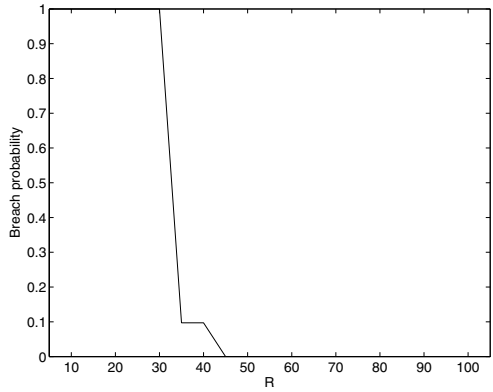


Figure 5. The effect of the number of sensor nodes on the breach probability for $y_v \sim \text{Uniform}(0, M-1)$.

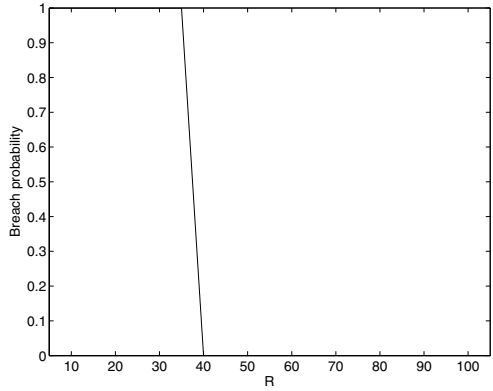


Figure 6. The effect of the number of sensor nodes on the breach probability for $y_v \sim \text{Normal}(M/2, N/10)$.

db is sufficient for satisfactory breach probability performance. Note that η and γ display a duality in that if one is fixed, the performance breaks down when the other parameter is below or above some value.

While analyzing the required number of sensor nodes for a given breach probability, we consider two cases of random deployment. In the first case, we assume that the sensor nodes are uniformly distributed along both the vertical and horizontal axes. In the second case, the sensor nodes are deployed uniformly along the horizontal axis and normally distributed along the vertical axis with mean $M/2$ and a standard deviation of 10% of the width of the field. The latter represents cases where the sensor nodes are deployed from an aircraft of a vehicle. In the

simulations, the sensor nodes that are deployed outside the field are not included in the computations of the detection probabilities.

Considering the uniformly distributed y-axis scheme, the required number of sensor nodes for a given breach probability is plotted in Fig. 5. A breach probability of 0.01 can be achieved by utilizing 45 sensors. Changing the false alarm rate to $\alpha = 0.1$, the requirement becomes 30 sensor nodes. The rapid decrease in the breach probability at $R = 16$ in Fig. 5a can be justified by the fact that most of the grid points are covered with high detection probabilities (saturated) for $R = 15$, and adding one more sensor node decreases the breach probability drastically. Once saturation is reached, placing more sensors in the field has marginal effect.

Analyzing Fig. 6, the above-mentioned saturation is seen more clearly for the normal-distributed y-axis scheme. For this kind of deployment, since the sensor node may fall outside the field, the breach probability decreases slower compared to the uniformly distributed y-axis scheme.

4. Data Processing Architecture for Target Tracking

In this section, we first define the process and observation models for target tracking. Then the foundations of the distributed data fusion architecture are presented.

Process Model

The target process is a four dimensional vector that consists of the two dimensional position of the target, (ξ, η), and the velocity of the target, $(\dot{\xi}, \dot{\eta})$, at each of these dimensions. The target process state vector is defined by

$$\mathbf{x} = [\xi \, \eta \, \dot{\xi} \, \dot{\eta}]^T, \tag{4}$$

and it evolves in time according to

$$\mathbf{x}(k+1) = \mathbf{F}\mathbf{x}(k) + \mathbf{v}(k)$$

where $\mathbf{x}(k)$ is the real target state vector at time k as given in (4), $\mathbf{F}$ is the process transition matrix, and $\mathbf{v}$ is the process transition noise.

Observation Model

Sensors can only observe the first two dimensions of the process. The velocity of the target is not observable by the sensors. Furthermore, sensors collect range and bearing data, but they cannot observe the coordinates of the target directly. Because sensors observe the target state

in polar coordinates, linear filtering formulations do not help. There are two implementation alternatives to remedy this problem: (1) by using the inverse transformation, obtain directly a *converted measurement* of the target position; (2) leave the measurement in its original form. The former yields a purely linear problem, allowing for linear filtering. The latter leads to a *mixed coordinate filter* [2]. In [21], the mean and covariance of the errors of Cartesian measurements, which are obtained by converting polar measurements, are derived. This conversion provides better estimation accuracy than the Extended Kalman Filter (EKF), in which the nonlinear target state measurements are utilized without conversion [21].

The measured range and bearing are defined with respect to the true range r and bearing θ as

$$\begin{aligned} r_m &= r + \tilde{r} \\ \theta_m &= \theta + \tilde{\theta} \end{aligned}$$

where the errors in range $\tilde{r}$ and bearing $\tilde{\theta}$ are assumed to be independent with zero mean and standard deviations σ_r and σ_θ, respectively.

The target mean state observed after the unbiased polar-to-Cartesian conversion is given by

$$\varphi^c = \begin{bmatrix} \xi_m^c \\ \eta_m^c \end{bmatrix} = \begin{bmatrix} r_m \cos\theta_m \\ r_m \sin\theta_m \end{bmatrix} - \mu$$

where μ is the average true basis

$$\mu = \begin{bmatrix} r_m \cos\theta_m (e^{-\sigma_\theta^2} - e^{-\sigma_\theta^2/2}) \\ r_m \sin\theta_m (e^{-\sigma_\theta^2} - e^{-\sigma_\theta^2/2}) \end{bmatrix}.$$

The covariances of the observation errors are [2, 21]

$$\begin{aligned} \mathbf{R}_{11} = r_m^2 e^{-2\sigma_\theta^2} &\left[\cos^2\theta_m(\cosh 2\sigma_\theta^2 - \cosh\sigma_\theta^2)\right. \\ &\left. + \sin^2\theta_m(\sinh 2\sigma_\theta^2 - \sinh\sigma_\theta^2)\right] \\ + \sigma_r^2 e^{-2\sigma_\theta^2} &\left[\cos^2\theta_m(2\cosh 2\sigma_\theta^2 - \cosh\sigma_\theta^2)\right. \\ &\left. + \sin^2\theta_m(2\sinh 2\sigma_\theta^2 - \sinh\sigma_\theta^2)\right], \end{aligned}$$

$$\begin{aligned} \mathbf{R}_{22} = r_m^2 e^{-2\sigma_\theta^2} &\left[\sin^2\theta_m(\cosh 2\sigma_\theta^2 - \cosh\sigma_\theta^2)\right. \\ &\left. + \cos^2\theta_m(\sinh 2\sigma_\theta^2 - \sinh\sigma_\theta^2)\right] \\ + \sigma_r^2 e^{-2\sigma_\theta^2} &\left[\sin^2\theta_m(2\cosh 2\sigma_\theta^2 - \cosh\sigma_\theta^2)\right. \\ &\left. + \cos^2\theta_m(2\sinh 2\sigma_\theta^2 - \sinh\sigma_\theta^2)\right] \end{aligned}$$

$$\mathbf{R}_{12} = \sin\theta_m \cos\theta_m e^{-4\sigma_\theta^2} \left[\sigma_r^2 + (r_m^2 + \sigma_r^2)(1 - e^{\sigma_\theta^2})\right].$$

Distributed Data Fusion Architecture

Information state $\mathbf{y}$ and the information matrix $\mathbf{Y}$ associated with an observation estimate $\hat{\mathbf{x}}$, and the covariance of the observation estimate $\mathbf{P}$ at time instant k are given by [16]

$$\begin{aligned} \hat{\mathbf{y}}(k) &= \mathbf{P}^{-1}(k)\hat{\mathbf{x}}(k), \\ \mathbf{Y}(k) &= \mathbf{P}^{-1}(k). \end{aligned}$$

In [16], it is shown that by means of sufficient statistics, an observation φ contributes $\mathbf{i}(k)$ to the information state $\mathbf{y}$ and $\mathbf{I}(k)$ to the information matrix $\mathbf{Y}$ where

$$\begin{aligned} \mathbf{i}(k) &= \mathbf{H}^T\mathbf{R}^{-1}(k)\varphi(k), \\ \mathbf{I}(k) &= \mathbf{H}^T\mathbf{R}^{-1}(k)\mathbf{H} \end{aligned} \tag{5}$$

and $\mathbf{H}$ is the observation matrix of the sensor.

Instead of sharing the measurements related to the target state among the collaborating sensors, sharing the information form of the observations results in a simple additive fusion framework that can be run on each of the tiny sensing devices. The distributed data fusion equations are

$$\hat{\mathbf{y}}(k \mid k) = \hat{\mathbf{y}}(k \mid k-1) + \sum_{i=1}^{N} \mathbf{i}_i(k), \tag{6}$$

$$\mathbf{Y}(k \mid k) = \mathbf{Y}(k \mid k-1) + \sum_{i=1}^{N} \mathbf{I}_i(k) \tag{7}$$

where N is the total number of sensors participating in the fusion process and $\hat{\mathbf{y}}(k \mid k-1)$ represents the information state estimate at time k given the observations up to end including time $k-1$.

Just before the data at time k are collected, if we are given the observations up to the time $k-1$, the predicted information state and the information matrix at time k can be calculated from

$$\begin{aligned} \hat{\mathbf{y}}(k \mid k-1) &= \mathbf{Y}(k \mid k-1)\mathbf{F}\mathbf{Y}^{-1}(k-1 \mid k-1)\hat{\mathbf{y}}(k-1 \mid k-1), \\ \mathbf{Y}(k \mid k-1) &= [\mathbf{F}\mathbf{Y}^{-1}(k-1 \mid k-1)\mathbf{F}^T + \mathbf{Q}]^{-1} \end{aligned}$$

where $\mathbf{Q}$ is the state transition covariance.

State estimate of the target at any time k can be found from

$$\hat{\mathbf{x}}(k \mid k) = \mathbf{Y}^{-1}(k \mid k)\hat{\mathbf{y}}(k \mid k). \tag{8}$$

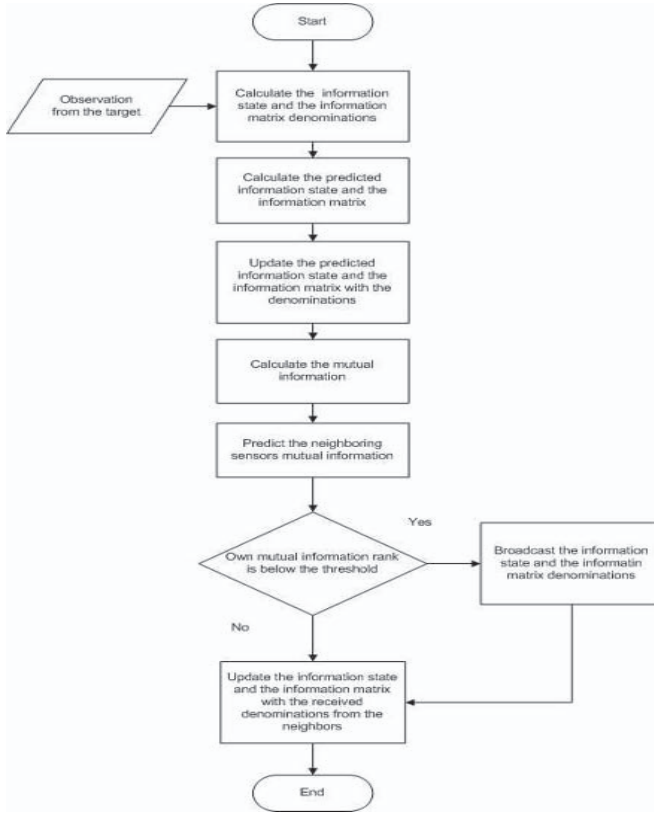


Figure 7. Target tracking algorithm for a sensor.

5. Maximum Mutual Information Based Sensor Selection Algorithm

Mutual information measures how much information one random variable tells about another one. In target localization and tracking applications, the random variables of interest are the target state and the observation obtained about the target state. By measuring the mutual information between the target state and the measurement, one can gain insight as to how much the current observation tells about the current target state.

The algorithm employed by a sensor for tracking targets in a collaborative manner within the distributed data fusion framework is depicted in Fig. 7. The information state and the information matrix are defined by (5). The predicted information state and the information matrix are computed by (7). The sensor's current belief is updated by its own

sensory observation according to

$$\begin{aligned} \hat{\mathbf{y}}(k \mid k) &= \hat{\mathbf{y}}(k \mid k-1) + \mathbf{i}(k), \\ \mathbf{Y}(k \mid k) &= \mathbf{Y}(k \mid k-1) + \mathbf{I}(k). \end{aligned}$$

Active participation to the current cycle is decided based on the mutual information gained with the last observation. This event can be formulated as

$$J(k, \varphi(k)) = \frac{1}{2} \log \left[\frac{|\mathbf{Y}(k \mid k)|}{|\mathbf{Y}(k \mid k-1)|} \right]. \tag{9}$$

Where $\mathbf{Y}(k \mid k)$ is the information matrix at the time instant k after the target state is observed. If the mutual information gain J of the sensor is sufficiently high to participate in the current cycle, the sensor shares its own information about the target state with the neighboring nodes. Otherwise, the sensor does not transmit during the current cycle. In (9), $\mathbf{Y}(k \mid k-1)$ denotes the predicted information matrix at the time instant k, given the observation up to the time instant $k-1$. Thus, the sensor has an estimate about the target state information that it will have at time instant k, before the observation of the target state at time instant k.

The mutual information in (9) measures the improvement in the target state estimate achieved with the observation. To decide if the mutual information is adequately high to participate in the current cycle, a sensor needs to know the mutual information values of its neighboring sensors. This information is hard to predict ahead of time. To tackle this problem, we design each sensor to hold a list of its neighboring sensors. The elements of this list are the sensor characteristics such as the standard deviation of the target range observations, standard deviation of the target angle observations, and the communication transmission power. Knowing the communication signal's transmission power of the neighboring sensor, it is easy to estimate the relative position of the neighboring sensor. This position estimation is done in a sliding window average of the last eight communications received from the neighboring sensor. With the sensors' own observation about the target state, it is again easy to estimate the $\mathbf{Y}(k \mid k)$ value of the neighboring sensor. $\mathbf{Y}(k \mid k-1)$ is the estimation of the cooperated information matrix. Given this information, the mutual information J for the neighboring sensors is estimated. All the neighboring sensors and the sensor itself are sorted according to the decreasing mutual information order. If the sensor detects the target, and the rank of its mutual information is lower than the maximum allowed number of sensors to communicate then the

sensor broadcasts its information state and the information matrix denominations to the network. The current belief is updated with the received information from the sensors in the vicinity according to (6). A current state estimate for the target can be found from (8).

6. Simulation Results

We run Monte Carlo simulations to examine the performance of the sensor selection algorithm based on the maximization of mutual information for the distributed data fusion architecture. We examine two scenarios: first is the sparser one, which consists of 50 sensors which are randomly deployed in the 200 m $\times$ 200 m area. The second is a denser scenario in which 100 sensors are deployed in the same area. All data points in the graphs represent the means of ten runs. A target moves in the area according to the process model described in Section 4. We utilize the Neyman-Pearson detector [20, 30] with $\alpha = 0.05, L = 100, \eta = 2$, 2-dB antenna gain, -30-dB sensor transmission power and -6-dB noise power.

The sensor tracks the target locally using the information form of the Kalman filtering [19] as described in Section 4. If the sensor does not detect a target, it updates its belief about the target state just by setting the Kalman filter gain to zero, which means that the sensor tracks the target according to the track history.

In collaborative information fusion, if a sensor is eligible to share its belief about the target state with other sensors, it broadcasts its information state and the information matrix. Sensors that receive these data according to the shadow-fading radio propagation model update their belief about the target state as described in Section 4. The shadow-fading radio propagation model assumes that the antenna heights are 10 cm., the shadow-fading standard deviation is 4, and the carrier frequency is 1.8 GHz. If the received communication signal from a sensor is below 15 dB, then the signal is treated as garbage.

In the simulations, we compare the mean squared error about the target localization for the collaborative tracking framework described in Section 4. We achieve maximum tracking accuracy when all sensors detecting the target participate in the distributed data fusion task. As the number of sensors allowed to participate in the fusion task is reduced, tracking quality deteriorates. This yields higher localization errors about the distributed target position estimations. However, a reduced number of sensors allowed to communicate yields a lower number of communication packets traveling in the network. Reduction in the number of sent packets affects the energy expenditure of the wireless sensor devices

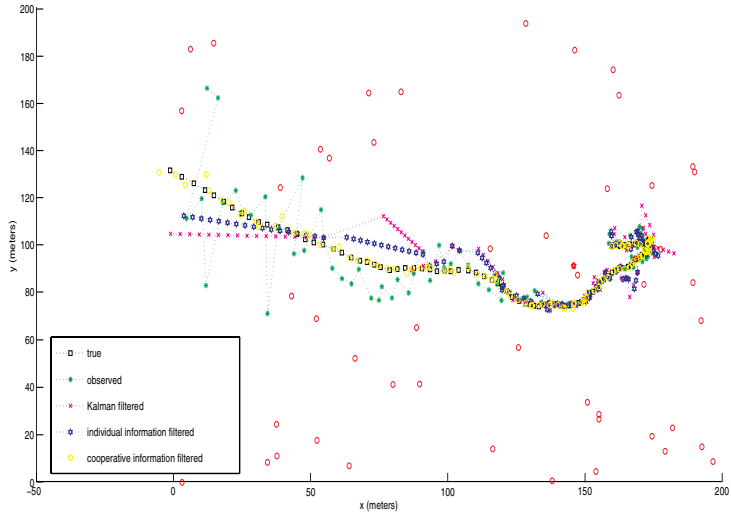


Figure 8. Illustration of the 50-sensor scenario.

directly. Selecting the sensors to actively participate in the fusion task in an intelligent manner improves tracking quality while allowing the same number of sensors to communicate. Figure 8 depicts the 50-sensor scenario, target location observation errors, Kalman-and-information-filtered target localization errors, and cooperative information-filtered target localization errors from the viewpoint of the sensor that is marked with a star inside it.

Selecting the participating active sensors randomly means that a sensor detecting the target broadcasts its information immediately if the maximum number of sensors to participate has not yet been reached. The minimum Mahalanabis distance-based sensor selection algorithm selects the closest sensors to the target location in terms of the Mahalanabis distance. Mahalanabis distance takes into account the correlations of the data. If the covariance matrix is the identity matrix, then Mahalanabis distance is the same as Euclidian distance. Figures 9 and 10 show, for the sparse and dense scenarios respectively, that as the maximum number of sensors allowed to communicate in the vicinity of the current cycle increases, total Mean Squared Error occurring throughout a hundred seconds-scenario decreases for all three sensor selection algorithms. Target localization errors are calculated each second. For the cases studied, selecting sensors which improve the global belief about the target position according to the mutual information measure results

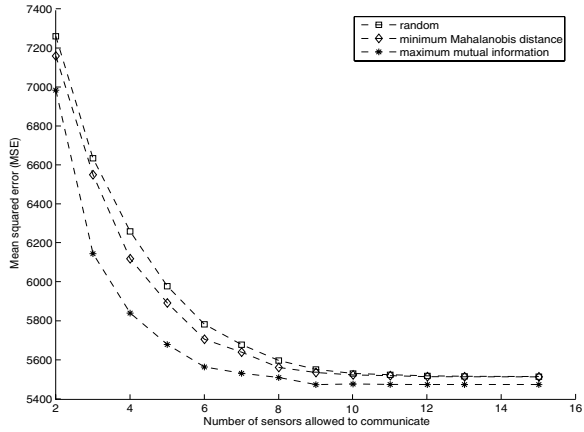


Figure 9. Mean squared error comparison for the sparse scenario.

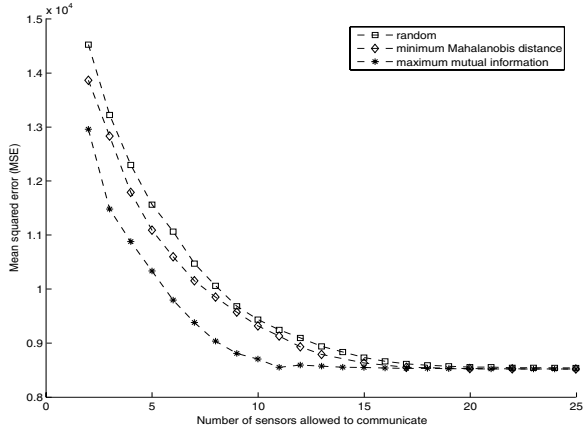


Figure 10. Mean squared error comparison for the dense scenario.

in an average of 4.07% improvement in tracking quality with respect to random sensor selection. 2.86%. A tracking quality improvement of 2.86% is achieved with respect to the maximum Mahalanabis distance based sensor selection for the sparse scenario. For the dense scenario of 100 sensors, these improvements with the mutual information-based sensor selection algorithm go to 9.65% and 6.32% with respect to the random and the Mahalanabis distance-based sensor selection algorithms, respectively.

Figures 11 and 12 depict the total energy exhausted in the network for all three sensor selection algorithms during the hundred second-scenario. Consumed energy increases as the maximum number of sensors that are

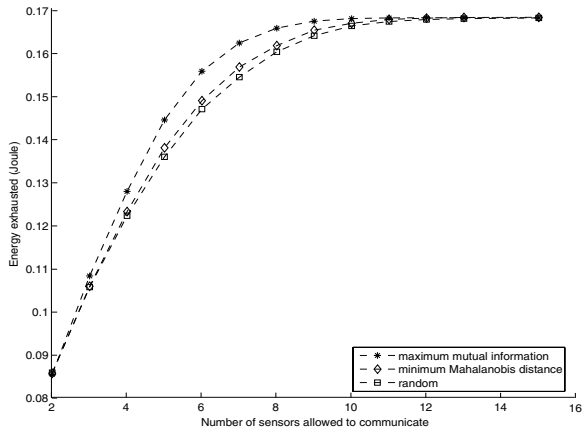


Figure 11. Comparison of the consumed energy for the sparse scenario.

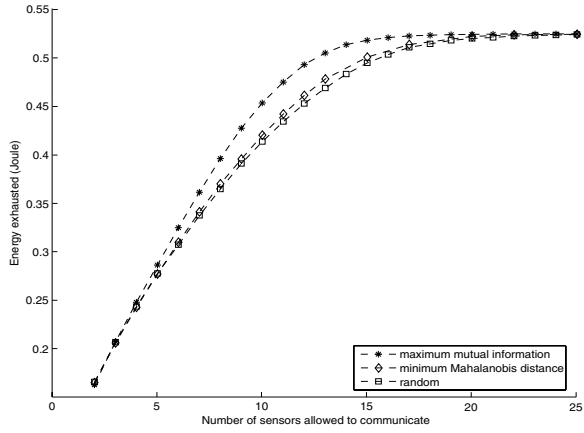


Figure 12. Comparison of the consumed energy for the dense scenario.

allowed to communicate for the current cycle increases. This is a natural result of the increasing number of communication packets in the network. However, the sensor selection algorithm does not have a significant effect on the exhausted energy of the network for any number of sensors allowed to communicate.

7. Conclusion

We employ the Neyman-Pearson detector to find the sensing coverage area of the surveillance wireless sensor networks. In order to find the breach path, we apply Dijkstra's shortest path algorithm by us-

ing the negative log of the miss probabilities as the grid point weights. The breach probability is defined as the miss probability of the weakest breach path. The false alarm rate constraint has a significant impact on the intrusion detection performance of the network, which is measured by the breach probability.

The model and results developed herein give clues that link false alarms to energy efficiency. Enforcing a low false alarm rate to avoid unnecessary response costs implies either a larger data-set (L) and hence a greater battery consumption, or a denser sensor network, which increases the deployment cost. Similar qualitative and/or quantitative inferences about the relationships between various other parameters can also be made.

Wireless sensor networks are prone to failures. Furthermore, the sensor nodes die due to their limited energy resources. Therefore, the failures of sensor nodes must be modeled and incorporated into the breach path calculations in the future. Simulating the reliability of the network throughout the entire life of the wireless sensor network is also required. Lastly, especially for perimeter surveillance applications, obstacles in the environment play a critical role in terms of sensing and must be incorporated into the field model.

A mutual information based information measure is adopted to select the most informative subset of sensors to actively participate in the distributed data fusion framework. The duty of the sensors is to accurately localize and track the targets. Simulation results show 36% energy saving for a given tracking quality can be achievable by selecting the sensors to cooperate according to the mutual information metric.

In all tests, we assumed that all the sensor nodes send reliable data to the network. In future work, detection of faulty and outlier sensors in the network must be investigated, and precautions need to be taken against them. We considered the effect of sensor selection algorithms in the context of distributed data fusion for tracking a single target. Existence of multiple targets introduces challenges with track-to-track association and track-to-sensor association, as well as issues related to access control and routing.

Acknowledgments: This work was supported by the State Planning Organization of Turkey under grant number 03K120250, and by the Boğaziçi University Research Fund under grant number 04A105.

References

[1] S. Appadwedula, V. V. Veeravalli, and D. L. Jones, "Robust and locally optimum decentralized detection with censoring sensors", in *Proceedings of the International Conference on Information Fusion*, Annapolis, USA, July 2002, pp. 56-63.

[2] Y. Bar-Shalom, X. R. Li, and T. Kirubarajan, *Estimation with Applications to Tracking and Navigation*, Wiley, 2001.

[3] M. S. Bazaraa and J. J. Jarvis, *Linear Programming and Network Flows*, Wiley, 1977.

[4] R. Bejar, B. Krishnamachari, C. Gomes, and B. Selman, "Distributed constraint satisfaction in a wireless sensor tracking system", in *Proceedings of the Workshop on Distributed Constraint Reasoning, International Joint Conference on Artificial Intelligence*, Seattle, USA, August 2001.

[5] R. R. Brooks, P. Ramanathan, and A. Sayeed, "Distributed target tracking and classification in sensor network", *Proceedings of the IEEE*, Vol. 91, No. 8, pp. 1163-1171, August 2003.

[6] R. R. Brooks and C. Griffin, "Traffic model evaluation of ad hoc target tracking algorithms", *International Journal of High Performance Computing Applications*, Vol. 16, No. 3, pp. 221-234, August 2002.

[7] R. R. Brooks, C. Griffin, and D. S. Friedlander, "Self-organized distributed sensor network entity tracking", *International Journal of High Performance Computing Applications*, Vol. 16, No. 3, pp. 207-219, August 2002.

[8] J. Carle and D. Simplot-Ryl, "Energy-efficient area monitoring for sensor networks", *IEEE Computer*, Vol. 37, No. 2, pp. 40-46, February 2004.

[9] M. Chu, H. Haussecker, and F. Zhao, "Scalable information-driven sensor querying and routing for ad hoc heterogenous sensor networks," *International Journal of High Performance Computing Applications*, Vol. 16, No. 3, pp. 293-313, August 2002.

[10] V. Chvatal, "A combinatorial theorem in plane geometry," *Journal of Combinatorial Theory*, Vol. B, No. 13, pp. 39-41, 1975.

[11] T. Clouqueur, V. Phipatanasuphorn, P. Ramanathan and K. K. Saluja, "Sensor deployment strategy for detection of targets traversing a region," *Mobile Networks and Applications*, Vol. 8, No. 4, pp. 453-461, August 2003.

[12] S. S. Dhillon and K. Chakrabarty, "Sensor placement for effective coverage and surveillance in distributed sensor networks," in *Proceedings of the IEEE Wireless Communications and Networking Conference*, New Orleans, USA, March 2003, pp. 1609-1614.

[13] A. Elfes, "Occupancy grids: a stochastic spatial representation for active robot perception," in *Autonomous Mobile Robots: Perception, Mapping and Navigation, Vol. 1*, S. S. Iyengar and A. Elfes, Editors, IEEE Computer Society Press, 1991, pp. 60-70.

[14] E. Ertin, J. W. Fisher, and L. C. Potter, "Maximum mutual information principle for dynamic sensor query problems", in *Proceedings of the 2nd International Workshop on Information Processing in Sensor Networks*, Palo Alto, USA, April 2003, pp. 405-416.

[15] Q. Fang, F. Zhao, and L. Guibas, "Counting targets: Building and managing aggregates in wireless sensor networks", Palo Alto Research Center (PARC), Tech. Rep. P2002-10298, June 2002.

[16] B. Grocholsky, A. Makarenko, and H. Durrant-Whyte, "Information-theoretic coordinated control of multiple sensor platforms," in *Proceedings of the IEEE International Conference on Robotics and Automation*, Taipei, Taiwan, September 2003, pp. 1521-1526.

[17] R. Jiang and B. Chen, "Decision fusion with censored sensors", *Proceedings of ICASSP*, Vol. 2, Montreal, Canada, May 2004, pp. 289-292.

[18] T. He, S. Krishnamurthy, J. A. Stankovic, T. Abdelzaher, L. Luo, R. Stoleru, T. Yan and L. Gu, "Energy-efficient surveillance system using wireless sensor networks," *Proceedings of the Second International Conference on Mobile Systems, Applications, and Services*, Boston, USA, June 2004, pp. 270-283.

[19] R. E. Kalman, "A new approach to linear filtering and prediction problems," *Transactions of the ASME-Journal of Basic Engineering*, Vol. 82, pp. 35-45, 1960, series D.

[20] D. Kazakos and P. Papantoni-Kazakos, *Detection and Estimation*, New York, USA: Computer Science Press, 1990.

[21] D. Lerro and Y. Bar-Shalom, "Tracking with debiased consistent converted measurements versus EKF", *IEEE Transactions on Aerospace and Electronic Systems*, Vol. 29, No. 3, pp. 1015-1022, July 1993.

[22] D. Li, K. D. Wong, Y. Hu, and A. M. Sayeed, "Detection, classification, tracking of targets in micro-sensor networks", *IEEE Signal Processing Magazine*, Vol. 19, No. 2, pp. 17-29, March 2002.

[23] X.-Y. Lin, P.-J. Wan and O. Frieder, "Coverage in wireless ad hoc sensor networks," *IEEE Transactions on Computers*, Vol. 52, No. 6, pp. 753-763, June 2003.

[24] J. Liu, P. Cheung, L. Guibas, and F. Zhao, "A dual-space approach to tracking and sensor management in wireless sensor networks", in *The First ACM International Workshop on Wireless Sensor Networks and Applications*, Atlanta, USA, September 2002.

[25] S. Megerian, F. Koushanfar, G. Qu, G. Veltri and M. Potkonjak, "Exposure in wireless sensor networks: theory and practical solutions," *Wireless Networks*, Vol. 8, No. 5, pp. 443-454, September 2002.

[26] D. P. Mehta, M. A. Lopez and L. Lin, "Optimal coverage paths in ad-hoc sensor networks," in *Record of the IEEE International Conference on Communications*, Anchorage, USA, May 2003, pp. 507-511.

[27] J. Moore, T. Keiser, R. R. Brooks, S. Phoha, D. Friedlander, J. Koch, A. Reggio, and N. Jacobson, "Tracking targets with self-organizing distributed ground sensors", in *Proceedings of the IEEE Aerospace Conference*, Vol. 5, Big Sky, USA, March 2003, pp. 2113-2123.

[28] E. Onur, C. Ersoy and H. Deliç, "Quality of deployment in surveillance wireless sensor networks", *International Journal of Wireless Information Networks*, Vol. 12, No. 1, pp. 61-67, January 2005.

[29] E. Onur, C. Ersoy and H. Deliç, "How many sensors for an acceptable breach probability level?", *Computer Communications*, Special Issue on Dependable Sensor Networks, in press 2005.

[30] E. Onur, C. Ersoy and H. Deliç, "Sensing coverage and breach paths in surveillance wireless sensor networks", in *Sensor Network Operations*, S. Phoha, T. F. La Porta and C. Griffin, Editors, IEEE Press, 2005.

[31] S. Pattern, S. Poduri and B. Krishnamacharie, "Energy-quality tradeoffs for target tracking in wireless sensor networks", in *Proceedings of Information Processing in Sensor Networks*, Palo Alto, USA, April 2003, pp. 32-36.

[32] N. Patwari and A. O. Hero, "Hierarchical censoring for distributed detection in wireless sensor networks", *Proceedings of IEEE ICASSP*, Vol. 4, Hong Kong, April 2003, pp. 848-851.

[33] C. Rago, P. Willett, and Y. Bar-Shalom, "Censoring sensors: A low communication-rate scheme for distributed detection", *IEEE Transactions on Aerospace and Electronic Systems*, Vol. 32, No. 4, pp. 554-568, April 1996.

[34] P. Ramanathan, "Location-centric approach for collaborative target detection, classification, and tracking", in *Proceedings of the IEEE CAS Workshop on Wireless Communications and Networking*, Pasadena, USA, September 2002.

[35] D. Tian and N. D. Georganas, "A coverage-preserving node scheduling scheme for large wireless sensor networks," *Proceedings of the First ACM International Workshop on Wireless Sensor Networks and Applications*, Atlanta, USA, September 2002, pp. 32-41.

[36] D. Tian and N. D. Georganas, "A node scheduling scheme for energy conservation in large wireless sensor networks", *Wireless Communications and Mobile Computing*, Vol. 3, No. 2, pp. 271-290, May 2003.

[37] S. Tilak, N. B. Abu-Ghazaleh and W. Heinzelman, "Infrastructure tradeoffs for sensor networks," *Proceedings of the First ACM International Workshop on Wireless Sensor Networks and Applications*, Atlanta, USA, September 2002, pp. 49-58.

[38] H. Wang, K. Yao, G. Pottie, and D. Estrin, "Entropy-based sensor selection heuristic for target localization," in *Proceedings of the Third Symposium on Information Processing in Sensor Networks*, Berkeley, USA, April 2004, pp. 36-45.

[39] X. Wang, G. Xing, Y. Zhang, C. Lu, R. Pless and C. Gill, "Integrated coverage and connectivity configuration in wireless sensor networks," in *Proceedings of the First International ACM Conference on Embedded Networked Sensor Systems*, Los Angeles, USA, November 2003, pp. 28-39.

[40] M. A. Weiss, *Data Structures and Algorithm Analysis in C++*, 2nd Edition, Addison-Wesley, 1999.

[41] T. Yan, T. He and J. A. Stankovic, "Differentiated surveillance for sensor networks," SENSYS 2003.

[42] F. Ye, G. Zhong, S. Lu and L. Zhang, "Peas: A robust energy conserving protocol for long-lived sensor networks," *Proceedings of the 23rd International Conference on Distributed Computing Systems*, Providence, USA, May 2003, pp. 28-37.

[43] H. Zhang and C.-J. Hou, "On deriving the upper bound of α-lifetime for large sensor networks," *Technical Report UIUCDCS-R-2004-2410*, University of Illinois at Urbana-Champaign, Department of Computer Science, February 2004.

[44] F. Zhao, J. Shin and J. Reich, "Information-driven dynamic sensor collaboration for tracking application," *IEEE Signal Processing Magazine*, Vol. 19, No. 1, pp. 61-72, March 2002.

[45] Y. Zou and K. Chakrabarty, "Sensor deployment and target localization based on virtual forces," *Proceedings of the IEEE INFOCOM*, San Francisco, USA, April 2003, pp. 1293-1303.

INTERNET-SCALE CHEMICAL SENSING: IS IT MORE THAN A VISION?

Dermot Diamond
AIC Adaptive Sensors Group,
National Centre for Sensor Research,
School of Chemical Sciences,
Dublin City University, Dublin 9, Ireland
dermot.diamond@dcu.ie

Abstract In order to realise scalability of chemical sensors in extensively deployed wireless sensor networks, considerable materials challenges must be overcome. Conventional devices are currently far too expensive and unreliable for massive long-term field deployment. Cost can be driven down by imaginative approaches to transduction and instrument design. For example, we have produced a complete instrument based on LED measurement of colour changes that has sub-micromolar detection limits for a number of heavy metals for around $1.2 In its current form, the device also has a short distance wireless communications functionality and very low power consumption. However, chemical sensors capable of long-term reliability will require imaginative solutions to the key issue — how can the sensing films/membranes in chemical sensors maintain predictable characteristics in long term deployment?

The vision of 'internet-scale sensing' will only be realised through advances in materials science and a complete rethink of how we do chemical sensing. For example, fully autonomous sensing platforms must be completely self-reliant in terms of power, communications, reagents and consumables. The sensor network must be self-sustaining, meaning that as individual nodes become unreliable, new nodes are established, for example through physical replacement or through devices capable of self-repair/regeneration. In this chapter, these issues are presented, along with some recent advances mentioned above.[1]

Keywords: chemical sensors; biosensors; CWA; BWA; sensor nets; wireless; ad-hoc networks.

[1]This is an extended version of an article previously published in the journal Analytical Chemistry, August 1, 2004, 75 (15), 278A-286A. Copyright 2004 American Chemical Society.

Jim Byrnes and Gerald Ostheimer (eds.),
Advances in Sensing with Security Applications, 121–146.

1. Introduction

Digital communications networks are at the heart of modern society. The digitisation of communications in general, the development of the internet, and the availability of relatively inexpensive but powerful mobile computing technologies have together established a global communications network capable of instantly linking billions of people, places and objects. Complex documents can be instantly transmitted to multiple remote locations, and web sites provide a platform for real-time notification, archiving, dissemination and exchange of information on a global basis. This technology, in its many guises, is rapidly becoming completely pervasive, and the average person now has multiple associations with this digital world on a daily basis. However, this is only the foundation for the next wave of development — one that will provide a seamless interface between the 'molecular' and 'digital' worlds. The crucial missing part is the gateway through which these worlds will communicate — how can the digital world look into the molecular world and become truly self-aware?

We are currently in the midst of a global technological revolution driven by the internet and the 'WEB'. Incredible advances in digital video and audio technologies, coupled with equally astonishing breakthroughs in digital communications and computer power, have changed our lives in a profound manner, affecting almost every aspect of modern society. But what lies ahead? How will science and technology converge to catalyse the next stage in societal change? Paul Saffo, interviewed by Michio Kaku, states that 'In the 21st Century, the next revolution will be driven by cheap sensors linked to microprocessors and lasers' [1].

The move from traditional analogue land-line to digital mobile phones has been an important part of this communications revolution. Inexpensive GSM and 3G mobile phones, and increasingly other wireless communications technologies such as 802.11 wireless LAN (Local Area Network), coupled with palmtop PCs and PDAs, provide individuals with communications capabilities that would have been almost unimaginable a decade ago. The exchange of data files containing text, graphics, and embedded video/audio in real time, using mobile communications platforms, is now standard practice.

In recent years, research into wireless networking has been dominated by the perceived need for high bandwidth access to data intensive files with extensive graphical, video and audio content. Bluetooth has been heralded as the low-power wireless standard of the future, but it is very much a high-bandwidth technology, designed to integrate portable devices into this communications infrastructure, and is relatively power

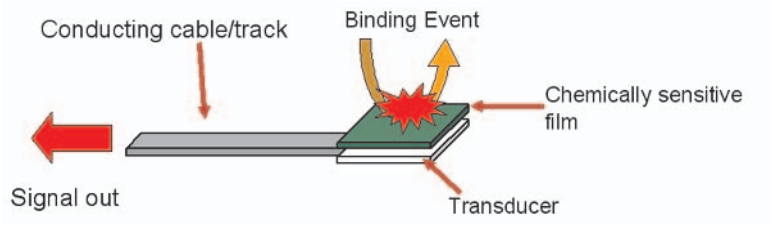


Figure 1. Stylised chemical sensor comprising a conducting cable or track to convey the electronic signal to the outside world, a transducer to sense the chemical signal and convert it into an electronic form, and a chemically sensitive film or membrane at which the molecular binding event occurs.

hungry. Similarly, wireless LAN (e.g. 802.11 standard) and particularly GSM mobile phones can provide personal gateways between portable computing and digital communications.

2. Chemical Sensing and Biosensing

Sensor research is driven by the need to generate a selective response to a particular analyte, for example, by a selective binding event such as occurs in host-guest complexation, enzyme-substrate reactions, antibody-antigen interactions, or other forms of biomolecular recognition. Hence many research papers are focused on developing a fuller understanding of the molecular basis for intra-molecular recognition, as this may ultimately lead to more selective devices that may find use in various futuristic applications. Coupled with this attention to selectivity is the need to provide a transduction mechanism, so that the binding event can be 'observed' from the outside via an electronic signal. Researchers typically will look to electrochemistry (e.g. potentiometry or voltammetry/amerometry) or spectroscopy (e.g. visible absorbance or fluorescence) for this signal. Often this is achieved via the presence of appropriate redox active sites or chromophores/fluorophores either as part of the molecular sensor itself, or as part of a sensing 'cocktail'. Success is dependent on the molecular binding event triggering transduction of the signalling moiety without adversely affecting the overall selectivity of the binding process.

Figure 1 shows the main features of a generic chemical sensor. The molecular binding event happens at a film or surface that contains sites designed to selectively interact with specific molecular targets. The binding event results in an electronic or optical signal that can be detected remotely and signalled to the outside world, and this signal conveys information about the molecular events occurring at this surface. Biosensors

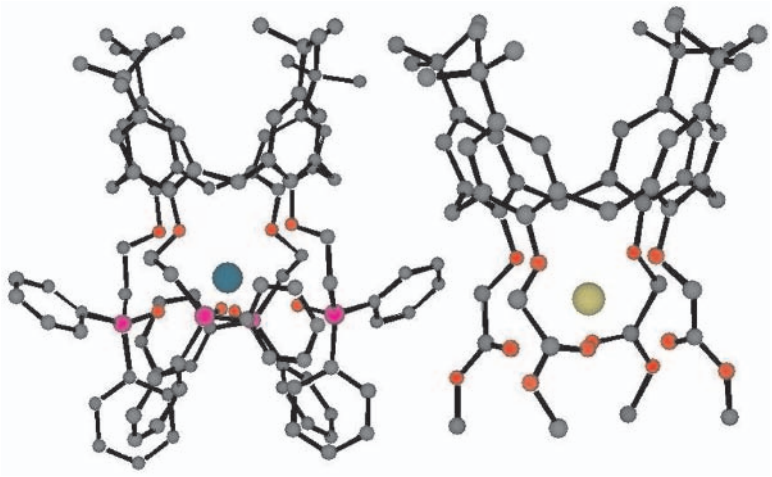


Figure 2. 3-d Energy minimised structures of t-butyl calix[4]arene molecular receptors with hydrogen atoms removed for clarity. Left is a sodium selective tetraester receptor and right is a calcium selective tetraphosphine oxide receptor. Carbon atoms grey, oxygen atoms red, phosphorus atoms pink. Sodium and calcium ions are shown in the energy minimised position within the negatively polar cavity in each case.

are similar, except that the binding events at the sensor-sample interface are generated by bio-receptors such as enzymes or antibodies in place of the chemo-receptors.

The 3-d structures of two 'calixarene' molecular receptors are illustrated in Figure 2. They contain certain structural features that make them very attractive for use in chemical sensors [2]. On the left is a so-called 'tetraester', which is highly selective for sodium ions, and on the right, a calcium selective 'tetraphosphine oxide'. The 'tetra' part of the name tells us that there are four phenoxy repeat units in both cases that define the calixarene macrocyclic cavity. At the upper end of the structures there are t-butyl groups, and these, along with the phenyl groups render the receptors highly insoluble in water. The receptors differ in the binding groups substituted at the phenoxy oxygens at the bottom of the calixarene annulus, and these, together with the size of the cavity defined by the number of repeat units, to a large extent determine the selectivity of the ion binding behaviour of the resulting ligand. The tetraester (left) has four ester groups that define a relatively rigid cavity with the negatively polar oxygen atoms (and particularly the carbonyl oxygens) ideally suited for binding sodium ions. In contrast, substituting phosphine oxide binding groups results in a calcium selective receptor.

Sensors (ion-selective electrodes) incorporating these receptors are typically based on lipophilic membranes made from highly plasticized

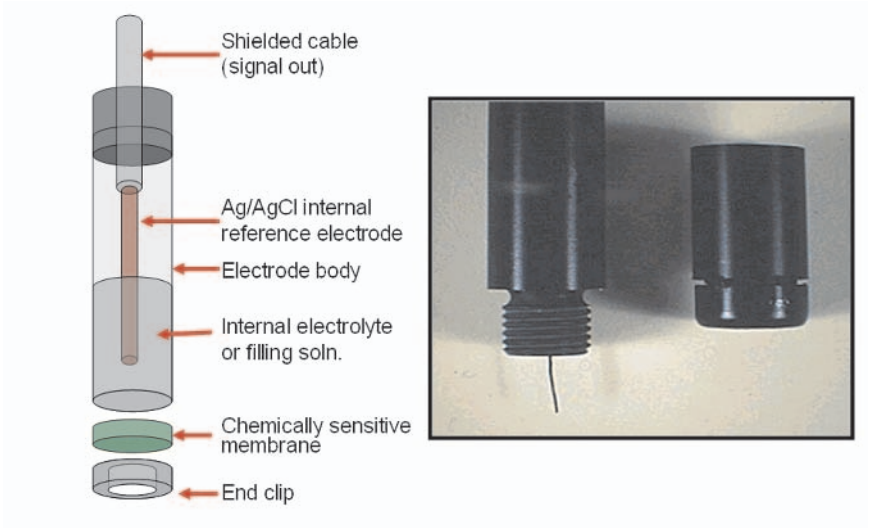


Figure 3. Components of an ion-selective electrode chemical sensor (left) and photographs of electrode body (right) showing electrode barrel with silver-silver chloride electrode, and screw-on electrode tip with end-clip for attaching the PVC membrane containing immobilised molecular receptors that will selectively bind specific target species.

PVC (60-70% plasticizer by weight), in which the receptors are effectively trapped in the membrane when exposed to aqueous samples. The mechanism of response is potentiometric; i.e. a galvanic cell potential is measured at zero current, and this is related to the ion-binding behaviour occurring at the membrane-sample interface. Selectivity is critical, as it is this that enables the user to relate the observed electrochemical signal to binding with the specific target ion. The sensors are made by attaching a membrane to the end of a hollow plastic tube which is filled with an internal reference electrolyte whose composition does not change, commonly a 0.1 M solution of the primary (target) ion chloride (see Figure 3). The membrane potential is sensed by a silver-silver chloride reference electrode (transducer) and the signal measured by a high impedance voltmeter. An external reference electrode completes the galvanic cell. The function of the reference electrode is to provide a stable, constant reference potential against which changes in the ion-selective electrode signal can be measured.

Until recently, it was accepted that the fundamental limit of detection of these sensors was at micromolar levels of the target ion in an aqueous sample, and the main application has been the determination of ions like sodium, potassium and calcium in blood samples, where the

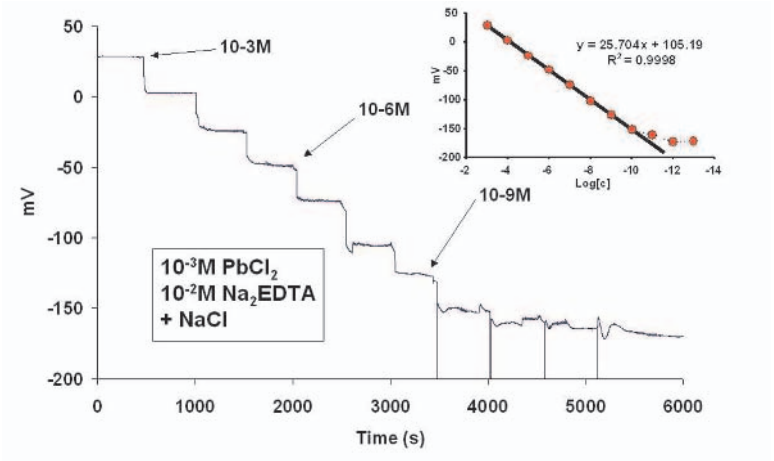


Figure 4. Response of an lead-selective electrode based on a calix[6]arene hexaphosphene oxide to sequential 10-fold dilutions of a sample solution demonstrating a very rapid Nernstian response down to sub-nanomolar concentrations of lead. The inset shows a linear Nernstian plot is obtained with almost theoretical slope (25.7 mV per decade) down to 10-10 M.

concentrations are relatively high. However, recent developments in the understanding of these sensors have led to breakthroughs in the operational characteristics, and sub-nanomolar measurements have been reported [3]. This has opened the door to many new potential applications for these low cost sensors, for example in environmental analysis, where metal ion concentrations are often sub-micromolar. Figure 4 shows results we have obtained with a t-butylcalix[6]arene hexaphosphine oxide based electrode in which the receptor is equivalent to the tetraphosphine oxide ligand illustrated in Figure 2, except that is has six repeat units that define a larger macrocyclic cavity, and results in a ligand highly selective for lead ions. Through careful optimisation of the inner filling solution (reference electrolyte) it is possible to inhibit transfer of primary ions from the typically highly concentrated inner solution to the outer membrane sample boundary region, and it is this process that masks the response of the sensor to sub-micromolar concentrations of the primary ion in aqueous samples. For example, in Figure 4 the response of a lead-selective electrode to sequential 10-fold dilutions of the primary ion is shown, and a Nernstian (theoretical) signal is obtained to sub-nanomolar lead ion concentrations (inset). In this case, the concentration of lead ions in the internal filling solution is reduced to 1.0 mM, and the free lead ion concentration further reduced using excess EDTA.

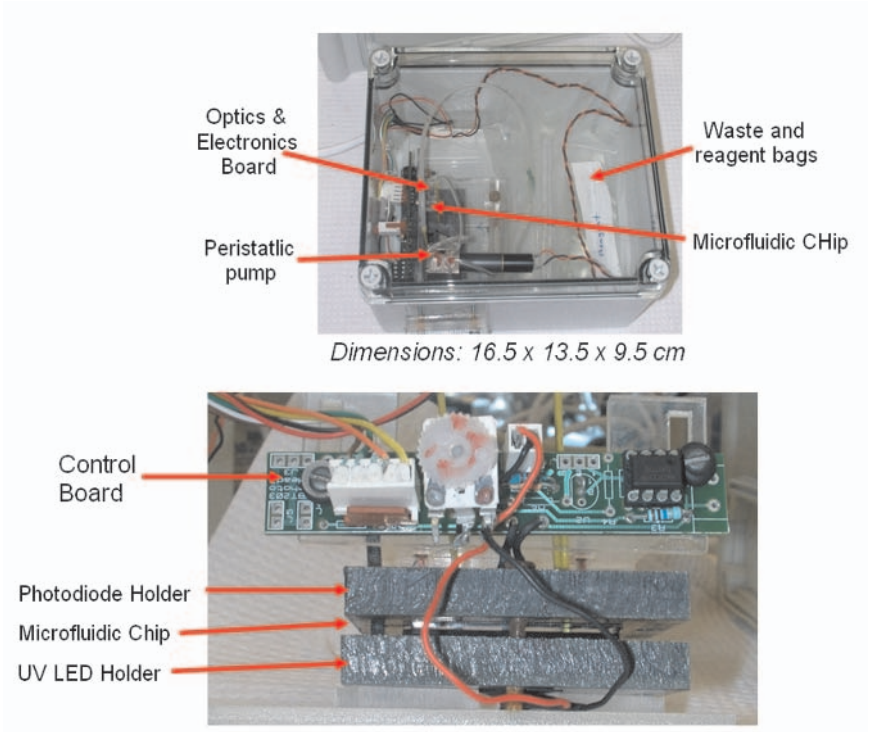


Figure 5. Photograph of an autonomous lab-on-a-chip system (top) configured for remote field monitoring of phosphorus in natural waters. Bottom is a closeup of the detector area of the system.

3. Miniaturised Analytical Instruments — Lab on a Chip Devices

In parallel with improvements in chemical sensor performance, analytical science has also seen tremendous advances in the development of compact, portable analytical instruments. For example, lab-on-a-chip (LOAC) devices enable complex bench processes (sampling, reagent addition, temperature control, analysis of reaction products) to be incorporated into a compact, device format that can provide reliable analytical information within a controlled internal environment. LOAC devices typically incorporate pumps, valves, micromachined flow manifolds, reagents, sampling system, electronics and data processing, and communications. Clearly, they are much more complex than the simple chemo-sensor described above. In fact, chemosensors can be incorporated into LOAC devices as a selective sensor, which enables the sensor to be contained within the protective internal environment. Figure 5

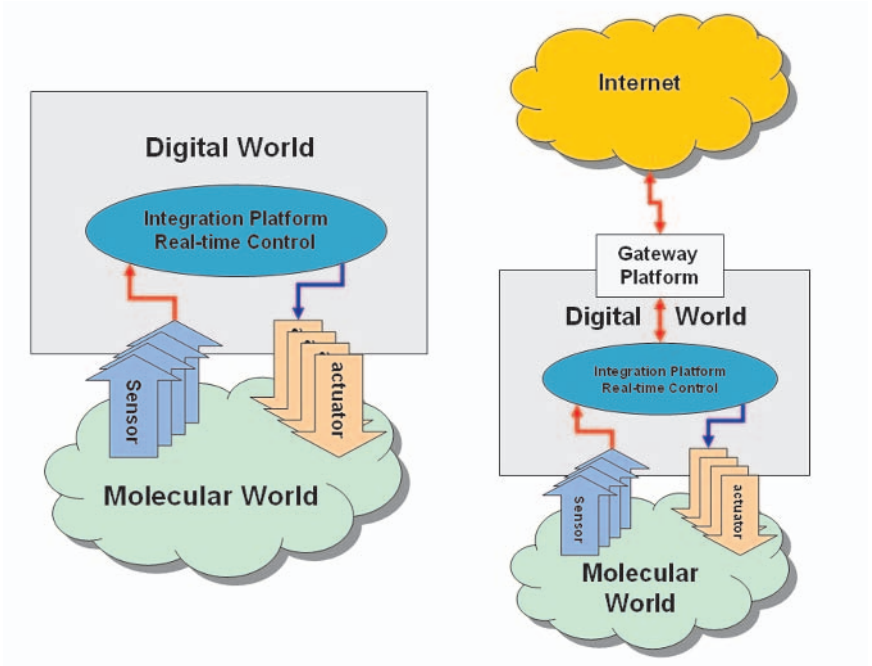


Figure 6. Left: Conventional control loops provide a localised interface between the real and the digital world. Sensors targeted at important control parameters feed information into digital control routines that can respond via actuators e.g. to maintain parameters within specified limits. Right: A vital step on the route to the realisation of the concept of internet scale sensing is to adopt the principle that all analytical measurements should be capable of being internet-linked. The localised control of important parameters is maintained, but the information is shared via the internet with external users [4]. (Reprinted with permission from Anal. Chem., August 1, 2004, 75 (15), 278A-286A. Copyright 2004 American Chemical Society.)

shows a relatively simple analytical device configured for performing autonomous optical (colorimetric) reagent based measurements. In this case, the sample is accessed via a microdialysis membrane and is drawn into a microfluidic manifold where it mixes with a reagent cocktail that generates a colour if a specific target species is present in the sample. The colour of the sample is measured with a photodiode using a light-emitting diode (LED) that has an emission maximum coinciding with the absorbance spectrum of the generated colour. The analytical literature is awash with papers on chemical sensors, biosensors and microfluidic (LOAC) instruments. However, for reasons that will be outlined below, these devices, unlike physical transducers, have not really been combined into extensive 'sensor networks'.

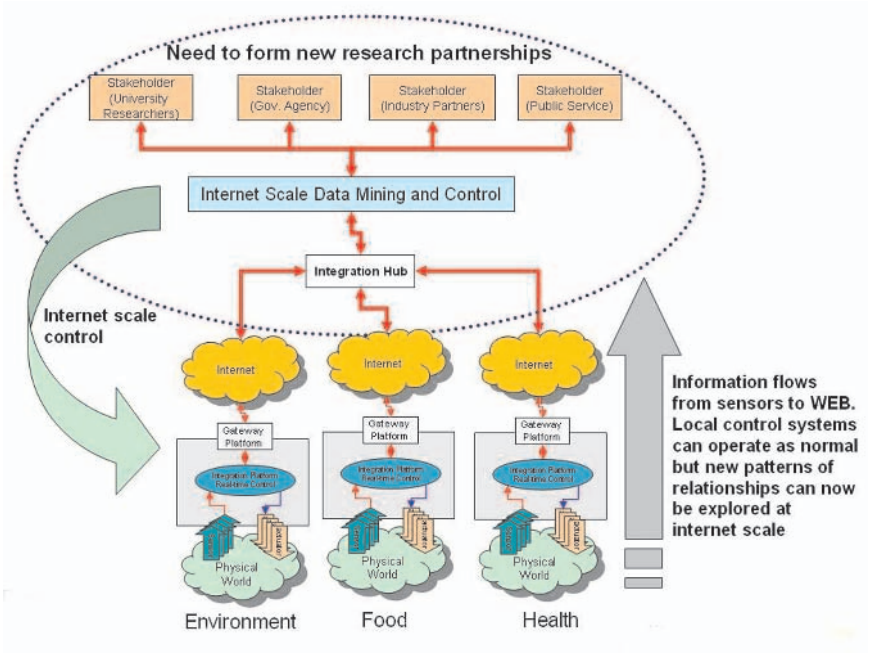


Figure 7. Widespread implementation of internet-enabled analytical measurements leads to Internet Scale Sensing, a new and powerful vision that links various application sectors (environment, security, health) and a wide variety of users [4]. (Reprinted with permission from Anal. Chem., August 1, 2004, 75 (15), 278A-286A. Copyright 2004 American Chemical Society.)

Internet Scale Sensing and Control

Internet scale sensing [4] conceptually involves large scale deployment of large numbers of sensors or sensing devices into wide area networks, sometimes referred to as sensor nets or sensor meshes. Figure 6 illustrates the basic principle. A conventional control loop consisting of one or more sensors and actuators (left) becomes part of a global information exchange through implementation of the basic principle of internet enabling every analytical measurement. This immediately enables external browsing of sensor status, external programming of control parameters, and feedback of information to individuals and to other devices. Widespread adoption of this principle leads to the emergence of 'internet-scale sensing and control systems', in which millions of sensing devices and actuators are linked in a seamless manner with a wide variety of users, ranging from individuals, to Government agencies, industrial users or public service providers, across many application sectors (Figure 7). However, the real value lies in the realisation that large-scale sensor networks provide much more information than is predictable from simple linkages between localised collections of individual sensors. In fact, there is tremendous potential for the discovery and use of entirely new types of relationships between information extracted from this data 'continuum', which will give rise to new business opportunities, and the emergence of completely new markets. This has been identified by Ron Ambrosio and Alex Morrow, from IBM's Watson Research Centre, as one of the key developments that will fuel rapid developments in the Information and Communications Technology (ICT) sector in the coming years [5]. Other major players are moving into this research space. Intel has established the Intel-Berkeley Research Lab to develop sensor 'mote' technology — motes are low power communications platforms with integrated sensors (temperature, light etc.), and built-in capabilities for adding other sensors.

Finding field demonstrators that are scientifically interesting, field deployable and available at a cost appropriate for the envisaged scale is a challenge. The Berkeley team has assembled a 50-node sensor network to monitor seismic activity across the campus, and a 32 node sensor network linked by satellite communications via a basestation has been used to study the microclimates associated with the nesting sites of storm petrels on Great Duck Island, Maine. Each node included a habitat monitoring kit that could monitor light levels, heat, temperature, barometric pressure and humidity [6, 7].

At UCLA, Deborah Estrin is heading up the 'Centre for Embedded Networked Sensors' or CENs, a $40 million, 10-year NSF-backed centre

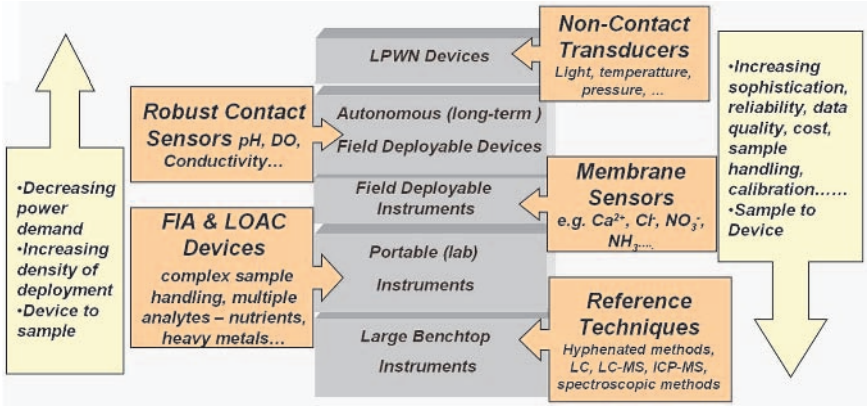


Figure 8. Analytical instruments can be arranged into an hierarchy in terms of highly correlated factors such as sophistication, capabilities, operational costs, and degree of autonomy. There will be a significant correlation between these factors and density of distribution throughout the networked world. Providing effective communications between these layers provides routes to validating data from low cost devices using more reliable data obtained from sophisticated devices [4]. (Reprinted with permission from Anal. Chem., August 1, 2004, 75 (15), 278A-286A. Copyright 2004 American Chemical Society.)

[8]. Parallel to this, the European Union has invested EUR 111 million in wireless research under the 6th Framework programme [9], and the indications are that this will be significantly increased in the forthcoming 7th Framework programme. The scale and extent of investment and research activity is a clear signal that this area is now receiving priority attention from academic researchers, research agencies and industry.

4. Analytical Device Hierarchy

Like the digital communications industry, analytical devices can be layered into an hierarchy in terms of their complexity, degree of autonomy and need for external services (Figure 8). Lab based instruments are already heavily integrated into conventional digital networks, usually as part of a site-based network. In principle, this information is readily available, but in practice, it tends to be restricted to the site, as it typically needs sophisticated work up and interpretation. Likewise, field based instruments, and particularly devices employed in an autonomous manner, have an immediate demand for integration into digital networks, but as these are placed increasingly at more remote and less serviced situations, conventional networking strategies become less feasible. As this trend ramps up, and the numbers of devices oper-

ated in an autonomous manner increases, the need for low power operation becomes rate determining. Clearly, the least sophisticated of these devices will be distributed in the largest numbers, and will essentially operate on almost zero bandwidth (e.g. they may send only a few bits information to indicate a threshold has been crossed). Low bandwidth is of course very unattractive to the communications industry, as their business model tends to be based on the volume of data transferred (hence the attractiveness of large audio and video files). In contrast to data quantity, the attractiveness of wireless sensor networks lies in the importance and value of the information they can provide.

At present, research into the most densely distributed layer (very low cost, autonomous devices) is dominated by the use of physical transducers such as pressure and temperature sensors that do not have to make an intimate contact with the sample/environment (i.e. they can be totally encased within a protective cladding and still function). Introduction of chemical sensing capabilities is happening through the use of 'old-reliables' such as pH and dissolved oxygen (DO) sensors, as these are known to be fairly robust. Sensors that depend on polymer membranes or surface films for response generation will be affected by exposure to the sample over time, so more sophisticated devices that incorporate calibration routines are typical, which drives up the cost base, making dense distribution economically unviable. This is the paradox at the heart of the chemical sensor/biosensor failure to demonstrate large scale sensor networks — we need reactive surfaces to generate the analytical signal and provide the molecular information, but we also want these surfaces to remain unchanged over long periods of exposure to the real world, in order to enable simple, low cost, calibration free measurements to be made.

Lab-on-a-chip (LOAC) approaches in which the analytical measurement is performed within a sheltered microfluidic environment is an attractive option, but these are also too costly to deploy in large numbers at present (see discussion below). The challenge for analytical scientists is to move devices towards the more densely distributed layers by driving down the cost base while simultaneously maintaining reliability and quality of the analytical data. Clearly, there will be increasing use of physical transducers and reliable chemical sensors in networked systems. Relatively simple measurements such as turbidity, colour, pH and conductivity can provide important general information about water quality throughout a complex distribution network, enabling contamination to be detected at an early stage, and corrective action to be taken before contaminants spread throughout the entire system. The success of these

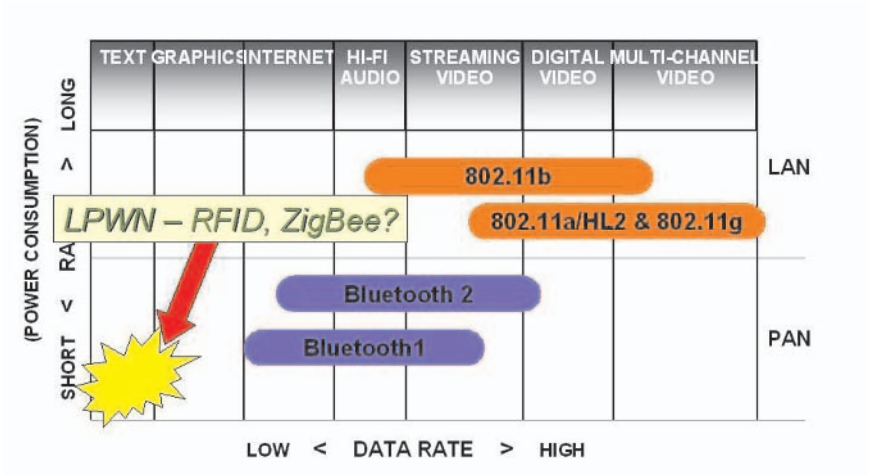


Figure 9. A variety of wireless options are becoming available. The 802.11 technologies are high bandwidth and relatively high power consumption, and are best suited for serviced areas. It is the basis of the emerging 'hotspot' services provided at hotels and airports. While Bluetooth has a lower power demand, in its current form, it is more targeted at 'Personal Area Networks' or PANs of peripherals. ZigBee appears to be the most attractive option for low power wireless sensor networks although the short distance between nodes is a serious limitation for field deployed devices. The ideal specification (starred area) would extend the distance of the ZigBee protocol while keeping the power consumption as low as possible [4]. (Reprinted with permission from Anal. Chem., August 1, 2004, 75 (15), 278A-286A. Copyright 2004 American Chemical Society.)

sensors will in turn drive demand for more complex measurements to be integrated into distributed sensor networks.

5. Networking Options

Network technologies have already made a major impact on analytical science. It is now standard practice to network analytical laboratories, and specialist services such as Laboratory Information Management Systems have been developed specifically to integrate laboratory information with conventional administrative networks in large organisations. Every instrument, down to the humble pH meter has a PC interface, and can be easily networked. The explosion in demand for mobile access to communications, has driven the rapid development of wireless networks such as the 'wireless hot spots' based on IEEE 802.11 standards that are appearing at airports, hotels, cafes and universities. These offer high bandwidth access, typically to users with laptops and palmtops seeking email or WEB access, but could provide an infrastructure for

networking analytical devices. Bluetooth is another wireless network technology targeted mainly at wireless connectivity between peripherals such as mice, keyboards, headsets, printers, devices like mobile phones, palmtops and PDAs, and home appliances in a 'personal area network' or PAN [10]. However, both 802.11 and Bluetooth are designed for conventional network communications, as is evidenced by the high-bandwidth specification, and relatively high power consumption (Figure 9). Therefore, while these technologies will undoubtedly make it much easier to connect analytical devices within a building, they are unlikely in their present forms to be suitable for long-term autonomous operation in remote locations.

In contrast, the ZigBee Alliance, involving companies such as Honeywell, Mitsubishi Electric, Phillips and Motorola, is developing hardware and software communications standards focused on low bandwidth, low-power consumption applications [11], which is potentially more interesting for autonomous analytical devices, although the limitation at present is in terms of distance (less than 100 metres between nodes), and typically just a few metres (ZigBee is being promoted mainly for linking items within the home). The future is likely to comprise a number of interlinking technologies, like ZigBee, bluetooth, 802.11, mobile phone technologies such as GSM/GPRS/3G (Global System for Mobile telecommunications/General Packet Radio Service/third generation) and conventional cabled networks. These technologies will gradually integrate and appear seamless to the user, and analytical devices will make use of them all, with density of deployment correlating with low cost, low power consumption technologies, where the limitations will be low bandwidth and short distances between nodes.

6. Integrating Chemical Sensors and Biosensors with Wireless Networks

Clearly, there is considerable research activity targeted at incorporating sensors into low power wireless networks (LPWNs). However, the sensors employed in these studies are almost entirely transducers measuring physical parameters such as heat, light, humidity, pressure, position etc. While the information available from these sensors is very valuable, and important advances are being made through their use, they cannot fulfil the vision of bridging the digital and molecular worlds. This next step will require the introduction of chemical sensors and biosensors into LPWNs.

However, realisation of the LPWN-compatible chemical sensors/biosensors critical to the realisation of a pervasive sensing vision depends

on the development of devices that are capable of massive scale up on the one hand (and are therefore very low cost, suitable for mass production) and very reliable (capable of autonomous operation for at least one year is a typical target for environmental applications, five years for biomedical implants).

The key challenge therefore is to identify stable analytical chemistries and methods that can be incorporated into these devices, as this will be the basis for the underlying quality of the multitude of analytical signals that will be fed into the web-based networks for higher-level decision making. Therefore, the delivery of the overall vision depends on the ruggedness and reliability of the analytical devices, and the knowledge held within the analytical community becomes the key to unlocking the next revolution in communications.

7. Scale-up Issues for Densely Distributed Analytical Devices

Cost

The successful integration of chemical sensing and biosensing into the broader vision of the 'context-aware world' depends on the availability of widely distributed (pervasive) networked sensors that feed reliable data into the information layer. Electronics and communications substructures (nodes) are still too expensive (80−150). Motorola [12] is incorporating ZigBee wireless protocols into all their pressure and acceleration sensor chips, to facilitate the development of sensor networks for distributed monitoring, control, and automation, but the chips for more generalised use (i.e. with other sensors) are not commercially available. The Berekely Mote technology is now available commercially from Crossbow Technologies Inc., in a number of developer kit formats that typically include motes, sensor boards, and relatively simple user interface software that runs under TinyOS [13], [14]. The motes incorporate ad-hoc networking capabilities and enable demonstrator applications to be developed relatively easily. The latest manifestation is a mote platform around the size of a 1 euro coin (see Figure 10) that can transmit signals approximately 100 m.

Obviously, scale up and cost are reciprocally related, and while relatively small-scale deployment for research purposes is possible at current costs (dozens of devices), unit costs must be driven down by orders of magnitude before large scale deployment of devices is feasible. However, the vision from the engineering community is that 'everything that costs over $25 except food will be connected to the Internet' [15]. To quote Kris Pister, from UC Berkeley, 'by 2010, RF circuits capable of trans-

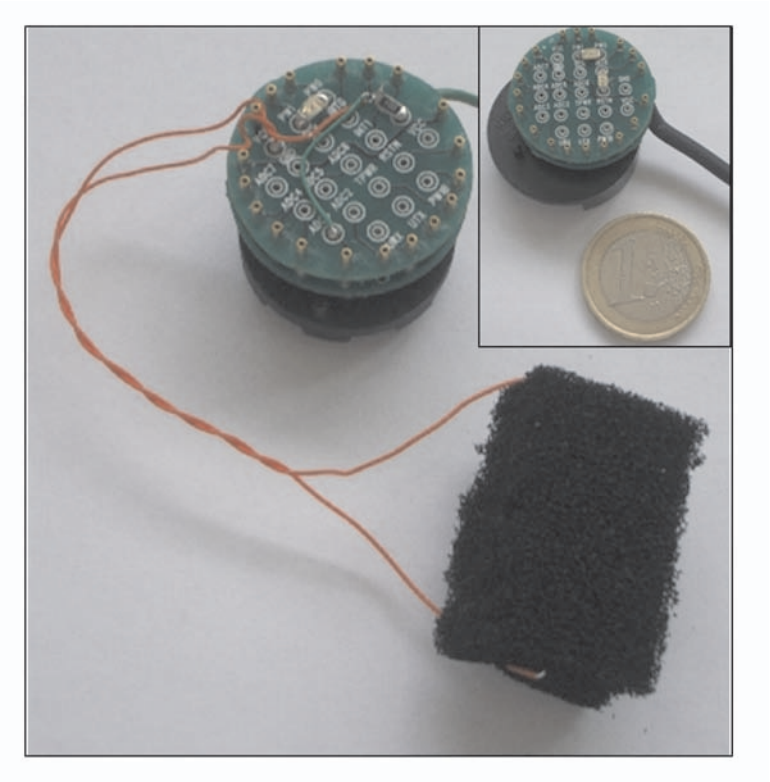


Figure 10. Wireless Mote platform available from Crossbow Technology Inc. The photo shows the platform attached to a small portion of smart foam we have developed for sensing compression e.g. in smart fabrics for monitoring breathing.

mitting 100 m at low bandwidth (up to 100 Kb/s) will cost $0.10, and consume very little power (10nJ/bit). The devices will have the capability of running for perhaps 10 years from a standard battery source' [16]. The implication is that it will cost virtually nothing to include wireless communications capability, and therefore it will become ubiquitous, and will be an almost universal standard feature.

Reliability

In order to realise their potential within this vision, a critical requirement for chemical sensors and biosensors will be stability over long-term deployment. For analytical measurements, the key component is the sensing membrane or film which, when exposed to the sample, generates a sensitive, selective signal that links a molecular binding event to the value of a digital number. Hence the ultimate functionality of a chemo/biosensor network depends on the nature of these molecular interactions. If these are subject to interference by non-specific binding, or changing characteristics due to instability over time, then a practical system cannot be realised, as it will impossible to distinguish real from spurious events, leading to unacceptable levels of false positives and negatives. A major challenge therefore, is for analytical scientists to demonstrate long-term stability of reagents, molecular receptors, films and membrane cocktails such that the chemistry/biology of the analytical measurement is not the limiting factor in terms of overall reliability.

The analyst's traditional strategy for coping with stability is to calibrate regularly. But calibration of autonomous devices involves additional complexity in terms of flow manifolds, valves and pumps, stable standards, and waste storage. Hence remote calibration and massive scale up are incompatible strategies in the medium term, although autonomous lab-on-a-chip devices with built-in calibration could be deployed on a moderate scale, perhaps as more sophisticated nodes against which simpler, more densely distributed but dumber devices targeted at the same analyte could be remotely calibrated. A starting goal is to find reagents that will be stable for up to one year, as this will help simplify the calibration issues.

We have looked in detail at the analysis of nutrients in natural water, as this is of considerable interest globally, in terms of the overall effect on water quality, and in particular, the prevention of algal blooms which have become an all-too familiar problem in many countries. Bearing in mind the need for inexpensive components and the requirement of low-power operation, we have focused on colorimetric detection using LED/photodiode detection in a microfluidic manifold as a generic

approach. So far two reagent methods have been identified as meeting the 12 month stability requirement (calibration slope should be better than 90% of the original slope after 12 months); the yellow method for orthophosphate [17], and the molybdenum blue or Bertholet method for ammonia [18]. Interestingly, the yellow method was found to be preferable to the more common indophenol method as the latter requires the use of ascorbic acid in the final reduction stage, which drastically limits the overall stability of the reagent cocktail. Note that the yellow method only became a viable option very recently with the commercial availability of UV-LEDs, as the absorbance maximum (380 nm) of the phosphate complex is too low for detection with pre-existing LEDs [19].

For ammonia, the commonly employed molybdenum blue method was examined. In this case, there were a number of issues. For example, the standard method requires the use of phenol and hypochlorite. Phenol is unsuitable for health, safety and environmental reasons, and hypochlorite is commonly regarded as unstable. We found that salicylate could be substituted for phenol, with little affect on sensitivity and a relatively small movement of the absorbance maximum, and hypochlorite is stable if stored carefully, and there is very low contamination by certain catalytic metals that accelerate decomposition, such as copper and iron [20].

An issue with reagent based methods is the amount of reagent consumed annually. For example, if 1 litre of reagent is consumed per year, then scale up to millions of devices is not feasible, from the point of view of re-supplying reagents, and of disposal of waste reagent generated. Scale up in numbers means equivalent (or better) scale down in reagent use. This is possible using LOAC devices. For example, at a flow rate of 1 uL/minute, 6 minute sample turnaround time, and a requirement of 1 analysis per hour, less than 50 mL of reagent is required for 1 year continuous operation. While this is still too much for massive scale up, it does fit into the scenario that LOAC instruments could be less densely distributed at more sophisticated nodes within the network.

Alternative strategies are to use arrays of single-shot sensors (ideally reagentless) manufactured to a very high level of reproducibility, and packaged to ensure the device surface is protected from change during storage. While this can reduce significantly the need for calibration, and provide reliable data for extended periods of time, it is not possible at present to produce the required instrumentation at a low enough price to make very large scale up economically viable.

Sampling

For LOAC devices, the use of dialysis sampling in aqueous environments is an attractive option for a number of reasons:

1 The dialysis membrane protects the microfluidic manifold from ingress of particulate matter that can block the narrow channels or damage valves/pumps.

2 A stopped flow approach can be employed using dialysis sampling which allows effective transport of low molecular weight components from the sample across the dialysis membrane and hence the dialysate will accurately reflect the composition of the sample.

3 The membrane produces a semi-sealed unit that enables waste reagents to be contained after use (waste bag fills as sample bag empties).

However, current forms of LOAC devices have many components external to the microfluidic chip such as valves, pumps, power supplies, electronic circuitry, and reagent/waste storage units. While these devices are a major advance on pre-existing autonomous instruments and could be deployed on a reasonable scale, they are typically too large, consume too much power and are too expensive for high-density deployment.

Nevertheless, LOAC devices hold considerable promise for short to medium term solutions to analytical applications that require medium or low density deployment, particularly for environmental and security monitoring of chemo/bio targets detectable via well-known reagent based analytical methods that can be transferred to a microfluidic platform. Perhaps the greatest barrier to more widespread deployment is the current use of scaled down conventional pumps and valves for controlling liquid movement throughout the manifold. Not only do these devices contribute greatly to the overall cost and size of the device, they are also prone to malfunction, and are usually the greatest source of power drain by the system. Research into novel methods for controlling liquid movement in microfluidic channels is therefore particularly important. One approach that appears to hold considerable promise is the integration of materials that will swell or contract under external photonic or electronic stimuli. A recent review summarises the many materials that are currently under investigation for this purpose [21]. Successful integration of such 'soft' valves and pumps would therefore be a significant advance for these devices, as they are inherently low power, low cost, much more tolerant of microparticulants, and would produce more compact microfluidic platforms.

8. Chemo- & Bio-warfare Agents

Sensor networks capable of providing advanced warning of the release of a biowarfare agent, and tracking its geographical spread is a very attractive vision. However, there are many issues and barriers inhibiting the successful widespread deployment of such devices. For example, in the case of anthrax, the requirement to detect down to single bacterium in 100,000 litres of air requires a sophisticated sampling and sample processing regime. To meet this type of need, it is critical that the presence of the target is detected when present at a very low threshold, with no false negatives or positives [22]. Such requirements pose significant challenges for the analytical community. The development of reliable, low-power, distributed networks of instruments capable of detecting trace levels of such targets is beyond the current state-of-the-art, and a sustained investment of R&D resources will be required over a considerable period of time to develop breakthrough technologies. In addition, the reporting of an event (e.g. release of a biowarfare agent in a public place) is a complex issue, as has been highlighted in a recent revue of the issues [23]. For example, there are temporal dynamics that underlie the layered information feedback loops to various user groups. Clearly, the information relayed to the public at rush hour in a busy city must be carefully managed to avoid triggering mass panic. The information management system must therefore recognise the potential effect of feedback on to various stakeholder groups, and how this effect varies with respect to time, and location. However, progress is happening, as demonstrated by the recent report of an autonomous pathogen detection instrument for the detection of aerosolized bacillus anthracis and yersinia pestis [24]. Such devices necessarily integrate multiple operations normally carried out in specialised laboratories such as sample processing, separation, amplification and orthogonal detection methods to ensure the analytical result is correct, and this leads to a device that is large and heavily serviced in terms of power requirements and reagents, and therefore unsuitable for large scale deployment in its current form.

Undoubtedly, further integration of analytical processes (and particularly sample handling) will happen, which will make denser deployment possible. The widespread availability of effective and widely distributed monitoring of chemo- and bio-warfare agents will reduce the incidence of false alarms and copy-cat events, and may reduce the effectiveness of these weapons, which often rely on the fear of what might be present, rather than the reality.

One idea recently reported in the media is to integrate sensors for specific threats into mobile phones [25]. These would essentially be pas-

sive until the sensor signal crosses a critical threshold, at which point a set message is sent to a pre-determined location. Coupled with GPS information inherent with every phone, this provides a very simple route to early detection of the source of the release of a potentially lethal agent, with signals from other phones providing cross-validation, reducing the potential for false negatives and false positives. Furthermore, as mentioned above, this would enable the dynamics of the event to be monitored, as a plume spreads through an area, while simultaneously providing a means to send information to individuals and emergency services. However, given the current status of biosensor development, it is unlikely that sensors with the required specifications will be available for some time. A more likely scenario for passive background monitoring of threats would be to incorporate threat detectors into vehicles as a standard component. This would enable more complex services and diagnostics to be provided and therefore reduce the operating constraints, making device deployment much more likely than in mobile phones. Vehicles are rapidly becoming integrated into communications networks through GPS-based route planners and GSM phone capabilities, and have sophisticated internal sensing and communications systems already in place. Threat detection could therefore be seen as a logical extension of these capabilities, which takes advantage of the pre-existing infrastructure.

9. Sensor communities and group behaviour

Moving from localised, lab based analytical methods to widely distributed sensor networks requires new strategies to deal with issues such as calibration, data validation and device diagnostics. In terms of the device hierarchy described above, the main challenge lies with the most densely distributed and least sophisticated devices, as incorporation of calibration, validation and diagnostic procedures requires a more sophisticated instrument and therefore less dense distribution. This is a critical issue, as scaling up to the use of large numbers of dumb devices inevitably will give rise to increasing instances of false positives/negatives and device malfunctions that must be identified and discarded from the overall data population. On the other hand, the use of sensor communities rather than single devices allows group behaviour strategies (e.g. collaborative reasoning) to be employed to help identify spurious signals and device malfunction. For example, in a densely distributed sensor network monitoring air or water quality, it is very unlikely that a real event will occur on a single device only. For real events, threshold crossing will normally occur in a number of devices clustered in a particular location,

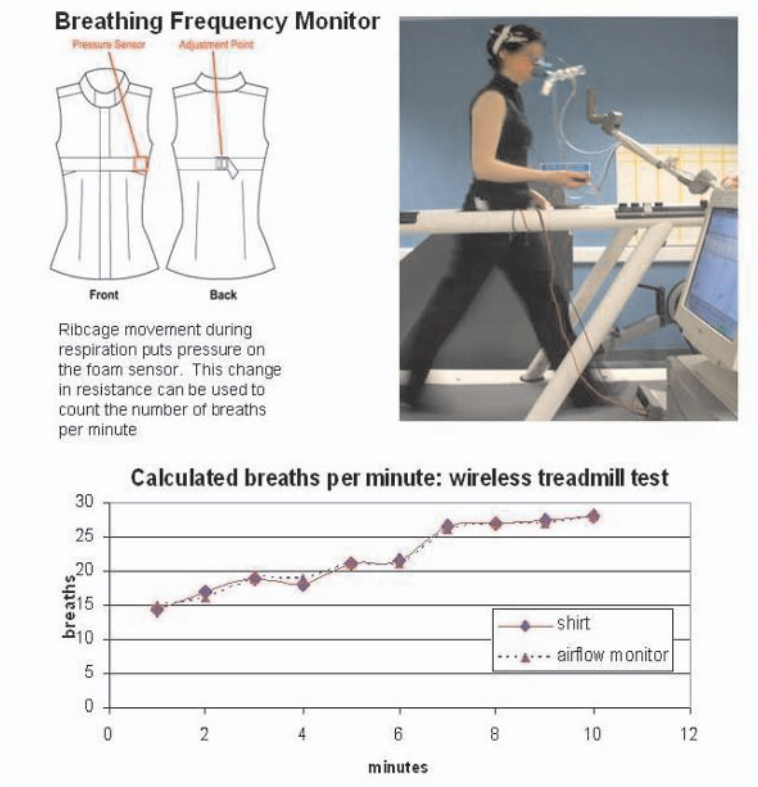


Figure 11. Comparison of a wearable foam sensor integrated into a shirt, and a reference airflow monitor (facemask) for monitoring breathing during treadmill experiments. The results (bottom) indicate that these types of innocuous wearable sensors can provide important information on general heath indicators such as breathing [27].

and the dynamics of the event can be followed by the temporal development of the signal pattern obtained from these simple devices. An additional level of functionality can be offered by using autonomous mobile (robotic) devices that move towards an event signalled by a device in order to locate the source, and to validate the event. Such swarming-like behaviour can be followed using GPS data, and the position and relative abundance of the devices around a particular location can in itself provide diagnostic information about the location and dynamics of an event.

10. pHealth

Personal Health or pHealth is an area of great potential for a number of reasons. Unlike remote environmental monitoring, the wireless

communications network is already in place, and individuals have their own personal gateway through the mobile phone. This means that the existing infrastructure can be used to provide pHealth support, and applications will be rolled out over the next few years. Examples already emerging include smart fabrics for monitoring personal heath indicators such as breathing, heartbeat, and movement (changes in gait or emergence of tremors) in which the sensing component is formed from a functionalised fabric, rather than a discrete sensor attached to the person [26, 27]. Materials like lycra and foam can be converted into stretch and compression sensing components and integrated into clothing (see Figure 11). This will be part of a more extensive move to develop pHealth products. Examples will range from electronic personal medical records, to fully integrated diagnostics and 'wellness' systems for elderly people (i.e. detecting movement, location, application of therapeutic agents, electronic monitoring of diagnostic data etc.). One of the main players in this area is Intel. Despite being an application agnostic company, Intel is not waiting for the potentially lucrative health related markets of widely distributed sensing to emerge spontaneously [28]. Rather, it is actively cooperating in pilot projects to demonstrate the tremendous benefits of this technology from the ground up, knowing that this will in turn drive growth in future markets.

11. Conclusions

The incorporation of analytical devices ranging from tiny sensors, to hand-held instruments, and autonomous field devices will happen during this decade. It is important for scientists to appreciate the implications of this process now, as it has the potential to generate truly disruptive technologies that will have a profound effect on the way people live. In medicine, it has the potential to enable people to remotely monitor disease markers and apply therapies, while still being in intimate contact with specialists who can access their data instantly via the web. In agriculture and food, it will facilitate quality tracking through the distribution chain, 'from harvest to home', and in the environment, it will allow access to data gathered from multiple locations by a variety of users, including local communities and individuals, as well as specialists. In fact, environmental sensing is already well advanced at this time, as distributed geo-sensing has been an area of very active research for many years, for example through satellite-based spectral mapping. Complementary to this information is that available from widely distributed surface-based sensors. Hence this community is one of the first to realise the potential of networked sensors. The first workshop on Geo-

sensor Networks was recently held in Portland, Maine and the report on the meeting touches on many of the issues highlighted in this article including programming sensor networks, device scalability, mobility of sensor nodes, and higher-level modelling and reasoning from large data sets [29]. Understandably, another area of intense research focus is security and threat detection, particularly for applications such as the early detection of air/water borne bioactive/chemoactive/radioactive warfare agents, and the associated approaches for alerting of agencies, emergency services, and the general public should this occur; i.e., personalisation of feedback information. In fact, the merging of data from this continuum of sources and its accessibility via the web will open up entirely new areas of research, such as linking of incidences of illness to environmental or food quality parameters, or correlating the effectiveness of community based health management with the individual's behaviour (was the marker measured and the drug taken at the correct intervals, was the correct dose taken?).

It is also clear that as this vision is gradually realised, there will be very significant ethical and moral issues related to the type of information stored, how long it is stored for, who gets access to it, and the conditions under which various users can access the data. As with any technology that generates information of direct impact to people and societies, there is the potential for beneficial use, and for abuse. Given the tremendous power of sensor networks to deliver highly personalised information, it will be vital that issues related to data security on the one hand, and the definition of an ethical framework for proper use on the other, are also addressed. Delivering the vision will clearly require new alliances of industries, Government agencies, public service providers, university research centres, and community groups. The scale of the opportunity is truly enormous, as is that of the research effort required to deliver it.

References

[1] Michio Kaku *Visions — How Science will Revolutionise the 21st* Century, Anchor Books, New York, 1997, p. 31.

[2] D Diamond and K Nolan Anal. Chem., 73 (2001) 22A–29A.

[3] A Ceresa, E Bakker, B Hattendorf, D Gunther and E Pretsch Anal. Chem., 73 (2001) 343–351.

[4] D Diamond Anal. Chem., 76 (2004) 278A–286A.

[5] Ron Ambrosio Internet-Scale Data Acquisition and Control Systems — Programming Paradigm Challenges, Paper presented at the conference, Creating An Expanded DER Industry, November 28–30, 2001, Loews L'Enfant Plaza Hotel, Washington, DC.

[6] Alan Mainwaring, Joseph Polastre, Robert Szewczyk, David Culler and John Anderson, Wireless Sensor Networks for Habitat Monitoring, in Proceedings of the first ACM international workshop on Wireless sensor networks and applications, Atlanta, Georgia, USA September 28 2002, 88–97.

[7] See http://www.greatduckisland.net

[8] See http://cens.ucla.edu/

[9] 'Remodeling the Wireless Landscape — Technologies of Massive Disruption,' a paper presented by Dr. J. Schwarz da Silva, Head Communications and Network Technologies, European Commission (DG-INFSO), at Wireless World Research Forum Meeting, New York, 27–28 October, 2003, see www.wireless-world-research.org for more details.

[10] An up to date list of applications can be found at www.bluetooth.com.

[11] See www.zigbee.org.

[12] 'Motorola goes with ZigBee Protocol for Wireless Networking,' press release, Motorola Inc., September 23rd 2003, Phoenix, Arizona.

[13] See http://www.xbow.com/Products/Wireless_Sensor_Networks.htm

[14] Latest TinyOS source code is available free at http://sourceforge.net/projects/tinyos/.

[15] Statement from Dan Rosen, formed head of Microsoft Advanced R&D, quoted by Alex Lightman in a presentation at the Wireless World Research Forum Meeting, New York, 27–28 October, 2003, see www.wireless-world-research.org for more details.

[16] www.eecs.berkeley.edu/ pister/SmartDust/in2010

[17] Michaela Bowden and Dermot Diamond, Sensors and Actuators B 90 (2003) 170–174.

[18] A Daridon, M Sequiera, G. Pennarun-Thomas, J Lichtenberg, E Verpoorte, D Diamond and NF de Rooij, Sensors and Actuators B, 76/1–3, (2001) 235–243.

[19] Margaret Sequeira, Michaela Bowden, Edel Minogue and Dermot Diamond, Talanta, 56, Issue 2, (2002) 355–363.

[20] Margaret Sequeira, Antoine Daridon, Jan Lichtenberg, Sabeth Verpoorte, N F de Rooij and Dermot Diamond, Trends Anal. Chem., 21 (2002), 816–827.

[21] Organic and Biomimetic Designs for Microfluidic Systems, Jaisree Moorthy, David Beebe; Anal. Chem.; 2003; 75(13); 292A–301A.

[22] Allen Northrup, CEO of Microfluidic Systems Inc., Pleasanton, California, personal communication.

[23] Biological Warfare Detection, David R. Walt and David R. Franz, Anal. Chem., 2000; 72(23); pp. 738 A–746 A.

[24] Mary T. McBride, Don Masquelier, Benjamin J. Hindson, Anthony J. Makarewicz, Steve Brown, Keith Burris, Thomas Metz, Richard G. Langlois, Kar Wing Tsang, Ruth Bryan, Doug A. Anderson, Kodumudi S. Venkateswaran, Fred P. Milanovich, and Bill W. Colston, Jr., Anal. Chem., 75 (2003) 5293–5299.

[25] See 'The Times' (London), March 28th 2004, 'Mobile to Serve as Dirty Bomb Detectors'.

[26] Sarah Brady, Dermot Diamond, King-Tong Lau, Sensors and Actuators A: Physical, Volume 119, Issue 2, 13 April 2005, Pages 398–404

[27] Lucy E Dunne, Sarah Brady, Barry Smyth, Dermot Diamond, J. Neuroengineering Rehabil. 2005 Mar 1;2(1):4.

[28] See http://www.intel.com/research/prohealth/.

[29] S Nittel, A Stefanidis, I Cruz, M Egenhofer, D Goldin, A Howard, A Labrinidis, S Madden, A Voisard and M Worboys, Report from the First Workshop on Geo-Sensor Networks, 33 (1) (2004) 141–144.

DATA ANALYSIS FOR CHEMICAL SENSOR ARRAYS

Corrado Di Natale, Eugenio Martinelli, Giorgio Pennazza, Andrea Orsini, Marco Santonico
University of Roma "Tor Vergata"
Dept. of Electronic Engineer, Via del Politecnico, 1 00133 Roma
dinatale@uniroma2.it, damico@eln.uniroma2.it, andrea.orsini@psm.rm.cnr.it

Abstract

Arrays were introduced in the mid-eighties as a method to counteract the cross-selectivity of gas sensors. Their use has since become a common practice in sensor applications [1]. The great advantage of this technique is that once arrays are matched with proper multivariate data analysis, the use of non-selective sensors for practical applications becomes possible. Again in the eighties, Persaud and Dodds argued that such arrays has a very close connection with mammalian olfaction systems. This conjecture opened the way to the advent of electronic noses [2], a popular name for chemical sensor arrays used for qualitative analysis of complex samples.

It is worth remarking that a gas sensor array is a mere mathematical construction where the sensor outputs are arranged as components of a vector. Arrays can also be utilized to investigate the properties of chemical sensors, or even better, the peculiar behaviour of a sensor as a component of an array. In this chapter, the more common sensor array methodologies are critically reviewed, including the most general steps of a multivariate data analysis. The application of such methods to the study of sensor properties is also illustrated through a practical example.

Keywords: electronic nose; principal component analysis; pattern recognition; chemical sensors; sensor arrays; olfaction system; multivariate data analysis.

1. Feature extraction

In pattern recognition, a "feature" is any direct or derived measurement of the entities to be classified that helps differentiate between classes. In chemical sensor arrays individual measurements are the entities assigned to classes while a measurement is a time sequence of sensor

Jim Byrnes and Gerald Ostheimer (eds.),
Advances in Sensing with Security Applications, 147–169.

signals collected during the exposure of the sensor to the sample. Consequently, feature extraction for chemical sensors is the evaluation, from a sensor signal stream, of a number of parameters that can, as much as possible, represent the sensor experience containing that information related to the classification objective.

Feature extraction is of fundamental importance because sensor features are utilized in any successive elaboration to produce the output of the sensor system in terms of estimation of the measured quantities.

To define a feature extraction procedure it is necessary to consider that the output signal of a chemical sensor follows the variation of the concentration of gases at which it is exposed with a certain dynamics. The nontrivial handling of gas samples complicates the investigation of the dynamics of the sensor response. Generally, sensor response models based on the assumption of a very rapid concentration transition from two steady states results in exponential behaviour.

A straightforward solution of the feature extraction problem disregards the dynamic transitions and considers only the signal shift between two stationary states before the application of gas stimuli and during the exposure to gas after the transitory phase. This quantity has a straightforward meaning related to the equilibrium conditions established between the molecules in gas phase and those interacting with the sensor. Although the direct chemical and physical meaning of the steady state signal shifts, it is worth investigating if the dynamic properties can provide features with an extended information content.

Simple evidence about the information content of the dynamic properties can be obtained by considering the Langmuir model of adsorption of molecules from gas phase to a limited number of interacting sites on a sensor surface [3].

According to this model, the total rate of adsorbed molecules (dn/dt) is given by the algebric sum of the two processes occurring at the interaction sites: adsorption and desorption. The equilibrium condition is reached once one equalizes the other.

In the case of N_s adsorbing sites, if P is the gas partial pressure held constant by some gas reservoir, and k_a and k_d are adsorption and desorption constants respectively, the following expression for adsorption and desorption rates holds:

$$\frac{dn}{dt} = \left(\frac{dn}{dt}\right)_{ads} - \left(\frac{dn}{dt}\right)_{des} = P \cdot k_a \cdot (N_s - n) - k_d \cdot n. \qquad (1)$$

Solution of this equation gives rise to the time behaviour of the number of adsorbed molecules. For most transducers, this quantity is then linearly converted into a sensor signal.

$$n(t) = \frac{k_a N P}{k_d + k_a P} \left(1 - e^{-(k_d + k_a P)t}\right). \tag{2}$$

The number of adsorbed molecules evolves exponentially towards a steady value. In figure 1 the qualitative behaviour of the steady state amount of adsorbed molecules and the time constant are plotted versus the gas concentration. It is evident that the feature representing the steady state signal reaches a saturation value as the number of adsorbed molecules equals the number of adsorbing sites while the dynamics represented by the time constant or, even better, its inverse, is linear with the gas concentration. In practice, when the concentration rapidly changes from zero to a value exceeding the saturation limit, the dynamics of the sensor continues to provide a linear response. From the point of view of sensitivity, namely the derivative of the feature versus the concentration, the steady-state feature results in a progressively decreasing sensitivity, while, in the limits of the model, the dynamic property provides a constant sensitivity over a theoretically unlimited range.

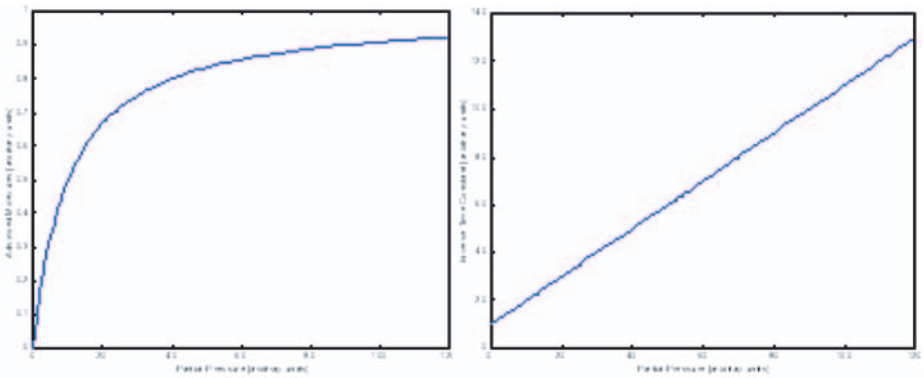


Figure 1. Qualitative behaviour with the gas concentration of the steady-state signal shift (a) and time constant inverse (b) for a pure Langmuir adsorption on a limited amount of equi-energetic adsorption sites over a sensor surface.

Some attempts to exploit sensor dynamics for concentration prediction were carried out in the past. Davide et al. approached the problem using dynamic system theory, applying non-linear Volterra series to the modelling of Thickness Shear Mode Resonator (TSMR) sensors [4]. This approach gave rise to non-linear models where the difficulty to discriminate the intrinsic sensor properties from those of the gas delivery systems limited the efficiency of the approach.

Other authors concentrated their attention on the evaluation of the time constants. It is well known that multiexponential fit is an ill-posed problem, and its solution, obtained with typical non-linear optimization

methods (e.g. the Levenberg-Marquardt algorithm), may be affected by large errors. Di Natale et al [5] applied an alternative method based on multiexponential spectroscopy, originally developed by Samitier et. al. [6], to the determination of the time constants in a tin-oxide sensor array. An opposite approach to feature extraction considered completely empirical methods. Eklöv et al [7] investigated a number of parameters taking into account several descriptors of the behaviour of CHEMFET sensor signals. Once a feature is chosen, the signals may be transformed into data, which can be analyzed to find the correlation between sensor response and relevant sample characteristics.

2. Data Pre-processing: Scaling

In data analysis, data are seldom used without some preprocessing. Such preprocessing is typically concerned with the scale of data. In this regard two main scaling procedures are widely used: zero-centered and autoscaling.

Zero-centered data means that each sensor is shifted across the zero value, so that the mean of the responses is zero. Zero-centered scaling may be important when the assumption of a known statistical distribution of the data is used. For instance, in case of a normal distribution, zero-centered data are completely described only by the covariance matrix.

Autoscaling means to scale each sensor to zero-mean and unitary-variance. This operation equalizes the dynamics of all sensor responses, preventing a sensor having a larger response range from hiding the contribution of other dynamically limited sensors. Further, autoscaling makes the sensor responses dimension-less. This feature becomes necessary when sensors whose signals are expressed in different units are merged in one array. This is the case for hybrid arrays (different sensor technologies in the same array) and when chemical sensors are fused with other instruments, e.g. the fusion of electronic noses and electronic tongues [11].

3. Normalization

Horner and Hierold [12] showed that the application of a simple normalization of sensor data can greatly help in preventing quantitative information from masking qualitative aspects of the data.

Figures 2 and 3 show the situation in the case of an array of TSMR sensors exposed to different compounds at various concentration levels [13]. The cross-selectivity of the sensors makes their individual responses ambiguous. Namely different samples, due to a combination of qualita-

tive and quantitative aspects, may give rise to similar sensor responses. In figure 2 the confusion among the data shows the ambiguity of the sensor response. Figure 3 shows a Principal Component Analysis (PCA) plot. PCA will be thoroughly discussed in a later section. It is a typical method to represent multivariate data in a bidimensional plot.

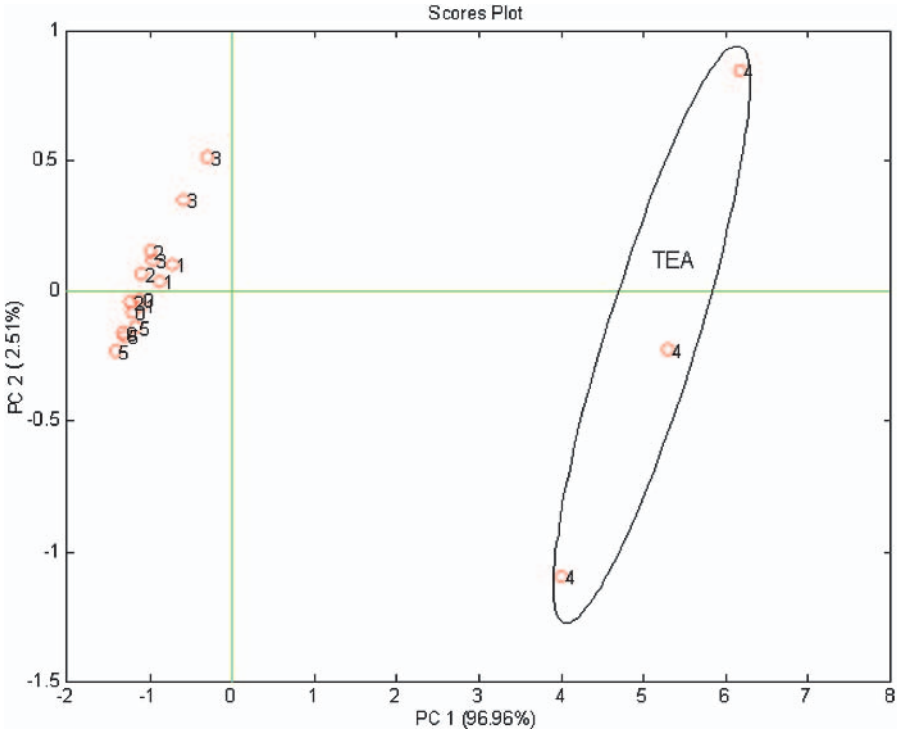


Figure 2. Examples of data characterized by strong concentration effects. Data are related to a quartz microbalance array exposed to six volatile compounds each measured three times at different concentrations. Only triethylamine (TEA in the plot) data emerge (experimental details in [10]).

A simple way to disentangle the information is obtained when the relationship between sensor responses and concentrations of analytes is linear, such as found in polymer coated TSMR:

$$\Delta f_i = K_{ij} \cdot c_j \tag{3}$$

where Δf_i is the response of the i-th TSMR, K_{ij} is the sensitivity of the i-th sensor towards the j-th compound, and c_j is the concentration of the j-th compound.

The normalization consists of dividing each sensor response by the sum of all the sensor responses to the same sample, so that the concen-

tration information may disappear.

$$\Delta f_i \Rightarrow \frac{K_{ij} \cdot c_j}{\sum_m K_{mj} \cdot c_j} = \frac{K_{ij}}{\sum_m K_{mj}} \tag{4}$$

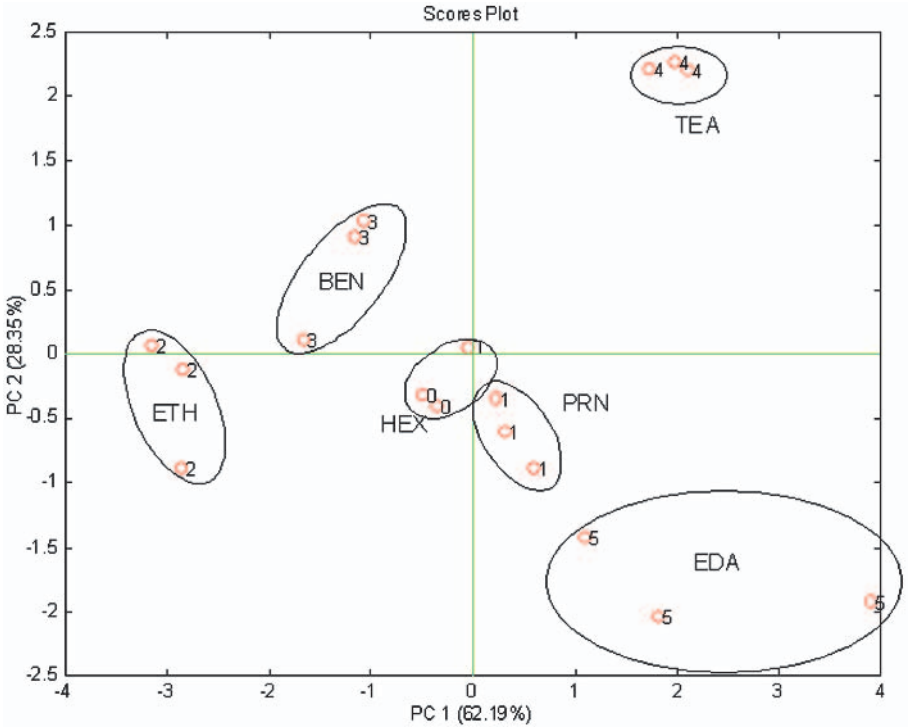


Figure 3. Data of figure 2 after the application of linear normalization of equation 4. Classes are now clearly separated. It is worth noting the condensation of TEA data (class 4) in a very restricted region. Meaning of the classes: HEX: hexane; PRN: propanal; ETH: ethanol; BEN: benzene; TEA: triethylamine; EDA: ethylendiamine.

Figure 3 shows the PCA score plot of the same data of figure 2 after the application of equation 4. The application of linear normalization to an array of linear sensors should produce, on the PCA score plot, one point for each compound, independent of its concentration, and achieve the highest possible recognition. Deviations from ideal behaviour, as shown in figure 3, are due to the presence of measurement errors, and to the non-linear relationship between sensor response and concentration.

In the previously quoted paper, Horner and Hierold also treated the case of sensors whose response is ruled by a power law relationship between sensor response and analyte concentration. This is the case for

metal-oxide semiconductor gas sensors. Eq.4 can be extended to sensors described by a power-law ($z = c^{\alpha}$) simply linearizing, through the logarithm, the sensor response.

Normalization is, in practice, also useful to counteract any possible fluctuations in the sample concentration. These fluctuations are, in practice, mostly due to sample temperature fluctuations, and to instabilities of the sampling system and they may lead to variations of the dilution factor of the sample with the carrier gas. Of course, normalization is of limited efficiency because the mentioned assumptions strictly hold for simple gases and they fail when mixtures of compounds are measured. Furthermore, it has to be considered that in complex mixtures, temperature fluctuations do not result in a general concentration shift, but since individual compounds have different boiling temperatures, each component of a mixture changes differently so that both quantitative (concentration shift) and qualitative (pattern distortion) variations take place.

4. Multivariate data exploration

Given a set of data related to a number of measurements, after the application of proper feature extraction, pre-processing and normalization, exploratory techniques aim at studying the intrinsic characteristics of the data in order to discover eventual internal properties.

Exploratory data analysis shows the aptitude of an ensemble of chemical sensors to be utilized for a given application, leaving to the supervised classification the task of building a model to be used to predict the class membership of unknown samples.

Two main groups of exploratory analysis may be identified: representation techniques and clustering techniques.

Representation techniques are a group of algorithms aimed at providing a representation of the data in a space of dimension lower than that of the original sensor space. The most popular of these methods are Principal Component Analysis, Self Organizing Map, and Sammon's mapping. Each of these techniques is based on specific hypotheses about the nature of the data and the sensor space. Each of them tends to preserve some particular characteristic of the data. It is worth remarking that the simplest, in terms of calculus and interpretation of results, is that based on the strongest assumption about the statistical distribution of the data. On the other hand, when assumptions about data distribution are removed a neural network is necessary for data representation.

Clustering techniques are mostly based on the concept of similarity expressed through the definition of a metric (distances calculus rule) in

the sensor space. The most trivial and common choice is to express the similarity as a Euclidean distance. Other definitions, such as the Mahalanobis distance, are also used [14].

5. Principal Component Analysis

The scope of Principal Component Analysis (PCA) is a *consistent* portrayal of a data set in a representation space. Mathematically, PCA is a linear transformation that may be described as $S = WX$. Here X is the original data set, W is the transformation matrix, and S are the data in the representation space. PCA is the simplest and most widely used method of multivariate analysis. Nonetheless, most users are seldom aware of its assumptions and sometimes results are badly interpreted.

The peculiarity of PCA is in a representation of the data set onto a subspace of reduced dimensionality where the statistical properties of the original data set are preserved.

Although the PCA concept is used in many disciplines it was strongly developed in chemometrics, where it was introduced at the beginning to analyze spectroscopic and chromatographic data, which are characterized by a higher correlation among the spectra channels [8].

The possibility of a reliable representation of a chemical sensor array data set in subspaces of smaller dimension lies in the fact that the individual sensors always exhibit a high correlation among themselves. PCA consists of finding an orthogonal basis where the correlation among sensors disappears.

As a consequence, in a sensor space of dimension N the effective dimension of the sub-space occupied by the data is less than N. This dimension can be precisely evaluated using algorithms developed to describe dynamic systems. An example is the correlation distance that allows evaluating the fractional dimensionality of a data-set [16]. Correlation distance provides an independent way to evaluate the expected reduction of dimension.

Statistical properties of a data set can be preserved only if the statistical distribution of the data is assumed. PCA assumes the multivariate data are described by a Gaussian distribution, and then PCA is calculated considering only the second moment of the probability distribution of the data (covariance matrix). Indeed, for normally distributed data the covariance matrix ($X^T X$) completely describes the data, once they are zero-centered. From a geometric point of view, any covariance matrix, since it is a symmetric matrix, is associated with a hyper-ellipsoid in N dimensional space. PCA corresponds to a coordinate rotation from the natural sensor space axis to a novel axis basis formed by the principal

direction of the hyper-ellipse associated with the covariance matrix. The reduction of an ellipsoid to its canonical form is a typical linear algebra operation performed by calculating the eigenvectors of the associated matrix.

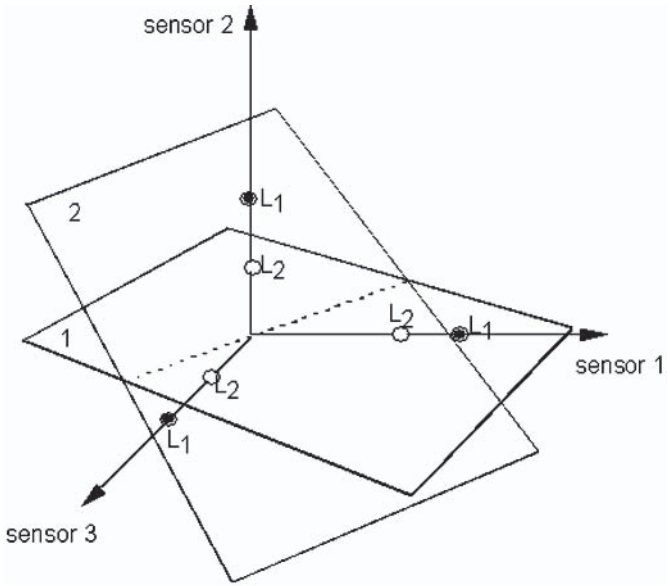


Figure 4. Sketch of the geometrical meaning of loading based feature extraction. Each measurement is represented by the loadings of a PCA decomposition limited to the most meaningful components. In the picture the case of three sensors is displayed limiting the feature to the first two principal components. Each measurement determines a plane. In terms of pattern recognition, the study of similarity and difference among measurements is translated into differently oriented planes.

In practice, PCA can be calculated with the following rule. Let us consider a matrix X of data, let $C{=}X^T X$ be the covariance matrix of X. The i-th principal component of X is $X^T\lambda(\mathrm{i})$, where $\lambda(\mathrm{i})$ is the i-th normalized eigenvector of C corresponding to the i-th largest eigenvalue. A sketch of the geometric meaning is shown in figure 4.

Since the eigenvalues of the matrix associated with the quadratic form describing a hyper-ellipse are proportional to the extension of the solid in the direction of the corresponding eigenvector, the eigenvalues of the covariance matrix are directly proportional to the variance along the corresponding eigenvector. Therefore considering the relative values of the eigenvalues $\lambda(\mathrm{i})$ it is possible to reduce the representation to only those components carrying most of the information.

Given a matrix of data PCA results in two quantities usually called scores and loadings. Scores are related to the measurements, and they

are defined as the coordinates of each vector measurement (a row of matrix X) in the principal components basis.

The loadings describe the contribution of each sensor to the principal components basis. A large loading, for a sensor, means that the principal component is mostly aligned along the sensor direction.

It is important to note that the highest eigenvalues correspond to components defining the directions of highest correlation among the sensors, while the components characterized by smaller eigenvalues are related to uncorrelated directions. Since sensor noises are uncorrelated the representation of the data using only the most meaningful components removes the noise of the sensors. In this way PCA is used to remove noise from spectral data [15].

When applied to electronic nose data the presence of various sources of correlated disturbances has to be considered. As an example, sample temperature fluctuations induce correlated disturbances, which may be described by principal components of highest order. When these disturbances are important the first principal component has to be eliminated in order to emphasize the relevant data properties. A set of algorithms called Minor Component Analysis (MCA) was introduced to take into account these phenomena mainly in image analysis [17].

The hypothesis of a normal distribution is a strong limitation that should be always kept in mind when PCA is used. In electronic nose experiments, samples are usually extracted from more than one class, and it is not always that the totality of measurements results in a normally distributed data set. Nonetheless, PCA is frequently used to analyze electronic nose data. Due to the high correlation normally shown by electronic nose sensors, PCA allows a visual display of electronic nose data in either 2D or 3D plots. Higher order methods were proposed and studied to solve pattern recognition problems in other application fields. It is worth mentioning here the Independent Component Analysis (ICA) that has been applied successfully in image and sound analysis problems [18]. Recently ICA was also applied to process electronic nose data results as a powerful pre-processor of data [19].

Nonetheless, a sub-set belonging to one class may very likely be normally distributed. In this case a PCA calculated on one class cannot work in describing data belonging to another class. In this way, the membership of data to each class can be evaluated. This aspect is used by a classification method called SIMCA (Soft Independent Modelling of Class Analogy). It is a clever exploitation of the limitations of PCA to build a classification methodology [20].

Non linear PCA algorithms have also been developed to provide a representation along principle curves rather than principal directions

[21]. Also neural networks were proposed to solve the problem of faithful representation of multidimensional data in representation spaces of lower dimensions [17].

Another limitation to the use of PCA comes from the fact that being a linear projection it may introduce mistakes. Indeed, in the projection, data separated in the original space may result in a similar score, a phenomenon like that producing a constellation in the starred sky.

A final consideration about PCA is concerned with its use as a preprocessor of non-linear methods such as neural networks [22]. The assumption of a normal distribution of the data requires all following analysis steps to adhere to this hypothesis. If positive results are sometimes achieved they have to be considered as serendipitous events.

6. Supervised Classification

Exploration analysis is not adequate when the task of the analysis is clearly defined. An example is the attribution of each measurement to a pre-defined set of classes. In these cases it is necessary to find a sort of regression able to assign each measurement to a class according to some pre-defined criteria of class membership selection. This kind of analysis is called supervised classification. The information about which classes are present have to be acquired from other considerations about the application under study. Once classes are defined, supervised classification may be described as the search for a model of the following kind:

$$\vec{c} = f(\vec{s}) \tag{5}$$

where $\vec{c}$ is a vector describing the class assignment, $\vec{s}$ is the vector of features of the sensors in the array and f is a generic function. This kind of problem is generally called pattern recognition. To solve pattern recognition problems using conventional algebra, class memberships must be encoded in a numerical form that allows treating the problem by numerical methods. The most common way to represent class memberships is the so-called "one-of-many" code. In this codification, the dimension of $\vec{c}$ is equal to the number of classes. The component corresponding to the class to which a sample belongs is assigned to 1, leaving to 0 the others. In chemical sensor array applications various sources of measurement errors may occur. As a consequence, equation 5 is written in a more realistic form as:

$$\vec{c} = f(\vec{s}) + \vec{e} \tag{6}$$

where the vector $\vec{e}$ contains anything not related to the classification scheme expressed by the vector $\vec{c}$. Equation 6 is formally similar to the general problem of regression where the scope is the determination

of the function f, in terms of functional form and parameters. Statistics provides the tools to estimate, from an experimental data set, the parameters of the function f, in order to approximate the measured experimental data. The classical approach is the Least Squares Method. It is important to reflect about the validity of least squares in chemical sensor array data. This method is based on the assumption that variables are normally distributed and that the quantity $\vec{e}$ of equation 6 is a normally distributed zero-mean variable . The assumption of zero-mean means that all the variables to which the sensor responses are sensitive, except those related to the classification of the samples, may fluctuate but do not bias the measurement. This may not hold for sensor array data where, except for sensor noise, the contributions to $\vec{e}$ (sensor drift, sample temperature variations, sample dilution changes,...) are not zero-mean quantities. Nevertheless, solutions based on least squares can be used to establish classification models, but it is important to be aware of the fundamental limitations of the methods. Before discussing practical solutions, it is necessary to detail the general framework. In pattern classification it is very important to estimate the expected error rate after the classifier model has been assessed on a calibration data set. The expected error rate determines the efficiency of the model by giving the probability of misclassifying future samples. Error rate estimation on the same set used for calibration induces large, optimistically biased estimates of performance. This effect is called "over-fitting". The importance of over-fitting grows with the order of the regression function. In particular, it is important when highly non-linear functions are used (e.g. in neural networks). In these cases the model, while almost perfect when estimating the data on which the model is trained, fails completely when generalizing other data. Over-fitting may be more important for classification scopes, where the samples are extracted from sets that may not be well defined, so that the assignment of samples to classes may be affected by errors due to the vagueness of the classification scheme. In this situation the possibility of generating models which are not able to predict unknown samples with sufficient accuracy is high. The straightforward solution for error rate estimation is to split the data set into two independent sets, and use one for calibration and the other to test the classifier and estimate the error rate. This method cannot be used when available sample sizes are small. Moreover, how to split samples is a non-trivial problem because the division should be done while keeping the distributions of the two sets as close as possible in order to avoid biasing performance evaluation. A more reliable model validation is achieved using the "leave-one-out" technique [34]. "Leave-one-out" repeats the model building n times for n measures, each time

leaving one measure out for testing and using the rest for training. The average test error rate over n trials is the estimated error rate. In case of small data sets, the bootstrap method has been proven to be more efficient than "leave-one-out". Both "leave-one-out" and bootstrap are kinds of re-sampling methods. The bootstrap method generates new samples (called "bootstrap samples") by drawing, with replacement, a number N of samples from the original samples [35]. Different methods to generate the bootstrap samples are available. A comparison of the four most popular is discussed in [36], where the efficiency of the methods is compared on a classification problem.

7. Linear Discrimination

The simplest way to estimate a supervised model is to consider that the descriptor of each class may be represented as a linear combination of the sensor responses. Considering N sensors and M classes the expressions can be written as:

$$\{c_1 = \sum_{j=1}^{N} k_{1j} \cdot s_j + e \quad \cdots \quad c_M = \sum_{j=1}^{N} k_{Mj} \cdot s_j + e\}. \tag{7}$$

Geometrically, this means sectioning the sensor space with straight lines, each bisecting the space. The result is a partitioning of the space into volumes, each defining one class. Considering a set P of experimental data, the previous set of equations can be written in a compact matrix form as:

$$c_{MxP} = k_{MxN} \cdot s_{NxP} + E_{NxP}. \tag{8}$$

The matrix k_{MxN} containing the model parameters can then be directly estimated using the Gauss-Markov theorem to find the least squares solution of generic linear problems written in matrix form [37]:

$$K_{MxN} = c_{MxP} \cdot s^{+}_{PxN}. \tag{9}$$

Here the matrix s^{+}_{PxN} is the generalized inverse, or pseudo-inverse, of the matrix s_{NxP}. The operation of pseudo-inversion generalizes the inversion of square matrices to rectangular matrices. This solution is often called Multiple Linear Regression (MLR).Once the model is assessed, it allows assigning any unknown samples to one class. Due to the presence of the above mentioned error matrix (E_{NxP}), the model provides a numerical estimation of the "one-of-many" encoding of class assignment. In practice, something of different from 0 and 1 is obtained.

The estimated class assignment vector is called "classification score" and the sample is assigned to the class represented by the component

with the higher value. This gives the possibility of evaluating a sort of goodness of the classification by considering either the ratio between the first and second values of the components of the estimated $\vec{c}$ or the difference between the highest value and 1 (target value). The components of the matrix K_{MxN} define the importance of each sensor in the classification of each class. This information can also be used, as the loadings of PCA, to design and optimize the sensor array composition.The pseudo-inversion, like the inversion of square matrices, is influenced by the partial correlation among the sensors. Chemometrics offers methods to solve problems with colinear sensors, such as Principal Component Regression (PCR) and Partial Least Squares (PLS) [15].

Although designed for quantitative analysis, PLS may be applied to solve classification problems. In this case, PLS offers not only a more robust solution of the classification problem, but by plotting the latent variables it is possible to graphically represent the class separation. It is worth mentioning that the PLS latent variables are strictly related to but different from the PCA factors. Geometrically, PCA components are rotated in order to maximize their correlation with the components of the matrix of classes. In linear discrimination only classes separable by straight lines may be correctly classified. Classification improves if non-linear boundaries between classes are used. Figure 5 shows an example where a parabolic function achieves the separation where a linear boundary fails. A simple non-linear discriminant analysis can be obtained by a simple modification of the method previously discussed. As an example, let us consider the following quadratic form:

$$\{c_1 = \sum_{j=1}^{N} k_{1j} \cdot s_j + \sum_{j=1}^{N} h_{1j} \cdot s_j^2 + e \cdots c_M = \sum_{j=1}^{N} k_{Mj} \cdot s_j + \sum_{j=1}^{N} h_{Mj} \cdot s_j^2 + e\}. \tag{10}$$

This system may be written in the compact form of equation 5 by defining a suitable sensor matrix. Considering p measures, let us define a matrix T from the sensor responses, by:

$$T_{Px2N} = \begin{matrix} \ulcorner \\ \\ \llcorner \end{matrix} \begin{matrix} s_{11} & s_{1N} & s_{11}^2 & s_{1N}^2 \\ \vdots & & \ddots & \\ s_{p1} & s_{pN} & s_{p1}^2 & s_{pN}^2 \end{matrix} \begin{matrix} \urcorner \\ \\ \lrcorner \end{matrix} \tag{11}$$

In the same way, the parameters k_{ij} and h_{ij} are joined to form a unique parameter matrix H. With these definitions a linear problem may be written like that of equation 5. The matrix H can then be estimated either by direct pseudo-inversion or by PLS. It is worth noting

that increasing the order of the function increases the colinearity of the sensor matrix and the use of a chemometrics methodology (e.g. PLS) becomes more advantageous. Increasing the order of the discriminant function may solve highly complex classes distribution. The extreme solution is to use a method where the choice of the function is not required. Neural networks offer the possibility of solving the classification problem disregarding the functional form. It is well known that optimized neural networks may reproduce any kind of non-linear function. Neural networks derive all the knowledge from the experimental data, so that increasing the size of the calibration data-set increases the accuracy of the neural network based classifier.

8. Application to the investigation of Chemical Sensors properties

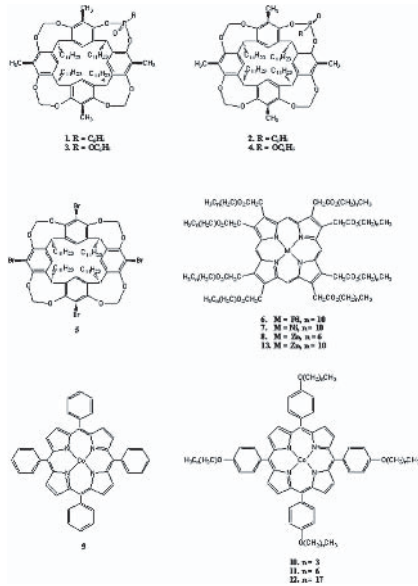


Figure 5.

As an example of the use of array methodology to study chemical sensor properties let us consider the thirteen molecular structures reported in Figure 5. To investigate the sensing properties of these molecules we studied the behaviour of the response of thickness shear mode resonators (TSMR) sensors, each coated with a molecular film, to different concentration of various volatile compounds (VOC). Analyte compounds were chosen in order to have different expected interaction mechanisms.

The sensors were exposed to the following VOC: n-pentane, methanol, benzene, triethylamine and acetic acid.

Data have been analyzed from a multivariate point of view. In this way the cooperative effects of the different materials is studied and the characteristics of each sensor are easily compared with those of the other sensors. PLS was used as a regression method for calculating the capability of the set of sensors to discriminate between the volatile compounds. Volatile compounds were checked at different concentrations in order to evaluate the response of sensors in a wide concentration range. Nevertheless, the concentration variation tends to shadow the reaction of sensors with analytes, since the sensor response contains both qualitative (sensor analyte interaction) and quantitative (analyte concentration) information. In order to remove the quantitative information, data have been normalized using the linear normalization discussed in section 3.

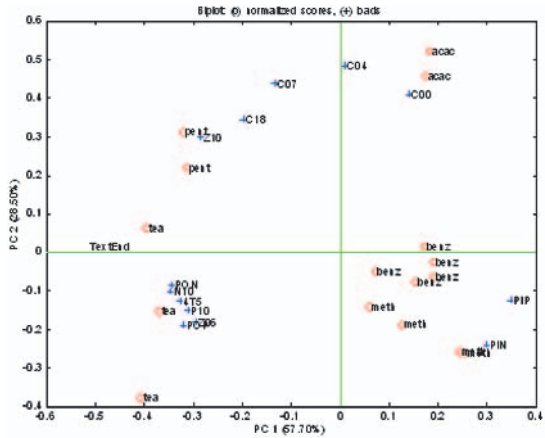


Figure 6.

Figure 6 shows the score plot of the first two latent variables of a PLS model aimed at discriminating the five substances. PLS is a data analysis method where the sensor signals are decomposed in latent variables obtained as linear combinations of the original variables, namely the sensors. The latent variables are chosen with the aim of maximizing the correlation between them and the scope of the regression. In the case treated here, the regression scope was the classification of the five substances. The original matrix decomposition gives place to the scores, the coordinates of the original sensors signals in the latent variable basis, and the loadings, the coordinates of the original variable axis (the sensors) in the new latent variable basis. These quantities can be plotted together in a so-called bi-plot. In a bi-plot the correlation between sen-

sors and data points can be investigated and some conclusions about the importance of some sensors in the detection of some particular species can be argued.

In Figure 6 it is possible to see that the individual species are quite distinct. In particular, methanol and benzene are close while triethylamine and pentane are displayed on the opposite side of the plot. Finally, the acetic acid lies in an orthogonal direction with respect to the others, which indicates that the interaction mechanism for acetic acid is completely different from the others.

The position of the sensors is also of great interest. As argued cavitands *1* and *3* lie in the same direction as the alcohol, in accordance with their selectivity towards alcohols. On the other hand, the co-linearity with benzene is not expected. This is a hint that in these cavitands the hydrogen bond interaction is improved but at the same time other kinds of interactions have a comparable magnitude.

Studying the loadings in Figure 6, the role of the length of the alkyl chain in porphyrins is also evident. *9* reveals in the plot a good affinity towards acetic acid, while the addiction of alkyl chains to the CoTPP skeleton increases the importance of sorption interaction. This is clearly visible in Figure 6 where the sensors shift from *9* up to *12*, covering the path leading from a high sensitivity towards acetic acid to a dominant sensitivity towards pentane, for which only sorption interaction is assumed to be present.

Other sensors are mostly grouped towards the triethylamine. In the case of porphyrins *6-8, 13* the coordinated metal is no longer able to drive the selectivity pattern and the presence of the peripheral alkyl chains completely shadows the coordination interactions. This result can explain the failure to observe the coordination interaction in the sensing mechanism of the metal complexes of the closely related alkyl chains functionalized phthalocyanines reported in the past by Göpel and coworkers [22].

In order to better investigate the relationship between sensor response and interaction mechanism it is useful to consider the way in which each volatile compound is expected to interact when in contact with a solid phase. These interactions can be modelled using the linear sorption energy relationship approach (LSER) [23].

According to this method and under the hypothesis of weak solubility interactions, the logarithm of the partition coefficient of a sorbent layer with respect to a certain volatile species is the linear combination of five terms expressing the intensity of five basic interaction mechanisms. They are: polarizability, polarity, two terms describe the hydrogen bonding considering the analyte acting as an acid and a base respectively, and

finally the solubility terms, a combination of dispersion and cavity interactions.

The relation can then be written as:

$$\log K = K_0 + r \cdot R + s \cdot \pi + a \cdot \alpha_2^H + b \cdot \beta_2^H + I \cdot \log L^{16}. \qquad (12)$$

Here K is the layer partition coefficient, R, $\pi, \alpha_2^H, \beta_2^H$, and $\log L^{16}$ are the coefficients of the volatile compounds and r, s, a, b, and l are the coefficients of the absorbing material.

LSER coefficients for various analytes are available in [23]. Figure 7 shows the values of the five parameters for the five volatile compounds considered here.

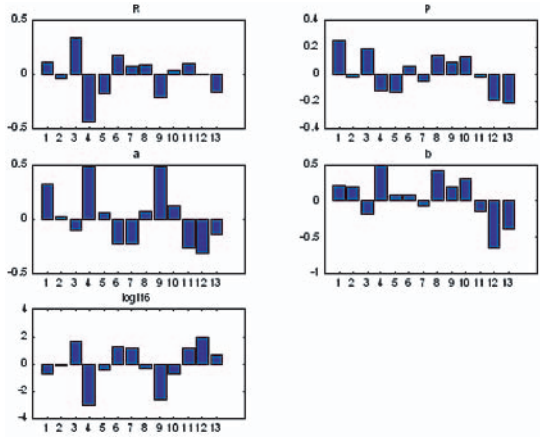


Figure 7.

As can be seen, all the compounds have a strong solubility interaction. Nonetheless Figure 6 shows that they are also well separated, so the more subtle difference between the other four interaction terms are important for their discrimination. It is worth noting that pentane interacts only via solubility, benzene has a very small hydrogen bonding term, while hydrogen bonding is present in methanol and acetic acid.

For a TSMR sensor, the partition coefficient turns out to be proportional to the overall sensitivity of the sensor ($S = \frac{\Delta f}{c}$) [24]. For some sensors a non-linear behaviour between sensor response and concentration is expected; in particular, for those cavitands functionalized to improve hydrogen bond interactions (labelled as *1* and *3* in Figure 8). In these cases, being limited by the number of interaction sites, the characteristic is expected to be steeper at low concentrations and then to reduce the slope at higher concentrations when the interaction sites

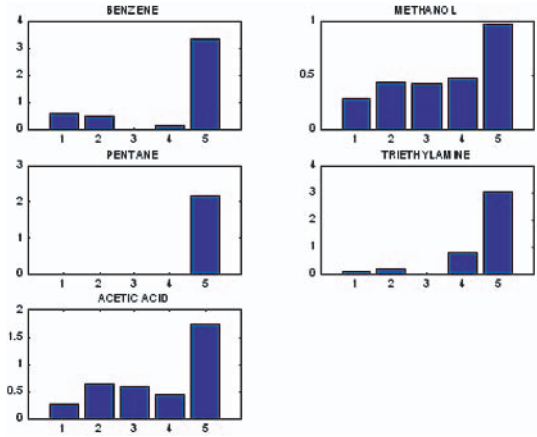


Figure 8.

may be considered completely occupied. These behaviours have been recorded several times once an absorbing layer passes some specific interaction sites [25] [26]. In order to take account of this fact and also to consider sensor sensitivity at its best, sensitivities and partition coefficients have been calculated at small concentration ranges where their values are higher.

The calculation of the partition coefficients for each sensor and for each volatile compound can then be used to estimate the value of the parameters (r, s, a, b, and l) characterizing the way in which each sensing layer may interact with volatile compounds.

Equation 12 can be written in matrix form as:

$$K_{nxm} = A_{nxm} \cdot S_{nxm} \tag{13}$$

where K is the partition coefficient matrix, A is the matrix of the volatile compound LSER parameters, and S is the matrix containing the sensor LSER parameters. n is the number of volatile compounds and m is the number of sensors (5 and 13, respectively, in this study).

K and A being known, the solution of equation 13 allows the determination of the LSER parameters characterizing the sensors studied here. Equation 12 was solved with the least squares method. In Figure 8 the LSER parameters for each sensor are shown.

It is interesting to compare the parameters exhibited by the cavitands with and without the functional group improving hydrogen bonding. The effect of the functionalization is expected to increase the parameter a; this effectively happens for *1* and *2*. The stronger effect of polarization is instead observed for the polarization term. This explains the

position of the benzene data in the plot of Figure 8. Another interesting effect is observed for compounds *9-12,* where alkyl chains of increasing length are introduced at the peripheral positions of the macrocycle; the decrease of the hydrogen bond (parameters a, b) and polar (parameter R) interactions are observed while at the same time the solution interaction grows continuously from negative to positive values. These effects are best observed in Figure 9 where a, b, P, and l are plotted versus the length of the alkyl chain. As the chain length increases the relative importance of the chain with respect to the porphyrin grows and the interaction with the porphyrin becomes quantitatively negligible.

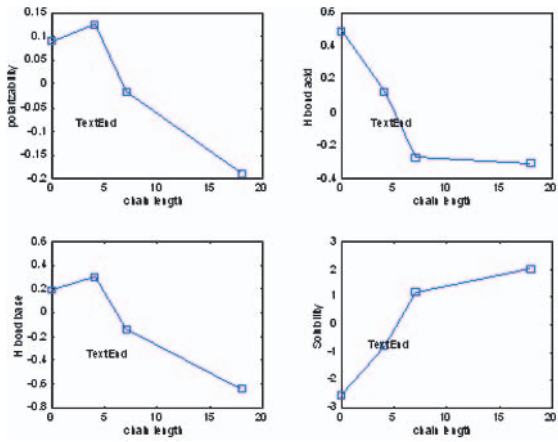


Figure 9.

This effect is complete in the case of porphyrins *6-8,13*, which behave in the same way, independent of metal present in the inner core of the macrocycle. In conclusion, the introduction of these alkyl chains from one side gives a higher porosity to the molecular film and, as a consequence, both a speed up of the response and an increase of the sensor response. On the other hand there is a significant decrease of the importance of the selective interactions in the sensing mechanism of the organic material and, as a consequence, a lowering of the selectivity of the sensor.

A good balance of these two opposite requirements is necessary in order to develop selective sensors.

9. Conclusions

Pattern recognition applied to several disciplines and practical problems produced a huge number of algorithms and techniques that are, in

principle, applicable to the classification of chemical sensor array data. Several of these techniques were actually utilized and frequently appeared in the literature. On the other hand, there is not yet an analysis approach based on natural olfaction paradigms. This makes the field almost totally dependent on the developments achieved in other fields. A comparison among the various techniques has also been attempted by some authors, but since the variables determining electronic nose performance are numerous (choice of sensors, samples, sampling systems,...) the results achieved are scarcely indicative of a general direction. Nonetheless some of these techniques became a classic of this field (such as scaling and normalization, PCA and discriminant analysis) and are often used without a deep understanding of the hypotheses on which they are based. The potentiality of these methods is not yet fully exploited. As for the other components of chemical sensor systems (e.g. sampling systems) a rethinking of the assumptions, implementations, and interpretation of the methods and solutions adopted is highly advised for a meaningful improvement in the field.

References

[1] K. Persaud, G. Dodd, Nature

[2] T. Pearce; Biosystems, 1997 41 pp. 43-67

[3] C. Di Natale, F. Davide, A. D'Amico; Sensors and Actuators B 1995 23 pp 111

[4] B.R. Kowalski and S. Wold; in Handbook of statistics Vol. 2, P.R. Krishnaiah and L.N. Kanal eds., North Holland Publ. (Amsterdam, The Netherlands), 1982 pp. 673 – 697

[5] Hierlemann A., Schweizer M., Weimar U., Göpel W.; in Sensors update Vol. 2, W. Göpel, J. Hesse, H. Baltes (eds.), VCH (Weinheim, Germany) 1995

[6] E.L. Hines, E. Llobet, J.W. Gardner; IEE Proc. –Circuits Devices Syst. 1999 146 pp. 297-310

[7] P.C. Jurs, G.A. Bakken, H.E. McClelland; Chemical Review 2000 100 pp. 2649-2678

[8] T. Eklöv, P. Mårtensson, I. Lundström; Analytica Chimica Acta 1997 353 pp. 291-300

[9] E. Martinelli, C. Falconi, A. D'Amico, C. Di Natale; Sensors and Actuators B 95 2003 132-139

[10] Martinelli E., Pennazza G., Di Natale C., D'Amico A.; Sensors And Actuators B 101 2004 346-352

[11] C. Di Natale, R. Paolesse, A. Macagnano, A. Mantini, A. D'Amico, A. Legin, L. Lvova, A. Rudnitskaya, Y. Vlasov; Sensors and Actuators B 2000 64 pp. 15-21

[12] G. Horner, C. Hierold; Sensors and Actuators B, 1990 2 p.p. 173-184

[13] C. Di Natale, R. Paolesse, A. Macagnano, V.I. Troitsky, T.S. Berzina, A. D'Amico; Analytica Chimica Acta 1999 384 pp. 249-259

[14] K. Fukunaga; Introduction to statistical pattern recognition, Academic Press, New York (NY, USA) 1992 H. Hotelling; J. Educat. Psych. 1933 24 pp. 498

[15] D.L. Massart, B.G. Vandegiste, S.N. Deming, Y. Michotte, L. Kaufmann; Data handling in science and technology vol.2: Chemometrics: a textbook, Elsevier (Amsterdam, The Netherlands) 1988

[16] R. Grossberg, I. Procaccia; Physica, 1983 9D pp.189-208

[17] E. Oja; Neural Networks, 1992 5 pp. 927-935

[18] J.F. Cardoso; Proc. of IEEE, 1998 9 pp. 2009-2025

[19] C. Di Natale, E. Martinelli, A. D'Amico; Sensors and Actuators B 82 2002 158-165

[20] O.M. Kvalheim, K. Oygard, O. Grahl-Nielsen; Analytica Chimica Acta, 1983 150 pp. 145

[21] T. Hastie, W. Stuetzle; Journal of the American Statistical Association, 1989 84 pp. 502-516 M.A. Kramer; AlChE Journal, 1991, 37 pp. 233-243

[22] M. Pardo, G. Sberveglieri, S. Gardini, E. Dalcanale; Sensors and Actuators B, 2000 69 pp. 359-365

[23] H. Kramer, M. Matthewes; IRE Trans. Inf. Theory, 1952 IT-2 pp.41

[24] T. Kohonen; Self Organising Map, 1995 Springer Verlag, Berlin Germany

[25] F. Davide, C. Di Natale, A. D'Amico; Sens. and Act. B, 1994 18 pp. 244 4.

[26] C. Di Natale, F. Davide, A. D'Amico, A. Hierleman, M. Schweizer, J. Mitrovics, U. Weimar, W. Göpel; Sens. and Act. B 1995 25 pp. 808

[27] G. Kraus, A. Hierleman, G. Gauglitz, W. Göpel: Technical Digest of Transducers '95 Conference, Stockholm (Sweden) 25-29 Jun. 1995, pp. 1675-678

[28] M.A. Kraajivels, J. Mao, A.K. Kain; Proc. Of 11th Int. Conf. On Pattern Recognition, 1992 IEEE Comp. Soc. Press, Los Alamitos (CA, USA) pp. 41

[29] C. Di Natale, A. Macagnano, A. D'Amico, F. Davide; Meas. Sci. and Techn. 1997 8 pp. 1-8

[30] C. Di Natale, J.A.J. Brunink, F. Bungaro, F. Davide, A. D'Amico, R. Paolesse, T. Boschi, M. Faccio, G. Ferri; Measurement Science and Technology 7 1996 1103-1114

[31] J.W. Sammons; IEEE Trans. Comp., 1969 C-18, pp. 401

[32] J.E.Dennis, R.B. Schnabel; Numerical methods for unconstrained optimization and non-linear equations, Prentice Hall Series in computational mathematics, New York, (NY, USA), 1983

[33] R.A. Johnson and D.W. Wichern; Applied multivariate statistical analysis, Prentice Hall, Englewood Cliffs (NJ, USA), 1982

[34] P.A. Lachenbruck, R.M. Mickey; Technometrics 1968 10 pp. 1-11

[35] B. Efron; Annual Statistics, 1979 7 pp. 1-26

[36] Y. Hamamoto, S. Uchimura, S. Tomita; IEEE Trans. on Pattern Analysis and Machine Intelligence, 1997 19 pp. 73-79

[37] S.L. Campbell, C.D. Meyer; Generalized inverses of linear transformations, Pitman (London, UK) 1979

[38] A.J. Maren (editor); Handbook of neural computing applications, J. Wiley and sons, (London, UK) 1991

[39] D.E. Rumelhart, J.L. McClelland; Parallel Distributed Processing: Explorations in the microstructure of cognition Vol.1: Learning internal representations by error propagation, MIT Press, Cambridge (MA, USA) 1986

[40] J. Hertz, A. Krogh, R.G. Palmer; Introduction to the theory if Neural Computation Vol.1, Addison Wesley, New York (NY,USA) 1991

[41] J.W. Gardner, E.L. Hines and M. Wilkinson; Measurement Science and Techn., 1990 1 pp. 446-451

[42] Neural Computing: a technology handbook for Professional II/plus© users, NeuralWare Inc., Pittsburgh (USA) 1993

[43] C.E. Martin, S.K. Rogers, D.W. Ruck; Proc. IEEE Int. Conf. Neural Network, 1994 pp. 305-308

[44] N. Ueda, R. Nakano; Proc. IEEE Int. Conf. Neural Network, 1995 vol. 1 pp. 101-105

FUNDAMENTALS OF TOMOGRAPHY AND RADAR

H.D. Griffiths and C.J. Baker
University College London
UK

Abstract Radar, and in particular imaging radar, has many and varied applications to security. Radar is a day/night all-weather sensor, and imaging radars carried by aircraft or satellites are routinely able to achieve high-resolution images of target scenes, and to detect and classify stationary and moving targets at operational ranges. Different frequency bands may be used, for example high frequencies (X-band) may be used to support high bandwidths to give high range resolution, while low frequencies (HF or VHF) are used for foliage penetration to detect targets hidden in forests, or for ground penetration to detect buried targets.

The techniques of tomographic imaging were originally developed in the context of medical imaging, and have been used with a number of different kinds of radiation, both electromagnetic and acoustic. The purpose of this presentation is to explore the application of tomographic imaging techniques at RF frequencies to a number of different applications in security, ranging from air defence to the detection of concealed weapons. Of particular interest is the use of ultra narrow band (UNB) transmissions with geometric diversity in a multistatic configuration to image moving targets. In the limit such transmissions could be CW, which would be particularly attractive for operation in a spectrally-congested environment. This arrangement effectively trades angular domain bandwidth for frequency domain bandwidth to achieve spatial resolution. Also of interest is the improvement in target classification performance afforded by multi-aspect imaging.

The presentation will review the theory of tomographic imaging, then discuss a range of applications to the overall security problem, the relevant system configurations in each case, the achievable performance and critical factors, and identify promising areas for future research.

Keywords: radar; radar imaging; tomography; high resolution; synthetic aperture radar; interferometry; polarimetry; Radon transform; projection slice theorem; backprojection.

Jim Byrnes and Gerald Ostheimer (eds.),
Advances in Sensing with Security Applications, 171–187.

1. Introduction

Radar, and in particular imaging radar, has many and varied applications to security. Radar is a day/night all-weather sensor, and imaging radars carried by aircraft or satellites are routinely able to achieve high-resolution images of target scenes, and to detect and classify stationary and moving targets at operational ranges. Short-range radar techniques may be used to identify small targets, even buried in the ground or hidden behind building walls. Different frequency bands may be used, for example high frequencies (X-band) may be used to support high bandwidths to give high range resolution, while low frequencies (HF or VHF) are used for foliage penetration to detect targets hidden in forests, or for ground penetration to detect buried targets.

In the notes that follow we consider the formation of high-quality radar imagery, and the means by which it is possible to extract useful target information from such imagery.

2. Imaging and Resolution

Firstly we can establish some of the fundamental relations for the resolution of an imaging system. In the down-range dimension resolution $\triangle r$ is related to the signal bandwidth B, thus

$$\triangle r = \frac{c}{2B}. \tag{1}$$

High resolution may be obtained either with a short-duration pulse or by a coded wide-bandwidth signal, such as a linear FM chirp or a step-frequency sequence, with the appropriate pulse compression processing. A short-duration pulse requires a high peak transmit power and instantaneously-broadband operation; these requirements can be relaxed in the case of pulse compression.

In the first instance cross-range resolution is determined by the product of the range and beamwidth θ_B. The beamwidth is determined by the size of the aperture d and thus cross-range resolution is given by

$$\triangle x = r\theta_B \approx \frac{r\lambda}{d}. \tag{2}$$

As most antenna sizes are limited by practical aspects (such as fitting to an aircraft) the cross range resolution is invariably much inferior to that in the down range dimension. However, there are a number of techniques that can improve upon this. All of these are ultimately a function of the change in viewing or aspect angle. Thus in the azimuth (cross-range) dimension the resolution $\triangle x$ is related to the change in aspect angle $\triangle\theta$

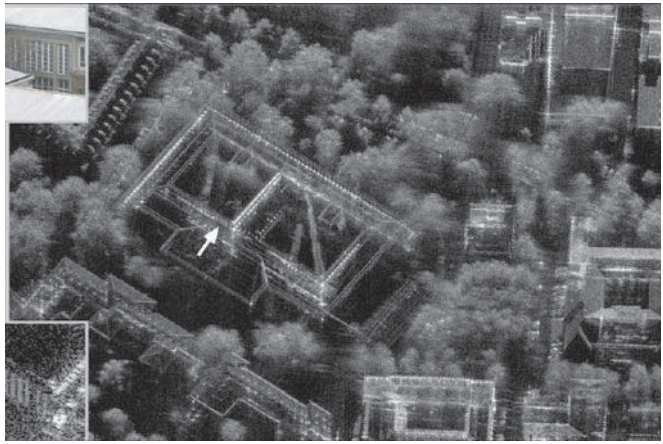


Figure 1. High resolution SAR image of a part of the university campus in Karlsruhe (Germany). The white arrow refers to a lattice in the left courtyard, which is shown in more detail in the small picture on the left bottom. The corresponding optical image is shown on the left top (after Brenner and Ender[4]).

as follows:

$$\triangle x = \frac{\lambda}{4 \sin\left(\frac{\triangle\theta}{2}\right)}. \tag{3}$$

For a linear, stripmap-mode synthetic aperture, equation (3) reduces to $\triangle x = \frac{d}{2}$, which is independent of both range and frequency. Even higher resolution can be obtained with a spotlight-mode synthetic aperture, steering the real-aperture beam to keep the target scene in view for a longer period, and hence forming a longer synthetic aperture.

Realistic limits to resolution may be derived by assuming a maximum fractional bandwidth $\frac{B}{f_0}$ of 100%, and a maximum change in aspect angle of $\triangle\theta = 30^\circ$ (higher values than these are possible, but at the expense of complications in hardware and processing). These lead to $\triangle r = \triangle x = \frac{\lambda}{2}$.

In the last year or so results have appeared in the open literature which approach this limit. Figures 1 and 2 show two examples from a recent conference of, respectively, an urban target scene and of aircraft targets. Critical to the ability to produce such imagery is the ability to characterise and compensate for motion errors of the platform, which can be done by autofocus processing [6]. Of course, motion compensation becomes most critical at the highest resolutions.

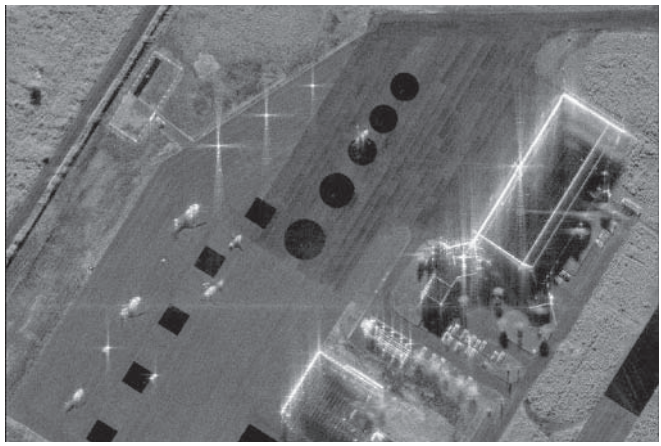


Figure 2. Example of 3-look image yielding 10 cm resolution (after Cantalloube and Dubois-Fernandez [5])

3. Tomographic Imaging

The techniques of tomography were developed originally for medical imaging, to provide 2D cross-sectional images of a 3D object from a set of narrow X-ray views of an object over the full 360° of direction. The results of the received signals measured from various angles are then integrated to form the image, by means of the Projection Slice Theorem. The Radon Transform is an equation derived from this theorem which is used by various techniques to generate tomographic images. Two examples of these techniques are Filtered Backprojection (FBP) and Time Domain Correlation (TDC). Further descriptions of these techniques may be found in [20].

In radar tomography the observation of an object from a single radar location can be mapped into Fourier space. Coherently integrating the mappings from multiple viewing angles enables a three dimensional projection in Fourier space. This enables a three dimensional image of an object to be constructed using conventional tomography techniques such as wavefront reconstruction theory and backprojection where the imaging parameters are determined by the occupancy in Fourier space. Complications can arise when target surfaces are hidden or masked at any stage in the detection process. This shows that intervisibility characteristics of the target scattering function are partly responsible for determining the imaging properties of moving target tomography. In other words, if a scatterer on an object is masked it cannot contribute to the imaging process and thus no resolution improvement is gained. However, if a higher number of viewing angles are employed then this

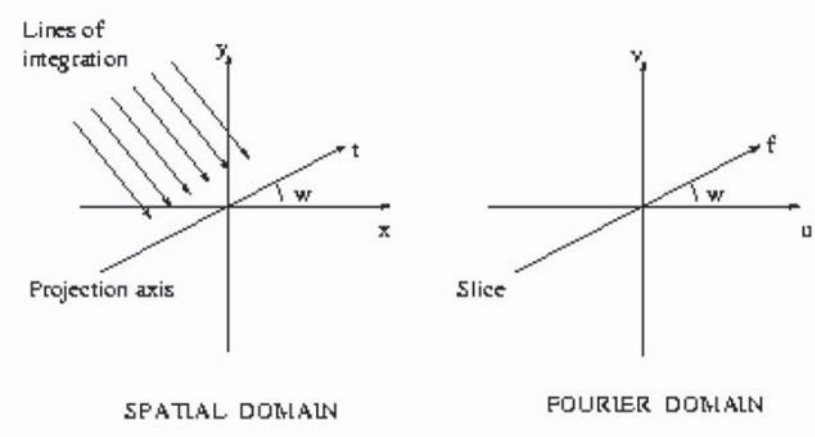


Figure 3. Tomographic reconstruction: the Projection Slice Theorem.

can be minimised. Further complications may arise if (a) the point scatterer assumption used is unrealistic (as in the case of large scatterers introducing translational motion effects), (b) the small angle imaging assumption does not apply and (c) targets with unknown motions (such as non-uniform rotational motions) create cross-product terms that cannot be resolved.

4. The Projection Slice Theorem

The Tomographic Reconstruction (TR) algorithm makes use of the Projection-Slice theorem of the Fourier transform to compute the image. The Projection-Slice theorem states that the 1D Fourier transform of the projection of a 2D function $g(x, y)$, made at an angle w, is equal to a slice of the 2D Fourier transform of the function at an angle w, see Figure 3. Whereas some algorithms convert the outputs from many radars simultaneously into a reflectivity image using a 2D Fourier transform, TR generates an image by projecting the 1D Fourier transform of each radar projection individually back onto a 2D grid of image pixels. This operation gives rise to the term Backprojection. The image can be reconstructed from the projections using the Radon transform. The equation below shows this:

$$\begin{aligned} g(x,y) &= \int_0^{\pi}\int_{-\infty}^{\infty} P(f)\cdot|f|\cdot e^{j2\pi f(x\cos w + y\sin w)}\,df\,dw \qquad (4)\\ \text{where}\quad w &= \text{projection angle}\\ P(f) &= \text{the Fourier transform of the 1-D projection } p(t). \end{aligned}$$

The Filtered Backprojection (FBP) method may be used to process by reconstructing the original image from its projections in two steps: Filtering and Backprojection.

Filtering the projection: The first step of FB Preconstruction is to perform the frequency integration (the inner integration) of the above equation. This entails filtering each of the projections using a filter with frequency response of magnitude $|f|$.

The filtering operation may be implemented by ascertaining the filter impulse response required and then performing convolution or a FFT/IFFT combination to correlate $p(t)$ against the impulse response.

Backprojection: The second step of FB Preconstruction is to perform the angle integration (the outer integration) of the above equation. This projects the 1D filtered projection $p(t)$ onto the 2D image by following these steps: place a pixel-by-pixel rectangular grid over the XY plane, then place the 1D filtered projection $p(t)$ in position at angle w for each pixel, then get the position of the sample needed from the projection angle and pixel position. Interpolate the filtered projection to obtain the sample. Add this backprojection value multiplied by the angle spacing. Repeat the whole process for each successive projection.

5. Tomography of Moving Targets

A development of these concepts has been the idea of imaging of moving targets using measurements from a series of multistatic CW or quasi-CW transmissions, giving rise to the term 'ultra narrow band' (UNB) radar. This may be attractive in situations of spectral congestion, in which the bandwidth necessary to achieve high resolution by conventional means (equation (1)) may not be available. Narrow band CW radar is also attractive as peak powers are reduced to a minimum, sidelobes are easier to control, noise is reduced and transmitters are generally low cost. Applications may range from surveillance of a wide region, to the detection of aircraft targets, to the detection of concealed weapons carried by moving persons. In general the target trajectory projection back to a given radar location will determine resolution. A random trajectory of constant velocity will typically generate differing resolutions in the three separate dimensions. However, even if there is no resolution improvement there will be an integration gain due to the time series of radar observations. A Hamming window or similar may be required to reduce any cross-range sidelobe distortions. The treatment which follows is taken from that of Bonneau, Bascom, Clancy and Wicks [3].

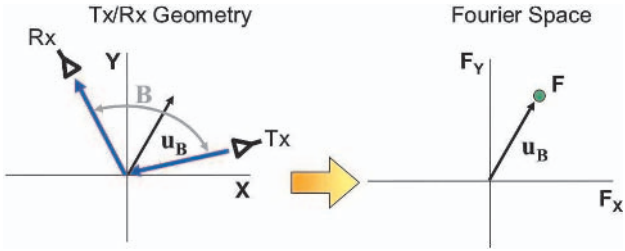


Figure 4. Relationship between bistatic sensor geometry and representation in Fourier space (after [3]).

Figure 4 shows the relationship between the bistatic sensor geometry and the representation in Fourier space. The bistatic angle is B and the bistatic bisector is the vector $\boldsymbol{u_B}$. The corresponding vector $\boldsymbol{F}$ in

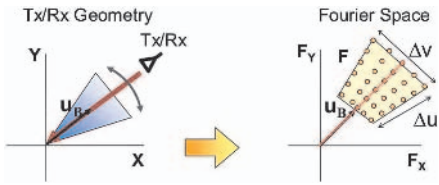


Figure 5. Fourier space sampling and scene resolution for a monostatic SAR (after [3]).

Fourier space is given by

$$\boldsymbol{F} = \frac{4\pi f}{c} \cos\left(\frac{B}{2}\right) \boldsymbol{u_B} \tag{5}$$

Figure 5 shows the equivalent relationship for a monostatic geometry. The resolutions are inversely proportional to the sampled extents $\triangle u$ and $\triangle v$ in Fourier space, thus

$$\triangle r = \frac{2\pi}{\triangle u} \qquad \triangle x = \frac{2\pi}{\triangle v} \tag{6}$$

which should be compared to equations (1),(2) and (3).

In an UNB radar the finite bandwidth of the radar signal limits the range resolution. However, this resolution can be recovered by multistatic measurements over a range of angles. Figure 6 shows four examples, and the Fourier space sampling corresponding to each.

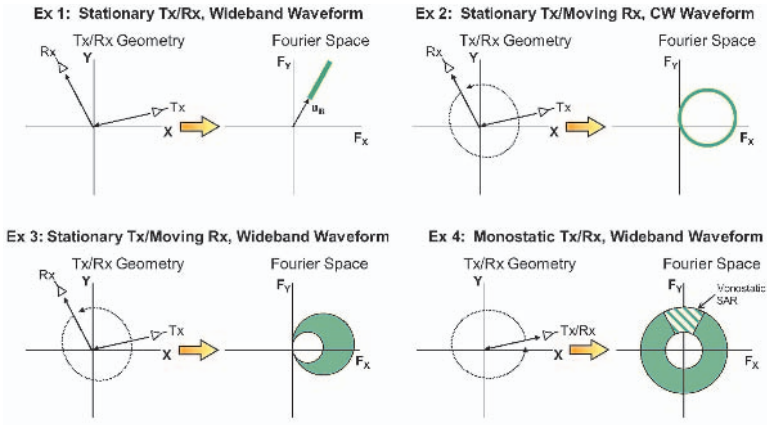


Figure 6. Fourier space sampling and scene resolution for four examples: (i) stationary tx/rx, wideband waveform; (ii) stationary tx, moving rx, CW waveform; (iii) stationary tx, moving rx, wideband waveform; (iv) monostatic tx/rx, wideband waveform (after [3]).

6. Applications

The applications of high resolution radar imagery are hugely varied and numerous. Invariably high resolution is used as a tool to improve the information quality resident in an electromagnetic backscattered signal. The resulting imagery may be used to gain information over extremely wide areas such as the earth's oceans, where data pertaining to sea state, current movements, etc. can be derived. Over the land, imagery is used for crop monitoring, evaluation of rain forest clearings, bio mass estimation and many other tasks. At the highest of resolution information on single objects is possible and it is here that the security applications are more likely. In particular improved detection and classification of objects such as vehicles, aircraft, ships and personnel, and at the very highest resolution, concealed weapons, are potentially possible. We consider a small sample here.

7. Automatic Target Recognition

These examples are illustrative of the potential of synthetic aperture imaging. However, it should be appreciated that the challenge is to extract useful information on the desired targets from such imagery.

The problem of determining the class to which a target belongs directly relies upon the amount of information available. ATRs are systems that contain an input sub-system that accepts pattern vectors from the

feature space, and a decision-maker sub-system that has the function of deciding the class to which the sensed attributes belong. Here we interchangeably refer to this process using the terms classification and recognition.

Pre-processing raw data is necessary in order to increase the quality of the radar signatures. Principal discriminating factors for classification purposes are Range Resolution, Side-Lobe Level (SLL) and Noise Level. Higher resolution means better point scatterers separation but the question of compromise regarding how much resolution is needed for good cost-recognition is difficult to resolve. Generally, high SLLs mean clearer range profiles but this also implies deterioration in resolution. Eventually, low noise levels mean high quality range profiles for classification. In this chapter we concentrate on the particular situation in which a single non-cooperative target has been previously detected and tracked by the system.

The improvement in performance due to the available multiplicity of perspectives is investigated examining one-dimensional signatures and the classification is performed on raw data with noise floor offset removed by target normalization. After generating a target mask in the range profile, the noise level is measured in the non-target zone and then subtracted from the same area. The result is a more prominent target signature in the range window.

Real ISAR turntable data have been used to produce HRR range profiles and images. In view of the fact that the range from the target is approximately constant, no alignment is needed. Three vehicle targets classified as A, B and C form the sub-population problem. Each class is described by a set of one-dimensional signatures covering 360 degrees of rotation on a turntable. After noise normalisation, a 28 dB SNR is achieved. Single chirp returns are compressed giving 30 cm of range resolution. The grazing angle of the radar is 8 degrees and $2''$ of rotation is the angular interval between two consecutive range profiles. Therefore, 10000 range profiles are extracted from each data file over the complete rotation of 360 degrees. The training set of representative vectors for each class is made by 18 range profiles, taken approximately every 20 degrees for rotation of the target. The testing set of each class consists of the remaining range profiles excluding the templates.

Three algorithms have been implemented in both single and multi-perspective environments. In this way any bias introduced by a single algorithm should be removed. The first is the statistical Naïve Bayesian Classifier. It reduces the decision-making problem to simple calculations of feature probabilities. It is based on Bayes' theorem and calculates the posterior probability of classes conditioned on the given unknown feature

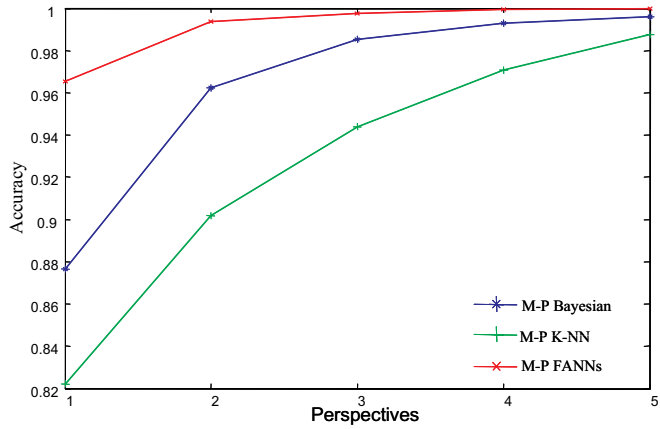


Figure 7. Multi-perspective classifier accuracies.

vector. The second is a rule-based method for classification: K-Nearest Neighbours (K-NN) algorithm. The rule consists of measuring and minimising the number of K distances from the object to the elements of the training set. The last approach involves Artificial Neural Networks (ANN), where the information contained in the training samples is used to set internal parameters of the network. In this work, Feed-forward ANNs (FANNs) supervised by a back-propagation strategy have been investigated and implemented.

We first consider classification based upon a multiplicity of viewing angles rather than using this multiplicity to form a single tomographic image. The combination of views of a target from a number of different aspects would be expected intuitively to provide an improvement in classification performance as clearly the information content should increase. Three different ways of combining the aspects are used here to illustrate possible performance improvements: These are the Naïve Bayesian Classifier, K-nearest neighbours (KNN), and Feed-forward Artificial Neural Networks (FANN). Details of these algorithms are provided in reference [21]. Figure 7 shows the improvement in classifier performance as a function of number of perspectives. In Figure 7, the classification performances of the three implemented classifiers are compared versus the number of perspectives used by the network. As anticipated, because of the nature of the data and the small available number of targets, the classifiers start from a high level of performance when using only a single aspect angle. It can be seen that there is a significant benefit in going from 1 to 2 perspectives, and a small additional benefit from 2 to 3, but rather less from further additional perspectives.

However, improved performance is achieved with increased radar number in the network. In particular, an improvement of 6.46% is shown comparing the single and the two-perspective classifier. The accuracy variation is then reduced to ±2.31% overall for two to three perspectives, ±1.2% for three to four and, finally, ±0.67% for four to five perspectives. In conclusion, the greatest improvement in performance can be observed with just a small number of radars. Since the number of perspectives, and therefore the number of radars, is strictly related to complexity, costs and execution time of the network, for classification purposes it might be a reasonable trade-off implementing networks involving a small number of nodes. However this analysis is against a small number of target classes and these conclusions require further verification.

We now examine the extent to which SNR affects classification and whether multi-perspective scenarios are effective at different SNR levels. The FANNs classifier has been applied for this particular task. The range profiles are corrupted with additive white Gaussian noise. The original data, after noise removal, has a 28 dB SNR. Subsequently, the classifier is tested with range profiles with 24, 20 and 16 dB SNRs. The object has a length of 6.2 metres (it falls into about 20 range bins). As can be seen, as the SNR decreases, some of the useful features become less distinct, making the range profile more difficult to be classified. In Figure 8 performance is plotted versus the number of perspectives used and SNR levels, showing how the enhancement in classification varies with different noise levels. The plot shows an increase in classification performance with numbers of perspectives in each case. The increase is greater at the lowest values of SNR. However below an SNR of 15 dB the performance quickly degrades indicating that classifiers will be upset by relatively small amounts of noise.

8. Bandwidth Extrapolation

Radar resolution in range is directly limited by the bandwidth of the received signal (1). High resolution can thus be achieved by transmitting a wideband signal but at the expense of high spectrum occupancy. However, the actual trend is to favor more efficient use of the electromagnetic spectrum due to a growing need for commercial applications. In this context several solutions are proposed to create smarter transmitters and receivers. For instance, there is hope that cognitive radios could improve the RF spectrum occupancy in time by using devices that can "dynamically adjust their RF characteristics and performance in real time to reflect what may be a rapidly changing local interference environment" [24]. In radar, one solution consists of transmitting

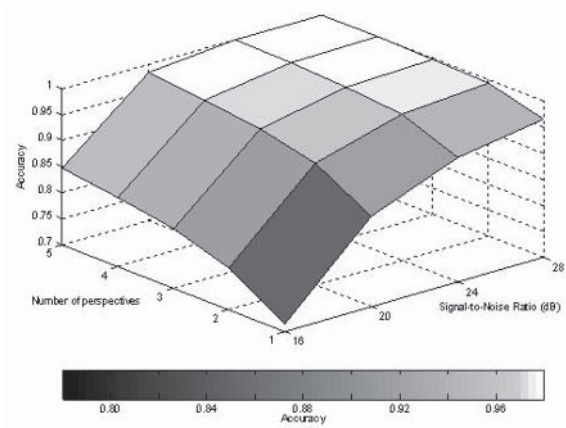


Figure 8. Variation with number of perspectives and signal-to-noise ratio.

narrow-band pulses and extrapolating the signal across a wider bandwidth. Bandwidth Extrapolation begins by fitting an *a priori* model to a measured radar signal. Auto-Regressive Models are commonly used for this purpose. These linear models are assumed to approximate scattering mechanisms that are often non-linear in practice. Model-parameter values can be obtained using super-resolution techniques such as MUSIC, Matrix Pencil and ESPRIT. Once the models have been fitted to the measured signal, they are utilised to predict the radar samples outside the band of measurements. Performances are affected in many ways by various parameters including Signal-to-Noise Ratio, target complexity and number of samples collected. In general, the models are deficient because part of the required information is corrupted by the noise. This particularly affects the extrapolation of a signal scattered by extended targets such as aircrafts [25]. Current research is focussed on the use of several bandwidths for building models that are more robust across a larger bandwidth. Such techniques can also be applied to ISAR image reconstruction when the initial signal is corrupted by interference.

In parallel with the development of bandwidth-extrapolation techniques is that of pattern recognition. Patterns created by the influence of the strongest scatterers on the target signature can be used to design additional knowledge-based methods. Patterns observed in time, frequency or angle provide information that can be used for classification and prediction. However, simple patterns are also associated with simple scattering mechanisms.

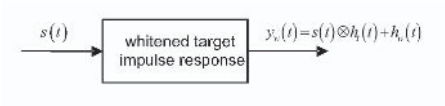


Figure 9. Whitening of the target impulse response in target-matched illumination.

9. Target-matched Illumination

The classical concept of the matched filter was further developed by Gjessing [9] [10] [11] and by Bell [2] to consider the optimum waveform for the detection of a target of a given range profile against a noise background. The target is characterized in terms of its impulse response as a function of delay time (i.e. range), which will also be a function of aspect angle (and therefore which in practice would require a library of target impulse responses versus aspect angle). The concept has been extended by Guerci, Pillai and co-workers [8] [13] [14] [19] to include the detection of a target against nonhomogeneous noise, and also to the problem of discriminating different targets. The problem is posed as follows (Figure 9) using the notation adopted by Guerci. The radar transmits a signal $s(t)$ towards a target, whose impulse response is $h_T(t)$. The echo signal $y(t)$ is the convolution of $s(t)$ with $h_T(t)$. To this is added noise $n(t)$, so the received signal is

$$r(t) = (s(t) \otimes h_T(t)) + n(t) \tag{7}$$

where $\otimes$ denotes the convolution operator.

The receiver is characterized by its impulse response $h_R(t)$. The problem is then to choose $s(t)$ and $h_R(t)$ to maximise the signal-to-interference ratio, which can be expressed in mathematical terms as follows:

$$y_0 = \max_s \max_h \rho(t_0) \tag{8}$$

where

$$SINR = \rho(t_0) = \frac{y_s^2(t_0)}{\langle y_0(t_0) \rangle^2}. \tag{9}$$

y_s is the signal component of the output and yo is the component contributed by interference and noise.

The first step is to maximise the SNIR working on the receiver. Once the optimal impulse response of the receiver, $H_{MF}(t)$, has been determined, it follows (Figure 9) that :

$$SNIR_0 \frac{1}{\sigma_w^2} \int_{T_i}^{T_f} |y_w(t)|^2 dt = f\left(s(t)\right) \tag{10}$$

where T_i and T_f are the time boundaries of the receiver and $y_w(t)$ is the signal echo after the whitening filter.

At this stage, the problem is to maximise SNIR at the instant of detection t_0 over the input signal $s(t)$ of finite energy and duration. Grouping the expressions for both whitening filter and matched filter:

$$h(t) = h_T \otimes h_w(t). \tag{11}$$

Using this, the integral in (10) can be written

$$\int_{T_i}^{T_f} |y_w(t)|^2 dt = \int_0^T s(\tau_1) \cdot \int_0^T s^*(\tau_2) K^*(\tau_1, \tau_2) d\tau_2 d\tau_1 \tag{12}$$

where

$$K(\tau_1, \tau_2) = \int_{T_i}^{T_f} h^*(t - \tau_1) h(t - \tau_2) dt. \tag{13}$$

The solution must satisfy a homogeneous Fredholm integral of the second kind with Hermitian kernel:

$$\lambda_{\max} S_{opt}(t) = \int_0^T S_{opt}(\tau) K(t - \tau) d\tau. \tag{14}$$

This principle can be extended to different models including signal dependent noise (clutter) [19]. In this case, one must take the non-linear term into account in the signal to interference plus noise equation:

$$SINR_0 = \frac{\left| \frac{1}{2\pi} \int_{-\infty}^{+\infty} H_R(\omega) H_T(\omega) S(\omega) e^{-j\omega T_f} d\omega \right|^2}{\frac{1}{2\pi} \int_{-\infty}^{+\infty} |H_R(\omega)|^2 \cdot \left(G_n(\omega) + G_c(\omega) \, |S(\omega)|^2 \right) d\omega} \tag{15}$$

where
$G_n(\omega)$ is the additive noise spectrum,
$G_c(\omega)$ is the clutter spectrum,
$H_T(\omega)$ and $H_R(\omega)$ are the transmitter and receiver spectrum respectively.

From the above model we can derive three main cases:

a) the clutter is not significant compared to the additive noise: $G_c(\omega) \ll G_n(\omega)$

b) the additive noise is not significant relative to the clutter: $G_c(\omega) \gg G_n(\omega)$

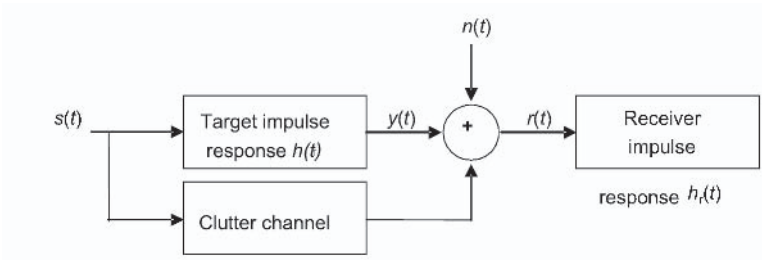


Figure 10. Target-matched illumination with signal-dependent noise (clutter).

c) clutter and noise are of equivalent power:
$G_c(\omega) \sim G_n(\omega)$

Unlike the first two cases, which can be solved by the previous method, the third one (clutter and noise) has been studied by Guerci using an iterative procedure [19].

Applications: Potential applications of matched illumination are:

- identical target resolution (Figure 10)
- target identification
- target tracking/tagging
- target aspect uncertainty.

10. Conclusion

The techniques described and the results presented demonstrate the value of radar imaging to security problems. In particular a novel multi-perspective approach to classification has been presented. High resolution data from turntable measurements have been processed and HRR range profiles and ISAR imagery from a number of radar targets have been successfully formed. Three algorithms for classification have been implemented using the radar signatures as the basis for recognition in both single and multi-perspective environments. Improvements in classification performance have been shown by using different information gathered by a network of radars whose nodes are placed around the target. In addition, the increase in recognition accuracy is not linear with the number of perspectives used. Greater positive variations can be seen for a small number of nodes employed in the network. Furthermore, the results obtained at lower SNR levels show valuable improvements in target recognition for more practical classification purposes. Alternatively the embedded information approach of target adaptive matched

illumination offers a means of directly implementing classification via exploitation of prior knowledge. Whilst encouraging, these conclusions should be treated with some caution as they are somewhat limited by the restricted available data.

11. Acknowledgements

We express our thanks to the students with whom we have worked on these subjects and whose results we have used, in particular Hervé Borrion, Shirley Coetzee and Michele Vespe, and to the organisations, including the UK Ministry of Defence, the US Air Force Office of Scientific Research, the UK Engineering and Physical Sciences Research Council, QinetiQ and its predecessors, BAE SYSTEMS, Thales Sensors and AMS, who have supported the various projects. We also thank Erik De Witte and Hervé Borrion for their help in rendering this document into LaTeX.

References

[1] Ausherman, D.A., Kozma, A., Walker, J.L., Jones, H.M. and Poggio, E.C., 'Developments in radar imaging', IEEE Trans. Aerospace & Electronics Systems, Vol. AES-20, pp 363-400, 1984.

[2] Bell, M., 'Information theory and radar waveform design', IEEE Trans. Information Theory, Vol.39, No.5, pp 1578-1597, 1993.

[3] Bonneau, R.J., Bascom, H.F., Clancy, J.T. and Wicks, M.C., 'Tomography of moving targets (TMT)'.

[4] Brenner, A.R. and Ender, J.H.G., 'Airborne SAR imaging with subdecimetre resolution', Proc. EUSAR 2004 Conference, pp 267-270.

[5] Cantalloube, H. and Dubois-Fernandez, P. 'Airborne X-band SAR imaging with 10 cm resolution—technical challenge and preliminary results', Proc. EUSAR 2004 Conference, pp 271-274.

[6] Carrara, W.G., Goodman, R.S. and Majewski, R.M., Spotlight Synthetic Aperture Radar: Signal Processing Algorithms, Artech House, 1995.

[7] Coetzee, S.L., 'Radar tomography of moving targets', MPhil. Transfer thesis, University College London, 2004.

[8] Garren, D.A., Osborn, M.K., Odom, A.C., Goldstein, J.S., Pillai, S.U. and Guerci, J.R., 'Enhanced target detection and identification via optimized radar transmission pulse shape', IEE Proc. Radar, Sonar and Navigation, Vol.148, No.3, pp 130-138, June 2001.

[9] Gjessing, D.T., 'Adaptive techniques for radar detection and identification of objects in an ocean environment', IEEE J. Ocean Engineering, Vol.6, No.1, pp 5-17, 1981.

[10] Gjessing, D.T., Target Adaptive Matched Illumination Radar: Principles and Applications, Peter Peregrinus, 1986.

[11] Gjessing, D.T. and Saebboe, J., 'Bistatic matched illumination radar involving synthetic aperture and synthetic pulse for signal to clutter enhancement and target characterization', Proc. 2001 CIE International Conference on Radar, Beijing, pp 20-24, 15–18 October 2001.

[12] Gjessing, D.T. and Saebboe, J., 'Bistatic matched illumination radar involving synthetic aperture and synthetic pulse for signal to clutter enhancement and target characterization', Proc. 2001 CIE International Conference on Radar, Beijing, pp 20-24, 15–18 October 2001.

[13] Grieve, P.G. and Guerci, J.R., 'Optimum matched illumination-reception radar', US Patent S517522, 1992.

[14] Guerci, J.R., 'Optimum matched illumination-reception radar for target classification', US Patent S5381154, 1995.

[15] Knaell, K.K. and Cardillo, G.P., 'Radar tomography for the generation of three-dimensional images', IEE Proc. Radar Sonar Navig., vol. 142, no. 2, pp. 54-60, 1995.

[16] Munson, D.C. Jr., O'Brien, J.D. and Jenkins, W.K., 'A tomographic formulation of spotlight-mode synthetic aperture radar', Proc. IEEE, Vol.71, No.8, pp 917-925, 1983.

[17] Oliver, C.J. and Quegan, S., Understanding SAR Images, Artech House, 1998.

[18] Pasmurov, A. Ya. and Zinoviev, Yu. S., Radar Imaging and Tomography, to be published by Peter Peregrinus, Stevenage, 2005.

[19] Pillai, S.U., Oh, H.S., Youla, D.C. and Guerci, J.R., 'Optimum transmit-receiver design in the presence of signal-dependent interference and channel noise', IEEE Trans. Information Theory, Vol.46, No.2, pp 577-584, March 2000.

[20] Soumekh, M., Synthetic Aperture Radar Signal Processing with MatLab Algorithms, Artech House, 1999.

[21] Vespe, M., Baker, C.J. and Griffiths, H.D., 'Multi-perspective target classification', Proc. RADAR 2005 Conference, Washington DC, IEEE Publ. No. 05CH37628, pp 877-882, 9–12 May 2005.

[22] Walker, J.L., 'Range Doppler imaging of rotating objects', IEEE Trans. AES, Vol. 16, pp 23-52, 1980.

[23] Wehner, D.R., High Resolution Radar, Artech House, 1987.

[24] Walko, J., 'Cognitive Radio', IEE Review, p36, May 2005.

[25] Borrion, H., Griffiths, H. Money, D., Tait P. and Baker C., 'Scattering centre extraction for Extended targets', Proc. RADAR 2005 Conference, Washington DC, IEEE Publ., pp 173-178, 9–12 May 2005.

REMOTE SENSING USING SPACE BASED RADAR

Braham Himed
Air Force Research Laboratory Sensors Directorate
26 Electronic Parkway Rome, NY 13441
Email: Braham.Himed@rl.af.mil

Ke Yong Li
C&P Technologies, Inc. 317 Harrington Avenue
Suites 9&10 Closter, NJ 07624
Email: kli@cptnj.com

S. Unnikrishna Pillai
Dept. of Electrical Engg. Polytechnic University
6 MetroTech Center Brooklyn, NY 11201
Email: pillai@hora.poly.edu

Abstract A Space-Based Radar (SBR) is a reconnaissance, surveillance, and target acquisition system capable of supporting a wide variety of joint missions and tasks simultaneously, including battle management, command and control, target detection and tracking, wide area surveillance and attack operations. SBR also supports traditional intelligence, surveillance and reconnaissance missions such as indications, warning, and assessment. These mission areas cover the strategic, operational, and tactical levels of operations of interest. SBR systems are also used for earth science projects. However, an SBR system, by virtue of its motion, generates a Doppler frequency component to the clutter return from any point on the earth as a function of the SBR-earth geometry. The effect of earth's rotation around its own axis is shown to add an additional component to this Doppler frequency. The overall effect of the earth's rotation on the Doppler turns out to be two correction factors in terms of a crab angle affecting the azimuth angle, and a crab magnitude scaling the Doppler magnitude of the clutter patch. Interestingly both factors depend only on the SBR orbit inclination and its latitude and not on the specific location of the clutter patch of interest. It is also shown that earth's rotation together with the range foldover phenomenon inherent in such systems; significantly degrade the clut-

Jim Byrnes and Gerald Ostheimer (eds.),
Advances in Sensing with Security Applications, 189–213.

ter suppression performance of adaptive processing algorithms. In this chapter, we provide analytical derivations of these phenomena and their impact on performance, and suggested ways to remedy for these effects are shown through computer simulations.

Keywords: SBR; STAP; earth rotation; range ambiguity; crab angle; crab magnitude; Doppler dispersion; range dependency; waveform diversity; Doppler warping.

1. Introduction

SBR because of its height can cover a very large area on earth for intelligence, surveillance and monitoring of ground moving targets. Once launched into orbit, the SBR moves around the earth while the earth continues to rotate on its own axis. By adjusting the SBR speed and orbit parameters, it is thus possible to scan various parts of the earth periodically and collect data. Such an SBR based surveillance system can be remotely controlled and may require very little human intervention. As a result, targets of interest can be identified and tracked in greater detail and/or images can be made with a very high resolution. In SBR systems, the range foldover phenomenon — clutter returns that correspond to previous/later radar pulses — contributes to the SBR clutter. Another important phenomenon that affects the clutter data is the effect of earth's motion around its own axis. At various points on earth this contributes differently to Doppler, and the modification to Doppler due to earth's rotation will be shown to induce a crab angle and a crab magnitude. These two components are shown to induce Doppler dispersion that is shown to be range-dependent. This range dependency causes the secondary data to be non independent and identically distributed (iid); an assumption that is required by most STAP approaches. The simultaneous presence of earth rotation and range foldover — a condition that generally applies — causes performance degradation in most STAP approaches. To mitigate these effects, we propose to use waveform diversity on transmit. Detailed performance analysis and methods that minimize these effects are given in great detail in this chapter.

2. Geometry

2.1 Radar-Earth Geometry

A space based radar (SBR) located at an orbital height H above its nadir point has its mainbeam focused to a point of interest D on the ground located at range R [1]–[4]. In general, the SBR can be in an orbit that is inclined at an angle to the equator. The inclination of

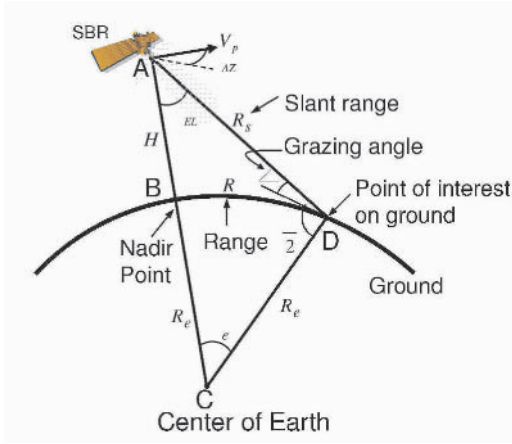


Figure 1. Parameters of an SBR pointing its mainbeam to a ground point D.

the SBR orbit is usually given at the point where it crosses the equator from which its local inclination at other latitudes can be determined. The range is measured from the nadir point B (that is directly below the satellite) to the antenna mainbeam footprint on earth (see Fig. 1). For example, a polar satellite at 506 km above the earth's surface has a period of 1.57 hours. While it completes a circle around the earth that is fixed with respect to the stars, the earth turns through 22.5^o or 1/6 of a revolution about its axis. Thus, every time the space craft crosses the equator the earth moves 2500 km eastward giving an 'automatic' scan of the surface below to the onboard radar. As a result, the radar is able to scan the earth in both latitude and longitude by virtue of the earth's rotation.

In Fig. 1, the SBR is located at A, and B represents the nadir point. The point of interest D is located at range R from B along the great circle that goes through B and D with C representing the center of the earth (see [5]). The main parameters of an SBR setup are as shown in Fig. 1 and are listed in the table in figure 2.

From Fig. 1, the core angle subtended at the center of earth by the range arc BD is given by

$$\theta_e = R/R_e \tag{1}$$

and from triangle ACD we get

$$R_s^2 = R_e^2 + (R_e + H)^2 - 2R_e(R_e + H)\cos\,\theta_e. \tag{2}$$

Thus, the slant range R_s equals

$$R_s = \sqrt{R_e^2 + (R_e + H)^2 - 2R_e(R_e + H)\cos(R/R_e)} \quad . \tag{3}$$

R	Actual ground range from the nadir point to the point of interest along a great circle on the surface of the earth
H	SBR orbit height above the nadir point
R_s	Radar slant range from the satellite to the antenna footprint at range R
	Grazing angle at the antenna footprint at range R (i.e., the angle at which the surface is illuminated by the radar beam)
R_e	Earth's radius (3,440 miles or 6,373 km)
$_{EL}$	Mainbeam elevation from the vertical line associated with range R.
$_{AZ}$	Azimuth point angle measured between the plane of the array (generally also the SBR velocity vector) and the elevation plane
V_P	Satellite velocity vector
$_e$	Core angle between the nadir point and the grazing point measured at the earth's center.

Figure 2. SBR Parameters

Similarly, the grazing angle ψ is also a function of range. To see this, referring back to the triangle ACD we have the grazing angle at range R to be

$$\psi = \cos^{-1}\left(\frac{R_e + H}{R_s}\sin\left(R/R_e\right)\right), \tag{4}$$

and the corresponding elevation angle is given by

$$\theta_{EL} = \sin^{-1}\left(\frac{R_e}{R_s}\sin\left(R/R_e\right)\right). \tag{5}$$

Notice that both the grazing angle ψ and the elevation angle θ_{EL} are range dependent.

From Fig. 1 we also have

$$\theta_{EL} = \sin^{-1}\left(\frac{1}{1 + H/R_e}\cos\psi\right). \tag{6}$$

Similarly from the triangle ACD we obtain the alternate formula

$$\theta_{EL} = \pi/2 - \theta_e - \psi = \pi/2 - \psi - R/R_e \tag{7}$$

for the elevation angle as well. The slant range, grazing angle and elevation angle as functions of range are shown in Fig. 3 and Fig. 4.

2.2 Maximum Range on Earth

The curvature of earth limits the maximum range achievable by a satellite located at height H as shown in Fig. 5.

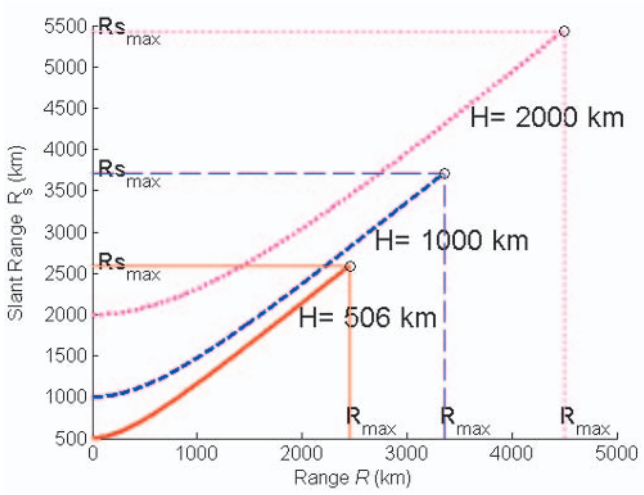


Figure 3. Slant range *vs.* range.

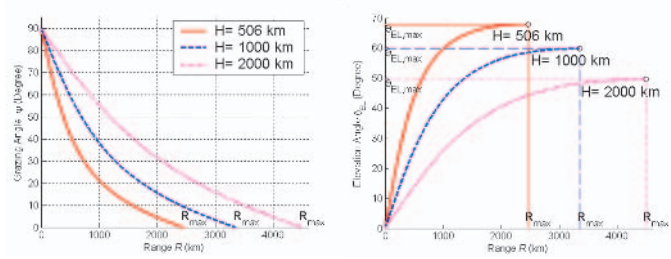


Figure 4. Grazing angle and elevation angle *vs.* range.

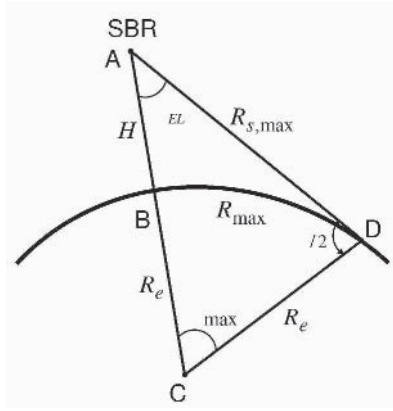


Figure 5. Maximum range on ground.

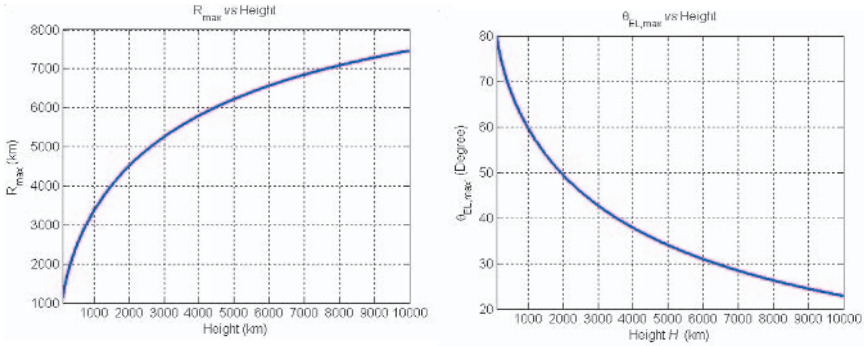


Figure 6. Maximum range and elevation angle *vs.* satellite height.

At maximum range, the slant range becomes tangential to the earth so that the grazing angle $\psi = 0$ and from Fig. 5

$$\theta_{\max} = \cos^{-1}\left(\frac{1}{1 + H/R_e}\right). \tag{8}$$

The maximum range on earth for an SBR located at height H is given by

$$R_{\max} = R_e\,\theta_{\max} = R_e \cos^{-1}\left(\frac{1}{1 + H/R_e}\right). \tag{9}$$

Similarly maximum slant range at the same height is given by

$$R_{s,\max} = (R_e + H)\sin\left\{\cos^{-1}\left(\frac{1}{1 + H/R_e}\right)\right\}, \tag{10}$$

and the maximum elevation angle equals

$$\theta_{EL,\max} = \frac{\pi}{2} - \theta_{\max} = \frac{\pi}{2} - \cos^{-1}\left(\frac{1}{1 + H/R_e}\right). \tag{11}$$

For a low-earth orbit (LEO) satellite located at 506 km above the ground, the maximum range is 2,460 km and $\theta_{EL,\max} = 67.9^o$.

2.3 Mainbeam Footprint Size

The mainbeam of the radar generates a footprint on the ground whose size will depend upon the actual range R. Let ϕ_{EL} represent the mainbeam width of the antenna pattern in the elevation plane. Further, let R_T and R_H denote the ranges of the 'toe' and 'heel' of the mainbeam footprint whose center is at range R, as shown in Fig. 7. Further, let ψ_T

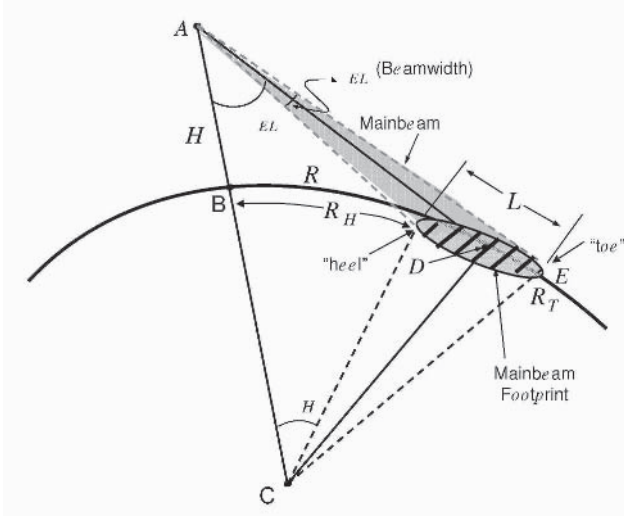


Figure 7. Mainbeam footprint at range R. Distances R_T and R_H correspond to ranges at the 'toe' and 'heel' of the footprint. Range R represents the curved distance BD to the center of the footprint.

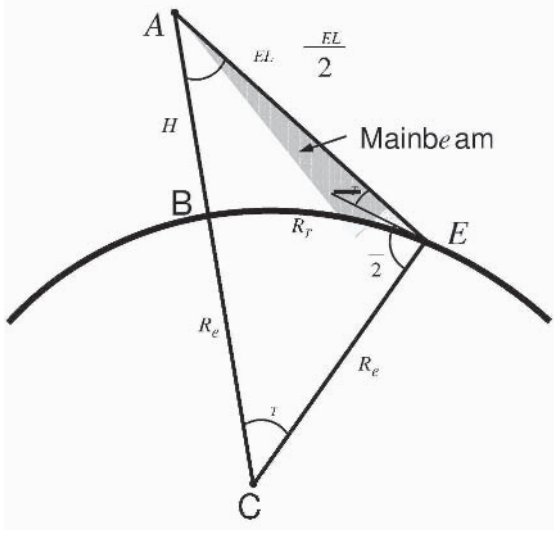


Figure 8. Range calculation at the 'toe' of the mainbeam footprint.

and ψ_H represent the grazing angles at the 'toe' and 'heel' of the mainbeam footprint. Thus, from triangle ACE in Fig. 8 that corresponds to the footprint 'toe', we have

$$\frac{\sin(\pi/2+\psi_T)}{R_e+H}=\frac{\sin(\theta_{EL}+\phi_{EL}/2)}{R_e} \tag{12}$$

where θ_{EL} represents the elevation at range R. This gives the grazing angle at the 'toe' to be

$$\psi_T=\cos^{-1}\left\{\left(1+\frac{H}{R_e}\right)\sin\left(\theta_{EL}+\frac{\phi_{EL}}{2}\right)\right\}, \tag{13}$$

and similarly the grazing angle at the 'heel' is given by

$$\psi_H=\cos^{-1}\left\{\left(1+\frac{H}{R_e}\right)\sin\left(\theta_{EL}-\frac{\phi_{EL}}{2}\right)\right\}. \tag{14}$$

Also, from Fig. 8, the core angle at the center of earth for the 'toe' equals

$$\theta_T=\frac{\pi}{2}-\theta_{EL}-\frac{\phi_{EL}}{2}-\psi_T \tag{15}$$

and the range to the mainbeam 'toe' equals

$$R_T=R_e\theta_T=R_e\left(\frac{\pi}{2}-\theta_{EL}-\frac{\phi_{EL}}{2}-\psi_T\right). \tag{16}$$

Similarly, the range to the 'heel' of the mainbeam equals

$$R_H=R_e\theta_H=R_e\left(\frac{\pi}{2}-\theta_{EL}+\frac{\phi_{EL}}{2}-\psi_H\right). \tag{17}$$

This gives the length of the footprint of the mainbeam at range R to be

$$L=R_T-R_H=R_e(\psi_H-\psi_T-\phi_{EL}). \tag{18}$$

Let ψ_{AZ} represents the beamwidth in the azimuth direction. Then the horizontal mainbeam beamwidth equals

$$W=R_s\,\phi_{AZ}. \tag{19}$$

As Fig. 9 shows, both the length and width of the footprint are functions of range and height. In summary, when the antenna mainbeam is focused along θ_{EL}, returns from the illuminated region of the corresponding mainbeam footprint will contribute toward clutter from that range [5].

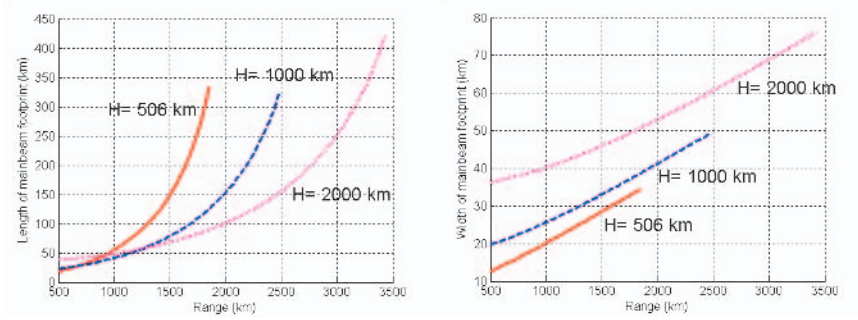


Figure 9. (a) Length and (b) width of mainbeam footprint *vs.* range. Mainbeam beamwidths in both elevation and azimuth directions are assumed to be 1^o.

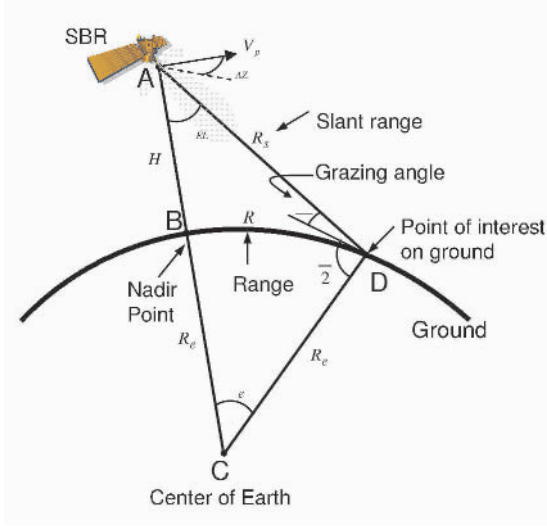


Figure 10. Parameters of an SBR pointing its mainbeam to a ground point D.

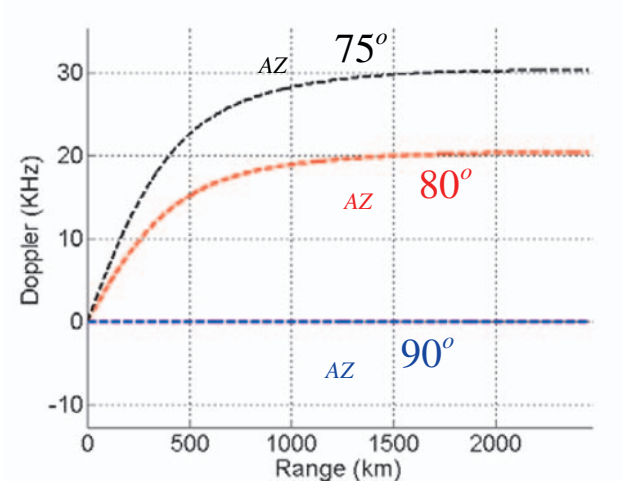


Figure 11. Doppler dependency on range *vs.* azimuth angle.

2.4 Doppler Shift

Consider a space based radar (SBR) at height H above the earth on a great circular orbit that is inclined at an angle η_i (with respect to the equator). By virtue of earth's gravity the SBR is moving with velocity

$$V_p = \sqrt{GM_e/(R_e + H)} \tag{20}$$

in a circular orbit and this contributes to a relative velocity of

$$V_p \cos\theta_{AZ} \sin\theta_{EL} \tag{21}$$

along the line-of-sight for a point of interest D on the ground that is at an azimuth angle θ_{AZ} with respect to the flight path and an elevation angle θ_{EL} with respect to the nadir line as shown in Fig. 10.

If T_r represents the radar pulse repetition rate and λ the operating wavelength, then the Doppler ω_d contributed by (21) equals [1] [7] [11] [12].

$$\omega_d = \frac{2V_pT_r}{\lambda/2} \sin\theta_{EL} \cos\theta_{AZ} \tag{22}$$

and (22) accounts for the Doppler frequency of the ground return due to the SBR motion. Fig. 11 shows the Doppler dependency on range as a function of the azimuth angle. Clearly for a given azimuth angle, the difference in Doppler along the range is minimum (zero) when the azimuth look direction coincides with the bore-side of the array $\theta_{AZ} = \pi/2$. If the earth's rotation is included as we shall see in section 3.2,

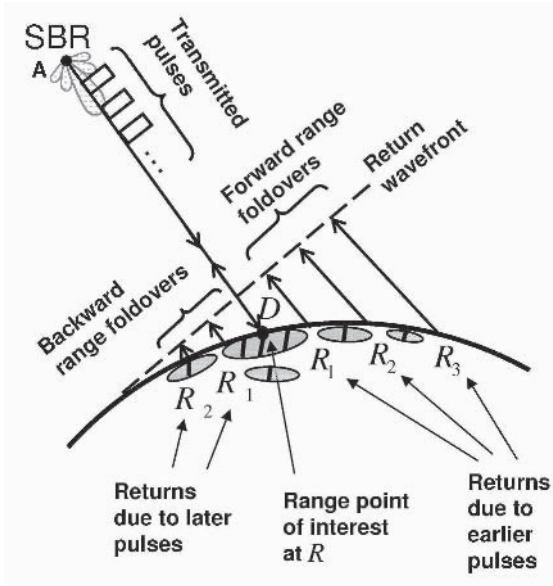


Figure 12. Common wavefront showing all range ambiguity returns corresponding to a point of interest at range R.

the Doppler difference due to range generates an undesirable 'Doppler filling' effect when data samples from different range bins are used to estimate the covariance matrix.

3. Range Foldover and Earth's Rotation

3.1 Range Foldover Phenomenon

To detect targets, the radar transmits pulses periodically. Range foldover occurs when clutter returns from previously transmitted pulses, returning from farther range bins, are combined with returns from the point of interest. Depending on the size of the mainbeam footprint, the 2-D antenna array pattern and the radar pulse repetition frequency, range foldover can occur both from within the mainbeam as well as from the entire 2-D region. The effect of mainbeam foldover is discussed first, followed by its extension to the entire 2-D region [1, 3].

Range Resolution. Let τ represent the *output* pulse length and T_r the pulse repetition interval. Pulses travel along the slant range and interact with the ground through the mainbeam as well as the sidelobes of the antenna array as shown in Fig. 13.

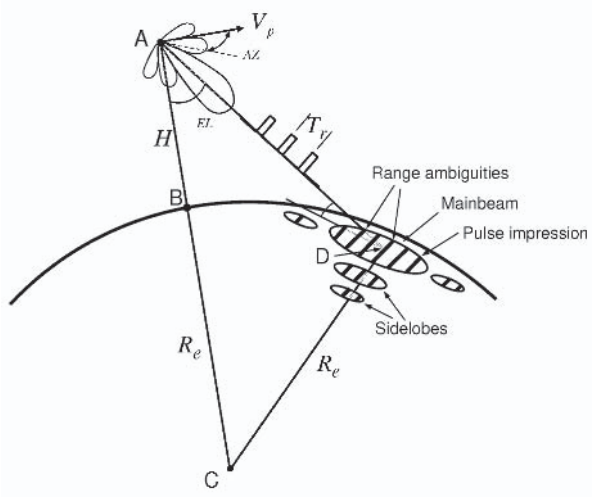


Figure 13. Mainbeam range ambiguities.

Each pulse travels along the slant range and hence the slant range that can be recovered unambiguously is of size $\frac{c\tau}{2}$. Thus, slant range resolution is given by

$$\delta_{SR} = \frac{c\tau}{2}. \tag{23}$$

Translating to the ground plane, since the pulse wavefront is perpendicular to the slant range direction, we get the range resolution on the ground to be

$$\delta_R = \frac{c\tau}{2\cos\psi} = \frac{c\tau}{2}\sec\psi. \tag{24}$$

Thus δ_R represents the ground-plane spatial resolution achievable by the SBR. Two objects that are separated by a distance less than δ_R will be indistinguishable by the radar. Notice that only the *output* pulse length contributes to the range resolution and it can be orders of magnitude smaller than the actual pulse length because of pulse compression effects. For example, using chirp waveforms it is possible to realize 1:100 or higher order compression. From (24) for short range regions where the grazing ψ is closer to $\pi/2$, the range resolution is very poor, and for long range the resolution approaches its limiting value δ_{SR} as $\psi \to 0$.

Total Range Foldover. Radar transmits pulses every T_r seconds and for high PRF situations, following (24), the distance Δ_R between range ambiguities on the ground (distance between consecutive pulse

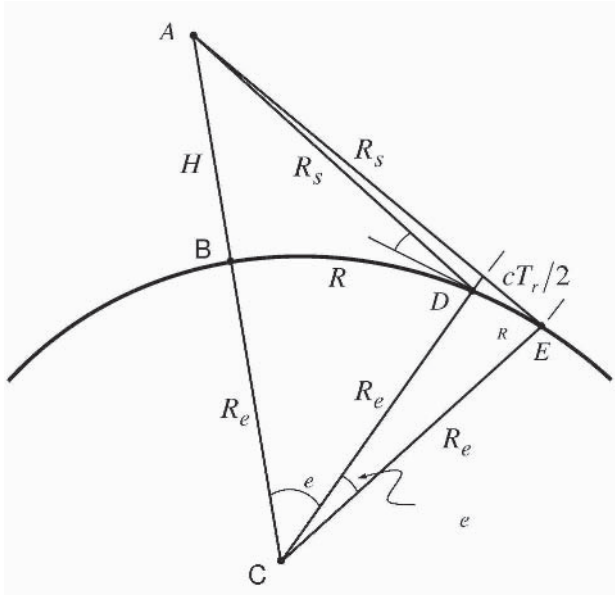


Figure 14. Distance between range ambiguities.

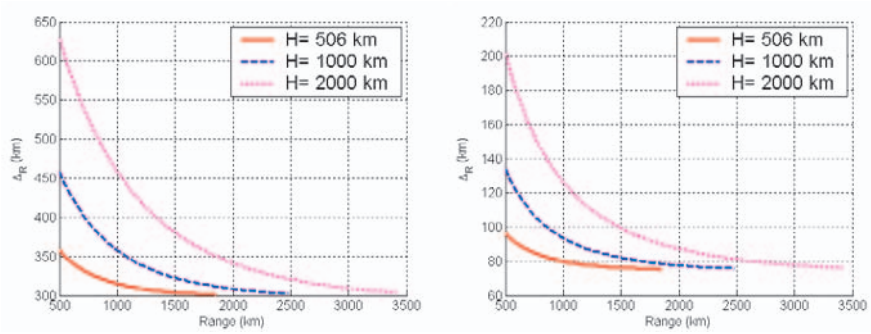


Figure 15. Δ_R vs. range for (a) $PRF = 500$ Hz and (b) 2 kHz.

shadows) is given by

$$\Delta_R = \frac{cT_r}{2} \sec \psi. \tag{25}$$

Equation (25) assumes a high PRF situation where the grazing angles at various range ambiguities are assumed to be equal. The general situation that takes the change in grazing angle into account is shown in Fig. 14.

If R_s represents the slant range at the end of one pulse (say at D), $R_s + cT_r/2$ is the new slant range at the end of the next pulse at E. Let

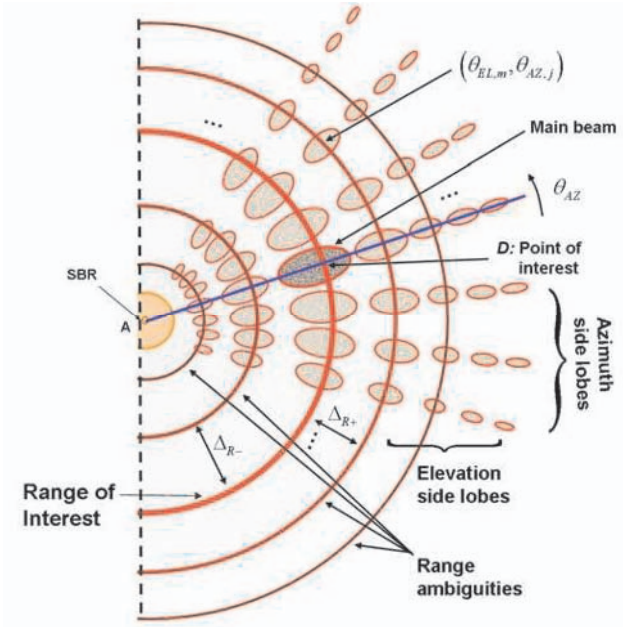


Figure 16. 'Range foldover' phenomenon.

$R_1 = R + \Delta_R$ represent the new range corresponding to the second pulse shadow on the ground at E. From triangle ACE in Fig. 14

$$(R_s + cT_r/2)^2 = R_e^2 + (R_e + H)^2 - 2R_e(R_e + H)\cos\left(\frac{R + \Delta_R}{R_e}\right) \quad (26)$$

or

$$\Delta_R = R_e \cos^{-1}\left(\frac{R_e^2 + (R_e + H)^2 - (R_s + cT_r/2)^2}{2R_e(R_e + H)}\right) - R. \quad (27)$$

From Fig. 15, interestingly Δ_R is a decreasing function of R, and when R is relatively small, the distance between the pulse shadows on the ground is large and is seen to decrease as R increases. However, for large values of range, Δ_R remains constant at its limiting $cT_r/2$. This also follows from (25) since for large R the grazing angle approaches zero.

To compute the total number of range foldovers for the entire range, we can make use of Fig. 16. In Fig. 16, the point of interest (D) is within the mainbeam, and the return of the radar pulse from there represents the main clutter. However because of the 2-D antenna pattern, previous pulse returns returning from adjacent 'range ambiguities points' — both

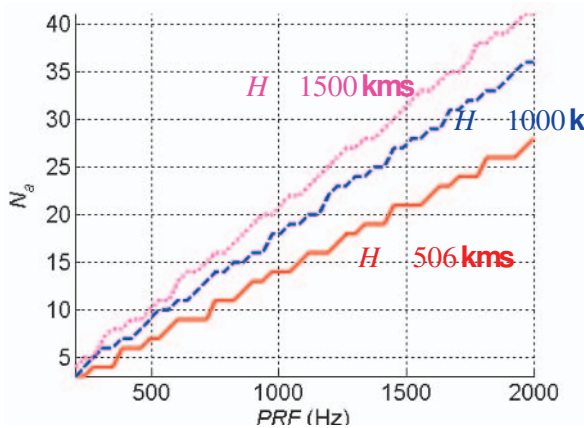


Figure 17. Number of range ambiguities as a function of SBR height and *PRF*.

forward and backward — that have been appropriately scaled by the array gain pattern get added to the mainbeam return causing additional range foldover.

To compute the immediate forward and backward range ambiguity points (E and F respectively), the geometry in Fig. 14 can be used. In general, the k^{th} forward and backward range ambiguity points are given by

$$R_{\pm k} = R_e \cos^{-1}\left(\frac{R_e^2 + (R_e + H)^2 - (R_s \pm kcT_r/2)^2}{2R_e(R_e + H)}\right), \quad k = 1,\ 2,\ \ldots$$

where $R_{+k} = R_k$. Fig. 12 shows the return wavefront from all range ambiguities corresponding to a point of interest D at range R.

Let N_a refer to the total number of range ambiguities (both forward and backward) corresponding to a range bin of interest. The clutter returns from forward and backward range ambiguities get scaled by the array gain corresponding to those locations and get added to the returns from the point of interest. Fig. 17 shows the total number of range ambiguities in the 2D region as a function of SBR height and *PRF*. From this figure, it is seen that the total number of range foldovers at 500Hz *PRF* is 7. Returns from the N_a range ambiguities contribute to the clutter at this particular range [1] [2] [3] [8].

3.2 Modeling Earth's Rotation for SBR

As we have seen in Section 3.1 the range foldover phenomenon — clutter returns that correspond to previous radar pulses — contributes to the SBR clutter. Another important phenomenon that affects the

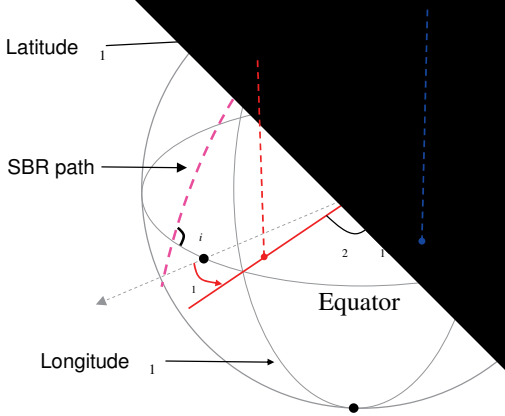


Figure 18. Doppler contributions from SBR velocity and

clutter data is the effect of earth's motion around its own
various locations on earth this contributes differently to Dopp
the effect is modeled here [1, 9].

For any point on earth at range R that is at elevation angle θ_{EL} a
azimuth angle θ_{AZ} from an SBR at height H, the conventional Doppler
shift relative to the SBR equals [5]

$$\omega_d = \frac{2V_pT_r}{\lambda/2}\sin\theta_{EL}\cos\theta_{AZ}, \tag{28}$$

as derived in Section 2.4. Let η_i denote the inclination of the SBR orbit with respect to the equator (see Fig. 18–Fig. 19).

As the SBR moves around the earth, the earth itself is rotating around its own axis on a 23.9345 hour basis in a west-to-east direction. This contributes an eastward motion with equatorial velocity of

$$V_e = \frac{2\pi R_e}{23.9345 \times 3600} = 0.4651\,\text{km/sec} \tag{29}$$

Let (α_1, β_1) refer to the latitude and longitude of the SBR nadir point B and (α_2, β_2) those of the point of interest D as shown in Fig. 18–Fig. 19.

As a result, the point of interest D on the earth at latitude α_2 rotates eastward with velocity $V_e\cos\alpha_2$, which will contribute to the Doppler

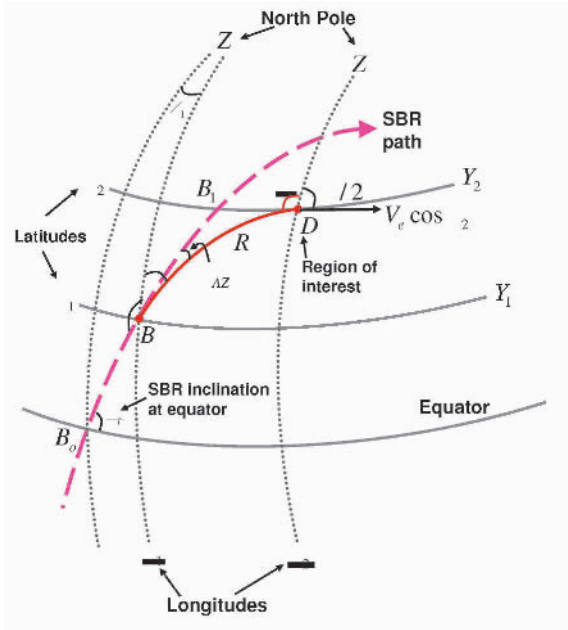


Figure 19. Effect of earth rotation on Doppler frequency.

in (22) as well. To compute this new component in Doppler shift, note from Fig. 19 that the angle BDY_2 between the ground range vector R and the earth velocity vector at D equals $(\pi/2 + \beta)$ so that [5]

$$V_o = V_e \cos\alpha_2 \cos(\pi/2 + \beta) = -V_e \cos\alpha_2 \sin\beta \tag{30}$$

represents the earth's relative velocity at D along the ground range direction towards B. Since the grazing angle represents the slant range angle with respect to the ground range at D (see Fig. 18), we have

$$V_o \cos\psi = -V_e \cos\alpha_2 \sin\beta \ \cos\psi \tag{31}$$

represents the relative velocity contribution between the SBR and the point of interest D due to the earth's rotation towards the SBR. Combining (21) and (31) as in (22), we obtain the modified Doppler frequency that also accounts for the earth's rotation to be

$$\omega_d = \frac{2T_r}{\lambda/2}\left(V_p \sin\theta_{EL}\cos\theta_{AZ} - V_e \cos\alpha_2 \sin\beta\cos\psi\right) \tag{32}$$

After some simplification, we obtain the modified Doppler frequency to be [11, 12]

$$\omega_d \ = \frac{2V_p T_r}{\lambda/2}\rho_c \sin\theta_{EL} \ \cos(\theta_{AZ} + \phi_c), \tag{33}$$

where

$$\phi_c = \tan^{-1}\left(\frac{\Delta\sqrt{\cos^2\alpha_1 - \cos^2\eta_i}}{1 - \Delta\cos\eta_i}\right) \tag{34}$$

and

$$\rho_c = \sqrt{1 + \Delta^2\cos^2\alpha_1 - 2\Delta\cos\eta_i}. \tag{35}$$

In (33)–(35), ϕ_c represents the crab angle and ρ_c represents the crab magnitude. In summary, the effect of earth's rotation on the Doppler velocity is to introduce a crab angle and crab magnitude into the SBR azimuth angle and modify it accordingly [1] [11] [12]. Interestingly both these quantities depend only on the SBR orbit inclination and its latitude, and *not* on the latitude or longitude of the clutter patch of interest.

Equation (33) corresponds to the case where the region of interest D is to the east of the SBR path as shown in Fig. 19. If, on the other hand, the region of interest is to the west of the SBR path, then

$$\omega_d = \frac{2V_pT_r}{\lambda/2}\rho_c\sin\theta_{EL}\ \cos(\theta_{AZ} - \phi_c) \tag{36}$$

with ϕ_c, ρ_c as defined in (33)–(35). Combining (33) and (36), we obtain the modified Doppler to be

$$\omega_d = \frac{2V_pT_r}{\lambda/2}\rho_c\sin\theta_{EL}\ \cos(\theta_{AZ} \pm \phi_c). \tag{37}$$

In (37), the plus sign is to be used when the region of interest is to the east of the SBR path and the minus sign is to be used when the point of interest is to the west of the SBR path.

Fig. 20 shows the crab angle and crab magnitude as a function of SBR latitude for different inclination angles. Once again about 3.77^o error can be expected for the crab angle in the worst case.

The effect of crab angle on Doppler as a function of azimuth angle for various range values is shown in Fig. 21. As (33) shows, the effect of earth's rotation is to shift the azimuth angle appearing in the Doppler by approximately $\phi_c = 3.77^o$ and simultaneously modify the Doppler magnitude as well. As a result, even for $\theta_{AZ} = 90^o$, the Doppler peak values occur away from $\omega_d = 0$ depending on the range. This shift in Doppler with and without the crab effect is illustrated in Fig. 21 for various azimuth angles.

4. Application of STAP for SBR

In this section, SBR data modeling is first carried out with appropriate Doppler parameters. By considering the two phenomena, with and

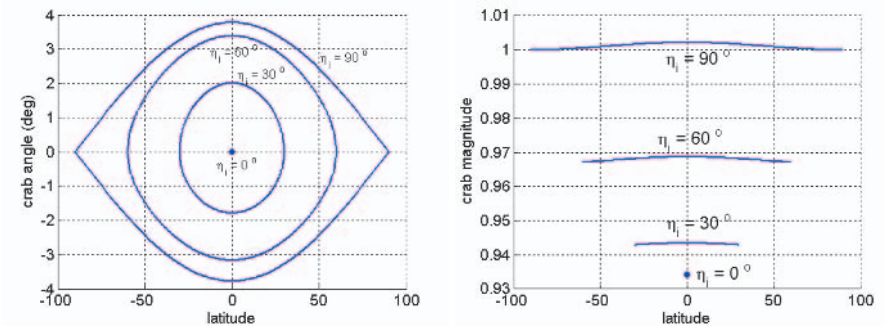


Figure 20. (a) Crab angle and (b) crab magnitude as functions of SBR latitude for different inclination angles.

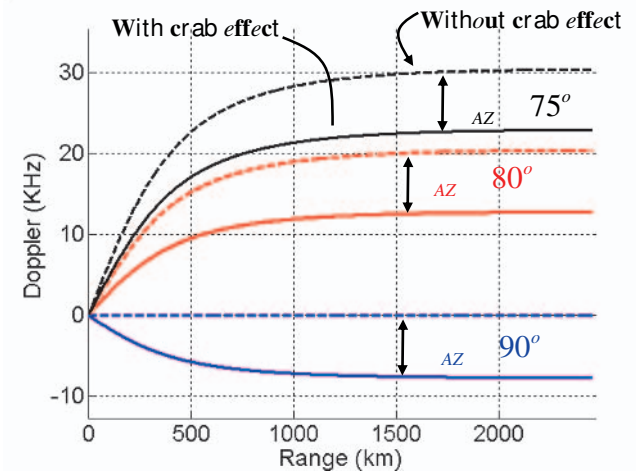


Figure 21. Effect of crab angle on range-Doppler profile.

without range foldover effect and with and without earth's rotational effect — four cases of interest can be generated.

4.1 SBR Data Modeling

Consider an SBR array with N sensors and M pulses. If the incoming wavefront makes an azimuth angle θ_{AZ} and elevation angle θ_{EL} with reference to the array, define the first sensor output to be $x_1(t)$ and

$$c = \sin\theta_{EL}\cos\theta_{AZ} \tag{38}$$

represent the 'cone angle' associated with the spatial point θ_{EL}, θ_{AZ} for the SBR array. Then the concatenated data vector due to the N sensors and M pulses is the MN by 1 vector given by [12]

$$\mathbf{x}(t) = \mathbf{s}(c, \omega_d)\, x_1(t), \tag{39}$$

where

$$\mathbf{s}(c, \omega_d) = \mathbf{b}(\omega_d) \otimes \mathbf{a}(c) \tag{40}$$

represents the spatial-temporal steering vector with $\otimes$ representing the Kronecker product. $\mathbf{a}(c)$ in (40) represents the spatial steering vector and is given by

$$\mathbf{a}(c) = \left[1, e^{-j\pi dc}, e^{-j2\pi dc}, \cdots, e^{-j(N-1)\pi dc}\right]^T \tag{41}$$

$\mathbf{b}(\omega_d)$ in (40) represents the temporal steering vector given by

$$b(\omega_d) + \left[1, e^{-j\pi\omega_d}, e^{-j2\pi\omega_d}, \cdots, e^{-j(M-1)\pi\omega_d}\right]^T \tag{42}$$

Let $\theta_{AZ_j} = \theta_{AZ} + i\Delta\theta,\ i = 0,\ \pm 1,\ \pm 2,\ \pm N_o$ represent the azimuth angles associated with the field of view, and $\theta_{EL_m},\ m = 0, 1, 2, \cdots, N_a$ the elevation angles corresponding to the total number of range foldover in the field of view. Further let

$$c_{m,\,i} = \sin\theta_{EL\,m}\cos\theta_{AZ_i} \tag{43}$$

represent the cone angle associated with the location $\theta_{EL_m}, \theta_{AZ_i}$. The total clutter return represents various range foldover returns that span over all azimuth angles. This gives the ensemble average clutter covariance matrix associated with range r_k to be

$$\mathbf{R}_k = E\left\{\mathbf{y}_k \mathbf{y}_k^*\right\} \tag{44}$$

where $\mathbf{y}_k$ represents the clutter data.

4.2 SINR With/Without Earth's Rotation and Range Foldover

In practice, the covariance matrix in (44) corresponding to range r_k is unknown and needs to be estimated from data using the expression

$$\hat{\mathbf{R}}_k = \sum_j \mathbf{x}_{k+j}\mathbf{x}_{k+j}^*. \tag{45}$$

In (45) the number of range bins over which the summation is carried out is chosen so as to maintain stationary behavior for $\hat{\mathbf{R}}_k$. The estimated adaptive weight vector corresponding to (45) is given by the sample matrix inversion (SMI) method as

$$\hat{\mathbf{w}}_k = \hat{\mathbf{R}}_k^{-1}\,\mathbf{s}(c_t,\,\omega_{d_t})\ . \tag{46}$$

A useful way of evaluating the performance of a particular STAP algorithm is the signal power to interference plus noise ratio (SINR) defined by

$$SINR = \frac{|\mathbf{w}^*\,\mathbf{s}|^2}{\mathbf{w}^*\,\mathbf{R}\,\mathbf{w}} \tag{47}$$

where $\mathbf{w}$ is the estimated adaptive weight vector and $\mathbf{R}$ is the ideal clutter plus noise covariance matrix defined in (44). For the SMI, (47) can be written as

$$SINR = \frac{|(s)^*\,\hat{\mathbf{R}}^{-1}\,(s)|^2}{(s)^*\,\hat{\mathbf{R}}^{-1}\,R\,\hat{\mathbf{R}}^{-1}(s)} \tag{48}$$

Clearly the performance of (48) is bounded by the ideal matched filter output $SINR_{ideal}$ obtained by letting $\hat{\mathbf{R}} = \mathbf{R}$ in (44). This gives

$$SINR_{ideal} = \mathbf{s}^*(c,\,\omega_d)\,\mathbf{R}^{-1}\,\mathbf{s}(c,\,\omega_d) \tag{49}$$

where $\mathbf{s}(c,\omega_d)$ is given by (40) and represents the space time steering vector for the desired point of interest located at $\theta = (\theta_{EL}, \theta_{AZ})$ that corresponds to the cone-angle

$$c = \sin\theta_{EL}\cos\theta_{AZ}, \tag{50}$$

and Doppler frequency ω_d for the SBR configuration under consideration. To quantify the performance deterioration due to earth's rotation and range foldover, the following four situations corresponding to four different clutter covariance matrices are identified:

1 No range foldover, no earth's rotation (ideal case);

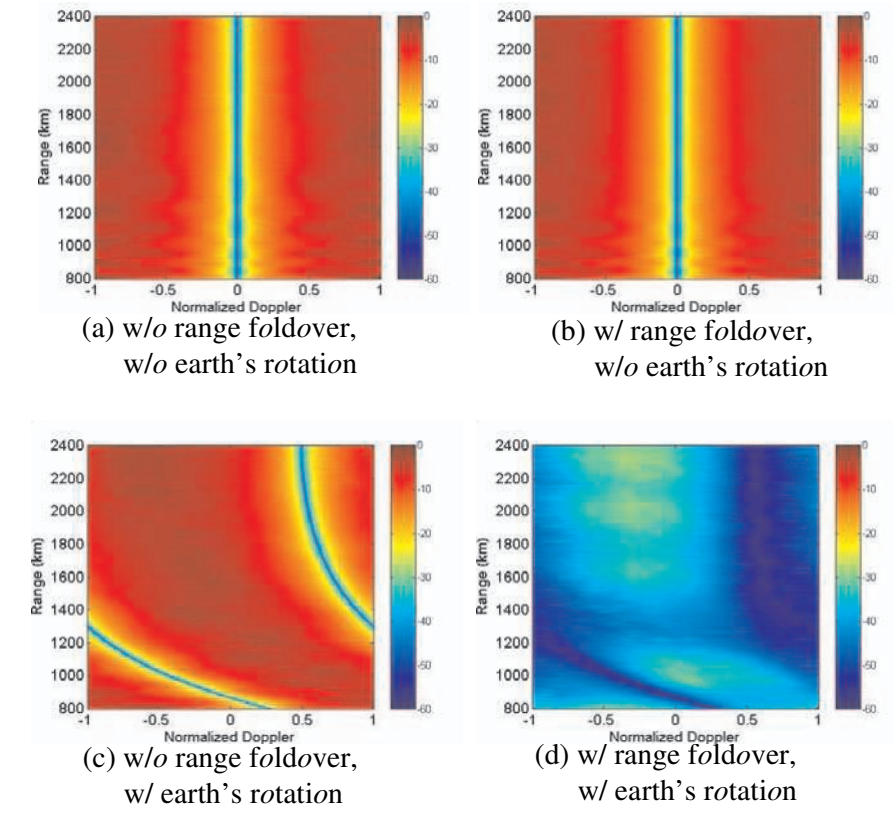


Figure 22. SINR loss with/without range foldover and earth's rotation as a function of range and Doppler

2 Range foldover present, no earth's rotation;

3 No range foldover, earth's rotation present;

4 Range foldover present, earth's rotation present.

Fig. 22 shows the SINR loss for the four cases above as a function of range and Doppler for an SBR located at height 506 km above the ground. The PRF is 500 Hz and $\theta_{AZ} = 90^o$. The output is normalized with respect to the noise only case. The performance is significantly degraded when both range foldover and earth's rotation are present at the same time.

Fig. 23 shows the SINR loss for two different ranges. The SINR loss is on the order of 20–40 dB when both range foldover and earth's rotation are present, depending on the actual range. The performance in terms of clutter nulling is inferior when these two effects are present.

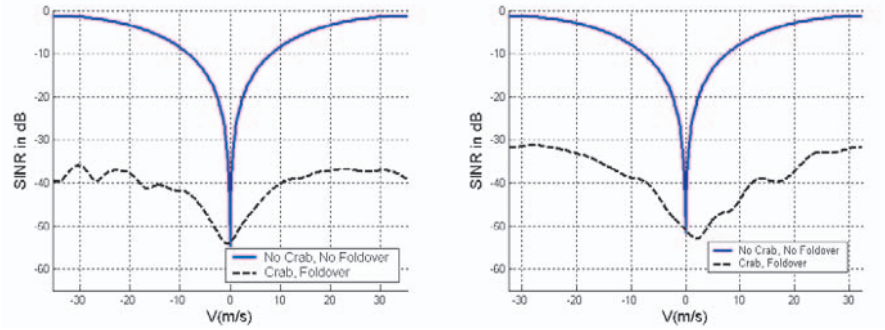


Figure 23. SINR loss with/without range foldover and earth's rotation for (a) 900 km and (b) 2000 km ranges.

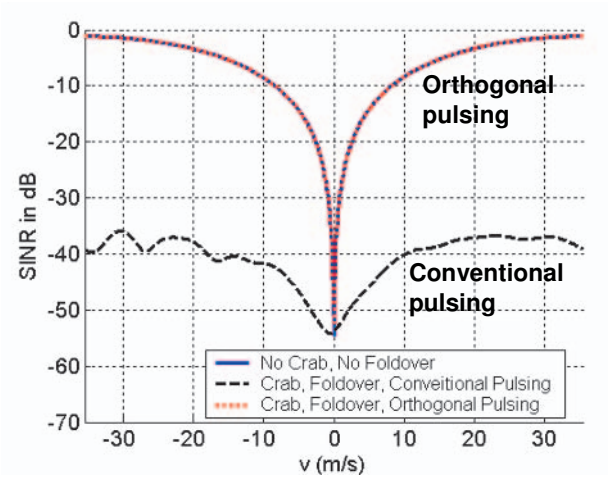


Figure 24. SINR improvement using orthogonal pulsing, range=900 kms.

Thus, having both range foldover and earth's rotation at the same time results in unacceptable performance degradation as shown in Fig. 23; whereas when either one is present separately, the effect can be rectified.

5. Orthogonal Pulsing Scheme

Waveform diversity can be used on the sequence of transmitted radar pulses to realize the above goal by suppressing the range foldover returns. In ordinary practice, a set of identical pulses are transmitted. To suppress returns due to range foldover, for example, individual pulses

$f_1(t),\ f_2(t),\ \cdots$ can be made orthogonal to each other so that

$$\int_o^{T_o} f_i(t) f_l(t)\, dt = \delta_{i,l}, \quad i, l = 1, 2, \cdots, N_a, \tag{51}$$

where N_a is the maximum number of distinct range foldovers present in the data and $\delta_{i,l}$ is the Kronecker delta. Then, with appropriate matched filtering [13], the range ambiguous returns can be minimized from the main return corresponding to the range of interest. Note that for range foldover elimination, waveform diversity needs to be implemented only over N_a pulses. For an SBR located at a height of 506 km and an operating PRF = 500 Hz, $N_a \approx 7$. This is the case since matched filtering will eliminate the superimposed range foldover returns since they correspond to waveforms that are orthogonal to the one in the mainbeam. Fig. 24 shows the SINR improvement using eight orthogonal waveforms [13]. The performance is restored to the ideal case when orthogonal waveforms are used.

In summary, using waveform diversity on transmit, it is possible to eliminate the effect of range foldover resulting in performance improvement as shown in Fig. 24. The resulting performance will be approximately the same as the performance shown in Fig. 22 (a), indicating that using waveform diversity on transmit, it is possible to achieve performance close to the ideal case even in the presence of both range foldover and earth's rotation. The results presented here correspond to the case where the ensemble averaged clutter covariance matrix is given. The case where the covariance matrix is estimated from secondary data is more challenging.

References

[1] K. Y. Li *et. al.*, STAP for Space Based Radar *Air Force Research Laboratory Final Technical Report*, AFRL-SN-RS-TR-2004-170, June 2004.

[2] Mark E. Davis and Braham Himed, L Band Wide Area Surveillance Radar Design Alternatives, *International Radar 2003 — Australia*, September 2003.

[3] Mark E. Davis, Braham Himed, and David Zasada, Design of Large Space Radar for Multimode Surveillance, *IEEE Radar Conference*, Huntsville, AL, pp. 1–6, May 2003.

[4] John Maher, Mark E. Davis, Robert J. Hancock, and Sidney W. Theis, High Fidelity Modeling of Space-Based Radar, *2003 IEEE Radar Conference*, Huntsville, AL, pp. 185–191, May 2003.

[5] Leopold J. Cantafio, *Space-Based Radar Handbook* Artech House, Boston, 1989.

[6] Troy L. Hacker, Performance Analysis of a Space-Based GMTI Radar System Using Separated Spacecraft Interferometry MS Thesis, Department of Aeronautics and Astronautics *Massachusetts Institute of Technology*, Lexington, MA, May 2000.

[7] J. R. Guerci, *Space-Time Adaptive Processing for Radar* Artech House, Boston, 2003.

[8] S.M. Kogon, D.J. Rabideau, R.M. Barnes, Clutter Mitigation Techniques for Space-Based Radar, *IEEE International Conference on Radar*, Vol. 4, pp. 2323–2326 March 1999.

[9] Peter Zulch, Mark Davis, Larry Adzima, Robert Hancock, Sid Theis, The Earth Rotation Effect on a LEO L-Band GMTI SBR and Mitigation Strategies, *IEEE Radar Conference*, Philadelphia, PA, April 2004.

[10] G.A. Andrews and K. Gerlach, SBR Clutter and Interference', Ch. 11, *Space-Based Radar Handbook*, Ed. Leopold J. Cantafio, Artech House, Boston, 1989.

[11] S. U. Pillai, B. Himed, K. Y. Li, Waveform Diversity for Space Based Radar, *Proc. of Waveform Diversity and Design*, Edinburgh, Scotland, Nov 8–10, 2004.

[12] S. U. Pillai, B. Himed, K. Y. Li, Modeling Earth's Rotation for Space Based Radar, *Asilomar Conference on Signals, Systems, and Computer*, Pacific Grove, CA, Nov 7–10, 2004.

[13] S. U. Pillai, B. Himed, K. Y. Li, Orthogonal Pulsing Schemes for Improved Target Detection in Space Based Radar, *2005 IEEE Aerospace Conference*, Big Sky, MT, March 5–12, 2005.

CONTINUOUS WAVE RADARS–MONOSTATIC, MULTISTATIC AND NETWORK

Krzysztof Kulpa
Warsaw University of Technology
00-665 Warsaw, Poland

Abstract Radar technology was designed to increase public safety on sea and in the air. Today radars are used in many fields of application, such as air-defense, air-traffic-control, zone protection (in military bases, airports, industry), people search and others. Classic pulse radars are often being replaced by continuous wave radars. Unique features of continuous wave radars, such as the lack of ambiguity, very low transmitted power and good electromagnetic compatibility with other radio-devices, enhance this trend. This chapter presents the theoretical background of continuous wave radar signal processing (for FMCW and noise radars), highlights the most important features of this type of radar and shows their abilities in the field of security.

Keywords: continuous wave radar; linear frequency modulation; noise radar; noise; modulation; correlation function; ambiguity function; synthetic aperture; radar; target identification.

1. Introduction

Today security technology migrates from strictly military areas to many different civil applications. On one hand this migration is caused by the high risk of terrorist threats, on the other one observes the growing interest of society in protecting personal life regardless of the costs of rescue operations. Many sophisticated technologies are in everyday use to protect people in airports, shopping centers, schools and other public places. Strong efforts have been made to locate people in buildings and ruins during rescue and peacekeeping missions. One of the most important security problems is to remotely detect and localize objects of interest (often referred to as targets) and to distinguish between the target and its environment. Another issue is to extract as much useful information about the object as possible and to classify it properly.

Jim Byrnes and Gerald Ostheimer (eds.),
Advances in Sensing with Security Applications, 215–242.

The process of detection can be performed by using different sensing technologies and different sensors.

One very common and mature technology is based on optical sensing, which is used in many military and civilian areas such us "city surveillance". Optical cameras convert the electromagnetic object signature into an image of the target, which can be further processed and stored by the computer. However, this technology faces many limitations. Working in the visual light region it is necessary to illuminate the target either by natural light sources (sunlight, moonlight) or artificial illuminators (street lights, reflectors etc). Some interesting results can be achieved by using other parts of the electromagnetic spectrum. Deep infrared cameras can provide images in a completely dark environment, but they require a temperature difference between the target and its surroundings. Their detection range is usually very limited and detection is impossible or very difficult in bad weather conditions or behind fire, smoke and obstacles.

For a long time radio frequency emission has been used for target detection and tracking. Classical radar senses the object by emitting short, powerful electromagnetic pulses towards the target and listening to the return echo. The distance to the target is calculated from the delay of the echo signal. It is also possible to estimate the direction of the target, its velocity and, using more sophisticated synthetic aperture radar (SAR)/Inverse SAR (ISAR) technology, the target size and shape. Having all this information, it is possible to recognize and track different objects. This radar technology is independent of weather and day/night conditions, and may be used to detect targets hidden by obstacles, walls or even buried in the ground. However, pulse radar technology has many disadvantages. To achieve large detection range, high power transmitters are required. Due to the pulse nature of such radars, there are many ambiguity problems. Pulse radars can be easily detected by very simple electronic support measurement devices, which is a big disadvantage for covert security missions.

A present trend in radar technology is to decrease the emitting peak power. Instead of emitting a train of high power pulses, it is possible to emit a low-power continuous signal with appropriate modulation. Advanced signal processing of the received signal allows radars to detect target echoes far below the surrounding noise level and to extract all required target parameters. It is also possible to produce high-resolution target pictures. In some applications it is even not necessary to emit an illuminating signal, but to exploit existing radio, TV or other emissions.

A single sensor usually has limited range and accuracy — especially in the cross-range direction. The use of bistatic ideas, where the transmit-

ter and receiver are separated in space, leads to a new technology with increased sensitivity and accuracy in selected regions. Increasing the number of sensors (multistatics) yields much bigger surveillance volume, much higher probability of detection and proper target classification. This technology is also much more robust to target shadowing, multipath signal fading and jamming. To fully use the information produced by the set of sensors, it is necessary to exchange information between sensors and generalize this information. This leads to the net-centric sensor concept.

Radar devices were primary designed to increase public safety and now are used in many security applications. The most obvious application fields are air traffic control (ATC) and air defense. Battlefield short or middle-range radar can now detect, track and identify slow moving targets such as tanks, cars, pedestrians and animals. Using micro-Doppler analysis, it is possible to identify vehicles, and using vibration analysis, to even distinguish between two cars of the some brand. Heartbeat analysis is used for people identification and, together with breathing analysis, can even provide information on the emotional state of each person. Walk analysis based on Doppler processing and walk modeling are being used for person tracking even in crowded areas. Ground-penetrating and wall-penetrating radars are used to search for people in rescue operations, when it is necessary to search big areas of ruins or avalanches. One of the most important security issues is to distinguish between people and animals, in order to save human life first. A big area of security is connected with trespass control as well. It is very important for border protection, airport and military base protection and elsewhere.

2. Radar fundamentals

The concept of radar was discovered in the beginning of the 20th century. The "father" of radar was Christian Huelsmeyer, who applied for a patent for his "telemobiloscope" on 30 April 1904. He was motivated in his work by a ship accident and his intention was to construct a device to boost the level of safety. His device worked reasonably well, detecting ships at ranges up to 3 km in all weather conditions, but he had no success in selling telemobiloscopes and that early radar concept faded from memory. The reinvention of radar was done almost simultaneously in many countries in the 1920's and 30's. In Great Britain, work on radar technology was carried out by the physicist Sir Edward Victor Appleton, who used radio echoes to determine the height of the ionosphere in 1924. The first demonstration of aircraft detection was done by Watson-Watt

and Wilkins, who had used radio-wave transmissions from the powerful BBC short-wave station at Daventry and measured the power reflected from a Heyford bomber. Detection was achieved at at distance of up to 8 miles. As a matter of fact, this was the first demonstration of a bistatic, continuous wave passive radar. The first British radar patent was issued in April 1935. In the United States radar research was carried out in the early 1920's by Dr. A. Hoyt Taylor at the Naval Research Laboratory in Washington, D.C. Radar research was also carried out in France and in 1934 the French liner "Normandie" was equipped with a "radio obstacle detector". Work on radar technology was also carried out in many other countries, but usually such research was classified.

Early radars were non-coherent pulse radars. To obtain long range detecting capability, the radar emits short pulses with very high peak power — up to several megawatts. The detection range limitation usually comes from the inadequacy of the maximum peak power. Available microwave valves and waveguides limit this parameter. To decrease the required peak power pulse compression techniques have been worked out. The first radars were used for air defense to indicate moving targets. All ground echoes were unwanted, and to distinguish between ground clutter and moving targets Doppler processing was introduced. Long distance pulse radar suffered from range or Doppler ambiguity. It is not possible to measure instantaneously both range to the target and target radial velocity, without the ambiguity given by the sampling theorem. The pulse radar concept was relatively simple and pulse radars could be designed using only analogue components. The problem with signal storage, required for Doppler processing, was solved by using acoustic delay lines (e.g. mercury tubes) or memo-scope bulbs.

Rapid progress in digital signal processing (DSP) hardware and algorithms enabled designers to exploit more complicated ideas. One such idea was to use continuous wave (CW) instead of high power short pulses. The idea was not new — the first Daventry experiment was based on CW radio emission, but practical implementation of CW radars was impossible without digital technology. The first practical continuous wave radars were constructed as Doppler-only radars. The well known police radar belongs to that group. Further development in CW radars lead to linear frequency modulated CW radars (LMCW), in which both target range and range velocity can be measured. But again, due to the periodicity in modulation, the ambiguity problem was still there.

CW radars have a very big advantage — very low peak power. As the transmitted power can be below 1 W (many of them work with 1 mW power), they belong to the low probability of intercept (LPI) class of radars, which can detect targets while they remain undetected.

The search for ambiguity-free waveforms has lead to the concept of random-waveform radars, often referred to as noise or pseudo-noise radars. In this kind of radar, the target is illuminated by continuous noise-like radiation. The reflected power is collected by the radar receive antenna, and detection is based on match filtering of the received signals. The single filter is matched to the particular value of range and range velocity. Because the target's distance and speed are unknown, it is necessary to apply a set of filters matched to all possible range-velocity pairs. The computational power required by the noise radars is much higher than for other radar types, so to date noise radars have not been used very much. There are also other drawbacks of noise radars. Due to the fact that the noise radar has to emit and receive signals simultaneously, good separation between transmitting and receiving antennas is required. Furthermore, this radar simultaneously receives strong echoes originating from close targets, buildings and ground, and weak echoes from distant targets. Thus, a large receiver dynamic range is needed.

The noise-radar concept may be used for moving target detection and also for many other applications. There are several works showing possible implementation of noise technology for imaging radar working in SAR or ISAR mode, passive detection and identification of targets and space radiometric applications. Noise radars will also be used in the future in other fields, including air traffic control, pollution control, and especially security applications.

2.1 Radar range equation

Classical pulse radar emits high power (P_T) short electromagnetic pulses using a directional transmitting antenna of gain G_T. The power density at the target at distance R from the transmitter is equal to [44]

$$p(R) = \frac{P_T G_T}{4\pi R^2}. \tag{1}$$

The total power illuminating the target of effective cross-section S_o is equal to

$$P_S = \frac{P_T G_T}{4\pi R^2} S_o. \tag{2}$$

Assuming that the target reflects all illumination power omni-directionally, the power received by the receiving radar antenna with effective surface S_R is equal to

$$P_R = \frac{P_T G_T}{16\pi^2 R^4 L} S_o S_R, \tag{3}$$

where L stands for all losses in the radar system, including transmission, propagation and receiving losses. Substituting the antenna gain for the

antenna effective surface in (3) one obtains the classical radar equation in the form

$$P_R = \frac{P_T G_T G_R \lambda^2}{(4\pi)^3 R^4 L} S_o. \tag{4}$$

The receiver's equivalent noise can be expressed as

$$P_N = kT_R B, \tag{5}$$

where T_R is the effective system noise temperature (dependent on the receiver's temperature, receiver's noise figure, antenna and outer space noise), B is the receiver's bandwidth (assuming that at the receiver's end match filtering is used) and k is the Boltzmann's constant (1.380 6505 x 10^{-23} $[JK^{-1}]$). Using the Neyman-Pearson detector it is possible to assume that there is a target echo in the signal when the echo power is higher than the noise power multiplied by the detectability factor D_o, usually having a value of 12–16 dB, depending on the assumed probability of false alarm. The radar range equation can be written in the form

$$\frac{P_T G_T G_R \lambda^2}{(4\pi)^3 R^4 L} S_o > kT_R B D_o, \tag{6}$$

and the maximum detection range is equal to

$$R_{max} = \sqrt[4]{\frac{P_T G_T G_R \lambda^2}{(4\pi)^3 L k T_R B D_o} S_o}. \tag{7}$$

For pulse radar, the receiver's bandwidth B is inversely proportional to the pulse width t_p. Substituting the receiver's bandwidth by pulse time in (7), one obtains the equation

$$R_{max} = \sqrt[4]{\frac{E_T G_T G_R \lambda^2}{(4\pi)^3 L k T_R D_o} S_o}. \tag{8}$$

The detection range depends on transmitted pulse energy E_T, transmitter and receiver antenna gains and wavelength, and does not depend on the pulse length or receiver's bandwidth. Thus, Equation 8 is very general and can be used to predict the detection range for all kinds of radars, including continuous wave ones. For continuous emissions the energy E_T is equal to the product of the transmitter power and the time of target illumination or coherent signal integration.

An example of required radar peak power for different pulse/illumination times, for X band radar equipped with a 30 dB antenna, is presented in Figure 1. For X band radar, emitting a 1 μs pulse, the peak

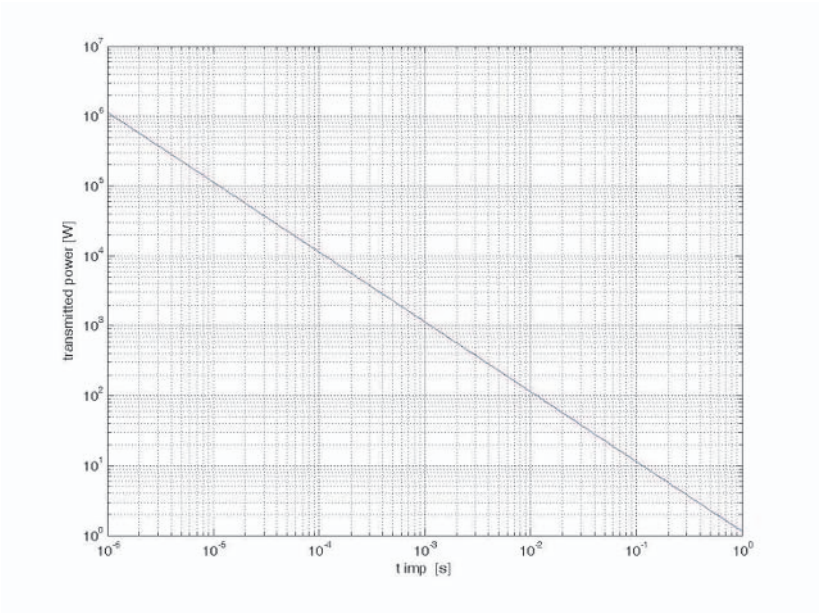


Figure 1. Transmitted power versus transmitted pulse length for X-band radar (30 dB antenna gain) — detection of 1dBsm (1 square meter) target at distance 50 km.

power must be at the level of 1 MW. A continuous wave radar having 1 s integration time requires only 1 W emission.

A large group of radars are surveillance radars which have to search a certain surveillance solid angle α_s in time t_s. Let us assume that a radar is equipped with a transmitter having mean transmitted power P_{tm}. The ideal transmission antenna with gain G_T emits electromagnetic radiation in the solid angle $\alpha_T = 4\pi/G_T$. To scan the whole surveillance space it is necessary to scan $n_s = \alpha_s/\alpha_T = \alpha_s G_T/4\pi$ directions, so the time of a single direction illumination (time on target) is equal to $t_t = \frac{t_s}{n_s} = \frac{4\pi t_s}{\alpha_s G_T}$, and the total energy associated with the single scan direction is equal to $E_t = t_t P_{Ts} = \frac{4\pi t_s P_{Ts}}{\alpha_s G_T}$. The detection range of surveillance radar can be calculated using the formula:

$$R_{max} = \sqrt[4]{\frac{P_{Ts} t_s G_R \lambda^2 S_o}{(4\pi)^2 \alpha_s L k T_R D_o}} = \sqrt[4]{\frac{P_{Ts} t_s S_R S_o}{4\pi \alpha_s L k T_R D_o}}. \tag{9}$$

It is easy to observe that the detection range depends on the receiving antenna gain, mean transmitter power, scanning angle and scanning time. The gain of the transmitting antenna does not contribute to the detection range, but influences the time-on-target.

In many radar systems a single antenna is used both for transmitting and receiving signals, and thus $G_T = G_R$. To extend effective radar range it is possible to use a set of highly directional, high gain antennas instead of a single receiving antenna. This leads directly to the concept of multi-beam or stack-beam antenna, very popular in 3-D surveillance radar. The multi-beam radar idea can be further extended. It is possible to develop a system with an omni-directional transmitter and a circular multi-beam receiving antenna set.

2.2 Radar range measurement and range resolution

The range to the target is determined in radar by measuring the time in which a radar pulse is propagated from the radar to the target and back to the radar. The time delay between the transmitted and received signal is equal to $\tau = 2R/c$, where R is the radar-target distance and c is the velocity of light, equal to 299,792,458 $[ms^{-1}]$. The radar-target distance can be easily calculated knowing the received signal delay by the formula

$$R = \frac{\tau c}{2}. \tag{10}$$

The time delay between two signals can be estimated using different methods; among them the most popular is finding the maximum of the cross-correlation function

$$r(\tau) = \int x_R(t) x_T^*(t-\tau) dt \tag{11}$$

between transmitted signal x_T and received signals x_R (* denotes complex conjugation of the signal). Using pulse radar, two objects can be separated if their distance is greater than the distance corresponding to pulse width $\Delta R = t_T c/2$. The theoretical spectrum of a boxcar shaped pulse signal is described by the sinc function. The width of the main lobe of the spectrum is equal to $B = 1/t_T$. For other types of transmitted waveforms, such as pulses with phase or frequency modulation or continuous waves with frequency or noise modulation, the range resolution depends on the width of the main lobe of the transmitted signal autocorrelation function. The typical range resolution for pulse-coded and continuous waveforms is equal to

$$r(\tau) = \frac{c}{2B}. \tag{12}$$

2.3 Radar range velocity measurement and range velocity resolution

The radar return signal is a delayed copy of the transmitted one only when the signal is reflected from a stationary (not moving) target. In most practical cases, the radar should detect moving targets. The reflection from moving targets modifies the return signal, which can now be written in the form

$$x_R^{HF}(t) = A(r(t))x_T^{HF}(t - \frac{2r(t)}{c}), \tag{13}$$

where x_T^{HF} is the transmitted (high frequency) signal, x_R^{HF} is the received signal, $r(t)$ is the distance between the radar and the target and A is the amplitude factor, which can be calculated using Equation (4). Most radars emit narrow band signals, which can be described by the formula

$$x_T^{HF}(t) = x_T(t)e^{j(2\pi Ft+\phi)}, \tag{14}$$

where $x(t)$ is the transmitted signal complex envelope, F is the carrier frequency and ϕ is the starting phase. For a constant velocity target ($r(t) = r_o + v_o t$) the narrowband received signal can be written in the form

$$x_R^{HF}(t) = A(r(t))x_T(t - \frac{2r(t)}{c})e^{j2\pi(F-2v_oF/c)t+j\phi_R}, \tag{15}$$

where $\phi_r = \phi - 4\pi r_o F/c$ is the received signal starting phase, and $4v_oF/c = 2v_o/\lambda$ is a Doppler frequency shift. For short pulse radars the Doppler shift is usually much smaller than the reciprocal of the time duration of the transmitted pulse ($4v_oF/c << 1/t_T$), and that Doppler shift has a very limited effect on a single pulse — it changes only the phase of the received pulse. Doppler frequency, and thus target radial velocity, can be estimated using a train of pulses and analyzing the phase difference between consecutive pulses. For long pulses, or for continuous waveforms, target movement introduces a Doppler shift of the carrier frequency, which can be measured by analyzing the received signal.

Velocity resolution is limited by the coherent signal processing time, which is smaller or equal to the time the target is illuminated by the radar (time-on target). Using classical filtering it is possible to separate two targets in velocity when the Doppler frequency difference is greater than the reciprocal of the coherent integration time. The velocity resolution is then equal to

$$\Delta v = \frac{2\lambda}{t_i} = \frac{2c}{t_i F}. \tag{16}$$

For example, for a 10 cm wavelength and integration time 10 ms the velocity resolution is 20 m/s, while for 1 s integration the resolution is 0.2 m/s.

3. Linear Frequency Modulated Continuous Wave Radar

Frequency modulated continuous wave (FMCW) radar is most commonly used to measure range R and range (radial) velocity of a target [42, 43]. The most common structure of a homodyne FMCW radar is presented in Figure 2.

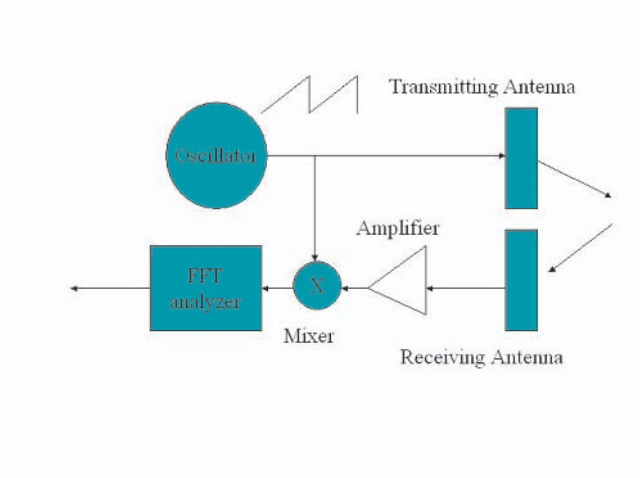


Figure 2. FMCW homodyne radar diagram

The microwave oscillator is frequency-modulated and serves simultaneously as transmitter and as the receiver's local oscillator. To estimate both the target range and velocity the triangle or sawtooth modulation is used. The echo signal is a delayed and Doppler shifted copy of the transmitted signal. After mixing the received signal with the transmitted one, the video (beat) signal is filtered and processed. Let us assume that the signal transmitted by the FMCW radar in time interval $[0, T_s]$, having linear frequency modulation, is in the form

$$s(t) = e^{j\phi(t)}. \tag{17}$$

The signal phase can be described by the second order time polynomial

$$\phi(t) = a_o + a_1 t + a_2 t^2. \tag{18}$$

The instantaneous frequency $f(t) = \frac{d\phi(t)}{dt}\frac{1}{2\pi}$ of the signal (17) is equal to

$$f(t) = \frac{a_1 + a_2 t}{2\pi}, \tag{19}$$

where $a_1/2\pi$ is the starting (carrier) frequency f_o, and $2a_2/2\pi$ is the slope of the frequency modulation. The signal bandwidth is equal to $2a_2T_s/2\pi$. After mixing the signal $s(t)$ with the return echo originating from a stationary target at distance R (which is equivalent to time delay $\tau = 2R/c$) one obtains

$$y(t) = s(t)s^*(t-\tau) = Y_o e^{j(b_0 + b_1 t)}, \tag{20}$$

where Y_o depends on the return signal amplitude and

$$b_0 = a_1\tau - a_2\tau^2, \tag{21}$$

$$b_1 = 2a_2\tau^1. \tag{22}$$

It is easy to notice that the video signal is the harmonic one, with frequency proportional to target range. To detect targets at different ranges, the Fast Fourier Transform (FFT) is often used [39, 13]. For moving targets, the range and time delay are continuously changing with time. This changes the starting phase of the video signal for each saw-tooth and shifts the video frequency. The video frequency shift, caused by the target movement, is equivalent to a change of the measured target range. The moving target range and velocity can thus be calculated using the 2-dimensional FFT. The dimension connected with "fast time" (time within each saw-tooth) gives a pseudo-range, while dimension connected with "slow time" (sawtooth count) gives information on the target velocity. The true range to the target is calculated by correcting the pseudo-range using velocity information.

LFMCW radar has many limitations [45, 41]. Minimum sawtooth length is limited by the maximum target distance. Usually the sawtooth length is 3-9 times longer than the maximum signal delay. This leads to a relatively low frequency of sawtooth modulation, which consequently limits the maximum unambiguous Doppler frequency (range velocity). It is possible to measure target velocity far beyond this ambiguity limitation, but this requires sawtooth modulation frequency diversity and use of the Chinese Reminder Theorem.

Classical FMCW radar can also be used for target acceleration estimation [45]. Constant target acceleration introduces a second order time polynomial to the starting phase of each sawtooth video signal. Using the Generalized Chip Transform (GCT) or Polynomial Phase Transform (PPT) it is possible to estimate not only range to target and target range velocity, but also target range acceleration.

4. Noise Radar

The name "noise radar" refers to a group of radars using random or pseudo-random waveforms for target illumination. In many papers this type of radar is referred to as a random signal radar (RSR) [5]. It can be used in a wide range of applications. It is possible to construct surveillance, tracking, collision warning, sub-surface and other radars using noise waveforms. Noise radars have several advantages over classical pulse, pulse-Doppler and FMCW radars. Noise waveforms guarantee the absence of range or Doppler ambiguity, low peak power and very good electromagnetic compatibility with other devices sharing the same frequency band. Due to the low peak power, noise radars also have very good electronic counter-countermeasures capability and very low probability of interception. This type of radar also has several disadvantages. Signal processing in noise radar is much more difficult and requires much higher computational-power than in traditional radar. Noise radars suffer from the near-far object problem. The received power changes with the reciprocal of the fourth power of the range, so for long-range radar very high effective dynamic range (usually much higher than 100 dB) is required. For smaller effective dynamic range the masking effect will be clearly visible; strong and close objects will mask weak and distant ones [38].

The first paper on noise radar was published in 1959 by B.M. Horton [1], who presented the concept of a range measuring radar. Further papers on that subject were published in the 1960s and 1970s [2] [15] [21]. At this time, the concept of noise radar had not attracted radar engineers, due to the fact that correlation signal processing was very difficult to implement using analog circuits. In the last decade, the noise radar concept has been "rediscovered" [4] [6] [18]. High-speed Digital Signal Processors (DSP) and Programmable Logic Devices (PLD), equipped with hardware multipliers, make it possible to calculate transmitted and received signal cross-ambiguity functions in real time [12, 23] and to exploit Low Probability of Intercept (LPI) properties of noise radars fully. At present, the noise waveform concept is applied in many different types of radars. Many papers have been published on short-range surveillance radar, on imaging radar working in both SAR [9] [16] [17] [19] [20] [30] [31] and ISAR mode [10] [14], ground penetrating radars [3] [7] and others [22]. Noise radar may be used with a mechanically or electronically scanning antenna as well as with a multi-beam antenna. To avoid strong cross talk between transmitter and receiver, separate transmitting and receiving antennas are usually used. A noise signal is transmitted continuously, and the received signal, which is a delayed and Doppler-shifted

copy of the transmitted signal, is divided into blocks and processed in a correlation processor. The ranges and radial velocities of the targets are evaluated directly by the correlation processor while the targets' azimuths are estimated using sigma-delta antenna angle estimation, beam power ratio or other techniques.

The detection process in noise radar is based on a correlation process [11]. The correlation-based radar diagram is presented in Figure 3.

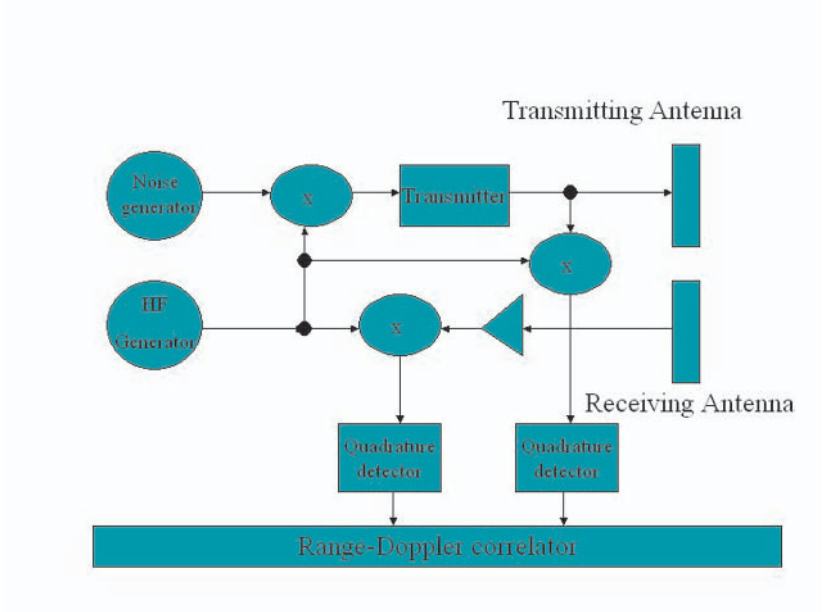


Figure 3. The structure of correlation-based noise radar

At the receiver the cross-correlation function between the transmitted and the received signal is calculated by

$$y_r(\tau) = \int_{t=0}^{t_i} x_T(t) x_T(t-\tau) dt, \tag{23}$$

where x_T is the transmitted signal complex envelope, x_R - received signal complex envelope, and t_i - integration time. While radar is a device that should estimate range to target, Equation (23) can be rewritten in the form

$$y_r(r) = \int_{t=0}^{t_i} x_T(t) x_T(t-\frac{2r}{c}) dt. \tag{24}$$

A correlation receiver enhances the signal-to-noise ratio (S/N) by the factor $t_i B$, where B is the bandwidth of the transmitted noise signal. In Fig. 4, the output of a correlator with a $t_i B$ factor of 100 is presented. To detect the useful signal in most radar systems the constant false alarm Rayleigh detector is being used. This detector compares the signal with the threshold. The hypothesis H_0 (there is only thermal noise and no useful signal) is assumed when the signal is below the threshold, and hypothesis H_1 (there is thermal noise plus target echo in the received signal) is assumed when the signal exceeds the threshold. It can be easily found that for a 10^{-6} probability of false alarm, the threshold level D_o should be at least 12 dB over the correlator output noise rms voltage.

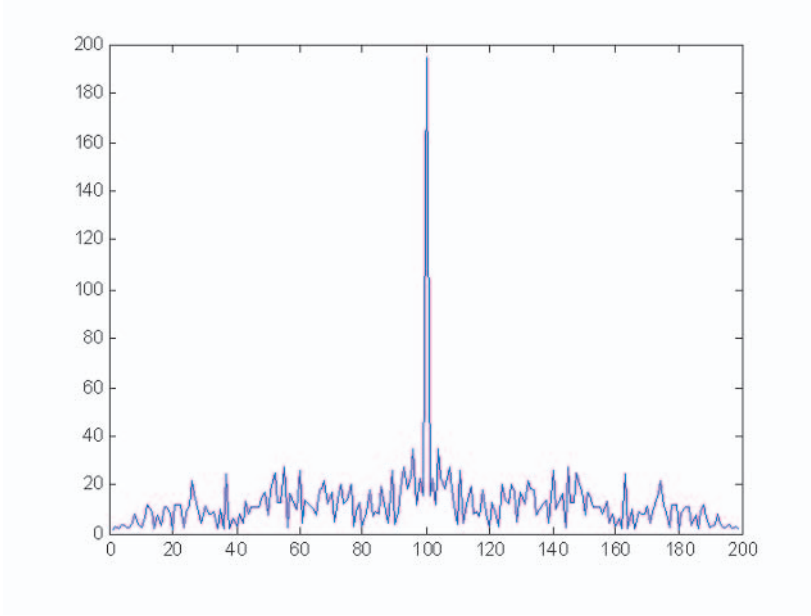


Figure 4. Example of noise correlation function for $Bt_i = 100$

This simple correlation processing can be used only for a short integration time radar. For longer correlation times Equation (24) can only be used for motionless targets. To detect moving targets, the Doppler frequency shift must be incorporated into the detection process. Assuming that the transmitted signal can be treated as a narrow-band one, then the received signal, reflected from the moving target, is a delayed and Doppler shifted copy of the transmitted signal. The complex envelope

of the received signal can be expressed by the formula

$$x_R(\tau) = Ax_T(t - \frac{2r}{c})e^{j2\pi(-\frac{2vF}{c})t+j\phi}. \tag{25}$$

The optimal (in the mean-square sense) detector is based on the matched filter concept. The output of the filter matched to the signal echo described by Equation (23) can be calculated as an integral in the form

$$y = \int_{t=0}^{t_i} x_R x_T^*(t - \frac{2r}{c})e^{-j2\pi(-\frac{2vF}{c})t}. \tag{26}$$

The single matched filter can be used only when the target's position and velocity are known. To detect a target at an unknown position, it is necessary to utilize a bank of filters matched to all possible target ranges and velocities. This approach leads directly to the range-Doppler correlation function, described by the formula

$$y(r,v) = \int_{t=0}^{t_i} x_R x_T^*(t - \frac{2r}{c})e^{-j2\pi(-\frac{2vF}{c})t}. \tag{27}$$

The above equation is very similar to the cross-ambiguity function, but here the time-delay is introduced only in the transmitted signal. This form of the transform is more convenient for digital implementation and will be referred to below as a range-Doppler correlation function.

Equation (27) can be treated as a set of correlation functions of the received signal and the Doppler shifted transmitted signal, or as a set of Fourier transforms of products of the received signal and the complex conjugate of the time shifted transmitted signal. An example plot of the absolute value of the range-Doppler correlation function for a single target at distance 10 km and radial velocity 100 m/s, calculated directly from Equation (27), is presented in Figure 5. It is easy to see the presence of very high side lobes along the frequency dimension, caused by the boxcar window introduced by integration limits. To decrease side lobe levels, more elaborate time windows (Hamming, Blackman and others) can be used. The windowing can be applied either at the transmission side (by changing the transmitted signal amplitude) or during signal processing. The second approach leads to the concept of unmatched filtering, described by the formula

$$y(r,v) = \int_{t=0}^{t_i} w(t)x_R x_T^*(t - \frac{2r}{c})e^{-j2\pi(-\frac{2vF}{c})t}, \tag{28}$$

where $w(t)$ is a time windowing function [13]. The result of applying the Hamming window to range-Doppler correlation processing is presented in Figure 5. The first side-lobe level is decreased from -13 dB to -60 dB.

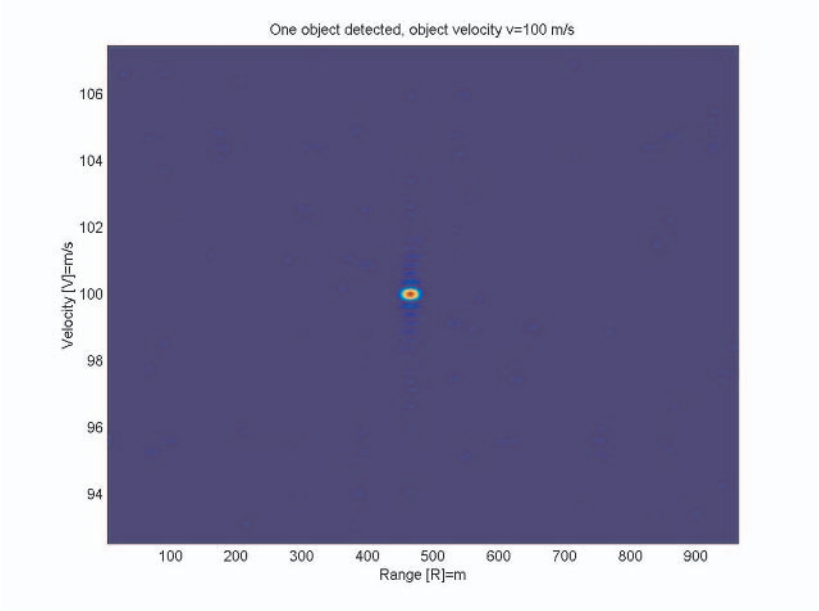


Figure 5. Range-Doppler correlation function, boxcar window, integration time t_i=0.05 s.

5. Noise radar range equation

Noise radar, equipped with a transmitting antenna with gain G_T and a receiving antenna with gain G_R, transmitting microwave power P_T, receives a reflected noise signal from a target having a cross section S_o at distance R. The received power is equal to

$$P_R = \frac{P_T G_T G_R S_o \lambda^2}{(4\pi)^3 R^4 L}. \tag{29}$$

To detect the reflected signal in the presence of thermal white noise, the correlation process (matched filtering) is used according to the fundamentals given above. The received signal will be detected when its power is higher than the thermal noise power P_N multiplied by detectability factor D, e.g.

$$P_R \geq P_N D = kT_R BD. \tag{30}$$

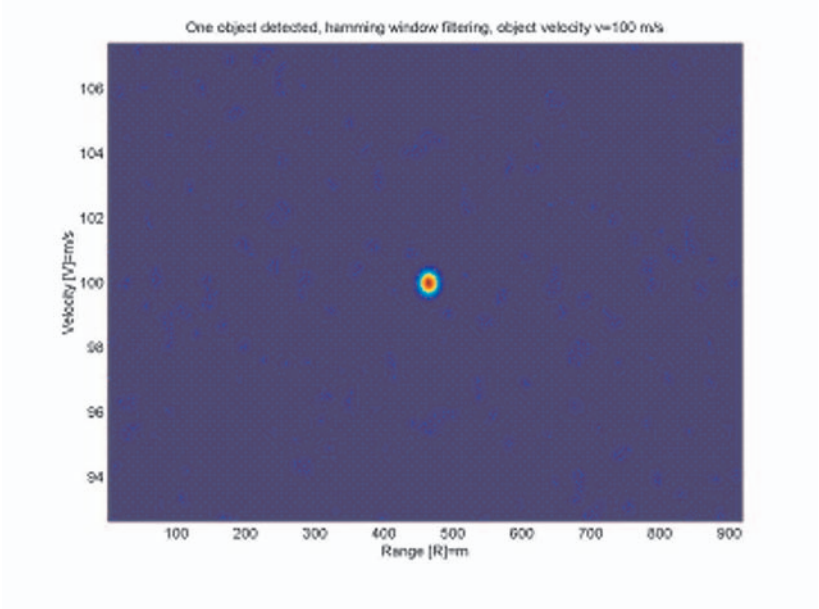


Figure 6. Range-Doppler correlation function, Hamming window, integration time $t_i = 0.05s$.

The correlation receiver's property described above shows that the detectability factor is equal to

$$D = \frac{D_o}{t_i B}. \tag{31}$$

For medium integration time the time-bandwidth product is limited by the range migration effect [40]:

$$t_i B \leq \frac{c}{2V_{max}}. \tag{32}$$

Assuming that the maximum target velocity is equal to 3M (1000m/s), the maximum value of the time-bandwidth product is limited to 150,000. The maximum processing gain is then equal to 51.7 dB. The use of a windowing function will decrease this value by a few dB. The maximum detection range for the noise radar is given by the formula

$$R_{max} = \sqrt[4]{\frac{P_T G_T G_R S_o \lambda^2 t_t}{(4\pi)^3 L D_o k T}}. \tag{33}$$

To increase the detection range one can increase the transmitted power, antenna gain or integration time.

To achieve a longer correlation time it is necessary to incorporate range and Doppler cell migration effects into the detection process. Let us first consider the range cell migration problem. Assuming constant target radial velocity v_o, the range to target can be expressed as $r = r_o + v_o t$. In Equation (27) the target's velocity influence is limited only to the Doppler effect. For a longer integration time it is necessary to take into consideration the fact that the time delay of the signal changes considerably during the integration time. Incorporating time delay changes the detection process; one can obtain a modified correlation expression in the form

$$y(r, v) = \int_{t=0}^{t_i} w(t) x_R x_T^*(t - 2\frac{r + vt}{c}) e^{-j2\pi(-\frac{2vF}{c})t}. \tag{34}$$

The computational complexity required to compute the range-Doppler correlation plane is now much higher than the computational cost of calculating results using equation (27), where it is necessary to time-scale the transmitted function. Time scaling may be performed in the time domain by resampling the transmitted signal or by using a chirp transform. For very long integration times it is hard to justify an assumption of a target's constant range velocity motion. The target's range velocity changes (range acceleration) can cause both velocity cell migration and additional range cell migration. Proper target detection can be achieved only when the target's velocity remains in the velocity resolution cell, which leads to the constraint

$$t_i < \sqrt{\frac{2\lambda}{a_{max}}}, \tag{35}$$

where a_{max} is the target's maximum range acceleration. For example, for a 10 cm wavelength and target acceleration of 1g (9.81 m/s), the coherent integration time is limited to 0.14 s, while for 10 g it is limitated to 0.045 s. Acceleration may also cause additional range migrations, if the integration time exceeds the limit

$$t_i < \sqrt{\frac{2\Delta r}{a_{max}}} = \sqrt{\frac{c}{Ba_{max}}}. \tag{36}$$

To extend the integration time further, it is necessary to introduce acceleration into the target's motion model. The range to the target should now be expressed as $r = r_o + v_o t + a_o t^2/2$. The matched filter concept now leads to the three-dimensional range-Doppler-velocity

correlation function in the form

$$y(r,v,a) = \int_{t=0}^{t_i} w(t)x_R x_T^*(t - \frac{2r+2vt+at^2}{c})e^{-j2\pi(-\frac{(2vt+at^2)F}{c})}. \quad (37)$$

Long integration time is important in multibeam radar. The transmitting antenna in this type of radar is either sector or omni-directional, and the transmitting gain is at the level of a few dB. The receiver antenna should be designed as a multi-beam antenna. Signals from each beam are passed to a multi-channel receiver where a correlation process according to Equation (37) is performed.

6. Bi-static and multi-static continuous wave radars

Monostatic radar is a radar which has the transmitter and receiver in one place. This idea has many advantages. For example it is easy to use a single clock source for both the transmitter and receiver and thus it is relatively easy to make a fully coherent device. There is no need to transmit the reference signal to the remote site; the entire processing can be done locally. The radar can send data to the command and control center at the track level, which requires very low transmission throughput (1-20 kb/s), or on the plot level (10-100 kb/s). For continuous wave radar, the mono-static configuration also has several disadvantages. For example, very good separation (usually better than 60 dB) between transmitter and receiver antennas is required. Echo power changes as the reciprocal of the fourth power of the range. If the ratio between the maximum and the minimum targets' distance is at the level of 1000 (e.g. maximum distance 100 km, minimum distance 100 m), then the required dynamic range exceeds 120 dB. Additionally, there exist stealth targets, which reflect energy in different directions than the direction towards the radar. It is very difficult or even impossible to detect such targets using mono-static radars. The accuracy of the target's position estimation in monostatic radar is limited. The range to the target is calculated usually with good accuracy (estimation error 1-30 m), but cross-range accuracy is usually very poor (error of a few kilometers at 100 km distance).

Using the bi-static concept can reduce all of the above mentioned problems. The spatial separation of the transmitter and receiver antennas leads to significant attenuation of the direct path signal. In addition, the near-far target problem is reduced, while the target signal can be

expressed by the formula

$$P_R = \frac{P_T G_T G_R S_o \lambda^2}{(4\pi)^3 R_1^2 R_2^2 L}, \tag{38}$$

where R_1 is the transmitter-target distance and R_2 is the target-receiver distance. The required dynamic range for the maximum and minimum targets' distance ratio equal to 1000 (e.g. the maximum target's distance from receiver 100 km, the minimum distance 100 m, the transmitter-receiver distance 10 km) is now reduced to 80 db (40 dB smaller than for the mono-static case). The maximum target detection range is described by the formula

$$R_{1max} R_{2max} = \sqrt{\frac{P_T G_T G_R S_o \lambda^2 t_t}{(4\pi)^3 L D_o k T}}. \tag{39}$$

The bi-static theoretical coverage diagram forms the so-called Cassini curve, presented in Figure 7. In practice it is not possible to detect targets in the direction of the transmitter line-of-sight, and practical bi-static radar coverage is presented in Figure 8.

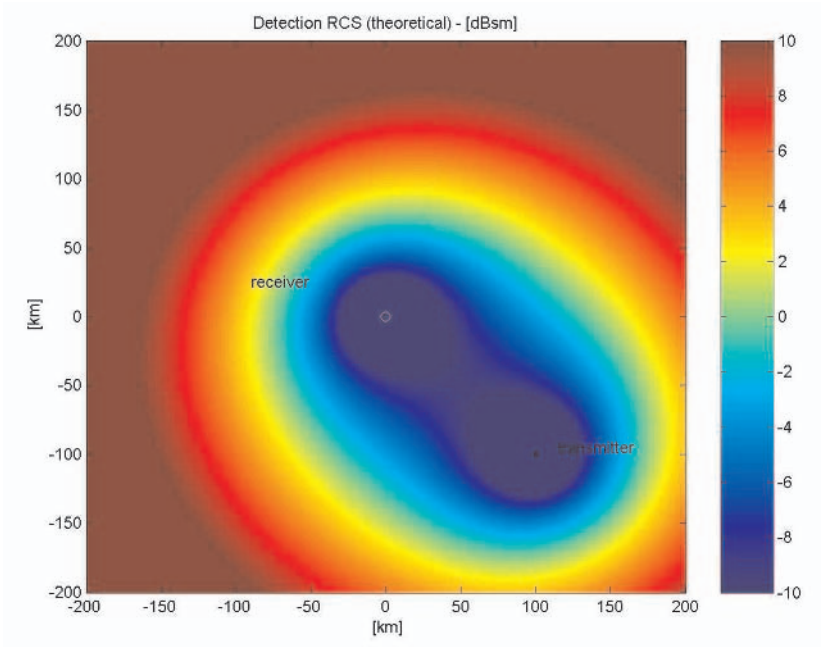


Figure 7. Bistatic theoretical coverage diagram, detected RCS [dBsm].

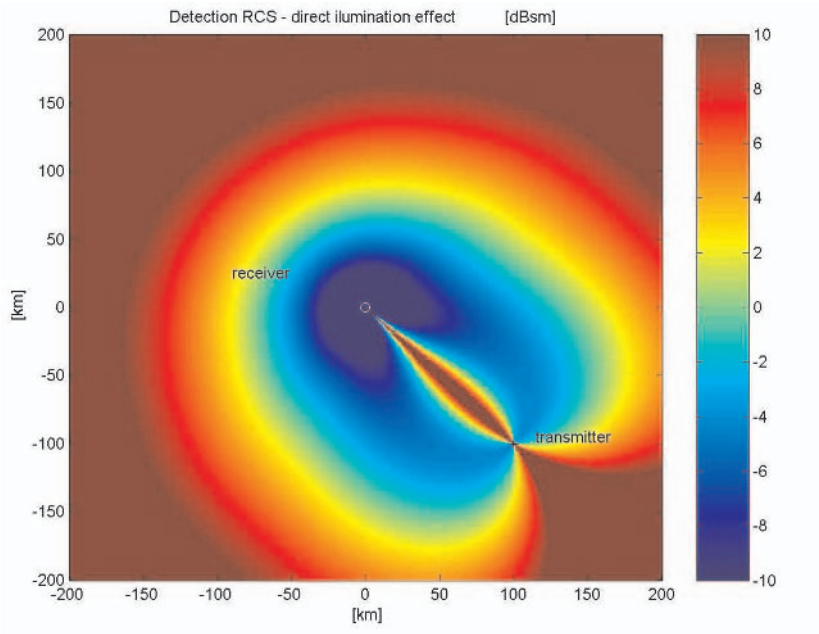


Figure 8. Bistatic coverage diagram — direct illumination attenuation effect, detected RCS [dBsm].

With the receiving antenna in a different place than the transmitting one it is possible to increase the effective radar cross-section of stealth targets. Increasing the number of receiving antennas and forming a multi-static constellation, it is possible to decrease the detection losses caused by signal fading, multi-path effects and target scattering directivity. It is also possible to locate and track targets very accurately using the multi-static concept. Expected errors in almost all surveillance spaces are at the level of 1–20 m. Using an adequate number of sensors (4 or more), it is possible to estimate the target's height, even in the case when the sensors do not have 3D measurement capabilities.

To use a multi-static configuration, it is necessary to have the reference signal at each receiving site. The reference signal can be received by the special reference channel or can be sent from the transmitter using high throughput (10-100 Mb/s) digital links.

Multi-static radar is usually combined with the net-centric approach to data exchange. All sensor sites have to be connected by high-throughput, self-configuring data links. The data links should also provide a very stable clock to make all processing coherent, and very accurate time data to synchronize all events in the distributed system.

7. Target identification in continuous wave radars

Identification of the target is the final stage of radar signal processing. Parameters of the identified target can be displayed on the sensor screen or can be sent to a remote command center. There is no one superior identification procedure, so identification processes are designed using different approaches. Almost all of them are based on Doppler processing of the received radar signal.

One widely-used identification method is based on the SAR concept. The radar is mounted on a moving platform. Platform motion is perpendicular to the line-of-sight of the fixed radar antenna, and the antenna beam is scanning the target during platform motion. The distance from the radar to the selected scatterer is almost a hyperbolic function of time. The Doppler frequency of the target's signal is then a linear function of time, and as a result, each echo signal has a chirp form. Applying match filtering, it is possible to obtain a high-resolution image of the target. The classical angular resolution of the radar is equal to the product of the target's range and the antenna beam width. For a distance of 10 km and antenna beam width 50 [mrad] the angular resolution is equal to 500 m. Using SAR technology, it is possible to improve resolution to $D/2$, where D is the antenna length. Using a low-gain, small size antenna it is possible to obtain a sub-meter cross-range resolution and a very detailed image of the target. A SAR image example is presented in Figure 9. The raw radar image used for the SAR processing is presented in Figure 10.

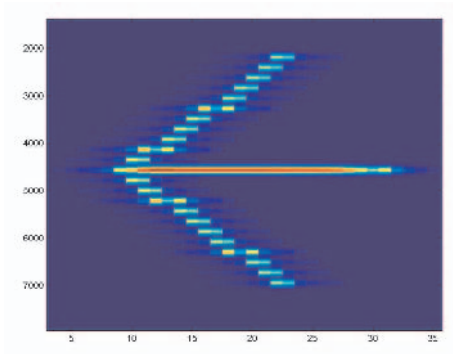


Figure 9. SAR image of arrow shaped object.

In this example, the SAR compression factor was 1000. It is possible to achieve even higher compression ratios and a better cross-range resolu-

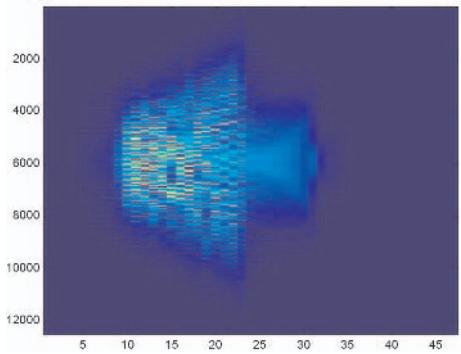


Figure 10. Raw radar data of arrow shaped object.

tion by using spotlight-SAR. In this mode the radar antenna tracks the target and observation time (time-on-target) is significantly enlarged. SAR is often used for air-borne and space-borne remote sensing. It is also used for ground penetrating radar and for imaging hidden targets.

An interesting modification of the above is ISAR (Inverse Synthetic Aperture Radar). The ISAR scenario is opposite to the SAR scenario. The radar is now placed in a fixed position and the target changes its position in time. The Doppler history of each scatterer on the target's surface is different, and the target image is reconstructed using a bank of digital filters matched to the signal originating from each target's scatterer. This technology is used for ship, airplane and space satellite imaging. Very often the radar, placed on a moving platform, is observing a moving target. For such a scenario it is possible to combine both of the above technologies.

Doppler processing is also used for vehicle and people identification. The identification can be achieved using micro-Doppler analysis. Micro-Doppler vibration analyses are based on detecting Doppler frequency modulations caused by vibration of the vehicle body resulting from engine or gearbox frequencies. Analyzing the Doppler signal originating from the vehicle it is possible to detect all characteristic frequencies, and even to make a mechanical diagnosis of the state of the engine and shafts.

Micro-Doppler analyses can also be used for people identification. In that case, it is possible to detect Doppler frequency modulation caused by heartbeats, breathing and body motion (walk, run, head turns etc.).

For people identification, it is also possible to combine micro-Doppler analyses with ISAR processing. During walk and run, there is a very

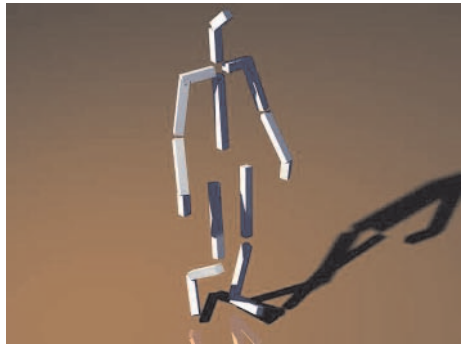


Figure 11. Wire model of human body used for walk Doppler signature analysis.

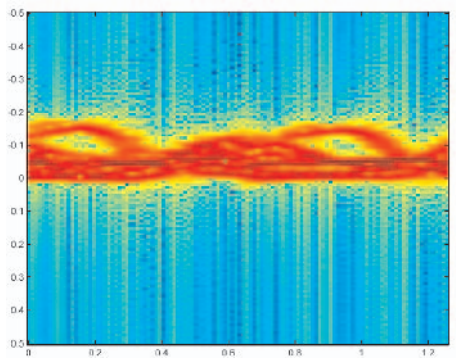


Figure 12. Simulated Doppler signature of human walk - x-axis: time, y-axis: Doppler frequency.

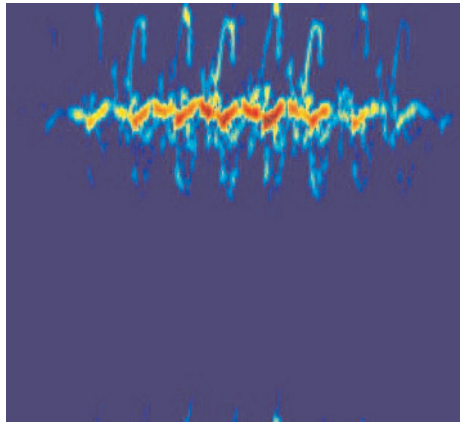


Figure 13. Registered Doppler signature of human walk - x-axis: time, y-axis: Doppler frequency.

characteristic movement of legs and hands. Making even a very simple "wire-model" of the human body (see Figure 11), it is possible to predict a human Doppler signature. The modeled signature of human walk is presented in Figure 12. The signature obtained using a real-life signal, registered in X-band radar, is presented in Figure 13. By comparing the predicted signature with the recorded one it is possible to recognize individual behavior, and combining it with heartbeat, walk and breathing analysis it is possible to identify people and even estimate their emotional state.

References

[1] B.M. Horton, "Noise-modulated distance measuring system", Proc. IRE, V0147, pp. 821–828, May 1959.

[2] G.R. Cooper and C.D. McGillem, "Random signal radar", School Electr. Eng., Purdue Univ., Final Report, TREE67-11, June 1967.

[3] R.M. Naryanan et al., "Design and performance of a polarimetric random noise radar for detection of shallow buried targets", Proc. SPIE Meeting on Detection Techn. Mines, Orlando, April 1995, vol. 2496, pp. 20–30.

[4] I.P. Theron et al., "Ultra-Band Noise Radar in the VHF/UHF Band", IEEE AP-47, June 1999, pp. 1080–1084.

[5] S.R.J. Axelsson, "On the Theory of Noise Doppler Radar", Proc. IGARSS 2000, Honolulu, 24-28 July 2000, pp. 856–860.

[6] R. M. Narayanan, Y. Xu, P. D. Hoffmeyer, and J. O. Curtis, "Design, performance, and applications of a coherent ultra wide-band random noise radar", Opt. Eng., vol. 37, no. 6, pp. 1855–1869, June 1998.

[7] Y. Xu, R. M. Narayanan, X. Xu, and J. O. Curtis, "Polarimetric processing of coherent random noise radar data for buried object detection", IEEE Trans. Geosci. Remote Sensing, vol. 39, no. 3, pp. 467–478, Mar. 2001.

[8] R. M. Narayanan and M. Dawood, "Doppler estimation using a coherent ultra wide-band random noise radar", IEEE Trans. Antennas Propagat., vol. 48, pp. 868–878, June 2000.

[9] D. Garmatyuk and R. M. Narayanan, "Ultrawide-band noise synthetic radar: Theory and experiment", in IEEE Antennas Propagat. Soc. Int. Symp. 1999, vol. 3, Orlando, FL, July 1999, pp. 1764–1767.

[10] D. C. Bell and R. M. Narayanan, "ISAR turntable experiments using a coherent ultra wide-band random noise radar", in IEEE Antennas Propagat. Soc. Int. Symp. 1999, Orlando, July 1999, pp. 1768–1771.

[11] D. J. Daniels, "Resolution of ultra wide-band radar signals", Proc. Inst. Elec. Eng.-Radar, Sonar Navig., vol. 146, no. 4, pp. 189–194, Aug 1999.

[12] M. E. Davis, "Technical challenges in ultra wide-band radar development for target detection and terrain mapping", in Proc. IEEE 1999 Radar Conf., Boston, MA, April 1999, pp. 1–6.

[13] F. J. Harris, "On the use of windows for harmonic analysis with the discrete Fourier transform", Proc. IEEE, vol. 66, no. 1, pp. 51–83, Jan. 1978.

[14] B. D. Steinberg, D. Carlson, and R. Bose, "High resolution 2-D imaging with spectrally thinned wide-band waveforms", in Ultra Wideband Short-Pulse Electromagnetics 2, L. Carin and L. B. Felsen, Eds. New York: Plenum , 1995, pp. 563–569.

[15] Craig. S.E, Fishbein, W., Rittenbach, O.E., "Continuous-Wave Radar with High Range Resolution and Unambiguous Velocity Determination", IRE Trans. Mil Electronics, vol. MIL 6. No. 2. April 1962, pp. 153–161.

[16] D. S. Garmatyuk and R. M. Narayanan, "SAR imaging using acoherent ultrawideband random noise radar", in Radar Processing, Technology, ond Applications IV, (William I. Miceli, Editor), Proceedings of SPIE Vol. 3810. pp. 223–230, Denver, CO, July 1999.

[17] M Soumekh, "Reconnaissance with ultra wideband UHF synthetic aperture radar", in IEEE Signal Proc. Mag.. Vol. 12, No. 4, pp. 21–40, July 1995.

[18] L. Y. Astanin and A . A. Kostylev, "Ultrawideband Radar Measurements, Analysis and Processing", The Institution of Electrical Engineers, London, 1997.

[19] Garmatyuk, D.S.; Narayanan, R.M., "SAR imaging using fully random bandlimited signals", Antennas and Propagation Society International Symposium, 2000. IEEE Vol. 4 (2000), pp. 1948–1951.

[20] Mogila, A.A.; Lukin, K.A.; Kovalenko, N.P.; Kovalenko, R.P., "Ka-band noise SAR simulation", Physics and Engineering of Millimeter and Sub-Millimeter Waves, 2001. The Fourth International Kharkov Symposium, 4–9 June 2001, Volume 1, pp. 441–443.

[21] M. P. Grant, G. R. Cooper, and A. K. Kamal, "A class of noise radar systems", Proc. IEEE, vol. 51, pp. 1060–1061, July 1963.

[22] R. M. Narayanan, R. D. Mueller, and R. D. Palmer, "Random noise radar interferometry", in Proc. SPIE Conf. Radar Processing, Technol. Appl., vol. 2845, W. Miceli, Ed., Denver, CO, Aug. 1996, pp. 75–82.

[23] R. M. Narayanan, Y. Xu, P. D. Hoffmeyer, and J. O. Curtis, "Design, performance, and applications of a coherent ultrawideband random noise radar", Opt. Eng., vol. 37, no. 6, pp. 1855–1869, June 1998.

[24] R. M. Narayanan and M. Dawood, "Doppler estimation using a coherent ultrawide-band random noise radar", IEEE Trans. Antennas Propagat., vol. 48, pp. 868–878, June 2000.

[25] I. P. Theron, E. K. Walton, and S. Gunawan, "Compact range radar cross-section measurements using a noise radar", IEEE Trans. Antennas Propagat., vol. 46, pp. 1285–1288, Sept. 1998.

[26] I. P. Theron, E. K. Walton, S. Gunawan, and L. Cai, "Ultrawide-band noise radar in the VHF/UHF band", IEEE Trans. Antennas Propagat., vol. 47, pp. 1080–1084, June 1999.

[27] L. Guosui, G. Hong, and S. Weimin, "Development of random signal radars", IEEE Trans. Aerosp. Electron. Syst., vol. 35, pp. 770–777, July 1999.

[28] J. D. Sahr and F. D. Lind, "The Manastash Ridge radar: A passive bistatic radar for upper atmospheric radio science", Radio Sci., vol. 32, no. 6, pp. 2345–2358, Nov. 1997.

[29] M. A. Ringer and G. J. Frazer, "Waveform analysis of transmissions of opportunity for passive radar," in Proc. ISSPA, Brisbane, Australia, Aug. 1999, pp. 511–514.

[30] D. S. Garmatyuk and R. M. Narayanan, "Ultra wide-band continuous-wave random noise arc-SAR", in IEEE Transactions on Geoscience and Remote Sensing, Volume 40, Issue 12, Dec. 2002, pp. 2543–2552.

[31] Xu Xiaojian and R. M. Narayanan, "FOPEN SAR imaging using UWB step-frequency and random noise waveforms", IEEE Transactions on Aerospace and Electronic Systems, Volume 37, Issue 4, Oct. 2001, pp. 1287–1300.

[32] S. R. J. Axelsson, "Noise radar using random phase and frequency modulation", Proc. of IEEE International Geoscience and Remote Sensing Symposium (IGARSS) 2003, Volume 7, 21–25 July 2003, pp. 4226–4231.

[33] S. R. J. Axelsson, "Suppressed ambiguity in range by phase-coded waveforms", Proc. of IEEE International Geoscience and Remote Sensing Symposium (IGARSS) 2001, Volume 5, 9–13 July 2001, pp. 2006–2009.

[34] K.S. Kulpa, Z. Czekala, "Ground Clutter Suppression in Noise Radar", Proc. Int. Conf. RADAR 2004, 18–22 October2004, Toulouse, France, p. 236.

[35] M.Nalecz, K. Kulpa, A. Piatek, "Hardware/Software Co-designin DSP-Based Radar and Sonar Systems", International Radar Symposium 2004 19-21 Maj, Warsaw, Poland, pp. 137–142.

[36] K. Kulpa, "Adaptive Clutter Rejection in Bi-static CW Radar", International Radar Symposium 2004 19–21 Maj, Warszawa Polska, pp. 61–68.

[37] M. Nalecz, K. Kulpa, R. Rytel-Andrianik, S. Plata, B. Dawidowicz, "Data recording and processing in FMCW SAR system", International Radar Symposium 2004 19–21 Maj, Warsaw, Poland, pp. 171–177.

[38] K. Kulpa, Z. Czekala, "Short Distance Clutter Masking Effects in Noise Radars", Proceedings of the International Conference on the Noise Radar Technology. Kharkiv, Ukraine, 21–23 October 2003.

[39] A. Wojtkiewicz, M. Nalecz, K. Kulpa, R. Rytel-Adrianiuk, "A novel Approach to Signal Processing in FMCW Radar", Bulletin of the Polish Academy of Science, Technical Sciences, Vol. 50, No. 4, Warszawa 2002, pp. 346-359.

[40] K. Kulpa, Z. Czekala, M. Smolarczyk, "Long-Time-Integration Surveillance Noise Radar", First International Workshop On The Noise Radar Technology (NRTW 2002), Yalta, Crimea, Ukraine, September 18–20, 2002, pp. 238–243.

[41] K.Kulpa, A.Wojtkiewicz, M.Nalecz, J.Misiurewicz, "The simple analysis method of nonlinear frequency distortions in FMCW radar", Journal of Telecommunications and Information Technology, No. 4, 2001, pp. 26–29.

[42] A. Wojtkiewicz, M. Nalecz, K. Kulpa, "A novel approach to signal processing in FMCW radar", Proc. Int. Conf. on Signals and Electronic Systems ICSES'2000, Ustron, Poland, 17–20 Oct. 2000, pp. 63–68.

[43] Stove A.G., "Linear FMCW radar techniques", IEE Proceedings-F, Vol. 139, No. 5, Oct. 1992, pp. 343–350.

[44] M. J. Skolnik, "Radar Handbook", McGraw-Hill Professional; 2nd edition, January 1990.

[45] A.Wojtkiewicz, M.Nalecz, K.Kulpa, W.Klembowski, "Use of Polynomial Phase Modeling to FMCW Radar. Part C: Estimation of Target Acceleration in FMCW Radars", NATO Research and Technology Agency, Sensors and Electronics Technology Symposium on Passive and LPI (Low Probability Of Intercept) Radio Frequency Sensors, Warsaw, Poland, April 23–25, 2001, paper #40C.

[46] K. Kulpa, "Novel Metchod of Decreasing Influence of Phase Noise on FMCW Radar", 2001 CIE International Conference on Radar Processing, Oct. 15–18, 2001, Beijing, China, pp. 319–323.

TERAHERTZ IMAGING, MILLIMETER-WAVE RADAR

R. W. McMillan
U.S. Army Space and Missile Defense Command
Huntsville, Alabama, USA

Keywords: terahertz imaging; subwavelength imaging; millimeter-wave radar; atmospheric effects; terahertz sources; terahertz detectors.

Abstract The millimeter wave (MMW) band of frequencies extends from 30 GHz to 300 GHz, with some fuzziness on both ends of this spectrum. The terahertz (THz) band extends from about 200 GHz to about 30 THz, despite the fact that the lower frequencies in this range are not strictly 10^{12} Hz or higher. These bands are also variously called submillimeter, far-infrared, and near-millimeter. In recent years, there has been some degree of hype associated with the capabilities of systems operating in these bands. Sometimes exorbitant claims have been made relative to the ability of these systems to see through walls, detect buried structures, and detect cancer cells, for example. In this chapter we shall examine some of these clams and assess their validity. We shall find that MMW and THz systems can do some amazing things, some of them not related to the above claims, and that there is substantial promise of even more interesting results. In this chapter we begin by discussing these atmospheric limitations, since they permeate the whole technology of MMW, sub-MMW, and THz technology. We then discuss MMW and THz sources, detectors, optics, and systems in separate sections. Finally, we present some results obtained using sensors operating in these bands. Perhaps the most interesting of these results demonstrate the capability to image objects at resolutions as good as $\lambda/100$, where λ is wavelength. These measurements show the connection between this sensor technology and applications to security.

1. Introduction

Electromagnetic radiation in the range of wavelengths from 1 cm to 1 mm is characterized as millimeter-wave (MMW) radiation, while the range extending from 1 mm to 0.3 mm is called sub-millimeter wave (sub-MMW or sub-mm), and that of shorter wavelengths extending to the infrared is terahertz (THz) radiation. From the opticist's point of

Jim Byrnes and Gerald Ostheimer (eds.),
Advances in Sensing with Security Applications, 243–268.

view, the latter range is also known as the far-infrared (far-IR) range. From a very practical aspect, these ranges have traditionally overlapped. For example, many workers in the field consider that the MMW band begins at 40 GHz (7.5 mm) because of the similarities in systems and components in Ka-band (26.5–40 GHz) to those of the microwave bands. Similarly, because of the difficulty in generating and detecting THz signals, this band is sometimes considered to begin at frequencies variously extending from 200 to 600 GHz, depending on the worker and application involved. Indeed, most of the papers presented at THz conferences discuss systems, techniques, and results obtained in this latter band of frequencies. The far-IR band generally extends from about 20 microns (15 THz) to 0.3 mm (1 THz).

System development and operation at frequencies above about 300 GHz are usually confined to passive applications such as radio astronomy and remote sensing using radiometry because of the difficulty in generating usable amounts of power at these frequencies. It is not particularly difficult to generate the small amount of power useful for pumping a THz mixer, however, so that many passive applications are found in this range. This limitation is being overcome to a large extent by continuing research in this area. Another limitation on performance of systems at these higher frequencies is the absorption in the atmosphere, mostly due to water vapor, which begins to limit transmission severely above about 300 GHz.

In this chapter we begin by discussing these atmospheric limitations, since they permeate the whole technology of MMW, sub-MMW, and THz technology. We then discuss MMW and THz sources, detectors, and systems in separate sections. Finally, we present some results obtained using sensors operating in these bands. These results show the connection between this sensor technology and applications to security.

2. Atmospheric Limitations

The atmosphere is the most significant factor in limiting the performance of MMW and THz systems. The characteristics of the atmosphere that cause this limitation are attenuation and turbulence, and attenuation is much more significant than turbulence. There are regions of the spectrum in these bands that are attenuated by as much as 500 dB/km, which makes operation at ranges of even a few meters extremely challenging. Atmospheric turbulence causes fluctuations in signal intensity of 2–3 dB and changes in angle-of-arrival of several tens of microradians. Each of these effects is discussed in more detail in the following paragraphs.

Perhaps the most useful calculations of atmospheric attenuation have been done by Liebe [1], who used a variant of the Van Vleck-Weisskopf (VVW) [2] equation for the line shapes of water vapor and other atmospheric constituents. It is a common practice in calculating atmospheric attenuation to add a frequency-dependent empirical correction factor to the calculation because none of the analytic line shapes developed thus far give accurate results in the atmospheric window regions. Based on the VVW line shape, this correction factor, and his own measurements [3], Leibe has developed a computer program called MPM [4] that calculates attenuation as a function of a variety of factors including temperature, pressure, and relative humidity as well as for a number of other conditions such as rain and fog. MPM has been shown to give results accurate to about 0.2 dB/km in the atmospheric window regions of interest in the range 0–1000 GHz. Figure 1 shows the results of calculating the atmospheric attenuation over the range 40–1000 GHz for relative humidities of 50 and 100 percent and rainfall rates of 5 mm/h and 20 mm/h. The curve for 100 percent relative humidity includes attenuation due to 0.5 g/m3 of condensed water vapor, corresponding to a fog that would give only 100 m visibility in the visible spectrum. Note that this thick fog does not greatly affect propagation, especially at the lower frequencies, because attenuation due to fog results mainly from Rayleigh scattering in these bands. Rainfall is another matter, however. Raindrop size distribution depends strongly on rain rate; larger drops occur at higher rain rates. Since raindrops are on the order of a few millimeters in diameter, strong attenuation due to Mie scattering occurs. The rainfall curves of Figure 1 show that the Mie resonance occurs in this region, corresponding to maximum Mie scattering. After the frequencies due to Mie resonance have been exceeded, Mie scattering remains almost constant well into the visible range.

It is a bit surprising that atmospheric turbulence affects propagation at MMW and THz frequencies, since it has been shown theoretically [5] and verified experimentally [6] that the log amplitude variance of amplitude fluctuations varies as f7/6 where f is frequency. These effects are strong at visible and infrared wavelengths, where frequencies are two orders of magnitude higher, so one would expect that they would be minimal at 1 THz. The answer to this question lies in the fact that the atmospheric turbulence structure parameter, commonly denoted as Cn2, is much larger, sometimes as much as two orders of magnitude, at MMW and THz frequencies because of the contribution of water vapor at these latter frequencies [7]. This parameter, which includes contributions from temperature and water vapor, as well as the cross-correlation of these two parameters, occurs in all computations of atmospheric tur-

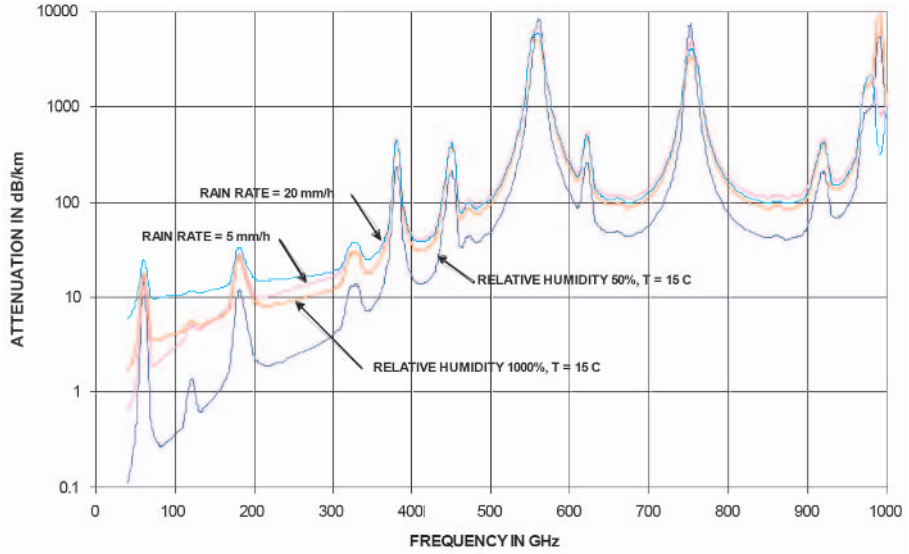


Figure 1. Atmospheric attenuation in the range 40–1000 GHz.

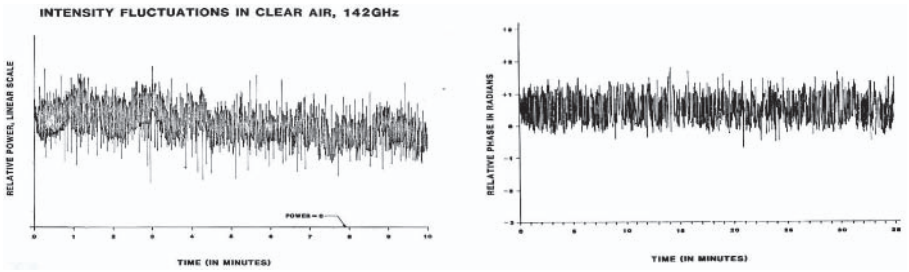


Figure 2. Intensity and phase fluctuations measured over a 1.3 km path at the same time.

bulence effects including both intensity and phase fluctuations. Figure 2 shows the results of measuring fluctuations in the intensity of a 140 GHz signal propagated over a range of 1.3 km at a very flat site near Champagne-Urbana, IL [6]. Figure 2 shows the corresponding phase fluctuations. These results were obtained on a hot and humid summer day, which is the worst-case condition for turbulence effects at MMW frequencies. Observations in snow, rain, and fog all gave smaller fluctuations, even though the signals propagated through rain showed strong, comparatively long-term changes due to variations in rain rate during the observation time.

3. Millimeter-Wave and Terahertz Sources of Radiation

As frequency increases, it becomes increasingly difficult to generate radiation at usable power levels. This limitation is due to the decreasing size of the frequency-sensitive elements in most sources of radiation. In general, the frequency-sensitive elements decrease in size linearly with wavelength. At the higher frequencies, it becomes more difficult to fabricate these small structures with accuracy good enough to ensure good performance. This problem occurs with both solid-state and tube-type sources of radiation.

Exceptions to this general rule are the so-called beam wave tubes and optically-pumped lasers. The gyrotron and the free-electron laser (FEL) are example of beam-wave tubes. The gyrotron uses stimulated cyclotron emission of electromagnetic waves by electrons [8]. This tube is an axially symmetric device having a large cathode, an open cavity, and an axial magnetic field. Electrons are emitted from the cathode with a component of velocity perpendicular to the magnetic field, so that they are caused to spiral as they are accelerated through the magnetic field to the collector. This spiraling occurs at the cyclotron frequency of the electrons, and it is this frequency that is radiated by the tube. The coupling between the electron beam and the MMW radiation allows the beam and microwave circuit dimensions to be large compared to a wavelength so that the power density and related circuit dimension problems encountered in almost all other MMW tubes are avoided. Power outputs of 22 kW CW at 2 mm and 210 kW pulsed at 2.4 mm have been obtained. Because of this high power, gyrotrons are used for such applications as plasma heating in fusion experiments and for extremely high power radars. Russian scientists have built a megawatt radar operating at Ka-band using an array of gyrotrons [9]. This system uses multiple power-combined gyrotrons and multiple Cassegrain antennas in a phased-array configuration.

FELs use relativistic electron beams propagated through a periodic structure of magnets, called an undulator, to generate radiation over a broad spectrum from the submillimeter wave to the x-ray region. These systems are extremely large and are placed in permanent installations. They are used for remote sensing and materials testing over a broad range of wavelengths.

Optically-pumped lasers (OPLs) offer very useful power levels at discrete frequencies well into the THz band. Most OPLs comprise some sort of gas cell, which is the active laser medium, that is pumped by a carbon dioxide laser. These devices are inherently inefficient because the

pump laser excites the active medium into a higher vibrational state, and the laser transitions occur between rotational levels within this state, an energy difference of several orders of magnitude. Both CW and pulsed OPLs have been built at thousands of different wavelengths within the MMW and THz bands. They are not useful for many applications because they operate only on discrete frequencies and because the pulsed versions have low duty cycle.

A family of vacuum-tube MMW sources is based on the propagation of an electron beam through a so-called slow-wave or periodic structure. Radiation propagates on the slow-wave structure at the speed of the electron beam, allowing the beam and radiation field to interact. Devices in this category are the traveling-wave tube (TWT), the backward-wave oscillator (BWO) and the extended interaction oscillator (EIO) klystron. TWTs are characterized by wide bandwidths and intermediate power output. These devices operate well at frequencies up to 100 GHz. BWOs, so called because the radiation within the vacuum tube travels in a direction opposite to that of the electron beam, have very wide bandwidths and low output powers. These sources operate at frequencies up to 1.3 THz and are extensively used in THZ spectroscopic applications [10] [11] [12]. The EIO is a high-power, narrow band tube that has an output power of 1 kW at 95 GHz and about 100 W at 230 GHz. It is available in both oscillator and amplifier, CW and pulsed versions. This source has been extensively used in MMW radar applications with some success [13].

A variety of solid-state sources operate in the MMW portion of the spectrum with varying power outputs and bandwidths. Gunn oscillators, which are bulk devices made from GaAs, InP, and GaN, are available at frequencies of up to 140 GHz in the InP version and may serve as oscillators in the THz region [14]. These oscillators operate based on the negative resistance caused when carriers are excited from a higher mobility to a lower mobility state by the application of an electric field. Gunn devices have outputs of a few milliwatts and are low-noise devices, so that they are useful as local oscillators in MMW receivers. Impact-Ionization Transit-Time (IMPATT) oscillators have higher output powers than Gunns, but have higher noise outputs as well because of the way in which carriers are generated. Carriers are generated in a semiconductor by impact ionization. These carriers traverse a drift region with some transit time. Because of the transit time, the current through the device lags the voltage, and when this lag exceeds 90 degrees, oscillation will occur. Both Gunns and IMPATTs have been supplanted for many applications, especially at lower frequencies, by field-effect transistors (FETs). The great advantage of FETs in circuit applications is that

they are three-terminal devices and this third terminal is a gate by which the gain of the device can be controlled. FETs are used in low-noise amplifier applications up to 100 GHz and in power amplifier applications to about 40 GHz. FETs can be integrated into circuits with other functions so that they are very useful in radar transmit-receive modules, but so far this application is limited to the microwave frequencies. FET oscillators have operated at frequencies up to 230 GHz. Solid-state devices suffer from the same decrease in size of the frequency-determining elements as do vacuum tubes, so that solid-state device operation is limited to about 230 GHz.

The above discussion indicates that the only fundamental source available in the THz spectral region is the BWO. For many years before these tubes became available, molecular spectroscopists relied on generating power by multiplying the output of a lower-frequency source. This technique is still used extensively today. Frequency multiplication is based on irradiation of a nonlinear device, such as a varactor, with a lower frequency. The varactor generates multiple harmonics because of its nonlinearity, and powers at the desired frequencies are enhanced by careful cavity design. Since the input frequency is multiplied, the bandwidth of the input signal source is multiplied by the same factor. In the THz band, power outputs of only a few microwatts are generated, but these levels are enough for many spectroscopic applications. Figure 3 is a photograph of a varactor-based multiplier capable of output in the range 1.26–1.31 THz at a level of 10 (W with an input of 1 mW at 18 GHz [15]. This power level is useful for spectroscopic measurements.

4. Millimeter-Wave and Terahertz Detectors and Receivers

The earliest detectors of MMW and THz radiation, and indeed of microwave radiation in general, were simple point-contact diodes made by carefully contacting a crystal with a sharpened wire, or "cat's whisker". This technique is still used for the higher frequencies today and is still a very useful and effective method of detecting radiation at frequencies into the visible range. It has been refined in recent years by the use of Schottky-barrier structures in which a GaAs semiconductor crystal is contacted through gold contacts set in a mask of insulating SiO2. These diodes have very low stray capacitance and operating frequencies extending well into the THz range. An interesting and perhaps surprising detector comprises a tungsten cat's whisker contacting a metal oxide layer on a metal post, such as aluminum [16]. This so-called metal-oxide-metal (MOM) diode has been used in mixing experiments for lasers

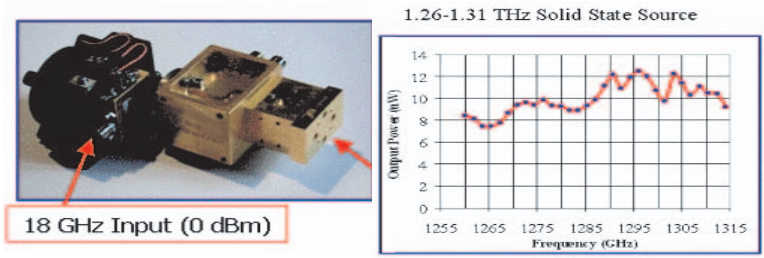


Figure 3. A varactor doubler with output in the THz spectrum [15].

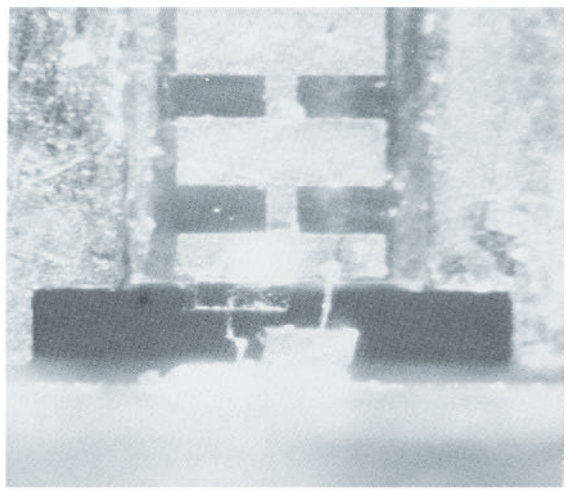


Figure 4. Back-to-back point contact diodes in a subharmonic mixer configuration.

operating into the visible spectrum and has been used for very precise measurements of the speed of light [17]. Figure 4 is a photograph of such a diode contact, and Figure 5 shows an array of gold windows in an insulating mask on GaAs that is used to form Schottky barriers.

An interesting variant of the point-contact mixer is the back-to-back diode configuration in which the fundamental frequency and all odd harmonics are cancelled, resulting in a mixer that operates at twice the fundamental [18], and in some cases at four times the fundamental [19]. This configuration has been used to extend the range of all operational solid-state receivers to 320 GHz and higher. Figure 6 [15] shows a 320 GHz receiver with a system noise temperature of 1360 K built using this technique.

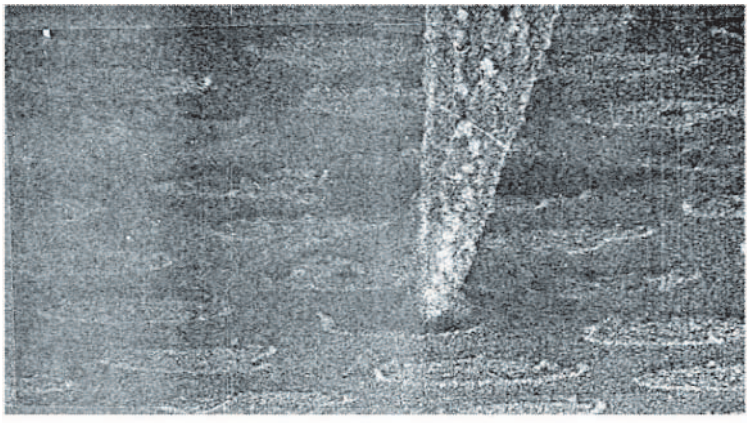


Figure 5. Photograph of a point-contact Schottky-barrier diode. The magnification is about 11,000 times.

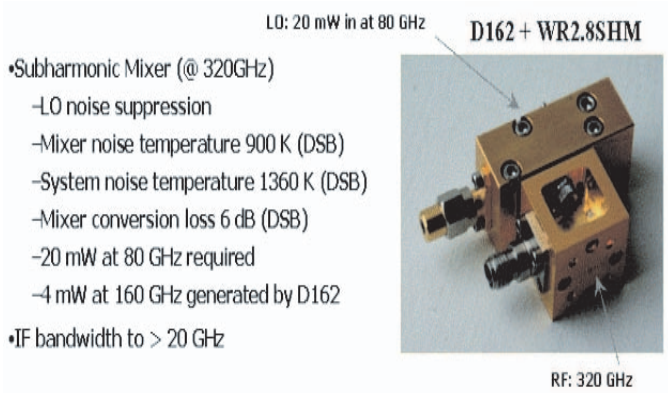


Figure 6. A 320 GHz receiver based on a subharmonic mixer [15].

The ability to make ever smaller solid-state devices by improved lithography techniques has led to the development of so-called beam lead Schottky-barrier diode detectors and mixers in which diodes are fabricated by the same techniques used to make integrated circuits, and for this reason, they can be included in these circuits. Figure 7 shows such a beam-lead detector/mixer made by Virginia Diodes of Charlottesville, VA [15]. This same configuration is used in fabricating the varactor devices used for frequency multiplication discussed in the preceding section.

A significant gap in detector coverage exists in the range extending from about 1 THz to 6 THz (300 m to 50 m wavelength). This band can be covered by the MOM devices mentioned above, but they are too noisy and fragile for most applications. Radio astronomers have long used cryogenically-cooled detectors such as Josephson junction devices for this region [20]. Another possibility for some needs is the uncooled bolometric detectors made from barium strontium titanate, for example, and designed for operation in the long-wave infrared band [21]. Significant progress has been made in the development of these detectors to the extent that they are extensively used for both civilian and military imaging purposes. They have demonstrated minimum detectable temperature differences of as low as 50 mK in the long-wave infrared band. Unfortunately, their use in the far-infrared and THz regions of the spectrum is limited because of the rapid rolloff in energy output from a room-temperature black body in the THz bands. These devices have the overwhelming advantage of being available in imaging arrays, so that numerous security related issues needing imaging sensors can be addressed.

5. Millimeter-Wave and Terahertz Optics

At frequencies above about 100 GHz, and at lower frequencies for many applications, metal waveguides become unacceptably lossy because of skin effect losses and our inability to make these waveguides with the precision required for low-loss operation. In many cases, these problems can be solved by using techniques developed for the visible and infrared portions of the spectrum. This approach to the propagation and handling of MMW and THz radiation has been called "quasi-optics" or "diffraction-limited optics". The latter term arises because the approximations of geometrical optics, namely infinitesimally small focal points and perfectly collimated beams, no longer apply. Instead, beams are focused to spot sizes depending on wavelength, distance, and aperture size, and collimated beams undergo significant diffractive spreading.

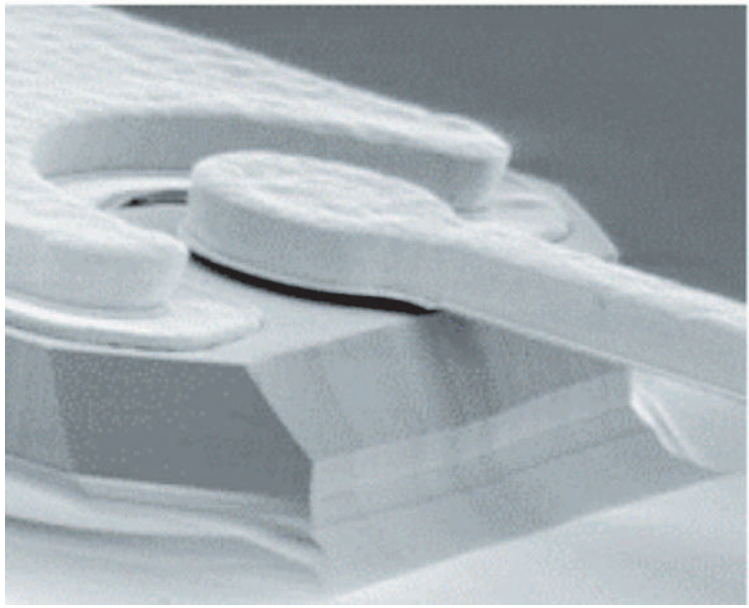


Figure 7. A beam-lead Schottky-barrier detector/mixer [15].

It is possible to build optical analogs of almost all waveguide components including waveguides, attenuators, polarization rotators, directional couplers, and antennas. For example, Figure 8 shows a section of optical beam waveguide that is analogous to metal waveguide. The difference is that the optical waveguide will operate at much higher power levels and at much lower loss than the metal waveguide. The only significant loss in the optical waveguide is the reflection loss at the lens interfaces caused by the index of refraction of the lens material. This loss can be corrected by machining radial grooves into the lenses that are one-quarter wavelength in depth and with an aspect ratio such that the average index of the machined area equals the square root of the index of the lens material. Essentially zero loss, together with high power handling capability, can be achieved if the beam waveguide is made from mirrors. Two disadvantages of beam waveguides, and of MMW and THz optical components in general, are that these devices are bulky, and they may require mode converters if it is necessary to convert to the fundamental waveguide mode.

Many common plastic materials can be used to make useful MMW and THz optical components. Lenses can be machined on a lathe, and the exact desired hyperbolic figure can be obtained to avoid spherical

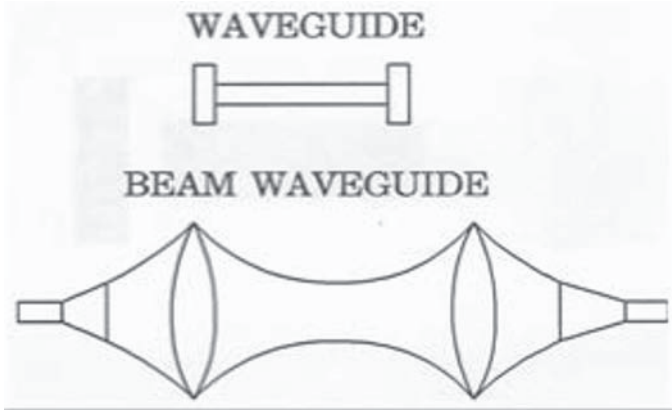


Figure 8. MMW/optical analog—waveguide/beam waveguide.

aberrations. Some plastics have the advantage of good transmission in the visible as well, so that visible images can be used for rough alignment of MMW optical systems. Crystalline quartz and sapphire are very useful materials in these ranges, since they have good transmission from about 20–50 microns to very long wavelengths. These materials can often be anti-reflection coated with plastics such as MylarR, since the indices of refraction of many plastics is roughly equal to the square root of the indices of quartz and sapphire. For example, MylarR, with an index of 1.7, is useful for AR coating sapphire, with an average index of 3.24. Both sapphire and quartz are birefringent, so that they are useful for fabricating wave plates at MMW and THz frequencies. The 225 GHz radar discussed in a later section used a quarter-wave plate to give a circularly-polarized output as part of a polarization duplexer. This radar used several other optical components including lenses and polarizing beam splitters.

The examples given in this section are just a few of the many applications of optical techniques in the MMW and THz bands. It would not be possible to build many systems in these bands without using optical techniques.

6. Millimeter-Wave and Terahertz Systems

The MMW and THz spectral bands are rich in phenomenology including those phenomena related to security applications. Although this region has long been the province of a few specialists seeking to learn more about physics-related problems, in recent years enough progress has been

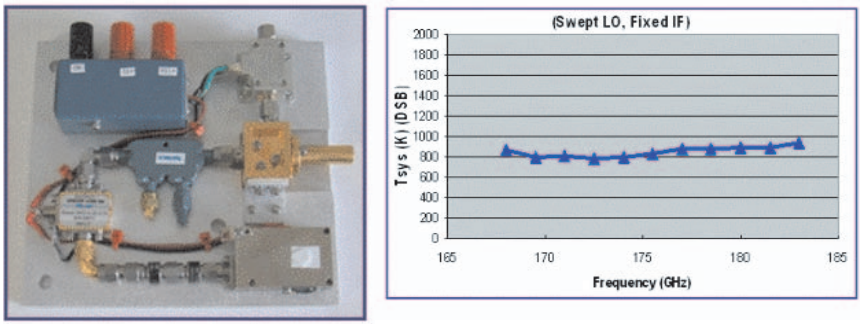


Figure 9. A 183 GHz receiver and its system noise temperature as a function of frequency [15].

made in sources and detectors that useful sensor systems are beginning to be available. In general, this progress has proceeded from the lower frequencies upward using extensions of the lessons learned by developing systems at the microwave and lower MMW frequencies. Recently, more progress has been made in moving downward from the visible/infrared spectral regions through the more extensive use of so-called quasi-optical or diffraction-limited optical techniques [22].

One of the earliest applications of MMW technology to practical needs was in the use of radiometers for atmospheric temperature and water vapor sounding and sea-ice detection. An early radiometer developed for the National Aeronautics and Space Administration (NASA) by Georgia Tech [23] operated in an atmospheric transmission window at 91.5 GHz for sea-ice sensing and on the strong atmospheric water-vapor absorption at 183 GHz for water vapor sounding. This device used a single klystron local oscillator at 91.5 GHz for both receivers. Its output was doubled for the water vapor sensor. A later version of the 183 GHz radiometer used back-to-back point-contact Schottky-barrier diodes in a X2 subharmonic mixer configuration [18]. Figure 9 shows an all solid-state 183 GHz mixer that has a system noise temperature of only 800 K [15]. Figure 10 shows a 94 GHz MMW image obtained from a scanning radiometer that clearly shows two handguns hidden beneath a person's sweater. One of these guns is made primarily of plastic. This latter image illustrates one application of MMW technology to security needs [24].

There are few examples of mass produced MMW radars used for either commercial or military applications. An exception is the US Army's Longbow Apache attack helicopter that is equipped with the Northrop Grumman MMW Longbow radar. The Longbow fire control radar incorporates an integrated radar frequency interferometer for passive location

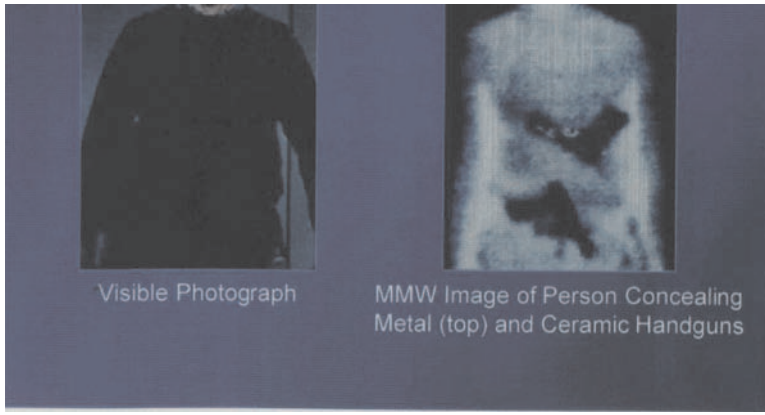


Figure 10. A 94 GHz image of a person concealing two handguns beneath a heavy sweater [24].

and identification of radar emitting threats [25]. The obvious advantages of MMW radar for this application are that the antenna can be made small, and long operating range is not needed.

MMW radars have been designed primarily for remote sensing or one-of-a kind evaluation in military applications. Such systems have been built in the atmospheric windows near 95, 140, and 230 GHz. Figure 11 is a photograph of a radar that operates at 225 GHz that was built for the US Army at Georgia Tech [19]. This radar used some unique techniques in its design, including an all solid-state receiver using an f/4 subharmonic mixer pumped by a 55 GHz Gunn oscillator, a 60 W pulsed extended interaction oscillator (EIO) klystron transmitter, and a quasi-optical polarization duplexer employing a sapphire quarter-wave plate. This radar employed the first all solid-state receiver built at 225 GHz and was the first radar at this frequency to be phase locked. A unique intrapulse feedback phase locking scheme was used to phase lock the EIO transmitter, making this radar the highest frequency microwave coherent radar built up to that time. Figure 12 is a chart showing Doppler returns from a tank measured with this system. The chart clearly shows Doppler returns from the tank body as well as from its treads.

The most common application of MMW systems to security needs is the MMW radiometric imager. Systems of this type are very useful because of their ability to detect a variety of concealed weapons hidden by clothing. Millivision Technologies of South Deerfield, MA has been a leader in this field by developing a family of imagers that operate in the 94 GHz atmospheric window and that use a superheterodyne receiver

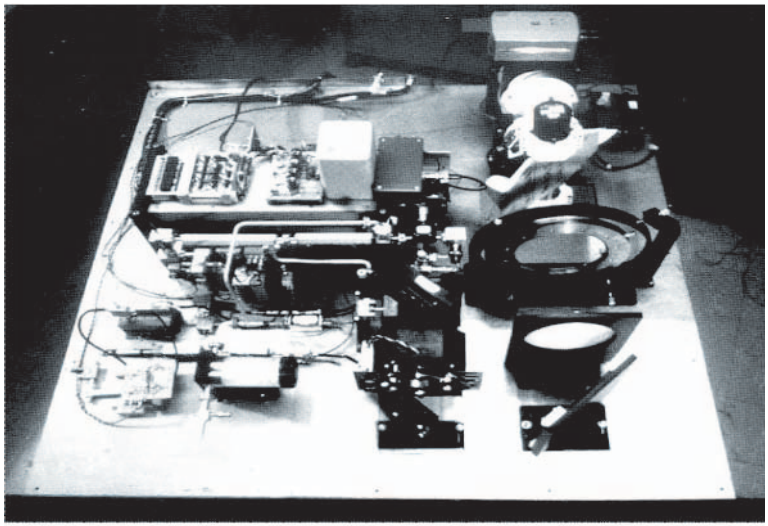


Figure 11. A 225 GHz pulsed coherent radar.

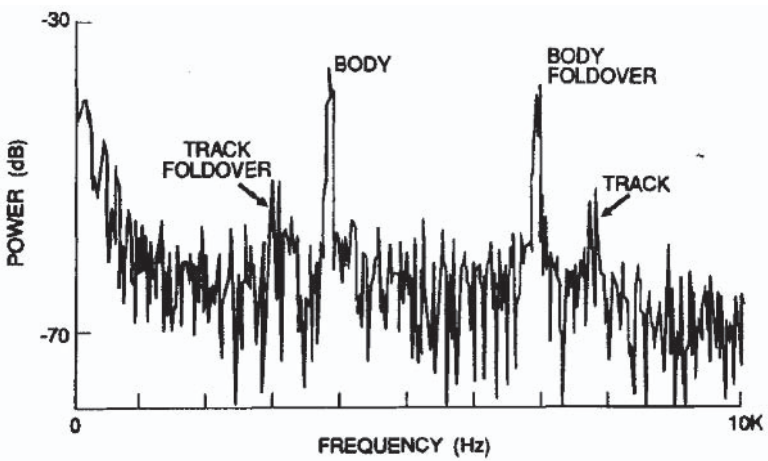


Figure 12. Doppler returns from a tracked vehicle measured by the 225 GHz radar described in the text.

Figure 13. The Millivision Vela 125 MMW imager and its specifications.

for improved sensitivity. A commercial imager developed by Millivision, the Vela 125, has a thermal sensitivity of 1 K at 1 Hz frame rate and 3 K at 10 Hz frame rate [26]. Such imagers operate on the principle that a metal weapon will reflect the temperature of the ambient surroundings, nominally at 300 K, and the human body is a good emitter at 310 K. This contrast of 10 K is easily detected by a radiometer such as the Vela 125 and weapons can be readily recognized if the radiometer has sufficient spatial resolution. Figure 13 is a photograph of the Vela 125 [26].

A MMW imager that uses a unique pupil plane array of sensors has been built by ThermoTrex of San Diego, CA. This device uses a vertical array of slotted waveguide antennas in an aperture of about one square meter. Horizontal resolution is achieved by the vertical array spacing, and vertical resolution by the characteristic of a slotted antenna whereby the sensitivity of the antenna at a given frequency is a function of angle. This sensor also operates at 94 GHz. Figure 14 [27] is an image of a group of persons, one of whom is concealing a weapon beneath clothing, made with this imager. A visible image is also shown.

A 94 GHz MMW imaging radar for aircraft landing operations in poor weather has been designed and built by BAE Systems (formerly Lear Astronics) of Ventura, CA [28]. This instrument has been extensively tested on Air Force airplanes and has been shown to provide useful landing information in weather conditions below normal minimums. This imager uses a scanned range-crossrange radar format with a single detector. It operates on the principle that radiation from the radar in-

Figure 14. Visible and MMW images of a group of persons, one of whom is concealing a simulated handgun [27].

cident on a runway is forward-scattered because of the oblique angle of incidence, while radiation incident on adjacent grassy surfaces is more strongly backscattered to the radar. Smooth surfaces such as runways and taxiways appear to be light-colored in the image, while surrounding areas are dark, thus providing good contrast.

A passive MMW imaging system using monolithic microwave integrated circuit (MMIC) amplifiers on its front end is being developed by Velocium, a subsidiary of Northrop Grumman, formerly TRW, of Redondo Beach, CA [29]. This instrument is also being developed for aircraft landing systems. It uses a single MMIC amplifier operating at a center frequency of 89 GHz with a 10 GHz bandwidth for each pixel of the target scene. The advantage of this imager is that its minimum detectable temperature is determined by the noise figure of the MMIC amplifier and not by the conversion loss of a mixer normally used on the front end of such an imager.

MMW and THz measurement systems have been developed to sense a wide range of features including the MMW turbulence sensing system described in the section on atmospheric limitations [7]. Other applications include the radiometer for sensing sea ice, described above in this section and other sensors designed to detect ice on the space shuttle external tanks. These measurement systems are necessarily limited in wavelength to that range of wavelengths that propagates more or less readily through the atmosphere. A sensor for the remote sensing of wind shear and clear-air turbulence has been proposed by McMillan [30]. This instrument would operate at the center of the group of atmospheric oxygen absorption lines centered at 60 GHz. Several radiometer channels can be processed to detect the temperature of the atmosphere remotely. Since atmospheric hazards are known to be associated with temperature fluctuations, they can be remotely sensed by this method. A similar

method is used to sense the atmospheric water vapor profile from high altitudes or from space. In this case, the sensor operates on the peak of the 183 GHz atmospheric water vapor absorption line.

THz radiometers have primarily been used for astronomical applications as noted above in the section on mixers and detectors. These devices are often based on superconducting mixers and oscillators. During the last few years, there has been some emphasis on the development of semiconductor-based THz radiometers for application in the feature-rich THz bands for remote sensing of chemical and biological agents. The development of these sensors is based on the fact that many chemical and biological agents have features in this spectral range. Efforts are being made to characterize these materials so that the proper THz frequencies can be used for sensing them. Figure 15 [29] shows the spectrum of velocities of HCO+ molecules, measured at 100 GHz, around the central black hole in the galaxy Centaurus A, superimposed on a visible image of the galaxy. These results were obtained by the Australian Radio Telescope using an InP amplifier developed by Velocium of Redondo Beach, CA.

A THz differential absorption radar has been proposed by Elliott Brown and coworkers at the University of California in Los Angeles [31]. This instrument is similar to the differential absorption lidar that has been used successfully to detect atmospheric species. Figure 16 is a block diagram of this system. A band of frequencies is transmitted through a region of interest and reflected back through this region by a retroreflector to a receiver. Differential absorption is measured as the transmitter sweeps through its range of frequencies. Despite the success of differential absorption lidar, differential absorption THz radar will have problems because of low transmitter power and relatively poor sensitivity of the available receivers. Progress is being made in the development of transmitters and receivers, however, as noted in other sections of this chapter.

A system capable of generating THz images would be of great interest for many applications because this spectral band combines the high resolution available from optical imagers with the ability to penetrate many common materials such as clothing and some building materials. An imager that would make use of the extensive uncooled focal plane array technology developed for the infrared bands has been proposed by McMillan, et al. [32]. This imager would operate in the wavelength range greater than 100 microns and would employ the bolometer detectors that have been used successfully in the long-wave infrared (LWIR) range. Among other problems with this approach is that the blackbody spectrum of a 300 K source has very little power beyond 100 microns, so

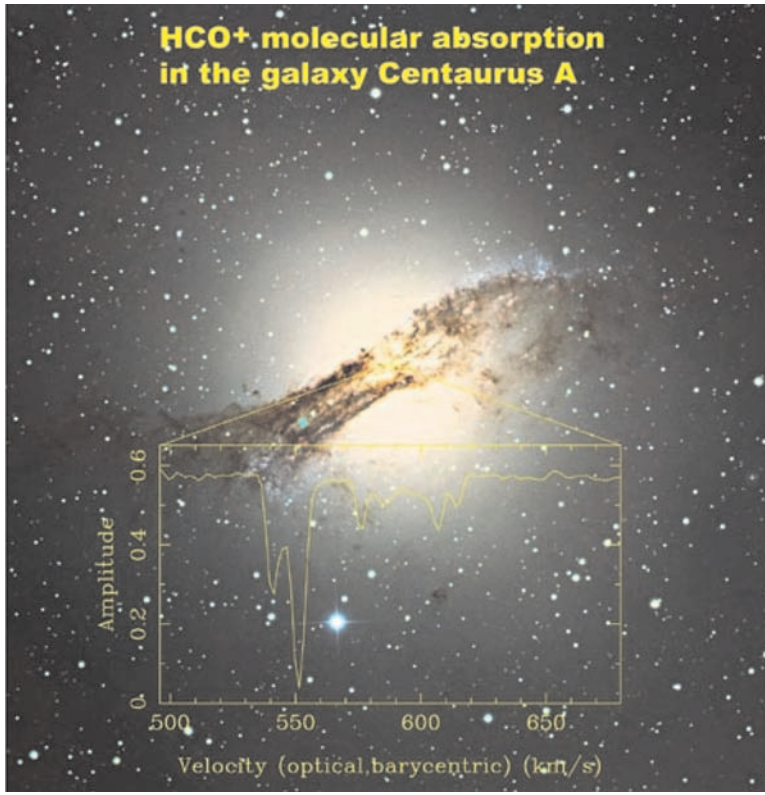


Figure 15. A spectrum indicating the velocity of HCO+ molecules around the central black hole in the galaxy Centaurus A (NGC5128) [29]. Centaurus A is the nearest galaxy containing a supermassive black hole.

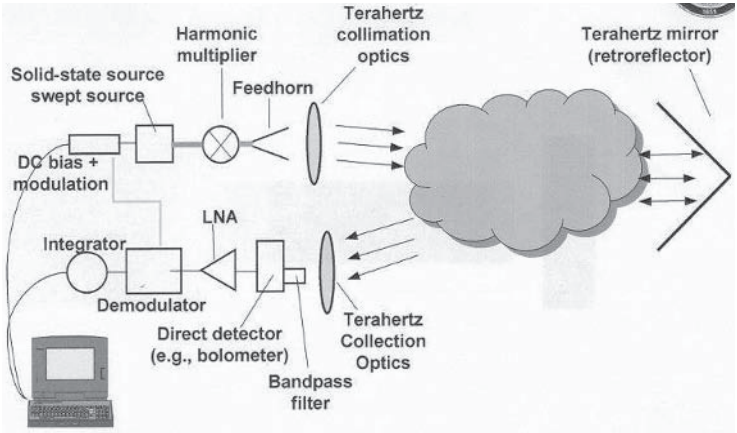


Figure 16. A differential absorption THz radar for detecting atmospheric species [31].

that some artificial means of illumination must be used. Careful analysis has shown that a liquid nitrogen cooled screen is probably the best illumination source, providing about 10 K contrast between the 300 K background and reflection of the 100 K temperature of the screen by metal objects. Another problem is the lack of lens and window materials with adequate transmission in these bands as will be discussed in the next paragraph. One of the first THz images was collected by DeLucia [33]. Figure 17 shows this image, which was collected by a scanned cooled bolometer.

In addition to the limitations imposed by low-power sources and receivers that are not sensitive, significant problems with THz materials exist. One of the most serious is the lack of window and lens materials as noted above. In the visible band, these elements can be made of quartz, glass, or clear plastic. Adequate windows for the IR bands are made of silicon or germanium. Unfortunately, these materials do not transmit well beyond about 20 microns wavelength, although quartz has good transmission beyond 100 microns and into the microwave bands. Figure 18 [34] shows the transmission of crystalline quartz in the 20–200 micron range and Figure 19 [35] shows transmission for roughly the same spectral range for high-resistivity silicon. Note that silicon is a promising material for lens and window fabrication in the far-IR spectrum. Figure 20 shows transmission of some common materials in the millimeter-wave spectrum.

Figure 17. A Terahertz image collected by a scanned cooled bolometer [33].

7. Summary

The utility of the MMW and THz spectral bands in sensing for security applications is limited by two major issues: (1) the atmospheric transmission in these bands limits their usefulness for remote sensing at ranges greater than a few meters except in a few window regions, and (2) adequate sources, detectors, and other components are not available. Paradoxically, the first issue is sometimes an advantage because absorption features in the atmosphere are used for remote sensing from high altitudes and space of water vapor and temperature profiles. For example, the 183 GHz water vapor absorption is used for water vapor profile sensing and the group of oxygen absorptions at 60 GHz is used for temperature sensing. Significant progress is being made in source and detector development driven by the need for these devices in the feature-rich MMW and THz bands. An example of source development is the discovery of a new impact ionization effect in heterojunction field effect transistors that could provide the means for developing higher-power sources in the THz spectrum [Dwight, PC]. Efforts at development of new systems and techniques in this very fruitful area of research will continue. The need for better understanding and exploitation of the MMW and THz bands is great.

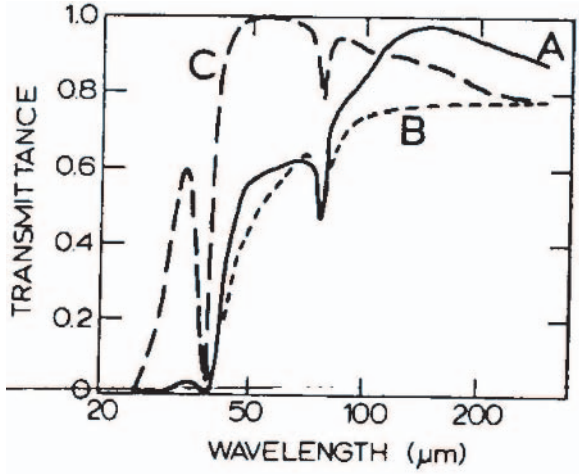


Figure 18. Transmission of crystalline quartz in the range beyond 20 microns. The A and C curves were obtained for anti-reflection coated samples and the B curve for an uncoated sample.

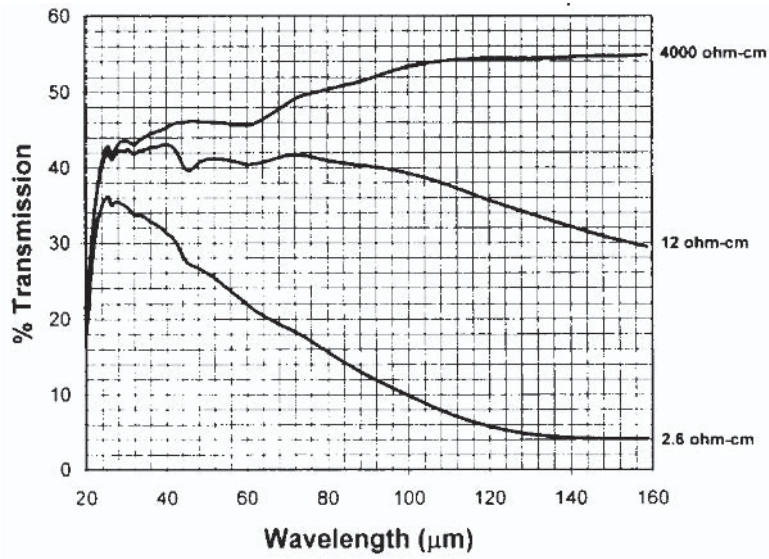


Figure 19. Transmission of high-resistivity silicon in the range beyond 20 microns. The samples are uncoated.

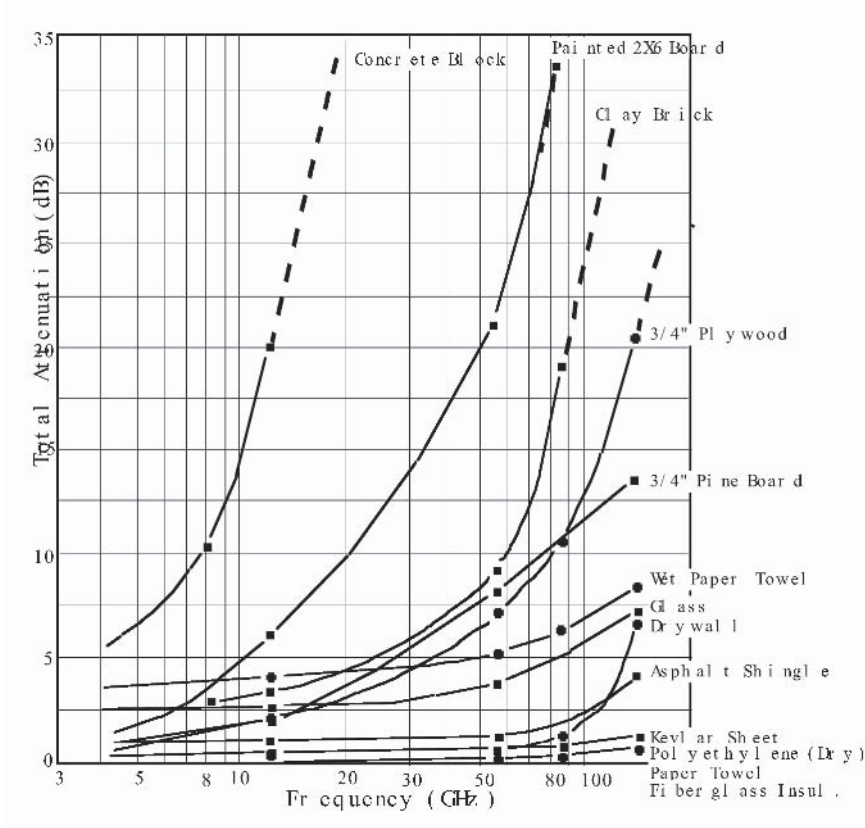


Figure 20. Attenuation of some common materials in the microwave frequency spectrum.

References

[1] K. C. Allen and H. J. Liebe. "Tropospheric Absorption and Dispersion of Millimeter and Submillimeter Waves" *IEEE Trans. Antennas Propagat.*, Vol. 31, pp. 221–223, January 1983.

[2] J. H. Van Vleck and V. F. Weisskopf. "On the Shape of Collision-Broadened Lines", Rev. Mod. Phys., Vol. 17, pp. 227–236, April–July 1945.

[3] H. J. Liebe, T. Manabe, G. A. Hufford. "Millimeter-Wave Attenuation and Delay Rates Due to Fog/Cloud Conditions", IEEE Trans. Antennas Propagat., Vol. 37, pp. 1617–1623, December 1989.

[4] H. J. Liebe and D. H. Layton. NTIA Report 87–224 National Telecommunications and Information Administration, Boulder, CO, 1987.

[5] R. S. Lawrence and J. W. Strohbehn. "A Survey of Clear-Air Propagation Effects Relevant to Optical Communications", Proc. IEEE, Vol. 58, pp. 1523–1545, 1970.

[6] R. W. McMillan, R. A. Bohlander, G. R. Ochs, R. J. Hill, S. F. Clifford. "Millimeter Wave Atmospheric Turbulence Measurements: Preliminary Results and Instrumentation for Future Measurements", Optical Engineering, Vol. 22, No. 1, pp. 32–39, January/February 1983.

[7] R. J. Hill, R. A. Bohlander, S. F. Clifford, R. W. McMillan, J. T. Priestley, W. P. Schoenfeld. "Turbulence-Induced Millimeter-Wave Scintillation Compared with Micrometeorological Measurements", IEEE Trans. Geosciences and Remote Sensing, Vol. 26, No. 3, pp. 330–342, May 1988.

[8] G.F. Brand. "Development and Applications of Frequency Tunable, Submillimeter Wave Gyrotrons", Int. J. Infrared and Millimeter Waves Vol. 16, pp. 879–887, 1995.

[9] A.A. Tolkachev, B. A. Levitan, G. K. Solovjev, V. V. Veytsel, V. E. Farber. "A Megawatt Power Millimeter-Wave Phased-Array Radar", IEEE AES Systems Magazine, pp. 25–31, July 2000.

[10] A. F. Krupnov, M. Yu. Tretyakov, Yu A. Dryagin, and S. A. Volokhov. "Extension of the Range of Microwave Spectroscopy up to 1.3 THz", J. Mol. Spetrosc., Vol. 170, 279–284 1995.

[11] V. L. Vaks, V. V. Khodos, and E. V. Spivak. "A Nonstationary Microwave Spectrometer", Review of Scientific Instruments, Vol. 70, Issue 8, pp. 3447–3453, August 1999.

[12] A. F. Krupnov and A. V. Burenin. "New Methods in Submillimeter Microwave Spectroscopy", Mol. Spectrosc: Mod. Research II, K. Narahar: Rao, ed. Academic Press, New York. (1976).

[13] Introduction to Extended Interaction Oscillators, Data Sheet 3445 5M 11/75, Varian Associates of Canada, Ltd. (Now CPI), Georgetown, Ontario, Canada, 1975.

[14] E. Alekseev and D. Pavlidis. "GaN-Based Gunn Diodes: Their Frequency and Power Performance and Experimental Considerations" www.eecs.umich.edu.

[15] www.virginiadiodes.com

[16] J. W. Dees. "Detection and Harmonic Generation in the Sub-Millimeter Wavelength Region", Microwave J., Vol. 9, pp. 48–55, 1966.

[17] K. M. Evenson, J. S. Wells, F. R. Petersen, B. L. Danielson, G. W. Day, R. L. Barger, and J. L. Hall, "Speed of Light from Direct Frequency and Wavelength Measurements of the Methane-Stabilized Laser", Phys. Rev. Lett. Vol. 29, 1346–1349 1972.

[18] R. E. Forsythe, V. T. Brady, and G. T. Wrixon. "Development of a 183 GHz Subharmonic Mixer", IEEE MTT-S International Microwave Symposium Digest, Orlando, FL, May 1978.

[19] R. W. McMillan, C. W. Trussell, Jr., R. A. Bohlander, J. C. Butterworth, R. E. Forsythe. "An Experimental 225 GHz Pulsed Coherent Radar", IEEE Trans. Microwave Theory and Techniques, Vol. 39, No. 3, pp. 555–562, March 1991.

[20] A. Vystavkin, D. Chouvaev, T. Claeson, D. Golubev, V., N. Kardashev, A., V. Kurt, L., M. Tarasov, A. Trubnikov, M. Willander. "Terahertz Andreev Reflection Based Normal Metal Hot-Electron Bolometer for the Cryogenic Telescope of the International Space Station", The 10th International Symposium on Space Terahertz Technology, Proceedings, pp 372–389, University of Virginia, March 16–18, 1999.

[21] "U3000 Uncooled Microbolometer Infrared Sensor", Data Sheet. The Boeing Company, Anaheim, CA, 1998.

[22] P. F. Goldsmith. Quasioptical Systems, IEEE Press, New York, 1997.

[23] J. M. Schuchardt, J. A. Stratigos, J. A. Gagliano, D. O. Gallentine, J. L. King. "Dual Frequency Multi-Channel Millimeter Wave Radiometers for High Altitude Observation of Atmospheric Water Vapor", 1979 MTT-S International Microwave Symposium Digest, pp. 540–542.

[24] P. F. Goldsmith, C.-T. Hsieh, G. R. Huguenin, J. Kapitzky, and E. L. Moore. "Focal Plane Imaging Systems for Millimeter Wavelengths", IEEE Trans. Microwave Theory and Techniques, Vol. 41, pp. 1664–1675, October 1993.

[25] www.army-technology.com/projects/apache/

[26] www.millivision.com

[27] Private Communication, Thermotrex Corporation, San Diego, CA.

[28] www.sae.org/aeromag/techupdate 12–99/05.htm

[29] http://www.st.northropgrumman.com/velocium/

[30] R. W. McMillan. "A Horizontal Atmospheric Temperature Sounder: Applications to Remote Sensing of Atmospheric Hazards", Int. J. Infrared and Millimeter Waves, Vol. 14, No. 5, pp. 931–948, 1993.

[31] E. Brown. University of California at Los Angeles, Private Communication, 2002.

[32] R. W. McMillan, Osborne Milton, Jr., M. C. Hetzler, R. S. Hyde, W. R. Owens. "Detection of Concealed Weapons Using Far-Infrared Bolometer Arrays", Proceedings of the 25th International Conference on Infrared and Millimeter Waves, Beijing, China, 12–15 September 2000.

[33] F. C. DeLucia. Ohio State University, Private Communication, 2004.

[34] K. R. Armstrong and F. J. Low. "Far-Infrared Filters Utilizing Small Particle Scattering and Antireflection Coatings", Applied Optics, Vol. 13, No. 2, pp. 425–430, February 1974.

[35] J. E. Peters and P. D. Ownby. "Far Infrared Transmission of Diamond Structure Semiconductor Single Crystals-Silicon and Germanium", Optical Engineering, Vol. 38, No. 11, pp. 1924–1931, November 1999.

SENSOR MANAGEMENT FOR RADAR: A TUTORIAL

Bill Moran
The University of Melbourne, Parkville, Vic 3010, Australia
and
Prometheus Inc. 21 Arnold Ave, Newport, RI 02840, USA

Sofia Suvorova
The University of Melbourne, Parkville, Vic 3010, Australia

Stephen Howard
Electronic Warfare and Radar Division
Defence Science and Technology Organisation
Edinburgh, SA5100,Australia

Abstract In this chapter we describe some of the ideas being pursued in sensor scheduling as they apply to radar. A modern phased-array pulse-Doppler radar has several different parameters available for scheduling: waveform, beam-shape, beam direction, pulse repetition interval, etc. Choice of different values for these parameters provides different transmit modes for the radar and these modes in turn provide a variety of "blurrings" of the image of the scene. The application of ideas in scheduling to the different possible modes of the transmit phase of such a radar, has been shown in simulation to improve many aspects of the performance in tracking and detection of targets. We give a quick introduction to the ideas of radar followed by a discussion of some of the theoretical ideas involved, and with results of some simulations. We end with a discussion of the theoretical problem of scheduling the measurements and tracking of a one-dimensional system.

Keywords: Radar; sensor scheduling; waveform; beam-shape; control; detection; tracking; revisit time; myopic; non-myopic.

1. Introduction

A radar system is a way of viewing a scene using electromagnetic radiation at wavelengths that can be processed using electronic equip-

Jim Byrnes and Gerald Ostheimer (eds.),
Advances in Sensing with Security Applications, 269–291.

ment. Since ambient radiation at these wavelengths tends to be low in power, typically radars provide the illumination as well as the viewing system. The control of the source of radiation leads to major advantages, as well as some disadvantages. The most important disadvantage is that the amount of illumination is limited. Most radar systems in use are *monostatic*; that is, their illumination source and receiver are collocated. This has the advantage of shared electronics and antennas. Much effort is currently going into *multistatic* radar systems, but in this chapter we will focus only on monostatic radars. For such radars the energy returning to the receiver from a scatterer is inversely proportional to the 4th power of the distance. This means that, to achieve significant range, radars have to rely on a mix of high transmission power, clever ideas in the use of waveforms, sophisticated antenna design to focus the energy, and high performance signal processing.

Our aim in this chapter is to describe ideas being explored for the control of radar systems. Since we are not assuming any expertise in radar, we begin with a short description of the ideas of radar theory. This description focuses on the most commonly used form of the technology, namely a pulse-Doppler radar system. After that we dicuss some of the basic ideas in sensor management and then give results of simulations that show the kind of improvement that the use of sensor scheduling might produce. We have focused on work we have been associated with, and have omitted much excellent work of other workers in this burgeoning subject. Finally we discuss a theoretical problem in sensor management.

2. Radar Fundamentals

In this first section we discuss the basic ideas in a pulse-Doppler radar system. Our treatment is brief and focuses on the underlying theory rather than on the important issues of implementation.

2.1 Ambiguity and Radar

Illumination of the scene is provided by a signal that is emitted from the radar system. This signal is usually a waveform that is relatively slowly varying superimposed on a rapidly oscillating sinusoidal carrier. Thus it can be expressed as

$$\mathbf{s}(t) = \mathbf{w}(t).\cos(2\pi f_c t). \tag{1}$$

Here $\mathbf{w}(t)$ is the slowly varying waveform, and f_c is the carrier frequency. It is important to make the rather obvious observation at this stage that all signals transmitted and received are real-valued. However, it is

possible to represent complex waveforms in such a way that they can be transmitted. Thus for a complex waveform $\mathbf{w}(t)$ we transmit the signal

$$\mathbf{s}(t) = (\Re\mathbf{w}(t)).\cos(2\pi f_c t) - (\Im\mathbf{w}(t)).\sin(2\pi f_c t). \tag{2}$$

On return, the "in-phase" or I component can be separated from the "quadrature" or Q component by demodulation against $\cos(2\pi(f_c t))$ and $\sin(2\pi f_c t))$ respectively. Much of the theory of radar processing takes place in the complex domain. It is convenient, and a powerful theoretical device, to replace the signal (2) by its complex version:

$$\mathbf{s}_c(t) = \mathbf{w}(t)).\exp(2\pi i f_c t), \tag{3}$$

so that $\mathbf{s}(t) = \Re\big(\mathbf{s}_c(t)\big)$. The carrier is often in the range 1–30GHz. The waveform will typically occupy a bandwidth that is less than 1/10 of that.

The superposition principle allows us to assume just a single scatterer in the view of the radar. The transmitted signal hits this scatterer whose distance (we measure distance and time in the same units) from the (collocated) transmitter and receiver is r. Assume that the scatterer is stationary. The return signal will be a delayed version of the original, delayed by the total round trip time from the radar to the scatterer. Specifically the signal voltage at the antenna of the receiver is

$$\mathbf{s}_u(t) = A\mathbf{s}(t - 2r) \tag{4}$$

where A represents the overall attenuation and includes a phase change (so is complex) due to reflection.

In the receiver some noise is added ("receiver noise"), arising from thermal activity generated within the components of the receiver. For distant scatterers the return signal is often so weak that this thermal noise can become a significant issue. We write

$$\mathbf{s}_r(t) = \mathbf{s}_u(t) + N(t),$$

where $N(t)$ is a white Gaussian process, for the signal after the initial stages of the receiver. Thermal noise is to a good approximation white and Gaussian.

Now we consider the possibility that the target is moving relative to the radar. The scattered waveform is modified by the Doppler effect. If this is done correctly it results in a "time dilation" of the return signal, so that, if the target has a radial velocity v, the return signal $\mathbf{s}_u(t)$ becomes

$$\mathbf{s}_u(t) = A\mathbf{s}(\alpha t - 2r),$$

where

$$\alpha = \frac{(1 - \frac{v}{c})}{(1 + \frac{v}{c})}.$$

When v is much smaller than c this is approximated by $\alpha = (1 - 2v/c)$. A further approximation is possible if, as is usually the case, the signal is "narrow band"; that is, if its (Fourier) spectrum is essentially in a range $(f_c - \delta, f_c + \delta)$ and its reflection in the origin, where δ is small compared to f_c. For most radar applications, this is a reasonable assumption since the signal modulating the carrier will have relatively low bandwidth. In this case, the return signal is approximated by shifting the frequency of the return from a stationary target at the same range by $f_d = (2v/c)f_c$, the so-called "Doppler frequency". This is best written in terms of the complex signal

$$\mathbf{s}_u(t) = \Re\big(\mathbf{w}(t - \frac{2R}{c}).e^{2\pi i f_c(1-2v/c)(t-\frac{2R}{c})}\big) \tag{5}$$

This equation is the standard one used in most radar calculations.

When the return is received, it is demodulated to strip off the carrier frequency. Typically, the return is "mixed with", that is multiplied by, $\cos 2\pi f t$ and then low-pass filtered to eliminate the high frequency component of the mixed signal. This is the demodulation phase refered to earlier.

In the complex domain, the demodulated signal is as described in (5). The signal is then filtered against another chosen signal $\mathbf{v}(t)$, often $\mathbf{v}$ is chosen to be the same as $\mathbf{w}$ (*match-filtering*); that is, it is correlated with that signal, resulting in

$$A\mathbf{w}, \mathbf{v}(x, f) = \int_{\mathbf{R}} \mathbf{v}(t)^* \mathbf{w}(t - x) e^{2\pi i f t}\, dt, \tag{6}$$

after a slight change of variable.

A general scene may be regarded as a function of range and Doppler, corresponding to a "reflectivity" assignment $\rho(t, f)$ to each value of range and Doppler. We include in this description of the scene the attenuation due to range of the scatterer. The superposition principle says that the resulting return is a convolution in range and Dopper of the scene with the ambiguity:

$$R(\tau, f) = \iint_{\mathbf{R}^2} \rho(\tau', f') A_{\mathbf{w},\mathbf{v}}(\tau - \tau', f - f')\, d\tau' df' \tag{7}$$

By varying the waveform, we are able to vary the shape of the ambiguity and thereby the kind of blurring that the radar process does to

the scene. Evidently it would be best if there were no blurring, that is, if the ambiguity were a "thumbtack" with a spike at the origin and zero elsewhere. Unfortunately, there is a fundamental limitation that prevents this. It is known in various forms, in particular, as (one form of) the Heisenberg Uncertainty Principle, and as Moyal's Identity. In the latter formulation, it is expressed as follows:

$$||A_{\mathbf{w},\mathbf{v}}||_{L^2(\mathbf{R}^2)} = ||\mathbf{w}||_{L^2(\mathbf{R})} \cdot ||\mathbf{v}||_{L^2(\mathbf{R})} \tag{8}$$

It states that the L^2 norm of the ambiguity function as a function on $\mathbf{R}^2$ is the product of the L^2 norms of the transmit signal and the filtering signal as functions on $\mathbf{R}$. Since signals have finite energy, the ambiguity must be an L^2 function, and have a lower bound on its L^2 norm. Accordingly a "thumbtack" is impossible. The range-Doppler must be "blurred" by the imaging process in radar.

2.2 Beam-forming

In addition to finding range and Doppler, a radar usually needs to estimate the direction of a target. This is done by pointing the illumination in particular directions and "filtering" the return according to which direction it comes from.

The classical way to form a beam in radar is to use a paraboloidal dish. The beam is pointed in a given direction by mechanically steering the dish. Both the transmit and return beams are "spatially filtered" by the dish. Returns from particular directions are emphasized and those from other directions are attenuated. More and more this approach is being replaced by an electronically steered array antenna. Typically, this is comprised of a multiplicity of small antenna elements to which the transmit signal is fed. By varying the phase of the signal across the array it is possible to steer the direction of the beam, and by varying the voltage applied to each element it is possible to reshape the beam. The direction and the shape of the transmit beam can be varied rapidly. This is particularly important in a situation where the radar is performing multiple functions such as tracking several targets while detecting new targets. As a receive antenna, such a system can simultaneously steer many beams by means of the processing of the returns at each antenna element.

In neither the mechanical nor the electronic approaches is the beam perfectly sharp. This is inevitable since the aperture of the system is finite in extent. In the case of the electronic array, this problem is compounded by the fact that the array has discrete elements, rather than a continuum. However, in the latter case it is controllable. As a result of this imperfection, again the scene is "blurred"; in this case

the directions of the scatterers are averaged over the response of the antenna. In the case of an electronic array, it is possible to change the "blurring" as well as beam-direction quickly. Thus in a phased-array system there is scope for the control of the illumination.

2.3 Doppler Processing and Pulse Compression

One way of copying with the ambiguity trade-off problem forced by Moyal's Identity (8) is to use a technique called *Doppler processing*. There are several issues associated with the accurate measurement of range and Doppler:

- A short pulse gives more accurate range measurement;
- A longer pulse has more energy in it, and the more energy used in illumination the more will be scattered back;
- The effect of the Doppler of typical targets on short pulses is essentially trivial.

An imperfect solution to the problems arising from the contradictory (to Moyal's Identity) requirements of good range and Doppler measurement is adopted by a *pulse-Doppler radar*. The solution involves the following mechanisms:

DP-1) Pulses of a length short enough to incur relatively little Doppler effect but long enough to individually give relatively high energy on target are chosen;

DP-2) These pulses are chosen in such a way that their auto-correlations are close to a spike with small side-lobes;

DP-3) A number of such pulses are transmitted with long gaps between them to give time for the Doppler to have effect across the whole sequence of pulses.

The effect of DP-2) is to produce a virtual pulse whose length is the width of the central lobe. Of course, this is never completely perfect since it does have side-lobes, but waveforms have been described for which the performance in this respect is excellent. DP-3) means that the Doppler frequency shift is being sampled at a discrete set of time points. If the sampling rate is faster than the Nyquist of the Doppler frequency shift, then the Doppler can be unambiguously extracted.

One might ask why Moyal's Identity does not cause problems here. Of course it does. Whatever the sampling rate, there are Doppler frequencies that are ambiguous and correspond to side-lobes in the overall

ambiguity of the series of pulses. It is important to choose the sampling rate to be high enough that this does not happen for targets of interest. On the other hand, if the sampling rate is high then returns of earlier pulses from distant targets can appear after later pulses have been transmitted. This *range-aliasing* also corresponds to side-lobes in the overall ambiguity. Thus Doppler processing also suffers the same problems as a single waveform. However, it provides a mechanism for control of the position of the side-lobes to best fit the context. Moreover, it is possible to view the sampling rate, as well as the number of pulses used in this processing, as control parameters in scheduling a sensor.

3. Sensor Management — Overview

Conventional radars typically employ the same waveform and beam-pattern over many pulses. The received signal can be, and often is, processed in several ways to extract different kinds of information, or in response to knowledge gained from the environment, but on the transmit side, the mode of operation of the radar system is essentially static. In these systems it may be possible to modify the waveform used offline but not during the processing period. Recent advances in hardware have made the possibility of changing transmit modes, and indeed most parameters quickly; if not between pulses then at least on a scale of a few tens of pulses. Moreover, as in the case of the receive-side adaptivity, these modifications can take into account the knowledge of the environment gained about the scene.

The key features of a managed sensor system are that it senses the environment and chooses an appropriate waveform, beam-pattern, pulse repetition interval (PRI), etc (collectively called the *sensor mode*) to best extract the required information. Any such system must have, at least, the following components in addition to the basic sensor and ancillary components:

SM-1) A method of estimating the current (that is at the time of transmission of next pulse) state of the environment. This is done on the basis of prior measurements together with some model of the dynamics of the environment. It may be important to estimate not only the scatterers of interest (*targets*) but also those that are not of interest (*clutter*), since knowledge of the latter may be useful for determination of an optimal radar mode.

SM-2) A measure of effectiveness of each potential sensor mode. This should be a function of both the mode (as defined above) and of the environment, or at least the estimate of it mentioned in

SM-1). Most importantly, it should be based on the operational problem at hand.

SM-3) A library of modes from which the optimal mode is chosen. This might be just a finite library, but also might be an infinite parameterized family of, say, waveforms.

SM-4) A method for finding the optimal choice of mode over one or more epochs, based on the measure of effectiveness.

We note that, at its simplest, the optimization will be on an epoch by epoch basis (the so-called "greedy" or "myopic" approach). In this case, the mode is chosen just to optimize for the next epoch and defer consideration of future behavior. A more sophisticated system would look several epochs ahead in applying the measure of effectiveness, though it would also update the scheduling policy on an epoch by epoch basis. Such an approach is, *a priori*, very computer intensive, and much work is needed to develop shortcuts to calculation of the optimal policy. Sometimes it may be appropriate to choose to measure the effectiveness of a policy only at the last epoch of application of that policy.

It should be noted that this regime allows the possibility that the sensor is spread over several platforms and/or is comprised of several physically different sensors within each platform. It can encompass trajectory control for platforms and even control of data rates in connecting platforms to each other and to a central node. In each case the system can be viewed as consisting of many real or virtual sensors, where a virtual sensor can be a particular mode of a sensor, a position of a platform, a particular bit of a measurement made by a sensor, etc. Thus the sensor management problem may seen in all of these cases as one of choosing to switch between many different sensors, where the choice is made on the basis on knowledge of the environment. This view is schematically represented in Figure 1.

The ultimate goal of research in this area is to "close the loop" in radar signal processing by producing algorithms for scheduling of beam-directions, beam-shapes, waveforms and other radar modalities so as to optimally extract information from the environment (targets and clutter). Several sub-objectives contribute to this. As we have already said, in order to choose the best modality for a given radar environment, an estimate of that environment needs to be available at the time of making the selection, a method of assessing the effectiveness of a given modality in a given environment is required, as well as an optimal scheduling algorithm to make the selection of an optimal modality for each of a number of future epochs. Because of space constraints, we limit our discussion to

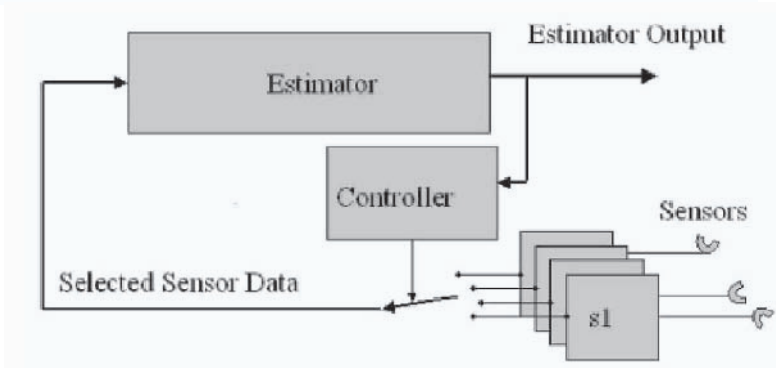


Figure 1. Schematic of Sensor Management.

simulations for just one- and two-step ahead scheduling. Before proceeding to the simulation work, we discuss the theory of waveform libraries. The choice of the library of modes between which the sensor can switch is, of course, an important consideration in the development of scheduled radar systems.

4. Theory of Waveform Libraries

With the advent of radars capable of waveform agility, the design of optimal waveform libraries comes into question. The purpose of this section is to consider the design of such waveform libraries for radar tracking applications, from an information theoretic point of view. We note that waveform libraries will depend in general on the specific applications in which the systems are to be used. Airborne radars will require different libraries from ship-borne ones. Radars used in a tracking mode will require different optimal libraries than radars in a surveillance mode.

The idea of selecting waveforms adaptively based on tracking considerations was introduced in the papers of Kershaw and Evans [3, 4]. There they used a cost function based on the predicted track error covariance matrix.

In designing or improving a waveform library certain questions arise. Firstly it is important to establish the measure of effectiveness (MoE) for individual waveforms (cost function) and then to extend this to an MoE for the library. If a particular set of waveforms is added, will this improve the library in these terms and, on the other hand, how much will removing some waveforms reduce the utility of the library? It is the purpose of this chapter to develop an information theoretic framework

for addressing such questions, at least from the target tracking point of view and to look at its application to specific waveform collections.

We use the basic sensor model proposed in [4]. While this has limitations, it is simple and therefore useful as a starting point for discussion of the problem. In this model, the sensor is characterized by a measurement noise covariance matrix which is waveform dependent

$$\mathbf{R}_\phi = \mathbf{T}^T \mathbf{J}_\phi^{-1} \mathbf{T}, \tag{9}$$

where $\mathbf{J}_\phi$ is the Fisher information matrix corresponding to the measurement using waveform $\phi \in L^2(\mathbf{R})$, and $\mathbf{T}$ is the transformation matrix between the time delay and Doppler measured by the receiver and the target range and velocity. The Fisher information matrix is given by an expression involving the normalized second order time and frequency moments of the waveform ϕ. It is also expressible in terms of the Hessian of the squared absolute value of the ambiguity function of the waveform at the origin of the range-Doppler plane. This calculation is done in [6].

It should be pointed out that the use of the Fisher matrix here is an approximation. It really corresponds to the Cramér-Rao lower bound on the estimator for the target from this measurement. It can be shown that the estimator here is *asymptotically efficient* (see[2], pp. 38–39) in that the covariance matrix approaches the Cramér-Rao lower bound over a large number of measurements (*loc. cit.*).

We note that the Hessian equivalence means that the Fisher matrix expresses purely local information about the ambiguity function at its peak. It says nothing about the structure of the ambiguity away from that peak. This local nature of the Fisher matrix is of some concern when considering its use in expressing a measure of effectiveness for a waveform. It can be argued, however, that this is a reasonable approach for tracking (where the return is "gated" in the vicinity of the predicted target position and Doppler) and in relatively low clutter situations. In a detection problem in a highly cluttered environment, the side-lobes will play a significant role and alternative measures of effectiveness ought to be considered.

In the context of our discussion in this chapter, we represent the measurement obtained using the waveform ϕ as a Gaussian measurement with covariance $\mathbf{R}_\phi$. The current state of the system is represented by the state covariance matrix $\mathbf{P}$. Of course, the estimated position and velocity of the target is also important for the tracking function of the radar, but in this context they play no role in the choice of waveforms. In a clutter rich (and varying) scenario, the estimate of the target parameters will clearly play a more important role. The *expected information* obtained from a measurement with such a waveform, given the current state of

knowledge of the target, is

$$I(X;Y) = \log\det(\mathbf{I} + \mathbf{R}_\phi^{-1}\mathbf{P}). \tag{10}$$

This is the mutual information between the target variable (range and Doppler) X and the processed (with a matched filter) radar return Y, resulting from the use of the waveform ϕ. $\mathbf{I}$ is the identity matrix. We use this expected information as the MoE of the waveform ϕ in this context. The more information we extract from the situation the better.

We assume a knowledge of the possible state covariances P generated by the tracking system. This knowledge is statistical and is represented by a probability distribution $F(\mathbf{P})$ over the space of all positive definite matrices.

We define the *utility* of a waveform library $\mathcal{L} \subset L^2(\mathbf{R})$, with respect to a distribution F, to be

$$G_F(\mathcal{L}) = \int_{\mathbf{P}>0} \max_{\phi\in\mathcal{L}} \log\det(\mathbf{I} + \mathbf{R}_\phi^{-1}\mathbf{P})\, dF(\mathbf{P}). \tag{11}$$

Thus we have assumed that the optimal waveform is chosen in accordance with the MoE defined in equation (10) and have averaged this over all possible current states, as represented by the covariance matrices $\mathbf{P}$ and in accordance with their distribution $F(\mathbf{P})$.

We consider two libraries $\mathcal{L}$ and $\mathcal{L}'$ to be *weakly equivalent*, with respect to the distribution F, if $G_F(\mathcal{L}) = G_F(\mathcal{L}')$, and *strongly equivalent* if $G_F(\mathcal{L}) = G_F(\mathcal{L}')$ for all F.

In what follows we will work in receiver coordinates, i.e., treat $\mathbf{T}$ above as $\mathbf{I}$. This amounts to a change in parameterization of the positive definite matrices in the integral in (11).

Having defined the utility of a waveform library we go on to investigate the utilities of a few libraries. Specifically, we consider libraries generated from a fixed waveform ϕ_0, usually an unmodulated pulse of some fixed duration, by *symplectic transformations*. Such transformations form a group of unitary transformations on $L^2(\mathbf{R})$ and include linear frequency modulation as well as the Fractional Fourier transform (FrFT) in a sense that we shall make clear.

Under such transformations $\phi = \mathbf{U}\phi_0$, the ambiguity function of the waveform ϕ_0, is modified according to the following equation.

$$|A_\phi(\mathbf{x})| = |A_{\phi_0}(\mathbf{S}^{-1}\mathbf{x})| \tag{12}$$

where $\mathbf{x} = (t, f)^T$ and $\det(\mathbf{S}) = 1$, and ϕ_0 ranges over all members of $L^2(\mathbf{R})$. Indeed, a reasonable definition of *symplectic transformation* in this context is any unitary operator on $L^2(\mathbf{R})$ that transforms the

ambiguity function according to equation (12). There is a technical problem here that requires resolution. A waveform is *not* determined by the absolute value of its ambiguity. Thus there may be more than one transformation $\mathbf{S}$ under which equation (12) is valid. It turns out that in this case the the transformation is unique.

It is not hard to see that such transformations form a group. Suppose that U_1 and U_2 are symplectic in this sense and S_1 and S_2 correspond to them. Then

$$|A_{U_1U_2\phi_0}(\mathbf{x})| = |A_{U_2\phi_0}(\mathbf{S}_1^{-1}\mathbf{x})| = |A_{\phi_0}(\mathbf{S}_2^{-1}\mathbf{S}_1^{-1}\mathbf{x})| = |A_{\phi_0}((\mathbf{S}_1\mathbf{S}_2)^{-1}\mathbf{x})|. \tag{13}$$

Furthermore, under symplectic transformations, it is relatively easy to show, using the Hessian formula for calculating the Fisher information matrix, that the measurement covariance matrix transforms as

$$\mathbf{R}_{U\phi_0} = \mathbf{S}^T\mathbf{R}_{\phi_0}\mathbf{S} \tag{14}$$

when S is associated with U.

An LFM ("chirp") waveform library consists of

$$\mathcal{L}_{\text{chirp}} = \{\exp(i\lambda\mathbf{t}^2/2)\phi_0 \mid \lambda_{\min} \le \lambda \le \lambda_{\max}\} \tag{15}$$

where ϕ_0 is an unmodulated pulse, $\lambda_{\min}$ and λ_{max} are the minimum and maximum chirp rates supported by the radar, and $\mathbf{t}$ is the (unbounded) operator on $L^2(\mathbf{R})$ defined by

$$\mathbf{t}\phi(t) = t\phi(t). \tag{16}$$

It follows that

$$(\exp(i\lambda\mathbf{t}^2/2)\phi)(t) = \exp(i\lambda t^2/2)\phi(t). \tag{17}$$

For this library the corresponding measurement covariance matrices are given by (14) with

$$\mathbf{S}(\lambda) = \begin{pmatrix} 1 & 0 \\ \lambda & 1 \end{pmatrix}. \tag{18}$$

It is relatively easy to see that

$$\mathcal{L}'_{\text{chirp}} = \{\exp(i\lambda_{\min}\mathbf{t}^2/2)\phi_0, \exp(i\lambda_{\max}\mathbf{t}^2/2)\phi_0\} \tag{19}$$

is strongly equivalent to $\mathcal{L}_{\text{chirp}}$. That is, we do just as well if we keep only the LFMs with the minimum and maximum rates. In range-Doppler coordinates, the error covariance matrix for each LFM can be represented by

$$R(\lambda) = \mathbf{S}(\lambda)^T R_0 \mathbf{S} S(\lambda), \tag{20}$$

where R_0 is a diagonal matrix with ρ_1, ρ_2 on the diagonal; that is, a covariance matrix for the rectangular pulse [1, 4]. Direct computations give the following expression for the mutual information $I(X;Y)$:

$$I(X;Y) = 4\frac{P_{11}}{\rho_2}\frac{\lambda^2}{4} - 4\frac{P_{12}}{\rho_2}\frac{\lambda}{2} + \frac{|P|}{|R|} + 1 + \frac{P_{11}}{\rho_1} + \frac{P_{22}}{\rho_2}. \tag{21}$$

This is a quadratic in λ with positive second derivative since P and R are both positive definite, and therefore achieves its maximum at the end points, i.e at maximum or minimum allowed sweep rate.

Another way to create a waveform library is to take an ambiguity and rotate it. In this case, the new waveform is a fractional Fourier transform of the old one.

$$\mathcal{L}_{\mathrm{FrFT}} = \{\exp(i\theta(\mathbf{t}^2 + \mathbf{f}^2)/2)\phi_0 \mid \theta \in \Theta\}, \tag{22}$$

where the set $\Theta \subset [0, 2\pi]$ can be chosen so as not to violate the bandwidth constraints of the radar, and $\mathbf{f}$ is the operator on $L^2(\mathbf{R})$ defined by

$$\mathbf{f}\phi(t) = i\phi'(t). \tag{23}$$

For this library the corresponding transformation in range-Doppler space is given by the rotation

$$\mathbf{R}(\theta) = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}. \tag{24}$$

It is possible to consider combinations of the rotation and chirping transformations applied to an unmodulated waveform ϕ_0; that is, we consider all transformations of the following form:

$$\mathcal{L}_{\mathrm{FrFT}} = \{\exp(i\theta(\mathbf{t}^2 + \mathbf{f}^2)/2)\exp(i\lambda\mathbf{t}^2/2)\phi_0 \mid \lambda_{\min} \le \lambda \le \lambda_{\max}, \theta \in \Theta\} \tag{25}$$

where the set Θ is chosen so as not to violate the bandwidth constraints of the radar, and $\mathbf{f}$ is the operator on $L^2(\mathbf{R})$ defined by

$$\mathbf{f}\phi(t) = i\phi'(t), \tag{26}$$

where $\cdot'$ denotes differentiation in time. Note that $\mathbf{f}$ and $\mathbf{t}$ commute up to an extra additive term (the "canonical commutation relations"). To be precise,

$$[\mathbf{t}, \mathbf{f}] = \mathbf{tf} - \mathbf{ft} = -i\mathbf{I}. \tag{27}$$

For this library the corresponding measurement covariance matrices are given by (14) with

$$\mathbf{S}(\theta, \lambda) = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} \begin{pmatrix} 1 & 0 \\ \lambda & 1 \end{pmatrix}. \tag{28}$$

In the case of a finite number of waveforms in the library, we observe that the utility of the rotation library improves with the number of waveforms in the library. We can show that there exists a unique θ which maximizes the mutual information $I(X;Y)$ and, in a similar fashion to the pure chirp library case,

$$\mathcal{L}'_{\text{FrFT-chirp}} = \{\exp(i\theta(\mathbf{t}^2+\mathbf{f}^2)/2)\exp(i\lambda\mathbf{t}^2/2)\phi_0 \mid \lambda \in \{\lambda_{\min},\lambda_{\max}\}, \theta \in \Theta\} \quad (29)$$

is strongly equivalent to $\mathcal{L}_{\text{FrFT-chirp}}$.

5. Sensor scheduling simulations and results

Here we discuss simulations for sensor scheduling problems over up to two epochs into the future. The difficulties here reside in the design of the cost function and tracking of the scene. Our aim here is to show that sensor scheduling does, at least in simulation, achieve performance improvement.

Several aspects are common to all of the simulations described here. The scenarios all involve multiple maneuvering and crossing targets in simulated clutter. The simulated targets move according to an interacting multiple models (IMM) method; that is, at each epoch one of a finite number of dynamical models is chosen. The choice changes from epoch to epoch according to a Markov chain. Each of the dynamical models is linear. Process noise is, in each case, white and independent from epoch to epoch. Measurement is made using a waveform from a small finite library of waveforms, that we specify in each case.

A brief description of the tracking and waveform scheduling aspects of the scheme is as follows:

Tracking Since we are tracking multiple maneuvering targets, we use an iterated multiple modes (IMM) based tracker. This assumes that each target assumes at each epoch one of a finite number of dynamical models, such as "constant velocity", "constant linear acceleration", "fast left turn", etc, and implements a filter for each such dynamical model. As is normal in IMM the dynamical model is assumed to evolve by means of a Markov chain. We remark that the models and transition matrices are not identical with those used in constructing the scene. All noise on the processes is assumed Gaussian and independent between epochs. Multiple targets and clutter are addressed by an integrated probabilistic data association tracker, specifically the LMIPDA-IMM algorithm described in [5]. This is a recursive algorithm combining a multi-target data

association algorithm (LMIPDA) with manoeuvring target state estimation implemented using IMM. Each track carries along with it a "probability of track existence" which is updated at each epoch along with the track. In addition the probability of each dynamical model is updated from the measurements.

Waveform Scheduling The choice of measurement is made using the control variable $n(k)$. In fact two choices are made at each epoch, the target to be measured and the waveform used. The waveforms impinge on the measurement process through the covariance matrix of the noise $\omega_n^t(k)$. In this model, the sensor is characterised by a measurement noise covariance matrix which is waveform dependent

$$R_\phi = T^T J_\phi^{-1} T, \tag{30}$$

where J_ϕ is the Fisher information matrix corresponding to the measurement using waveform ϕ and T is the transformation matrix between the time delay and Doppler measured by the receiver and the target range and velocity. It is assumed that N different *measurement modes* are available for each target, each given by a measurement matrix H_n^t $n = 1, 2 \ldots, N$.

In order to determine which target to measure and which waveform to use, for each existing target and each waveform the track error covariance $P_{k-1|k-1}^t$ is propagated forward using the Kalman update equations. In the absence of measurements, as will be the case in the study of revisit times, the best we can do is to use current knowledge to predict forward and update the covariance matrix, dynamic model pdf and probability of track existence. The tracking and scheduling algorithms now becomes as follows:

- *IMM mixing* as in [5] is conducted as usual;
- *Forward prediction* is then performed separately for each dynamical model.
- *Covariance update:* this is normally done with the data, but since we are interested in choosing the best sensor mode at this stage the following calculations are required. If the target does not exists there will be no measurements originating from the target and the error covariance matrix is equal to the *a priori* covariance matrix, if the target exists, is detected, and the measurement is received then the error covariance matrix is updated using the Kalman equation.

- The dynamic model and track existence pdfs are updated. If the target does not exist it produces no measurement; if it does and is detected the expected measurement pdf,dynamical model and track existence pdfs are using the LMIPDA-IMM filter.

- The next step is to combine the estimates for all dynamics models $j = 1, \ldots, M$ into one, using the standard "IMM combination" formulae [5]. We refer the interested reader to this paper for details.

5.1 One- and Two-Step Ahead Scheduling

Our first aim is to do a simple comparison of one-step and two-step scheduling of waveforms and other radar parameters. The modes of the radar system (beam-direction and waveform) are chosen for the next one or two PRIs based on the predicted scene over that time. We note that in the two-step case the choice of radar mode is updated on a PRI by PRI basis. A comparison between one and two-step ahead scheduling is an important issue, since if it is shown that the improvement achieved by two step ahead optimal scheduling over just one-step ahead scheduling is slight, it is reasonable to guess that one-step ahead scheduling is for practical purposes optimal. Since multi-step scheduling is inherently much more computationally intensive, it is best avoided if it results in only a marginal improvement. We emphasize that, of course, results of this kind are very likely to be scenario dependent unless there is some inherently mathematical reason why optimal multi-step ahead scheduling is achievable by a myopic approach. That would appear unlikely. We emphasize too that this work has been done on a simulator. The structure of the scene is highly artificial and the clutter models very simplistic.

We have compared one-step and two-step ahead scheduling using two performance measures. The first is the root mean square error of the track estimation; this is a fairly obvious measure of the performance of the tracker. The second measure was the number of track updates. Since the sensor is managed in such a way that track updating is done only when the predicted track error exceeds a threshold, this also gives a measure of how far the estimation process is diverging from the actual target state.

We refrain here from giving detailed descriptions of the experiments. Their outcome suggests that, in the presence of clutter, the tracking performance can be improved with multiple step ahead scheduling as opposed to one step ahead. The results are represented in Figures 2 and 3. One observes that for two steps ahead the tracking accuracy is improved, albeit slightly, while the number of times the track had to be

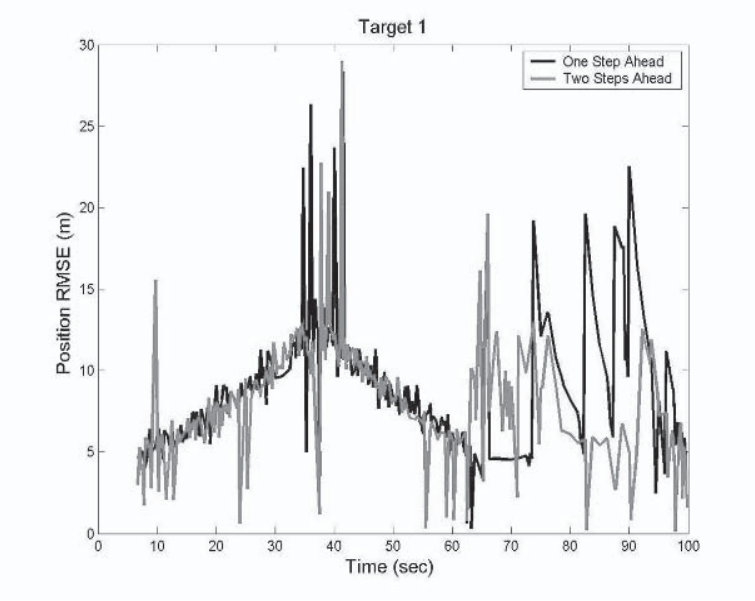


Figure 2. Root Mean Square Error (RMSE).

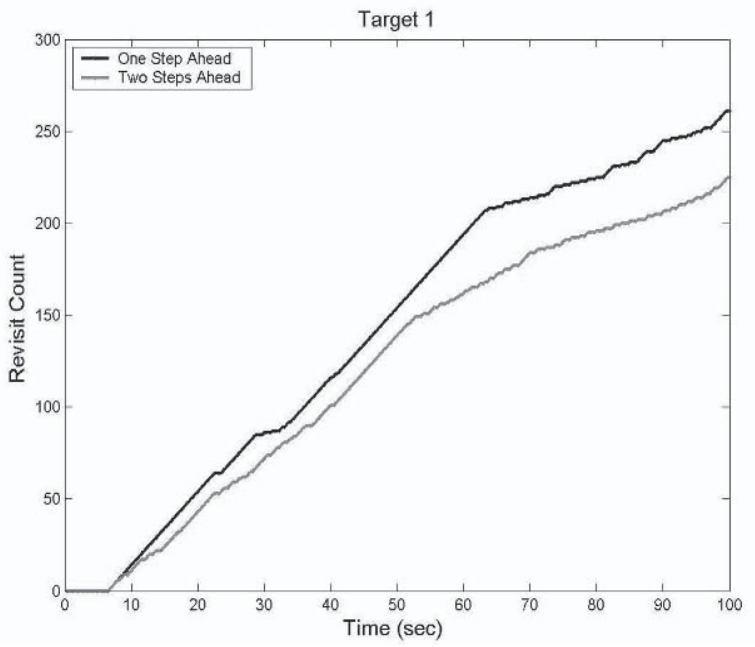


Figure 3. The number of track updates.

updated is reduced. In both cases the improvement is not large, and is worse immediately after the aircraft has maneuvered. Once the aircraft has settled back into a linear model again the two-step ahead scheduler does better.

5.2 Scheduling of Waveform Libraries

The next series of experiments is focused on how the choice of waveform libraries affect the problem of tracking of maneuvering targets.

As in the previous experiments, at each epoch we would like to select a waveform (or really the error covariance matrix associated with a measurement using this waveform) so that the measurement will minimize the uncertainty of the dynamic model of the target. We study two possible measures: entropy of the *a posteriori* pdf of the models and *mutual information* between the dynamic model pdf and measurement history. Both of these involve making modifications to the LMIPDA-IMM approach that are described in [5]. Since we want to minimize the entropy *before* taking the measurement, we need to consider the *expected* value of the cost. To do this we replace the measurement $\mathbf{z}$ in the IMM equations by its expected value. In the case of the second measure, for a model we have

$$I(\Gamma; Z) = -\sum_{\gamma=1}^{M} P\{\gamma\} \log P\{\gamma\} + \int P\{\mathbf{z}\} \sum_{\gamma=1}^{M} P\{\gamma|\mathbf{z}\} \log P\{\gamma|\mathbf{z}\} d\mathbf{z}, \tag{31}$$

where $P\{\gamma\}$ is the *a priori* probability of the model $\gamma \in \Gamma$, and $\mathbf{z}$ is the measurement.

Simulations were performed for both cost functions. Target trajectories in range and Doppler were randomly created. The maneuvers for the trajectories were generated using a given transition probability matrix. We identified four maneuvers: 0 acceleration; 10m/s^2 acceleration; 50m/s^2 acceleration; -10m/s^2 acceleration.

In the experiments we considered rotation-LFM waveform libraries with 1 waveform (max upsweep chirp), 2 waveforms (max upsweep and max downsweep chirps), and 6 waveforms (maximum upsweep, maximum downsweep chirps and 2 rotations 0.2π and 0.4π as defined in equation (22) to the left for the maximum upsweep and maximum downsweep chirps).

The results are presented in Figures 4, 5, 6, and 7. Clearly, for either cost function, waveform scheduling using the six-waveform library outperforms waveform scheduling using the two-waveform library, which in turn outperforms no scheduling (one waveform) in both estimation ac-

curacy (Figures 4 and 6) and correct identification of target maneuver (Figures 5 and 7).

5.3 Re-visit Time Scheduling

Finally in this section on simulations, we briefly describe a project that includes many of the ideas we have presented already. The crucial problem is to use scheduling to reduce the amount of time spent on tracking known targets while retaining a given level of track accuracy. By doing this we permit the sensor to spend more time in surveillance for new targets.

We postulate a radar system tracking T targets where T is a random variable $0 \leq T \leq T_0$ and the tth target is in state $x^t(k)$ at epoch k. In addition the radar undertakes surveillance to discover new targets. This surveillance is assumed to require a certain length of time, say T_{scan} within every interval of length T_{total}. The remainder of the time is spent measuring targets being tracked. We aim to schedule revisit times to targets within these constraints.

At each epoch a target track and a beam direction have to be selected. The scheduler has a list $\Delta = \{\delta_1, \delta_2, \ldots, \delta_K\}$ of "revisit intervals". Each of the numbers δ_k is a number of epochs representing the possible times

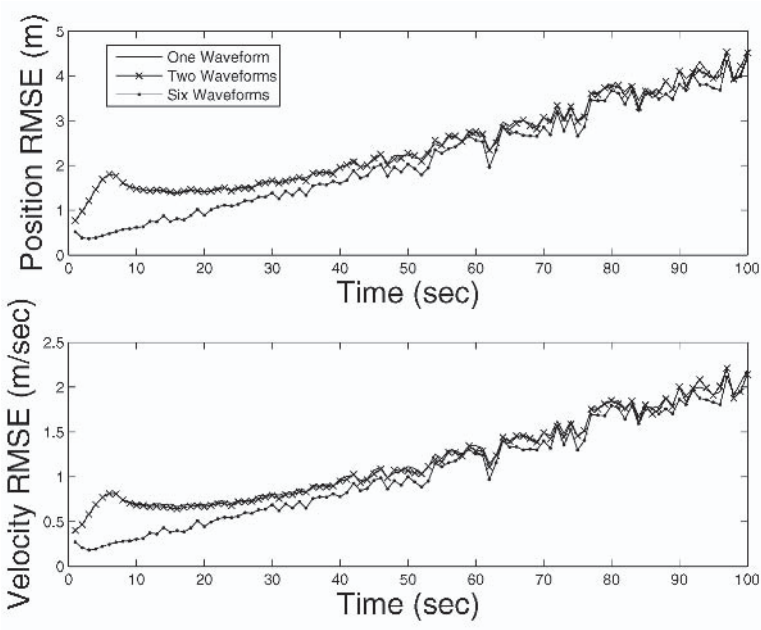


Figure 4. Root Mean Square Error for Entropy Cost.

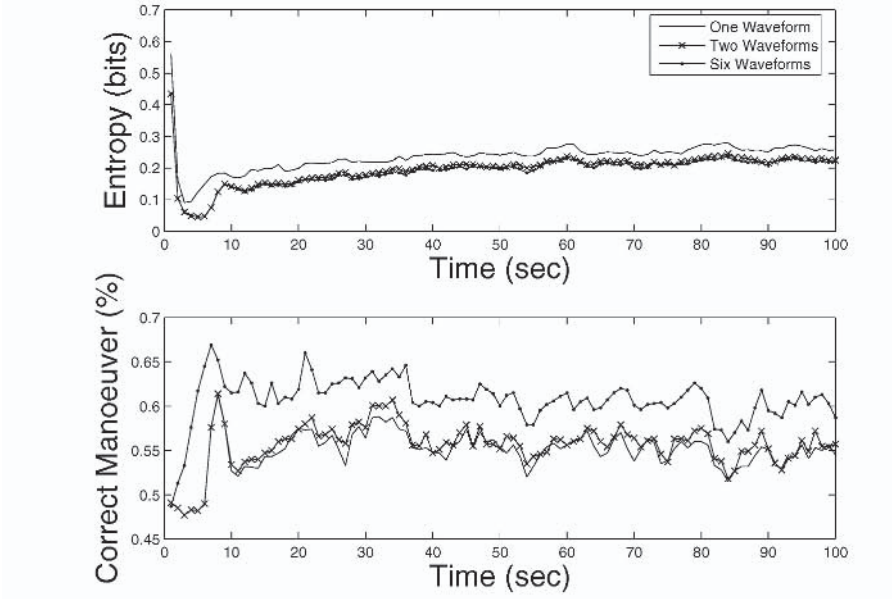
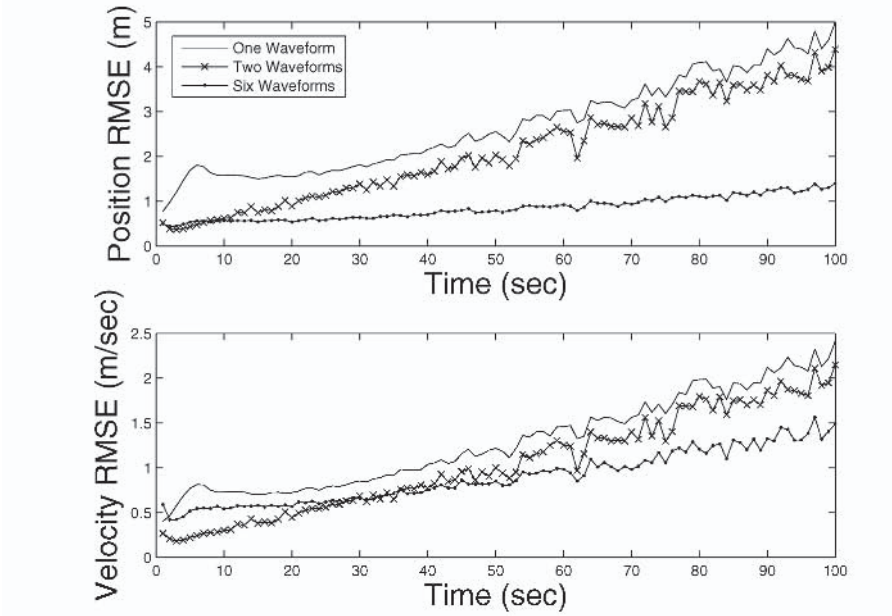


Figure 5. Cost Function and Correct Maneuver Identification for Entropy Cost.



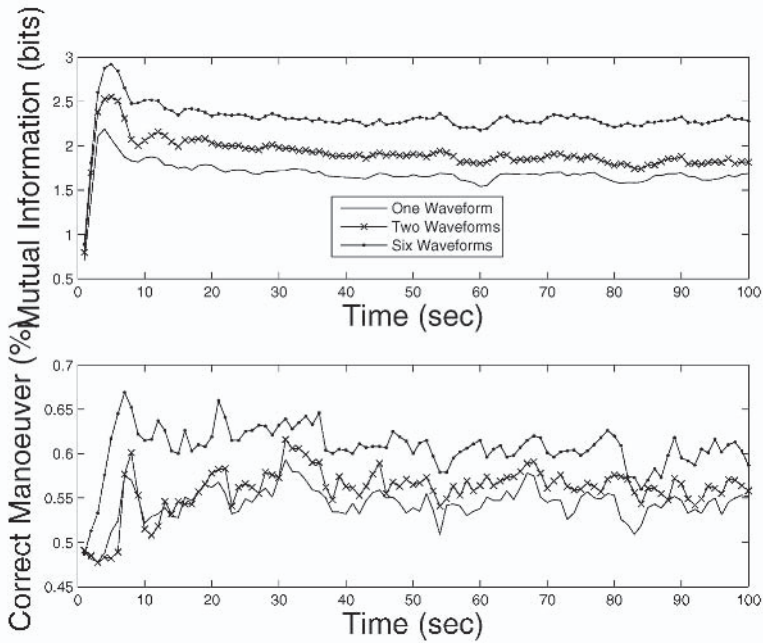


Figure 7. Cost Function and Correct Maneuver Identification for Mutual Information Cost.

between measurements of any of the existing targets. It is assumed for the purposes of scheduling and tracking that during any of these revisit intervals the target dynamics do not change, though the simulator permits target maneuvers on an epoch by epoch basis.

The LMIPDA-IMM calculations are performed for all combinations of revisit times in Δ and waveforms in the library. Evidently then the number of combinations grows exponentially in the number of steps ahead, and soon becomes impractical for implementation. Having obtained the error covariance matrix for all possible combinations of sensor modes, the optimal sensor mode (waveform) is then chosen for each target to be the one which gives the longest re-visit time, while constraining the absolute value of the determinant of the error covariance matrix to be smaller than the prescribed upper limit K. In other words, our objective is

$$\phi, \delta = \arg\max \Delta, \text{ subject to } |\det(P_{k|k})| \leq K. \tag{32}$$

Scheduling is then done to permit a full scan over the prescribed scan period while also satisfying the constraints imposed by the revisit times obtained by the sensor scheduler. Once a target is measured, its revisit time is re-calculated.

We note that for many manoeuvring targets there may be no solution to the scheduling problem that satisfies the constraints. However, we have not been able to simulate a situation in which this happens.

We have, on the other hand done simple simulations for the case of one-step ahead and two-step ahead scheduling. In the latter case, the revisit times and waveforms are calculated while the target states are propagated forward over two measurements, with the cost function being the absolute value of the determinant of the track error covariance after the second measurement. Only the first of these measurements is done before the revisit calculation is done again for that target, so that the second may never be implemented.

Simulations were performed to compare the effects of no scheduling with random choice of waveform against one-step and two-step ahead beam and waveform scheduling as described in the last section. All three simulations were performed 100 times on the same scenario. In the first case, measurements were taken at each scan with no further measurements beyond the scan measurements permitted. The waveforms were chosen at random from the three waveforms in the library. The simulated scene corresponded to a surveillance area of 15km by 15km contained two maneuvering land targets in stationary land clutter which had small random Doppler to simulate movement of vegetation in wind. The number of clutter measurements at each epoch was generated by samples from a Poisson distribution with mean ~ 5 per scan per sq.km. Target measurements were produced with probability of detection 0.9. The target state x^t consisted of target range, target range rate and target azimuth. The targets were performing the following maneuvers: constant velocity, constant acceleration, constant deceleration and coordinated turns with constant angular velocity. In these experiments we used the waveform library consisting of three waveforms: an up-sweep chirp, a down-sweep chirp and an unmodulated pulse. In the scheduling cases, surveillance time used approximately 80 percent of each scan period, the remaining 20% being allocated as described above to the maintenance of tracks of existing targets.

The outcome of experiments suggests that in the presence of clutter tracking performance can be improved with scheduling and even more with multiple step ahead scheduling as opposed to one step ahead. The results are represented in Figure 8. It should be observed in Figure 8 that RMS error was considerably worse especially during the early part of the simulation for the unscheduled case. In fact the RMS error in the unscheduled case is 5larger immediately after significant manoeuvres as can be expected. Of course, in this case the revisit time is fixed and is not plotted in the second subplot. One observes, that, for the two-step

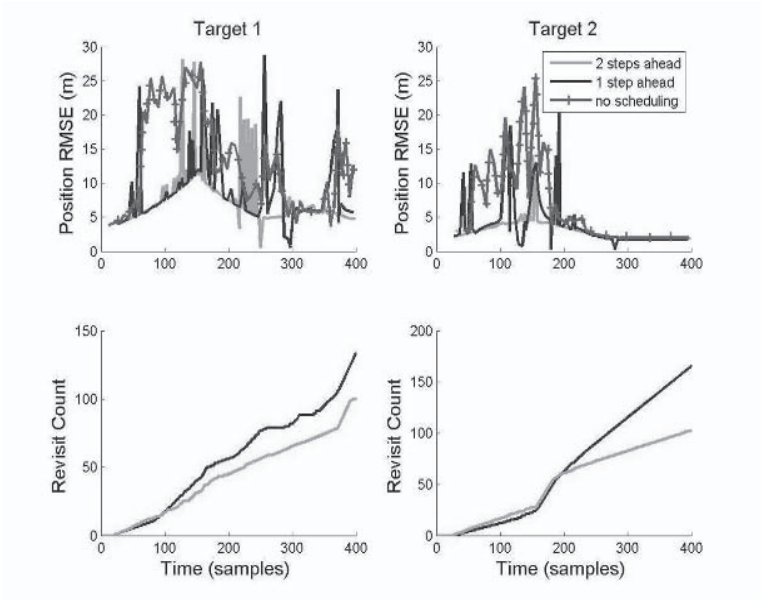


Figure 8. Root Mean Square Error (RMSE) and Revisit Count for one vs. two step ahead beam and waveform scheduling.

ahead case, tracking accuracy is improved (top plots) slightly over the one-step ahead case but with a significant reduction in revisit times to maintain those tracks.

References

[1] A. Nehorai A. Dogandžić. Cramér-Rao bounds for estimating range, velocity and direction with an active array. *IEEE transactions on Signal Processing*, 49(6):1122–1137, June 2001.

[2] Steven M. Kay. *Fundamentals of Statistical Signal Processing*. Prentice-Hall, 1993.

[3] D.J. Kershaw and R.J. Evans. Optimal waveform selection for tracking systems. *IEEE Transactions on Information Theory*, 40(5):1536–50, September 1994.

[4] D.J. Kershaw and R.J. Evans. Waveform selective probabilistic data association. *IEEE Aerospace and Electronic Systems*, 33(4):1180–88, October 1997.

[5] Darko Mušicki, Subhash Challa, and Sofia Suvorova. Multi target tracking of ground targets in clutter with LMIPDA-IMM. In *7th International Conference on Information Fusion*, Stockholm, Sweden, July 2004.

[6] H.L. van Trees. *Detection, Estimation and Modulation Theory, Part III*. Wiley, New York, 1971.

SOME RADAR TOPICS: WAVEFORM DESIGN, RANGE CFAR AND TARGET RECOGNITION

H. Rohling
Technical University Hamburg-Harburg
Hamburg, Germany
rohling@tu-harburg.de

Abstract The first RADAR patent was applied for by Christian Huelsmeyer on April 30, 1904 at the patent office in Berlin, Germany. He was motivated by a ship accident on the river Weser and called his experimental system "Telemobiloscope". In this chapter some important and modern topics in radar system design and radar signal processing will be discussed. Waveform design is one innovative topic where new results are available for special applications like automotive radar. Detection theory is a fundamental radar topic which will be discussed in this chapter for new range CFAR schemes which are essential for all radar systems. Target recognition has for many years been the dream of all radar engineers. New results for target classification will be discussed for some automotive radar sensors.

Keywords: automotive radar; continuous wave radar; CFAR processing; rank-order; filters; FSK modulation; matched filters; unmatched filters.

1. Introduction

The objective of this chapter is to discuss some important contributions to radar system design and digital radar signal processing. The focus is on waveform design in general and on automotive applications in particular. Target detection is an important issue for all radar systems. Therefore some range CFAR (constant false alarm rate) procedures will be discussed which can be applied, especially in multiple target situations, to avoid masking. Additionally some new results are discussed for target recognition systems which have been developed for automotive applications.

Jim Byrnes and Gerald Ostheimer (eds.),
Advances in Sensing with Security Applications, 293–322.

2. Combination of LFMCW and FSK modulation principles for automotive radar systems

High performance automotive radar systems are currently under development for various applications. Comfort systems like Adaptive Cruise Control (ACC) are already available on the market as 77 GHz radars. Target range and velocity are measured simultaneously with high resolution and accuracy even in multi-target situations, but the measurement and processing time to detect the relevant object is approximately 100 ms. Future developments will be more concentrated on safety applications like Collision Avoidance (CA) or Autonomous Driving (AD). In this case the requirements for reliability (extreme low false alarm rate) and reaction time (extreme short delay) are much higher compared with ACC systems.

To meet all these system requirements specific waveform design techniques must be considered. For ACC systems both radar types of classical pulse waveform with ultra short pulse length (10 ns) or alternatively continuous wave (CW) transmit signal with a bandwidth of 150 MHz are considered. The main advantage of CW systems in comparison with classical pulse waveforms is the low measurement time and low computational complexity.

This section describes a new waveform design for automotive applications based on CW transmit signals which lead to an extremely short measurement time. The basic idea is a combination of linear frequency modulation (LFM) and FSK CW waveforms in an intertwined technique. Unambiguous range and velocity measurement with high resolution and accuracy can be required in this case even in multi-target situations. After introductions to FSK and LFM waveform design techniques in sections 2.1 and 2.2 the combined and intertwined waveform is described in detail in section 2.3.

2.1 Pure FSK modulation

Pure FSK modulation (as shown in Figure 1 (a)) uses two discrete frequencies f_A and f_B (so-called two frequency measurement) [1] in the transmit signal. Each frequency is transmitted inside a so-called coherent processing interval (CPI) of length T_{CPI} (e.g. $T_{CPI} = 5$ ms). Using a homodyne receiver the echo signal is down converted by the instantaneous frequency into base band and sampled N times. The frequency step $f_{Step} = f_B - f_A$ is small and will be chosen to depend on the maximum unambiguous target range. The time-discrete receive signal is Fourier transformed in each CPI of length T_{CPI} and targets will be

detected by an amplitude threshold (CFAR). Due to the small frequency step a single target will be detected at the same Doppler frequency position in the adjacent CPI's but with different phase information on the two spectral peaks. The phase difference $\Delta\varphi = \varphi_B - \varphi_A$ in the complex spectra is the basis for the target range (R) estimation. The relation between the target distance and phase difference is given by the equation:

$$R = -\frac{c \cdot \Delta\varphi}{4\pi \cdot f_{Step}}. \tag{1}$$

To achieve an unambiguous maximum range measurement of 150 m a frequency step of f_{Step} =1 MHz is necessary. In this case the target resolution only depends on the CPI length T_{CPI}. The technically simple VCO modulation is an additional advantage of this waveform. But this FSK waveform does not allow any target resolution in the range direction, which is an important disadvantage of this measurement technique. Especially in the automotive traffic environment, more than a single fixed target will occur simultaneously inside an antenna beam. These fixed targets cannot be resolved by a FSK waveform.

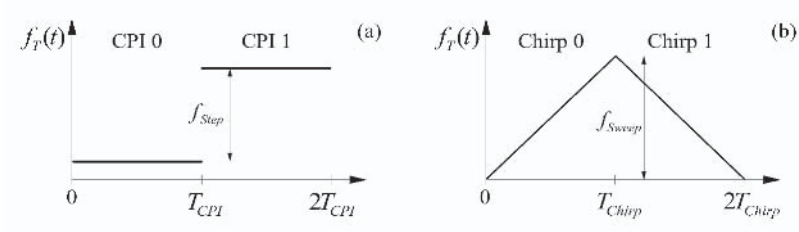


Figure 1. Two CW waveform principles: (a) FSK modulation, (b) LFM.

2.2 Pure Linear Frequency Modulation

Radars which use a pure LFM technique modulate the transmit frequency with a triangular waveform [6]. The oscillator sweep is given by f_{Sweep}. A typical value for the bandwidth is f_{Sweep} =150 MHz to achieve a range resolution of $\Delta R = \frac{c}{2 \cdot f_{Sweep}} = 1$ m. In general, a single sweep of the LFM waveform gives an ambiguous measurement in range and relative velocity. The down converted receive signal is sampled and Fourier transformed inside a single CPI. If a spectral peak is detected in the Fourier spectrum at index κ (normalized integer frequency) the ambiguities in target range and velocity can be described in an R-v-diagram by the equation:

$$\kappa = \frac{v}{\Delta v} - \frac{R}{\Delta R} \quad \Leftrightarrow \quad \frac{v}{\Delta v} = \frac{R}{\Delta R} + \kappa \tag{2}$$

where Δv gives the velocity resolution resulting from the CPI duration T_{Chirp} ($\Delta v = \frac{\lambda}{2 \cdot T_{Chirp}} = 0.8$ m/s, λ is the wavelength of 4 mm @ 77 GHz and $T_{Chirp} = 2.5$ ms).

Due to resulting range-velocity ambiguities, further measurements with different chirp gradients in the waveform are necessary to achieve an unambiguous range-velocity measurement, even in multi-target situations. The well-known up/down-chirp principle as it is depicted in Figure 1 (b) is described in detail in [5]. LFM waveforms can be used even in multi-target environments, but the extended measurement time is an important drawback of this LFM technique.

2.3 Combined FSK and LFM waveforms

The combination of FSK and LFM waveform design principles offers the possibility of unambiguous target range and velocity measurement simultaneously. The transmit waveform consists in this case of two linear frequency modulated up-chirp signals (the intertwined signal sequences are called A and B). The two chirp signals will be transmitted in an intertwined sequence (ABABAB...), where the stepwise frequency modulated sequence A is used as a reference signal while the second up-chirp signal is shifted in frequency by f_{Shift}. The received signal is down converted into base band and directly sampled at the end of each frequency step. The combined and intertwined waveform concept is depicted in Figure 2.

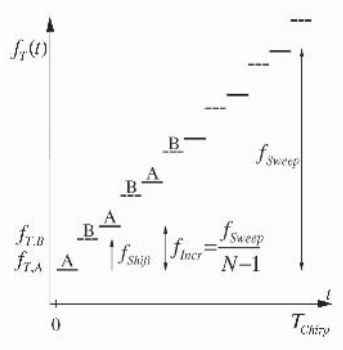


Figure 2. Combined FSK-LFMCW waveform principle.

Each signal sequence A or B will be processed separately by using the Fourier transform and CFAR target detection techniques. A single target with specific range and velocity will be detected in both sequences

at the same integer index $\kappa = \kappa_A = \kappa_B$ in the FFT-output signal of the two processed spectra. In each signal sequence A or B the same target range and velocity ambiguities will occur as described in equation 2. But the measured phases φ_A and φ_B of the two (complex) spectral peaks are different and include the fine target range and velocity information which can be used for ambiguity resolution. Due to the coherent measurement technique in sequences A, B the phase difference $\Delta\varphi = \varphi_B - \varphi_A$ can be evaluated for target range and velocity estimation. The measured phase difference $\Delta\varphi$ can be described analytically by the following equation:

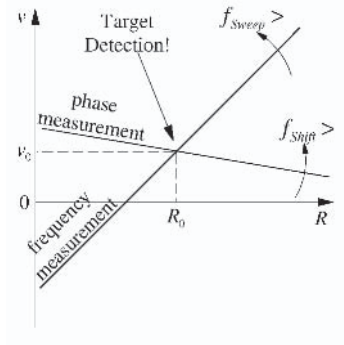


Figure 3. Graphical resolution principle of ambiguous frequency and phase measurements.

$$\Delta\varphi = \frac{\pi}{N-1} \cdot \frac{v}{\Delta v} - 4\pi \cdot R \cdot \frac{f_{Shift}}{c} \tag{3}$$

where N is the number of frequency steps (or receive signal samples) in each transmit signal sequence A, B. In this first step $\Delta\varphi$ is ambiguous but it is possible to resolve these ambiguities by combining the two measurement results of equations 2 and 3. The intersection point of the two measurement results is shown in Figure 3 in a graphical way. The analysis leads to an unambiguous target range R_0 and relative velocity v_0 :

$$R_0 = \frac{c \cdot \Delta R}{\pi} \cdot \frac{(N-1) \cdot \Delta\varphi - \pi \cdot \kappa}{c - 4 \cdot (N-1) \cdot f_{Shift} \cdot \Delta R} \tag{4}$$

$$v_0 = \frac{(N-1) \cdot \Delta v}{\pi} \cdot \frac{c \cdot \Delta\varphi - 4\pi \cdot f_{Shift} \cdot \Delta R \cdot \kappa}{c - 4 \cdot (N-1) \cdot f_{Shift} \cdot \Delta R} \tag{5}$$

This new intertwined waveform shows that unambiguous target range and velocity measurements are possible even in a multi-target environment. An important advantage is the short measurement and processing time.

2.4 System example

In this section a waveform design based on the new intertwined signal is developed as an example for automotive applications. The signal bandwidth is f_{Sweep} =150 MHz to fulfill the range resolution requirement of 1 m. The stepwise frequency modulation is split into $N = 256$ separate bursts of $f_{Incr} = \frac{150\text{MHz}}{255} = 588$ kHz each. The measurement time inside a single burst A or B is assumed to be 5 μs resulting in a chirp duration of the intertwined signal of $T_{Chirp} = 2.56$ ms. This results in a velocity resolution of $\Delta v = \frac{\lambda}{2 \cdot T_{Chirp}} = 2.7$ km/h.

The important waveform parameter f_{Shift} is optimized on the basis of high range and velocity accuracy. The highest accuracy occurs if the intersection point in the R-v-diagram results from two orthogonal lines.

For this reason the frequency shift between the signal sequences A and B is $f_{Shift} = -\frac{1}{2} \cdot f_{Incr} = -294$ kHz.

In this specific case equations 4 and 5 turn into

$$\frac{R_0}{\Delta R} = \frac{N-1}{2\pi} \cdot \Delta\varphi - \frac{\kappa}{2} \tag{6}$$

$$\frac{v_0}{\Delta v} = \frac{N-1}{2\pi} \cdot \Delta\varphi + \frac{\kappa}{2} \tag{7}$$

The proposed intertwined and stepwise CW waveforms show high performance in simultaneous range and velocity measurement accuracy. The main advantage is the short measurement time in comparison to classical LFM waveforms while the resolution and accuracy are unchanged. The properties of the new intertwined CW waveform technique are quite promising. This concept is a good basis for high performance automotive radar systems with different safety applications (e.g. pre crash) which require ultra short measurement and processing times.

3. Automotive Radar Network Based On 77GHz FMCW Sensors

Automotive radar systems need to have the capability to measure range, velocity and azimuth angle simultaneously for all point and extended targets inside the observation area. Short measurement time even in dense target situations, and high range accuracy and resolution, are required in all automotive applications. We will distinguish in this section between single radar sensors and radar network systems. So-called far distance single radar sensors use an observation area of $\pm 10°$ in azimuth angle and a maximum range of up to 200m.

If, in contrast, a large azimuth angle coverage and a short maximum range are required, radar network systems with e.g. four individual sensors mounted behind the front bumper are used instead of a single radar sensor. Typical automotive applications for radar networks with a large azimuth angle coverage but limited range are, for example, Collision Avoidance and Pre Crash Warning. Figure 4 illustrates the observation area considered in this section. The relevant system parameters for a short-range radar network are given in Table 1.

The developed near distance single radar sensor provides target range and radial velocity simultaneously with high accuracy and for all objects inside the observation area. It is characteristic of radar networks signal processing that the angular position of each target is calculated by means of multilateration techniques based on the sensor specific measured target ranges inside the network [7, 8]. This is to derive the desired target position by calculating the intersection point of all range measurements from different radar sensor positions.

Figure 4. Observation area of the radar network system.

Table 1. Requirements for a single sensor in a radar network system.

Parameter	Value
Observation area	120° in azimuth
Maximum range	40m
Range resolution	0.4m
Range accuracy (required by multilateration)	0.02m
Velocity resolution	1m/s
Velocity accuracy	0.3m/s
Target acquisition time	20-100ms

3.1 Radar network

Architecture. In this section we present a short-range radar network that consists of four distributed 77GHz radar sensors [9]. All sensors are mounted behind the front bumper of the vehicle in an invisible form. In contrast to already known radar networks in the 24GHz frequency domain, the presented sensors are working at a 77GHz carrier frequency. Due to European regulations, this frequency band provides a wider bandwidth for automotive radar applications. To achieve the desired high range accuracy of 2cm, which is required by the multilateration process, a frequency sweep of 1GHz has to be utilized within the Linear Frequency Modulated Continuous Waveform (LFMCW). Figure 5 shows an image of the 77GHz prototype sensor [10]. Time synchronization between the individual sensors is needed to avoid any interference situations between the radar sensors. With additional carrier synchronization between the sensors the radar network could even provide the capability of bistatic measurements. The measurement results described in this section are based on monostatic sensor measurements.

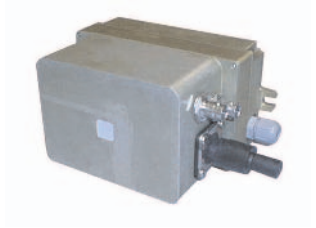


Figure 5. Close-up of a single 77GHz radar sensor.

Experimental system. To validate the analytical results some experimental cars, shown in Figure 6, have been equipped with a radar network. These test cars are used for data acquisition and recording to optimize the signal processing algorithms. In section 3.5 experimental results are presented to illustrate the theoretical results of developed and applied algorithms.

3.2 Single sensor signal processing

For data acquisition a classical linear FMCW waveform is used which consists of four individual chirp signals (see Figure 7) and covers a bandwidth of 1GHz, which corresponds to a range resolution of 0.4m under realistic signal processing conditions. This waveform combines high ac-

Figure 6. Two experimental cars equipped with 77GHz radar sensors.

curacy measurements in range and velocity and reliable target detection in multi-target situations. Furthermore, FMCW signal processing techniques compared with pulse radar waveforms lead to reduced computational complexity and hardware requirements [11].

Four individual chirps provide sufficient redundancy in multi-target or extended target situations [4] to suppress ghost targets in the range-velocity processing. For each individual chirp signal the beat frequencies $df_1, \ldots df_4$, shown in Figure 7, will be estimated by applying an FFT.

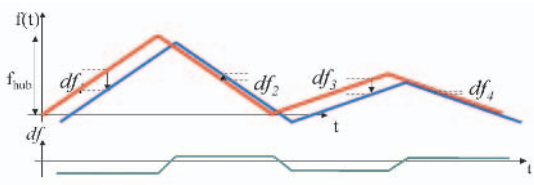


Figure 7. FMCW radar waveform (red) and the corresponding receive signal (blue) for a single target situation.

Table 2 gives a summary of the considered waveform parameters. It is a real technical challenge to handle, in multiple and extended target situations, the large number of detections in the data association, tracking as well as range, velocity and angle parameter estimation procedure. Therefore, a specific signal processing technique has been developed and is implemented in the experimental system.

The classical FMCW radar signal processing scheme is structured into the following different independent blocks: beat frequency estimation based on FFT, target detection (CFAR), range and velocity processing, multilateration for azimuth angle measurement and tracking; see Figure 8 and [12]. Even as an enhancement of classical signal processing a classification procedure could be used to derive additional information about the target object class [13].

Table 2. Waveform Parameters.

Parameter	Value
Center frequency	77GHz
Number of Chirps	4
Single chip duration	2 ms
First sweep bandwidth	1GHz
Second sweep bandwidth	500MHz

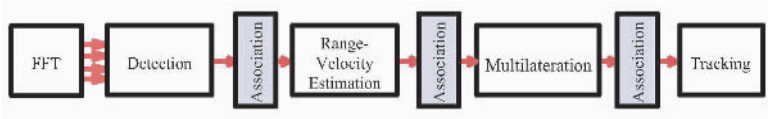


Figure 8. Classical FMCW radar signal processing.

After CFAR detection each signal-processing block contains an independent association procedure to combine measurements from different chirps and different radar sensors, which belong to a single point or even extended target. In the following, all signal processing steps are presented and will be discussed in detail.

The objective of this section is to optimize the classical signal processing structure especially for automotive applications. The main drawback of the classical signal processing structure is that in multiple and extended target situations each of the three independent association procedures (Figure 8) will induce some ghost targets. This network behavior has been observed in many sets of measured data. Based on these observations and results inside a classical radar network an alternative signal processing structure is proposed herein, which is based on a joint optimization of the three independent association schemes.

Target detection. Due to the large angular coverage considered it is characteristic for automotive applications that each radar sensor has many detections at the CFAR procedure output due to multiple and extended target situations. For reliable target detection in this multi target environment, an ordered-statistic (OS) constant-false-alarm-rate (CFAR) thresholding has been applied which showed the best experimental results [14].

As a result of the FMCW waveform with four individual chirp signals, a proper detection process results in four detected beat frequencies $f_{\mathrm{C,S}}$ per point target and sensor. In total a single target, detected by the

radar network, will lead to 16 detected beat frequencies at the FFT output. The vector $\overrightarrow{m}^f$ combines these beat frequencies and therefore describes all information which is available in the signal-processing scheme for each individual target.

$$\vec{m}^f = \left[\underbrace{f_{1,1}, f_{2,1}, f_{3,1}, f_{4,1}}_{Sensor1}, \cdots, \underbrace{f_{1,4}, f_{2,4}, f_{3,4}, f_{4,4}}_{Sensor4} \right]^T \tag{8}$$

Range-Velocity Estimation. Each measured beat frequency contains information about the target range and velocity in an ambiguous way. But each individual target with range r_S and velocity v_S leads to a deterministic beat frequency for each chirp signal of the waveform. This beat frequency and the relation to target range and velocity is given by the linear equation:

$$f_{C,S} = a_C \cdot r_S + b_C \cdot v_S. \tag{9}$$

Parameters a_C and b_C depend on chirp characteristics like chirp duration, bandwidth and carrier frequency [6]. Based on the 4 beat frequencies measured by a single sensor the point target range and velocity can be derived simultaneously by an intersection process. In this case and in a single point target situation the four measured frequencies are transformed into target range and velocity, unambiguously. But in multiple or even extended target situations this range velocity calculation could lead to some ghost targets.

Each sensor of the radar network has an individual position behind the front bumper. Therefore, each sensor will calculate individual values for target range and velocity based on the four measured beat frequencies, equation 8, inside the FMCW waveform. The measurement result is described by an eight-element parameter vector.

$$\vec{m}^t = \left[\underbrace{r_1, v_1}_{Sensor1}, \cdots, \underbrace{r_4, v_4}_{Sensor4} \right]^T \tag{10}$$

A set of linear equations can now be derived which describes the relation between 16 measured beat frequencies and sensor specific target range and velocity parameters.

$$\vec{m}^f = C \cdot \vec{m}^t \tag{11}$$

In multi-target situations the association of the detected beat frequencies at the FFT output to different targets is not trivial. Therefore, a

data association process has to be performed using the redundancy given by the four chirp signals inside a single waveform.

3.3 Radar network signal processing

The objective of radar network signal processing is to calculate the azimuth angle (or target position in Cartesian coordinates) of each target in multiple or even extended target situations based on the precise range measurement of each radar sensor. Furthermore the tracking procedure is part of the network processing.

Multilateration Procedure. To derive target position and velocity described by a target state vector in Cartesian coordinates

$$\vec{t} = (t_x, t_y, v_x, v_y)^T \tag{12}$$

a multilateration technique is used. Based on sensor specific range and velocity measurements for each individual point target the state vector can be estimated if the position of each sensor behind the front bumper is taken into account, which is described by the vector:

$$\vec{s} = (s_x, s_y)^T. \tag{13}$$

For each sensor the target range, see (9) , can be calculated by a nonlinear equation if target and sensor Cartesian positions are known.

$$r_S = \sqrt{(t_x - s_x)^2 + (t_y - s_y)^2}. \tag{14}$$

The target radial velocity in (9) can be processed as follows:

$$v_S = \frac{t_x - s_x}{r_S} \cdot v_x + \frac{t_y - s_y}{r_S} \cdot v_y. \tag{15}$$

Combining both equations, the relation between target state vector $\vec{t}$ and the measurement parameter vector $\vec{m}^t$, equation 10, can be formulated by the nonlinear equation

$$\vec{m}^t = h(\vec{t}). \tag{16}$$

The Jacobian matrix

$$H_{\vec{t}_0} = \left. \frac{\partial h(\vec{t})}{\partial \vec{t}} \right|_{\vec{t}_0} \tag{17}$$

is used in an iterative Gauss-Newton algorithm to estimate the target position and velocity in Cartesian coordinates based on the given measurements and a given initial position estimate [15].

At the beginning of the multilateration procedure an association process is integrated to select the sensor specific range and velocity measurements belonging to a single target. In multiple and extended target situations there is high risk of ghost targets due to errors in the association procedure. Furthermore, in situations with low target detection probability the multilateration procedure does not have sufficient information for calculating the target state vector. While the data association for range-velocity processing was provided by four individual chirp signals within the waveform, the data association for multilateration processing is provided by measurements of four individual radar sensors at different positions.

Target Tracking. Normally radar tracking is based purely on plot-to-track association. In this section we develop a radar network and FMCW waveform specific frequency-to-track association as an extension of the classical tracking procedure. For pulse radar networks a similar idea of range-to-track association schemes have been published in [12]. In this case the multilateration technique is not processed independently and explicitly but is integrated into the tracking procedure which is based on Kalman filtering.

For automotive applications with relative low velocities and a high update rate 1/T, a pure linear motion model with constant velocity can be considered. The respective state transition matrix A for a constant-velocity trajectory can be used to calculate the predicted target state vector for the next time step by the following equation:

$$\vec{t}_{k+1} = A \cdot \vec{t}_k + \vec{w}_k = \begin{bmatrix} 1 & & T & \\ & 1 & & T \\ & & 1 & \\ & & & 1 \end{bmatrix} \cdot \begin{bmatrix} t_x \\ t_y \\ v_x \\ v_y \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ w_x \\ w_y \end{bmatrix}. \tag{18}$$

Here $\vec{t}_k$ is the target state vector at time index k and $\vec{w}_k$ contains two random variables which describe the unknown process error, which is assumed to be a Gaussian random variable with expectation zero and covariance matrix Q. In addition to the target dynamic model, a measurement equation is needed to implement the Kalman filter. This measurement equation maps the state vector $\vec{t}_k$ to the measurement domain. In the next section different measurement equations are considered to handle various types of association strategies.

3.4 Jointly optimized radar network signal processing

In general, the tracking procedure starts with an association process to combine the established track parameter with the radar sensor or radar network measurements. Errors in the association process will always lead to ghost targets. But the general requirement for automotive applications is to keep the false alarm probability as low as possible, which underlines the importance of the association process for radar networks.

There are several possibilities in radar networks to design this association process in the tracking procedure:

- The target state vector $\vec{t}_k$ measured by the multilateration procedure can be considered directly as a target plot input of the association process. In this case, the input of the Kalman filter describes the same parameters that the internal state vector does. It is characteristic for the plot-to-track association procedure that the measurement equation contains directly the target state vector $\vec{t}_k$ which is influenced by noise $\vec{n}_k^s$ only:

$$\vec{y}_k^s = [t_x, t_y, v_x, v_y]_k^T + \vec{n}_k^s = \vec{t}_k + \vec{n}_k^s. \tag{19}$$

- Alternatively the radar sensor specific measured ranges and velocities $\vec{m}_k^t$ can be used for a track update. In this case the tracking procedure can even be applied in the low target detection situation where the multilateration process cannot be applied. In the range-velocity-to-track association scheme the corresponding measurement equation is based on range and velocity calculations and has a nonlinear analytical structure,

$$\begin{aligned} \vec{y}_k^t &= [r_1, v_1, \cdots, r_4, v_4]_k^T + \vec{n}_k^t \\ &= \vec{m}_k^t + \vec{n}_k^t \\ &= h(\vec{t}_k) + \vec{n}_k^t. \end{aligned} \tag{20}$$

  It has to be noted that the measurement values for range and velocity are not uncorrelated according to the LFMCW measurement described in section 8. As a consequence, the observed measurement errors $\vec{n}_k^t$ can also be considered as correlated random variables for a single sensor's data. For 24GHz pulse radar networks, developed also for automotive applications, a similar idea has been described by a range-to-track association scheme [12], because no velocity measurements are provided in such a radar network.

- Finally for an FMCW radar waveform all the measured single beat frequencies $\vec{m}^f$ can be used directly for the association process and track update. This technique will be called frequency-to-track association. In this case each radar detection and beat frequency measurement can be directly integrated into the tracking process. An explicit and independent calculation of the range-velocity parameter and multilateration processing is not necessary in this case. This joint association process reduces the ghost target probability dramatically and improves the radar network performance.

$$\begin{aligned} \vec{y}_k^f &= [f_{1,1}, f_{2,1}, \cdots, f_{4,1}, \cdots, f_{4,4}]_k^T + \vec{n}_k^f \\ &= \vec{m}_k^f + \vec{n}_k^f \\ &= C \cdot \vec{m}_k^t + \vec{n}_k^f \\ &= C \cdot h(\vec{t}_k) + \vec{n}_k^f \end{aligned} \tag{21}$$

  The vector $\vec{n}_k^f$ describes the unknown additive measurement noise, which is assumed in accordance with Kalman filter theory to be a Gaussian random variable with zero mean and covariance matrix R. Instead of the additive noise term $\vec{n}_k^t$ in equation (20), the errors of the different measurement values are assumed to be statistically independent and identically Gaussian distributed, so

$$R = E\{\vec{n}_k^f \cdot \vec{n}_k^{fT}\} = \sigma_f^2 \cdot I. \tag{22}$$

  The term $E\{\cdot\}$ denotes the expected value and I is the identity matrix. This covariance matrix can be derived from the radar sensor characteristics.

The respective Kalman filter equations for the position correction and prediction steps can now be formulated based on equations (18) and (19), (20) or (21) accordingly for the different mentioned association schemes. Since the measurement equation is nonlinear in case of range-velocity-to-track or frequency-to-track association, the Extended Kalman filter is used for this particular application [16].

In Figure 9 a block diagram of the frequency-to-track processing is given. The association procedures are no longer processed step–wise at three different places in the block diagram compared to the general classical radar network scheme described in Figure 8.

Even a small subset of the maximum possible sixteen beat frequencies is sufficient for track update processing based on the frequency-to-track association scheme. Almost all association errors could be avoided in multiple and extended target situations applying this procedure. This

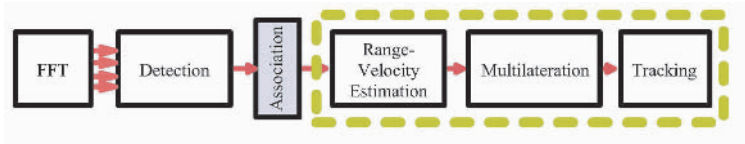


Figure 9. Block diagram of frequency-to-track association.

proposed joint optimization procedure shows increased performance and is additionally quite robust in all situations when some radar sensors in the radar network have low detection probability. The ghost target probability is dramatically reduced.

3.5 Experimental results

As already mentioned, the TUHH experimental car has been equipped with the described 77GHz radar network to validate and prove the efficiency of the derived algorithms. This radar network was used to record measurement data of typical scenarios in real street applications. Based on these recorded data, a comparison of the different signal processing strategies (classical or jointly) for radar network signal processing has been performed.

Typical targets for automotive radar networks are moving cars inside the observation area. Compared with the single radar sensor range resolution of 0.4m a common car cannot be considered any longer as a point but as an extended target. Therefore, each sensor will measure several echo signals in different but closely related range gates for this single car.

Measurements of such extended targets contain and describe all signal processing and association effects discussed in the previous sections.

In the classical signal processing case the target azimuth angle is calculated in the radar network based on multilateration techniques. In this case an extremely high range accuracy of 2cm is required due to the small baseline of radar sensor position inside the network and the required position accuracy. Furthermore, all sensors are observing the car from slightly different aspect angles and can therefore detect different reflection centers. It is obvious that the data association technique becomes very crucial in such situations and the risk of producing ghost targets caused by multiple detections and misassigned measurements in the three independent association schemes is rather high.

On the other hand, the high update rate compared to realistic velocities and accelerations of cars results in a high quality target prediction in the tracking procedure. Therefore, nearest neighbour and gating tech-

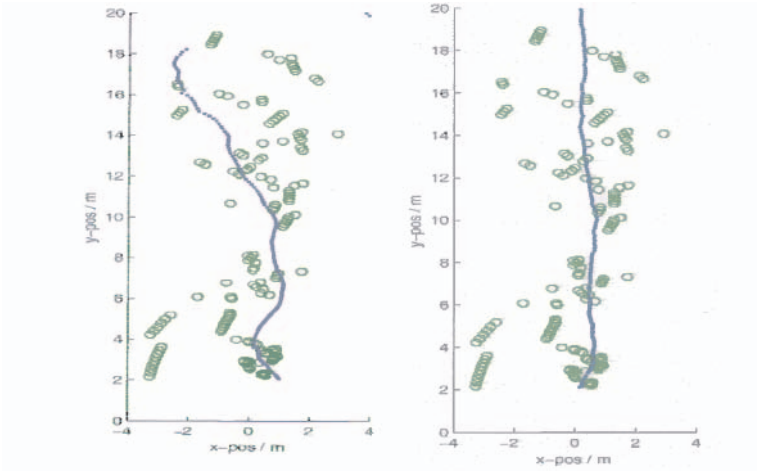


Figure 10. Range-velocity-to-track association with a range gate size of 0.5m and a velocity gate size of 1m/s (left) and frequency-to-track association with a gate size of 1FFT bin (right).

niques with rather small gate sizes can be used for signal processing and association procedures.

As an example, the measurement results of a car driving straight towards the radar-sensor network with a velocity of approximately 5m/s have been analyzed. Figure 10 shows the results of the different signal processing stages. The car can be considered from the single radar sensor point of view as an extended target, which leads to many echo signals with different ranges. Therefore the three association steps in the classical signal processing procedure must be considered quite carefully.

It can be seen that the target positions derived by a pure multilateration procedure (green circles) in the classical signal processing scheme have poor quality and accuracy due to many association errors in the range-velocity measurement inside each sensor and inside the multilateration step. Considering these green dots it is hopeless to establish a plot-to-track association in the tracking procedure.

The situation can be improved by a range-to-track association technique. The blue line in the left part of Figure 10 shows the result of a range-velocity-to track association procedure as described in section 3.4. In this case the target can be tracked over the complete measurement time but with limited accuracy, especially in the azimuth angle. The target position accuracy is increased for short-range positions.

The right part of Figure 10 shows the result of the joint optimization procedure based on a direct frequency-to-track association as de-

scribed in section 3.4. The green circles again show the results of a pure multilateration process as in the left part for the extended target measurement situation. The most obvious difference is improved accuracy in angular estimation of the target position. From Figure 10, the improved performance of the proposed joint optimization process and the frequency-to-track association can be seen.

4. Range CFAR Techniques

The general task of primary radars used in air or vessel traffic control is to detect all targets inside the observation area and to estimate their range, azimuth and radial velocity parameters respectively. The target detection scheme would be an easy task if the echo signal was observed before an empty or statistically completely known noise or clutter signal background. In this case all received echo signal amplitudes would be compared with a fixed threshold, which is based on the noise and clutter statistic only, and targets are detected in all cases when this threshold is exceeded by the echo signal inside the test cell.

But in real radar applications many different noise and clutter background signal situations can occur. The target echo signal practically always appears before a background signal, which is filled with point, area or even extended clutter and additional superimposed noise. Furthermore the location of this background clutter varies in time, position and intensity. Clutter is, in real applications, a complicated time and space variant stochastic process.

All these conditions call for an adaptive procedure in detection and signal processing, operating not with a fixed but with a variable threshold in the detection procedure, to be determined in accordance with the locally observed clutter situation with different range extension, intensity and fluctuation. In a first step of the detection procedure the unknown parameter of a certain statistical background signal is always estimated by analysing the signal inside a fixed window size, which is oriented in the range direction surrounding the radar test cell. The general detection procedure is shown in a block diagram in Figure 11 where the sliding range window is split into two parts, the leading and lagging part surrounding the test cell. Additionally some guard cells are introduced to reduce self-interferences in a real target echo situation.

All data inside the window will be used to estimate the unknown statistical parameters of the background clutter and to calculate the adaptive threshold for target detection. All the background signals, undesired as they are from the standpoint of detection and tracking, are denoted just as “clutter”. The detection procedure has to distinguish

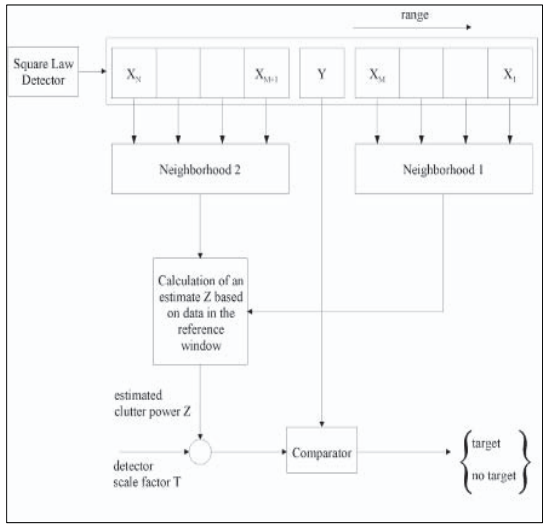


Figure 11. General architecture of range CFAR procedures.

between useful target echoes and all possible clutter situations. Clutter is not just a uniformly distributed sequence of random variables but can be caused in practical applications by a number of different physical sources.

Therefore, the length of the sliding window will be chosen as a compromise based on rough knowledge about the typical clutter extension. To get good estimation performance (low variance) in a homogeneous clutter environment the window size N should be as large as possible. But the window length N must be adapted to the typical range extension of homogeneous clutter areas to fulfill the statistical requirement of identically distributed clutter random variables. In typical air traffic control radars the number of range cells is e.g. between N=16 and 32.

4.1 Radar target detection in noise

In a first simple model for target detection it is assumed that the background clutter can be described by a statistical model in which the different range cells inside the sliding window contain statistically independent identically exponentially distributed (iid) random variables $\{X_1,\ldots,X_N\}$. The probability density function (pdf) of exponentially distributed clutter variables is fully described by the equation:

$$p_0(x) = \quad 1/\mu \; e^{-x/\mu}, \quad x \geq 0. \tag{23}$$

In this ideal case it is assumed that μ is a known parameter which describes the expected value of the exponentially distributed random variables. The variance of the random variable X in this case is $\sigma^2 = \mu^2$. The false alarm probability (P_{fa}) depends on the noise and clutter statistic only and a certain threshold S will be calculated analytically as follows:

$$P_{fa} = \int_{S}^{\infty} p(x)dx \tag{24}$$

$$S = T \cdot \mu. \tag{25}$$

The factor T of the threshold S can be described for a given false alarm probability (P_{fa}) in this case analytically by the equation:

$$T = \ln \frac{1}{P_{fa}}. \tag{26}$$

The non-fluctuating target amplitude statistic can be described by the Rician pdf:

$$p_1(x) = \begin{cases} \frac{x}{\sigma_0^2} e^{-\frac{x^2+c^2}{2\sigma_0^2}} I_0\left(\frac{xc}{\sigma_0^2}\right), & x \geq 0; \\ 0, & \text{otherwise.} \end{cases} \tag{27}$$

Therefore the detection probability (P_d) for a non-fluctuating target in homogeneous noise is:

$$P_d = \int_{T\mu}^{\infty} p_1(x)dx. \tag{28}$$

The general objective of all radar detection procedures is to get a constant false alarm rate (CFAR) due to the fact that the test cell almost always contains clutter and noise and only in a very few cases contains radar target echo signals. The statistical model and general detection procedure, in which the detector is fixed only with regard to the noise and clutter statistic and independently to the target statistic, has been developed by Neyman and Pearson.

But in real radar applications the average noise and clutter power level (μ) is unknown and must be estimated in the detection procedure first. This is done by several published CFAR procedures, which will be discussed in this section, where each specific CFAR technique is motivated by assumptions about a specific background signal or target signal model.

4.2 Range CFAR procedures with sliding window techniques

Each developed and published CFAR technique refers implicitly to a certain background clutter or even target model. Therefore, in the following these assumptions will be described explicitly and the different CFAR procedures considered will be compared in some clutter situations. The amplitude in each individual range cell is tested. Therefore a window of fixed size is applied to each range cell inside the full range coverage using a sliding technique. The amplitudes in the leading and lagging reference cells are used in a signal processing procedure to estimate the unknown statistical parameters of the clutter background signal. Based on these estimated parameters the detection threshold will be calculated. The different range CFAR techniques differ in the way to estimate the statistical parameters. In the following the background signal model and the resulting motivation for each CFAR technique will be described. Furthermore each range CFAR technique will be applied in four different but (for radar applications) characteristic signal situations which consist of: pure noise; local clutter; single target in noise; and two targets respectively. From these examples the characteristic behavior of each CFAR procedure can be seen clearly.

Cell averaging CA CFAR. In this first signal model it is assumed that the clutter and noise background at the output of a square law detector can be described by statistically independent and identically distributed (iid) exponential random variables with a single exception: the average clutter plus noise power level is unknown. The optimised signal processing technique in this situation, from a statistical point of view, is to calculate an estimation of the clutter power level just by applying the arithmetic mean to the received amplitudes inside the considered window.

$$Z = \left[\frac{1}{N} \sum_{i=1}^{N} X_i \right]. \tag{29}$$

The arithmetic mean Z has excellent estimation performance. The estimation Z is unbiased, which means

$$E\left[\frac{1}{N} \sum_{i=1}^{N} X_i \right] = \frac{1}{N} \sum_{i=1}^{N} E\left[X_i\right] = \mu$$

and shows additionally a minimum estimation variance.

$$Var\left[\frac{1}{N}\sum_{i=1}^{N}X_i\right] = \frac{1}{N^2}\sum_{i=1}^{N}Var\left[X_i\right] = \frac{1}{N}\mu^2. \tag{30}$$

The false alarm rate is given by [4] as:

$$P_{fa} = P(Y \geq T_{CA} \cdot Z) = (1 + T_{CA})^{-N}. \tag{31}$$

If this estimation procedure is applied to the random variables inside the range window this CFAR procedure is called "cell averaging," CA-CFAR. The statistical performance is excellent if the assumptions of homogeneous clutter inside the reference window are fulfilled in the statistical model and in the real world application. It is not clear to the author who first analyzed and published this CA-CFAR idea, but Nitzberg [17] published a paper in 1978 analyzing CA-CFAR for fluctuating targets. To demonstrate the general CFAR characteristic some typical signal situations are generated which are considered to be characteristic for radar applications. These are: a pure noise background signal inside the full range coverage; a local clutter area over 10 adjacent range gates superimposed with noise; a single target and a double target situation (20 dB SNR each) superimposed with noise. The clutter power was chosen 13 dB above noise power. Figure 12 shows the resulting threshold S in these noise, clutter and target situations when the CA-CFAR procedure is applied.

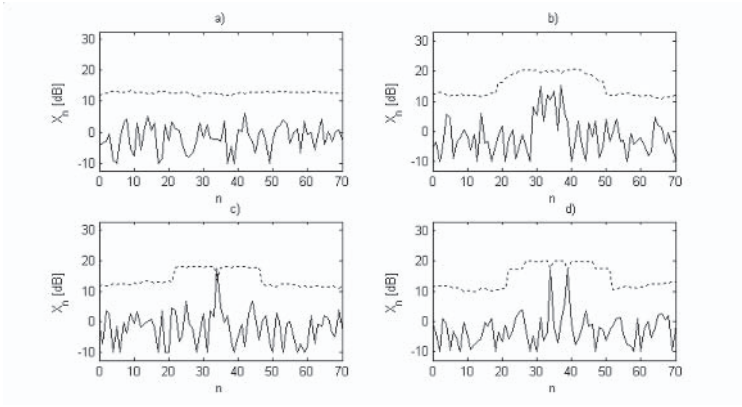


Figure 12. CA-CFAR (N=24), $P_{fa} = 10^{-6}$

Figure 12 shows the appropriately calculated detection threshold in the pure noise situation (12 a). This homogeneous noise model with unknown noise power was the motivation to develop CA-CFAR. The

detection procedure adapts quite well in the local clutter with some losses (increased P_{fa}) at the clutter edges (12 b). The characteristic behavior in target situations is not acceptable (12 c). In the two target situation both targets are not even detected by CA-CFAR due to the resulting masking situation (12 d).

Cell averaging with greatest of CAGO-CFAR. The clutter and noise signals are varying in time and position and the average clutter power level can fluctuate in different range areas and range cells. If a CA-CFAR is applied in the radar detector it may happen that the sliding window is located in the transition between a pure noise and strong clutter area with different average power level, as shown in Figure 12c for example. From a statistical point of view this means that the random variables inside the sliding window are no longer identically distributed but have different expected values μ in their individual statistics. CA-CFAR leads in such cases to an increase in false alarm rate (P_{fa}), which is unacceptable for practical applications and requirements.

Therefore the clutter model is extended and non-homogeneous clutter situations are integrated, such as in typical transition areas between noise and beginning clutter areas. Referring to such realistic clutter situations a new CFAR procedure has been designed by Vilhelm G. Hansen in his well-known paper [18]. In order to demonstrate the advantage of this CFAR technique it is important to recall the CA-CFAR procedure. In this case the arithmetic mean is calculated in the leading and lagging part of the range window and both values are summed up; see Figure 11.

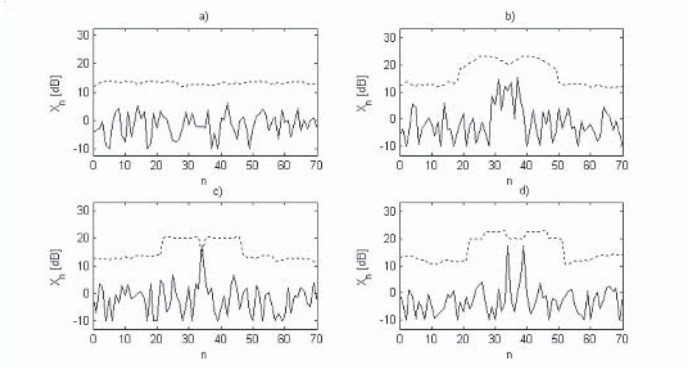


Figure 13. CAGO-CFAR (N=24), $P_{fa} = 10^{-6}$

In the extended CFAR case a cell averaging technique is used in each part of the range window but with a "greatest of" selection (CAGO-CFAR).

$$Z = Max\left(\left[\frac{2}{N}\sum_{i=1}^{N/2} X_i\right], \left[\frac{2}{N}\sum_{i=N/2+1}^{N} X_i\right]\right). \tag{32}$$

If CAGO-CFAR is applied in such typical clutter situations the false alarm rate is reduced at the clutter edges but the detection rate is simultaneously reduced slightly, see Figure 13.

CAGO-CFAR shows a clear advantage in typical transition areas between pure noise and strong clutter or between two different clutter regions. But it simultaneously reduces the sensitivity in target detection in a homogeneous clutter situation. This is a compromise, which always occurs in radar detection due to the variety of different noise and clutter background situations, which can occur in real radar applications. It is a useful extension of CA-CFAR but shows the same characteristics in multiple target situations.

Moshe Weiss [25] developed an extension of CAGO-CFAR to get better performance in multiple target situations and designed cell averaging CFAR with smallest of (CASO-CFAR) procedures and analysed the performance especially in multiple target situations.

These CFAR procedures suffer from the fact that they are specifically tailored to the assumption of uniform and homogeneous clutter inside the reference window. Based on these assumptions, they estimate the unknown clutter power level using the unbiased and most efficient arithmetic mean estimator. Improved CFAR procedures should be robust with respect to different clutter background and target situations. Also in non-homogeneous situations CFAR techniques should remain able to provide reliable clutter power estimations.

OS-CFAR. Applying CAGO-CFAR in the detection procedure brings several advantages in clutter transition areas. But the CA and CAGO-CFAR detection procedures behave very sensitively in multiple target situations and show pure performance. This observation has been described in [14]. It was shown that even in a two-target situation it is possible that both targets are masked by each other. Weak targets in the neighbourhood of strong targets are masked in almost all cases, which reduces the range resolution and is not acceptable in real radar applications. The "ordered statistic" OS-CFAR has in all these multiple target situations a much better performance compared to CA and CAGO-CFAR procedures.

OS-CFAR is not based on the assumption of homogeneous clutter inside the reference window. Therefore the window length can be extended in the OS-CFAR case without any disadvantages. For comparison, OS-CFAR with $N = 24$ will even outperform classical CA-CFAR with $N = 16$. This is an important advantage of the OS-CFAR procedure.

The general idea of an ordered statistic is technically simple. To estimate the average noise and clutter power a single rank $\mathrm{X}_{(k)}$ of the ordered statistic is used instead of the arithmetic mean. In this case a very few large amplitudes in the sliding window have a very small effect on the estimation result. OS-CFAR is robust in multiple target situations. The threshold is hardly influenced by a second or third target inside the window.

All amplitudes inside the sliding window are sorted according to increasing magnitude.

$$X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(N)}$$
$$Z = X_{(k)}. \tag{33}$$

The pdf of the kth value of the ordered statistic is given by

$$P_{X_{(k)}}(x) = p_k(x) = k \left(\begin{array}{c} N \end{array} \right) (1 - P_X(x))^{N-k} \, (P_X(x))^{k-1} \, p_x(x). \tag{34}$$

Thus the pdf of the kth value of the ordered statistic for exponentially distributed random variables is given by

$$P_{X_{(k)}}(x) = p_k(x) = k/\mu \left(\begin{array}{c} N \end{array} \right) \left(e^{-x/\mu}\right)^{N-k+1} (1 - e^{-x/\mu})^{k-1}. \tag{35}$$

The relation between P_{fa} and T_{OS} can be calculated by combining (2) and (13)

$$P_{fa} = k \binom{N}{k} \frac{(k-1)! \; (T_{OS} + N - k)!}{(T_{OS} + N)!}. \tag{36}$$

The importance of this case is that OS-CFAR can be analytically analysed without any approximations. Furthermore the resulting scaling factor T_{OS} is completely independent of μ. Figure 14 shows the typical behaviour of OS-CFAR in clutter edge and multiple target situations. The threshold follows the clutter contour with a certain safety distance. In two target situations the threshold is more or less unchanged compared with a pure noise situation.

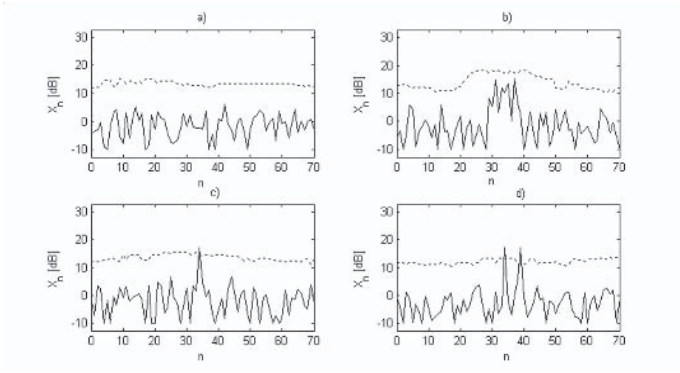


Figure 14. OS-CFAR (N=24), k=3/4*N, $P_{fa} = 10^{-6}$

Nadav Levanon [21, 22] has applied OS-CFAR to a Weibull distributed background signal and described the results analytically. Blake [26] analysed OS-CFAR in non-uniform clutter. Weber and Haykin [24] have extended OS-CFAR to a two parameter distribution with variable skewness.

In [28] the performance of OS-CFAR in a 77GHz radar sensor for car application is examined. In the automotive radar application case multiple target situations occur almost always.

OSGO-CFAR. An extension of OS-CFAR has been developed by He You in his paper [20]. In this case an ordered statistic is applied in the left and right window part separately followed by a greatest of selection. Therefore, the procedure is called OSGO CFAR. The computational complexity is reduced in this case and the detection performance shown in Figure 15 is good.

Gaspare Galati et. al. compared the performance of OS-CFAR and OSGO CFAR in the presence of different backgrounds [23]. He found that the OSGO method suffers only a small additional loss with respect to the OS. In a non-homogeneous background with clutter edges it even shows superiority in the control of the false alarm probability.

Censored CFAR. Richard and Dillard [19] have proposed a CFAR procedure which is based on CA-CFAR but is already close to the general OS-CFAR idea when they are calculating the largest m values inside the sliding window and excluding these values from the arithmetic mean calculation. This step makes modified CA-CFAR less sensitive in mul-

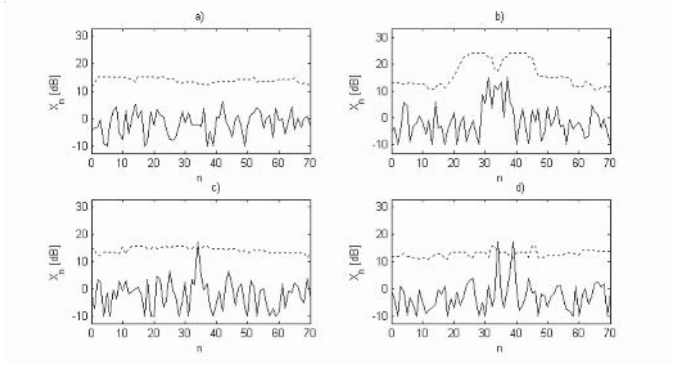


Figure 15. OSGO-CFAR (N=24), k=3/4*N, $P_{fa} = 10^{-6}$

tiple target situations. In their paper they analysed a Censored-CFAR detector for m equals 1 and 2.

The resulting false alarm probability for m=1 can also be calculated independently of the actual noise level:

$$P_{fa} = (N-1)\cdot\left\{\frac{N}{N-1}\sum_{k=1}^{N-1}\binom{N-1}{k}(-1)^{k+1}(1+k+T)^{-1}\right\}^{N-1} . \quad (37)$$

Figure 16 demonstrates the performance of a censored-CFAR detector with m=1. Compared with CA-CFAR the results here are better, because in all scenarios the targets are detected. But compared with OS-CFAR the threshold in the 2 target scenario is still too high in the neighborhood of the targets. Ritcey [29] studied the performance of this method for multiple target situations.

WCA-CFAR. If a-priori information about the target position is made available by the tracking system the adaptive threshold can be lowered. A CFAR method called weighted CA-CFAR uses this and was proposed by Barkat, Himonas and Varshney [30].

This method separates the window cells into a leading and a lagging part. Before the mean values of these parts are averaged, they are weighted by the factors α and β. Optimum values for α and β are calculated in accordance with the level of interference of present targets, so that a constant false alarm rate and a high detection probability can be guaranteed.

Other publications. Many additional papers have been published which are based on the four fundamental procedures – CA-, OS-, GO-

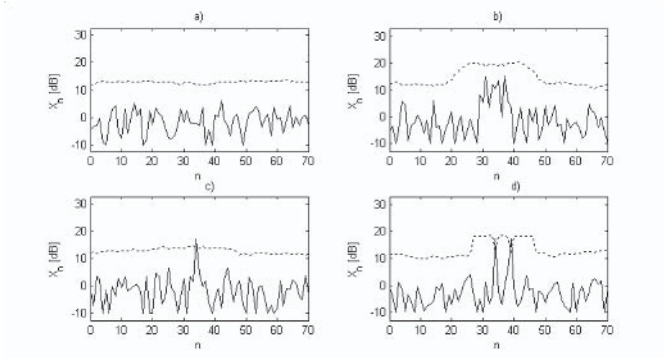


Figure 16. Censored-CFAR (N=24), k=1, $P_{fa} = 10^{-6}$

and SO-CFAR. In these publications either new combinations and modifications of existing procedures or the performance in new environments are analysed.

The main relevant models for radar detection are

- Pure clutter or noise situations with fully homogeneous statistic
- Transition between noise and clutter or between two clutter areas with different average power
- Non-homogeneous clutter
- Correlated clutter
- Different clutter amplitude statistics (Rayleigh, Weibull, ...)
- Different target amplitude statistics (Rice, Swerling models, ...)
- Single target situations
- Two or multiple target situations.

A simple quantitative comparison between different CFAR procedures in pure clutter and noise situations, which is the most important situation, can be calculated using the average detection threshold (ADT) [14].

In this case the expectation of all calculated thresholds is calculated and can be used for system comparison.

5. Conclusion

The objective of this chapter was to discuss some important contributions for radar system design and digital radar signal processing. The focus was on waveform design in general and on automotive applications in particular. Target detection is an important issue for all radar systems. Therefore some range CFAR procedures have been discussed which can be applied especially in multiple target situations to avoid any masking situations. Additionally some new results have been discussed for target recognition.

References

[1] Artis, Jean-Paul; Henrio, Jean-François: *Automotive Radar Development Methodology*, International Conference on Radar Systems, Brest, France, 1999.

[2] Klotz, Michael; Rohling, Hermann: *A high range resolution radar system network for parking aid applications*, International Conference on Radar Systems, Brest/ France, 1999.

[3] Klotz, Michael; Rohling, Hermann: *24 GHz Radar Sensors for Automotive Applications*, International Conference on Microwaves and Radar, MIKON2000, Wroçław/ Poland, 2000.

[4] Rohling, Hermann; Meinecke, Marc-Michael; Mende, Ralph: *A 77 GHz Automotive Radar System for AICC Applications*, International Conference on Microwaves and Radar, MIKON98, Workshop, Kraków/ Poland, 1998.

[5] Rohling, Hermann; Meinecke, Marc-Michael; Klotz, Michael; Mende, Ralph: *Experiences with an Experimental Car controlled by a 77 GHz Radar Sensor*, International Radar Symposium, IRS98, München, 1998.

[6] Stove, A. G.: *Linear FMCW radar techniques*, IEE Proceedings-F, Vol. 139, No. 5, Oct. 1992.

[7] H. Rohling, A. Höß and U. Lübbert: *Multistatic Radar Principles for Automotive RadarNet Applications*, IRS 2002 International Radar Symposium, Bonn, Germany, 2002, pp. 181–185.

[8] M. Schiementz, F. Fölster and H. Rohling: *Angle Estimation Techniques for different 24GHz Radar Networks*, IRS 2003 International Radar Symposium, Dresden, Germany, 2003, pp. 405–410.

[9] A. Hoess et al.: *The RadarNet Project*, 7th ITS World Congress, Torino, November 2000.

[10] Brian Ricket: *A Vision Of Future Applications For An Automotive Radar Network*, WIT 2004, 1st International Workshop on Intelligent Transportation, Hamburg, Germany, 2004, pp. 117-121.

[11] N. Levanon: *Radar Principles*, Wiley & Sons, New York, 1988.

[12] D. Oprisan and H. Rohling: *Tracking Systems for Automotive Radar Networks*, IEE Radar 2002, Edinburgh, UK.

[13] H. Rohling, F. Fölster, F. Kruse and M. Ahrhold: *Target Classification Based on a 24GHz Radar Network*, Radar 2004 Conference, Toulouse, France.

[14] H. Rohling: *Radar CFAR Thresholding in Clutter and Multiple Target Situations*, IEEE Trans. Aerosp. Electron. Syst. 19 No. 4, 1983, pp. 608-621.

[15] M. Klotz: *An Automotive Short Range High Resolution Pulse Radar Network*, PhD thesis, TU Hamburg-Harburg, 2002.

[16] S. Blackman and R. Popoli: *Design and Analysis of Modern Tracking Systems*, Artech House, Boston, 1999.

[17] Nitzberg R.: *Analysis of the Arithmetic Mean CFAR Normalizer for Fluctuating Targets*, IEEE Transactions on AES, 1, pp. 44-47, 1978.

[18] Hansen, V.G., Sawyer, J.H.: *Detectability loss due to greatest of selection in a cell-averaging CFAR*, IEEE Transactions on AES, 16, pp. 115-118, 1980.

[19] Richard, J.T., Dillard, G.M.: *Adaptive Detection Algorithms for Multiple Target Situations*, IEEE Transactions on AES, 4, pp. 338-343, 1977.

[20] He You: *Performance of Some Generalized Modified Order Statistics CFAR Detectors with Automatic Censoring Technique in Multiple Target Situations*, IEE Proceedings F, 131(4), pp. 205-212, 1994.

[21] Levanon, N.: *Detection loss due to interfering targets in ordered statistic CFAR*, IEEE Transactions on AES, (11/1988), pp. 678-681, 1988.

[22] Levanon, N., Shor, M.: *Order statistic CFAR for Weibull background*, IEE Proceedings, F, 137, 3 (6/1990), pp. 157-162, 1990.

[23] Di Vito, A., Galati, G., Mura, R.: *Analysis and comparison of two order statistics CFAR systems*, IEE Proceedings, F, 141, 2 (4/1994), pp. 109-115, 1994.

[24] Weber, P., Haykin, S.: *Ordered statistic CFAR processing for two-parameter distribution with variable skewness*, IEEE Transactions on AES, AES-21, 6 (11/1985), pp. 819–821, 1985.

[25] Weiss, M.: *Analysis of some modified cell-averaging CFAR processors in multiple-target situations*, IEEE Transactions on AES, 18, pp. 102-114.

[26] Blake, S.: *OS-CFAR theory for multiple targets and nonuniform clutter*, IEEE Transactions on AES, 24, 6 (11/1988), pp. 785-790, 1988.

[27] Gini, F., Farina, A., Greco, M.: *Selected List of References on Radar Signal Processing*, IEEE Transactions on AES, 37, 1 (1/2001), pp. 329-359, 2001.

[28] Rohling, H., Mende, R.: *OS CFAR performance in a 77GHz Radar Sensor for Car Application*, CIE International Conference on Radar, Peking, China, 1996.

[29] Ritcey, J. A.: *Censored mean-level detector analysis*, IEEE Transactions on Aerospace and Electronic Systems, AES-22, 3 (July 1986), pp. 443-454, 1986.

[30] Barkat, M., Himonas, S.D., Varshney, P.K.: *CFAR detection for multiple target situations*, IEE Proceedings, F, 136, 5 (10/1989), pp. 193-209, 1989.

TOMOGRAPHY OF MOVING TARGETS (TMT) FOR SECURITY AND SURVEILLANCE

Michael C. Wicks, Braham Himed
Air Force Research Laboratory Sensors Directorate
26 Electronic Parkway Rome, NY 13441

Harry Bascom, John Clancy
L3-Communications Analytics Corporation
Advanced Technologies & Law Enforcement Sector
1300 B Floyd Avenue Rome, NY 13440

Abstract In order to improve upon automated sensor performance for security applications in public and private settings, numerous alternative sensor designs have been developed to provide affordable and effective detection and identification performance. Radio frequency (RF) sensors offer a balanced approach to system design for a wide variety of geometries and threat targets. These threat targets include persons carrying weapons and explosives, portable containers with contraband including cargo boxes, suitcases, and briefcases and fixed structures including building or underground facilities harboring criminals, terrorist or enemy combatants. In order to achieve the resolution required for the detection and identification of threat targets, separation of interference from the target response is essential. High bandwidth offers a conventional approach to high resolution sensing of the threat. An alternative approach, one based upon wide angular bandwidth (spatial diversity), is presented here.

This chapter addresses the issue of spatial diversity in radar applications. There has been an increased need for information via radio frequency (RF) detection of airborne and ground targets while at the same time the electromagnetic spectrum available for commercial and military applications has been eroding. Typically, information concerning ground and air targets is obtained via monostatic radar. Increased information is often equated with increased bandwidth in these monostatic radar systems. However, geometric diversity obtained through multistatic radar operation also affords the user the opportunity to obtain additional information concerning these targets. With the appropriate

Jim Byrnes and Gerald Ostheimer (eds.),
Advances in Sensing with Security Applications, 323–339.

signal processing, this translates directly into increased probability of detection and reduced probability of false alarm. In the extreme case, only discrete Ultra Narrow Band (UNB) frequencies of operation may be available for both commercial and military applications. As such, the need for geometric diversity becomes imperative.

Keywords: tomography; moving targets; spatial diversity; ultra narrow band (UNB); monostatic radar; multistatic radar.

1. Introduction

The electromagnetic spectrum available for commercial and military applications is continuously being eroded while the need for increased information via RF detection of threat targets is increasing. Typically, military information concerning ground and air targets is obtained via monostatic radar. Increased information is often equated with increased bandwidth in these monostatic radar systems. However, geometric diversity obtained through multistatic radar operation also affords the user the opportunity to obtain additional information concerning these targets. With the appropriate signal processing, this translates directly into increased probability of detection and reduced probability of false alarm. In the extreme case, only discrete Ultra Narrow Band (UNB) frequencies of operation may be available for both commercial and military applications. As such, the need for geometric diversity becomes imperative. In addition, geometric diversity improves target position accuracy and image resolution which would otherwise remain unavailable.

However, coherent signal processing of a multitude of UNB radar signals emanating from and received by geometrically diverse sites requires more than the simple processing (multiplication and addition) which forms the basis for synthetic aperture radar (SAR) or even moving target indication (MTI). Classical tomographic signal processing offers one basic (albeit sub-optimum) approach to the processing of multi-frequency UNB radar data collected via geometrically diverse transmit and receive sites. Tomography is applicable to radar even though the basic mechanism for re-radiation is the back scatter and forward scatter of electromagnetic waves by moving objects, and not reflection and transmission as in tomography. When modest increases in bandwidth are permitted at each transmitter site, further complications arise in the coherent signal processing required for target detection and interference suppression. While classical tomography (designed to operate under the monochromatic assumption) is applicable as a baseline, this mathematical formulation has been modified and extended to optimize target detection and interference rejection in the Tomography of Moving Tar-

gets. It is the objective of this chapter to present a practical approach to target detection and interference rejection via tomographic processing of geometrically diverse, multi-frequency, multistatic, UNB radar data. The emphasis in this chapter is on the detection of slower moving weak target returns.

In classical radar, frequency diversity offers one method to obtain additional information about threat targets. With the most basic form of frequency diversity, namely increased bandwidth, high range resolution is afforded to the user. With high range resolution comes increased target-to-clutter ratio (assuming the target is not over-resolved), while target-to-noise is unavoidably reduced since increased bandwidth results in additional unwanted thermal noise competing with and potentially masking weak target returns. Geometric diversity also offers the potential for increased resolution, and is a dual to frequency diversity (increased bandwidth) in classical monostatic radar. In the extreme case, 360^o of geometric diversity (across a large number of sensor sites) offers sub-wavelength resolution, even under the monochromatic assumption. Operating with UNB radar signals permits a substantial reduction in thermal noise power as well, improving overall detection performance. Here, sophisticated tomographic signal processing is required to extract a moving target from clutter. In this chapter, the Tomography of Moving Targets is developed and demonstrated for geometrically diverse, multi-frequency, multistatic, UNB radar. Additionally, only moderately directive broad beam antennas (approximately 60^o) are used to radiate and receive UNB signals for the Tomography of Moving Targets, unless otherwise justified by analysis.

A baseline design and approach has been developed to demonstrate the Tomography of Moving Targets. Preliminary simulations and analysis have been performed indicating how well this approach addresses the stated goals of increased target detection/identification and improved target location via the analysis of geometrically diverse, multi-frequency, multistatic UNB radar data. The number and locations of transmitter and receiver sites, and the UNB frequencies are selected heuristically.

2. Tomography Concept and Framework

The TMT concept leverages the spatial or geometric diversity of a multistatic ground based 'netted' radar to deliver high resolution MTI. The TMT concept provides the resolution of conventional wideband MTI radars, while using UNB signals. These UNB signals are particularly attractive with consideration to the ongoing spectral erosion due to the escalation in wireless. In this TMT research, we are considering both

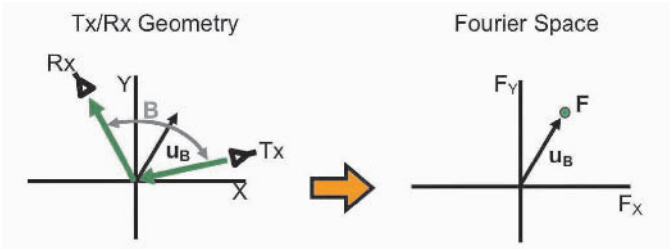


Figure 1. Sensor Geometry and Fourier Space Relationship.

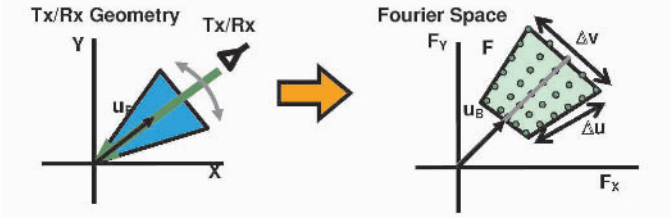


Figure 2. Fourier Space Sampling and Scene Resolution.

ground and airborne MTI applications. Some sites have collocated transmitters and receivers, while other sites are receive only. By locating the sites in a somewhat random manner, the geometric diversity is enhanced.

The radar data samples are mapped onto a polar grid in the spatial Fourier domain. The positions of the transmitter and receiver along with the signal's instantaneous frequency determines the Fourier space sample position, as given by equation 1. This relation is illustrated in figure 1. A bistatic sensor configuration is shown with a bistatic angle (B), $\mathbf{u}_B$ is the bistatic bisector. This geometry and signal frequency maps into the Fourier space sample given by the vector $\mathbf{F}$. As shown, the sample position lies along the bistatic bisector with a magnitude proportional to the instantaneous frequency scaled by $\cos B/2$.

$$\mathbf{F} = \frac{4\pi f}{c} \cos \frac{B}{2} \mathbf{u}_B \tag{1}$$

where f is the frequency, c is the speed of light, B is the bistatic angle, and $\mathbf{u}_B$ is the bistatic bisector unit vector.

Figure 2 illustrates typical Fourier space sampling as provided by monostatic SAR. As shown, the samples in the radial dimension straddle a term proportional to the carrier frequency and have an extent proportional to the signal bandwidth. Samples in the angular dimension correspond to pulse numbers in the coherent processing interval. In

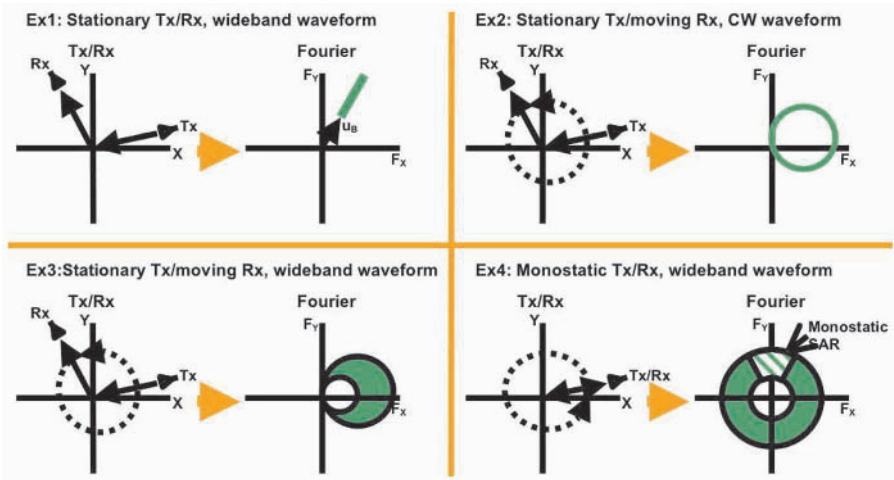


Figure 3. Sensor Geometry and Fourier Space Relationship Examples.

the monostatic case, the angular extent of the samples is the same as the angular aperture created by the synthetic aperture.

Recall that the image resolution, δ_{DOWN} and δ_{CROSS}, is inversely proportional to the size of the region of Fourier space sampled, as given by 2, where Δu and Δv are the sizes of the sides of the sampled region. The unambiguous scene size is inversely proportional to the Fourier space sampling frequency.

$$\delta_{\text{DOWN}} = \frac{2\pi}{\Delta u}; \quad \delta_{\text{CROSS}} = \frac{2\pi}{\Delta v} \quad . \tag{2}$$

In the spatial Fourier domain, radial band limiting is due to the finite bandwidth of the transmitted pulse while angular band limiting is due to the finite diversity of look angles. With variations of frequency and angular diversity, the spatial Fourier domain can be sampled in a variety of ways. This spatial Fourier domain sampling impacts the resulting image's resolution. Higher resolution is achieved with greater diversity, be it frequency, angular or some combination of both. Image resolution is inversely proportional to the size of the region of Fourier space sampled. Figure 3 illustrates, by way of four examples, how different bistatic geometries and waveforms map into Fourier space.

In example 1, a fixed bistatic geometry and a wideband waveform result in Fourier space sampling along a radial line, at the bistatic bisector. In example 2, a fixed frequency (CW) waveform is used as the receiver is moved in a circle about the origin. The resulting Fourier space sampling is a circle who's center is offset from the origin. In example 3, a

wideband waveform is used as the receiver is moved in a circle about the origin. The resulting Fourier space sampling is a combination of the results of examples 1 and 2. For completeness, example 4 shows the case of a wideband waveform and monostatic geometry. The resulting Fourier space sampling has a donut shape. The typical monostatic SAR doesn't fly a circular flight path around a scene, but instead flies a straight line path, the resulting Fourier space sampling is highlighted as a wedge of the donut.

SAR and tomography may be viewed in terms of image reconstruction from a bandlimited region in 2-D Fourier space. Resolution, for both SAR and tomography, is a function of the bandwidth available in the 2-D Fourier space. The conventional SAR resolution formulas are approximations to this, for the limited cases of small apertures and percent bandwidths. Narrowband, wide angle tomographic imaging achieves the resolution of wideband, narrow angle SAR systems by trading frequency for spatial diversity. The resolution limit for this system is practically about one third of a wavelength. A comparison of Fourier space sampling provided by mono, bi and multi-static SAR and their corresponding resolution is illustrated in figure 4. The colored area represents the region of Fourier space sampled. The simplified resolution formulas for small aperture and percent bandwidth are given for the mono and bistatic cases. For the bistatic case, assume a pseudo-monostatic geometry with the transmitter in a fixed position and the receiver forming the same aperture as in the monostatic case. The angular sampling is compressed in the bistatic case, compared to the monostatic case, due to sampling occurring on the bistatic bisector. This results in loss in cross range resolution. In the multistatic case, a circular region could potentially be sampled. The radius of this circle is proportional to the highest frequency used and the resulting image has a resolution approaching a third of a wavelength at this frequency.

It is interesting to note that the resolution result for the monostatic SAR case can be derived using the tomography/Fourier space or the radar range/Doppler principles. However, the resolution result for the multistatic SAR is easily understood using the tomography/Fourier space principles. To compare the resolution capabilities of these sensor configurations, consider a wideband monostatic SAR with a 50% bandwidth. It has a range resolution of λ and a pixel area of λ^2. Multistatic SAR with a range resolution of $\frac{\lambda}{3}$ has a pixel area of $\frac{\lambda^2}{9}$, yielding a 9.5 dB improvement over monostatic.

Consider a geometry where multiple transmitters and receivers are positioned on circle, surrounding the region to be imaged. Each transmitter with all receivers creates a multistatic geometry with Fourier

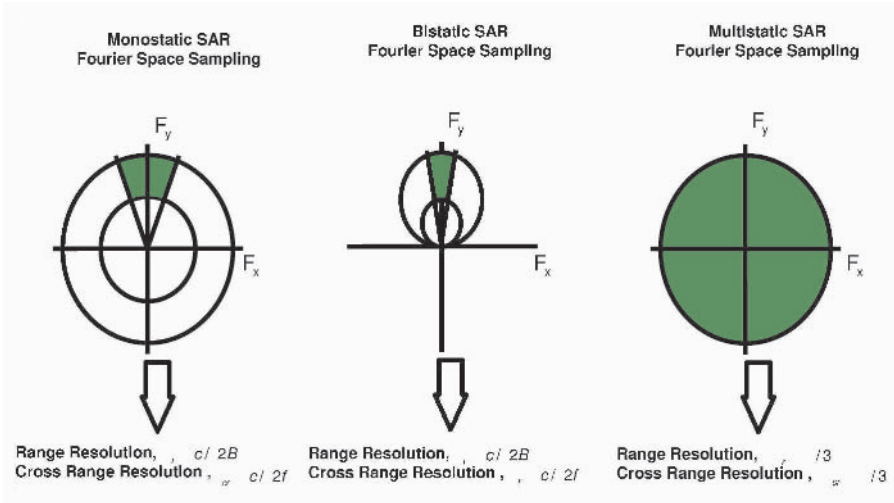


Figure 4. Comparison of Fourier Space Sampling.

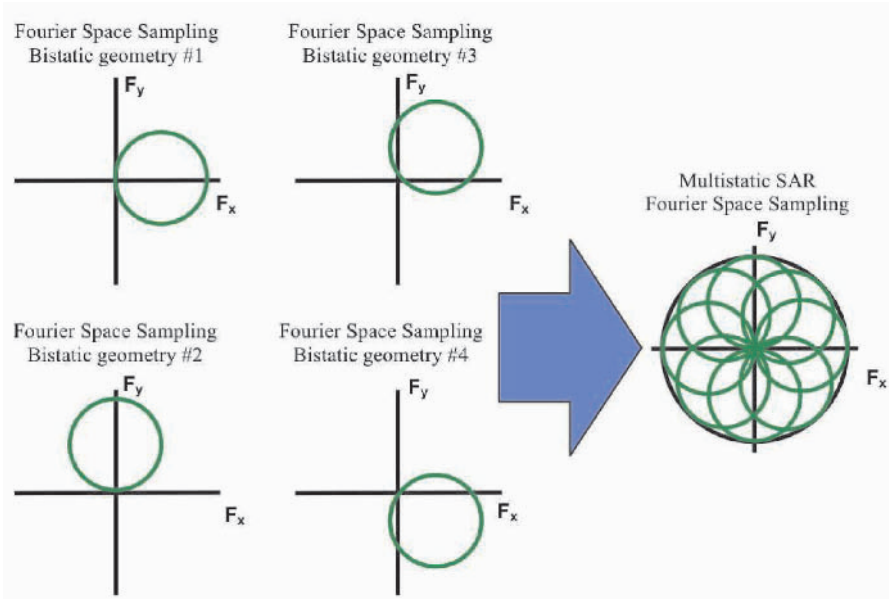


Figure 5. Multistatic Fourier Space Sampling.

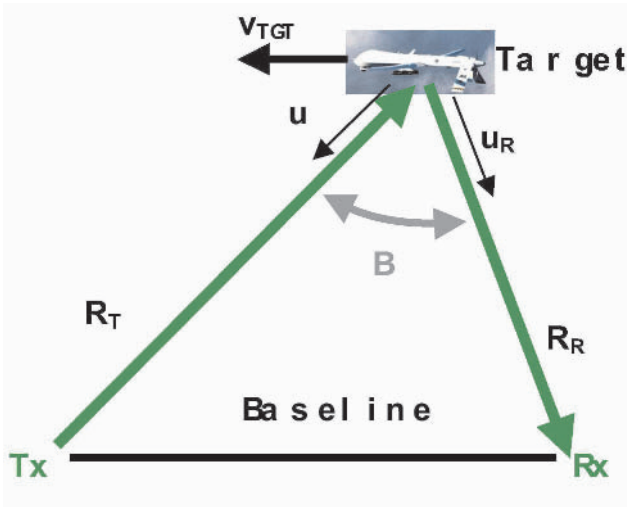


Figure 6. Bistatic Geometry.

space sampling as shown in figure 5. In this approach, it is assumed that the transmitted signal is narrowband. By piecing together the regions sampled by each multistatic transmitter/receiver combination, the result is a multistatic Fourier space sampling.

3. Bistatic Geometry and Observables

The bistatic geometry discussed in this section is shown in figure 6. The round trip path causes the received target return to be time delayed (t_d) from the transmitted signal, given by equation 3, where R_T and R_R are the distances from the target to the transmitter and receiver. For the case of stationary, ground based, transmitter and receiver, the Doppler frequency (f_D) of a target return is due to the motion of the target. The target Doppler is computed as the sum of the target's velocity vector (v_{TGT}) dot product with unit vectors pointing from the target to the transmitter (u_T) and receiver (u_R) as given by equation 4.

$$\mathbf{t}_d = \frac{\mathbf{R}_T + \mathbf{R}_R}{\mathbf{c}}, \tag{3}$$

$$\mathbf{f}_D = \frac{\mathbf{v}_{\text{radial}}}{\lambda} = \frac{(\mathbf{v}_{\text{TGT}} \cdot \mathbf{u}_T + \mathbf{v}_{\text{TGT}} \cdot \mathbf{u}_R) \cdot f}{\mathbf{c}}. \tag{4}$$

Previously, tomography applied to radar has been limited to SAR applications. In this chapter, the tomographic paradigm is extended to MTI. When considering multistatic geometries and moving targets, the issue arises of a target having a different Doppler for each of the trans-

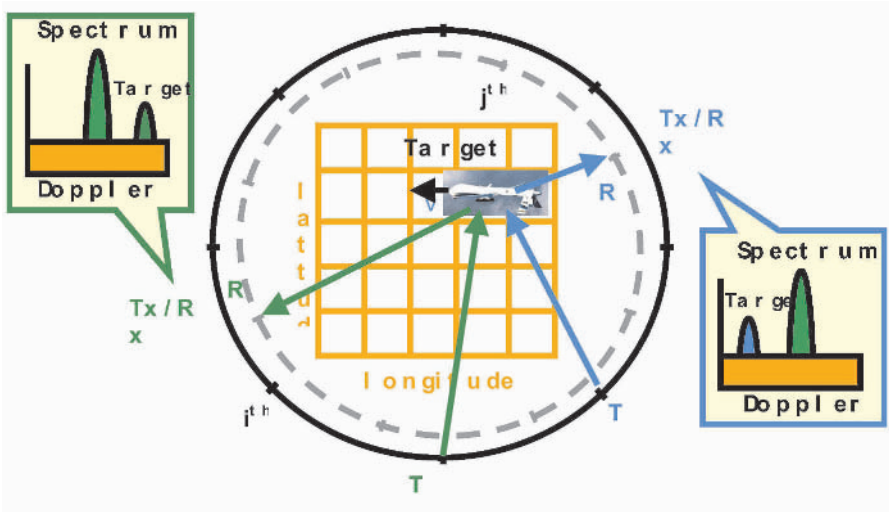


Figure 7. Multistatic MTI — Target Doppler Issue.

mit/receive pairs, as illustrated by figure 7. Conventional processing, employing FFTs for spectral analysis, results in an overwhelming confusion factor. In response to this issue, a matched filter processing (MFP) algorithm is developed.

4. Matched Filter Processing (MFP)

MFP has its origin in SAR image reconstruction and is considered a spatial domain image reconstruction technique. It implements a 'matched filter' for each pixel of a scene. This matched filter simply attempts to replicate the signal's expected delay and Doppler, which can be viewed as a 'steering' vector. The matched filter and received signal is tested for correlation. The extension to moving targets involves much more work. For each scene pixel, matched filters are built for a range of hypothesized target velocities (speeds and headings). In adaptive processing, a Doppler steering vector is used. In MFP, the Doppler steering vector is generalized to a velocity steering vector. Assume the transmitted signal (**T**) to be a CW tone, as given by equation 5, where f is the carrier frequency and **t** is a time vector of a coherent processing interval (CPI) at a sample rate required by the expected IF bandwidth. The receive signal (**R**) is the superposition of time delayed and Doppler shifted target signals plus noise (**N**) (equation 6). Each target has a velocity vector (speed and heading) that provides a unique Doppler for each transmit/receive pair. Likewise the target's time delay also varies for each transmit/receive pair.

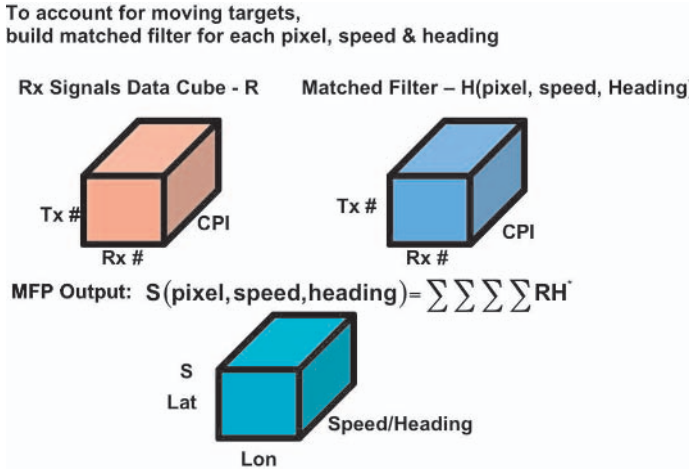


Figure 8. MFP Formulation.

$$\mathbf{T}_i(f) = e^{j2\pi \cdot f \cdot \mathbf{t}}, \tag{5}$$

$$\mathbf{R}_{il}(\mathbf{f}) = \sum_{k=1}^{K} e^{j2\pi \cdot (f + f_{D_{ilk}}) \cdot (\mathbf{t} - t_{d_{ilk}})} + \mathbf{N}, \tag{6}$$

where i is the Tx index (location), l is the Rx index (location), k is the Target index, $\mathbf{t}$ is the time sample vector, and $\mathbf{N}$ is the additive noise.

For MFP, a matched filter ($\mathbf{H}$) is computed for each scene pixel, time delay, and hypothesized target velocity, as given by equation 7. To cover all pixels and target velocities, a bank of filters is employed, using the velocity steering vector. The MFP output ($\mathbf{S}$) is computed as a conjugate inner product of the received signals and match filtered over all transmit/receiver pairs, frequency and time, as given by equation 8.

$$\mathbf{H}_{il}(f, pixel, \mathbf{v}_{TGT}) = e^{j2\pi \cdot (f + f_{D_{il}}) \cdot (\mathbf{t} - t_{d_{il}})} \tag{7}$$

$$\mathbf{S}(pixel, \mathbf{v}_{TGT}) = \sum_i \sum_l \sum_f \sum_t \mathbf{R}_{il}(f) \cdot \mathbf{H}^*_{il}(f, pixel, \mathbf{v}_{TGT}) \tag{8}$$

The MFP process is illustrated in figure 8. The received signal, for all transmit/receive pairs, over a CPI forms a data cube. The matched filter, for a particular scene pixel and target velocity, also forms a data cube. For multiple operating frequencies, additional cubes would be formed. A single MFP output, a pixel and velocity, is the inner product

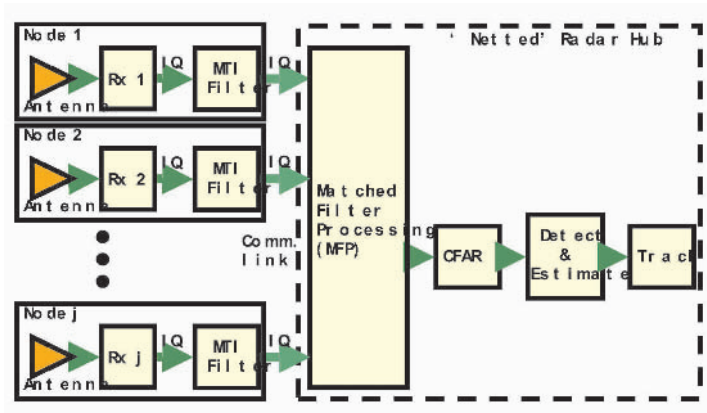


Figure 9. System Receive Processing Block Diagram.

of these cubes. The process is repeated for all pixels and hypothesized target velocities. Target detection is then performed on the **S** cube.

5. TMT Netted Radar System

A 'netted' radar receiver system block diagram is shown in figure 9. The various receiver positions or sites are designated as nodes. These receivers may employ multiple channels covering multiple bands. The in-phase and quadrature (IQ) data is MTI filtered to pass moving targets and filter out stationary ground clutter. The IQ data from each node is communicated to a central processing location, referred to as the netted radar 'hub'. The MFP is commenced by assembling the received signal data cube and matched filtering it. The MFP output is then sent to a CFAR detector, followed by detection and tracking stages.

6. TMT MFP Simulation

A simulation of the TMT process is being used to probe the various issues and projected performance. Consider the following parameters:

- Sensors, Targets are all in a plane (flat earth, no altitude)
  - 10 Transmitters
  - 30 Receivers
  - 5 Targets
  - Area Of Interest (AOI) = 1km × 1km (25 m pixel spacing)
- Waveform
  - 4 CW Tones

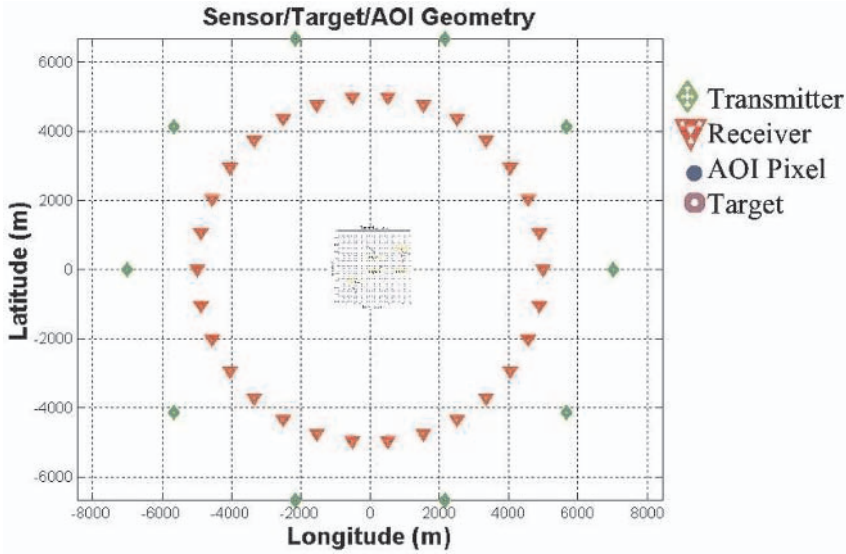


Figure 10. Test Geometry.

- Frequencies = 1, 2, 3 & 4 MHz
- CPI = 5 sec

The sensor geometry is shown in figure 10. The 10 transmitters and 30 receivers form concentric circles about the scene to be imaged. This geometry was used out of convenience and to be familiar to the typical tomographic geometry. The actual geometry may be more random and is accommodated by MFP. A small scene of 1km by 1km was used for computational reasons. Five targets were positioned within the scene. Each target has a unique velocity vector.

The Fourier space sampling, corresponding to the sensor geometries and frequencies used, is shown in figure 11. The left side shows the sampling for one transmitter and all receivers, while the right side shows the entire sampling. Based on this sampling, the expected spatial resolution is about 25 meters.

A zoom-in view of the scene, figure 12, shows the area of interest (AOI) pixel locations and target locations. The targets were spaced so that they should be spatially resolved, based on the expected 25 meter resolution.

The received signal spectrum for a single transmit/receive pair is shown in figure 13. For this pair the five target responses are clearly visible. The noise floor is approximately 10 dB, providing a signal to

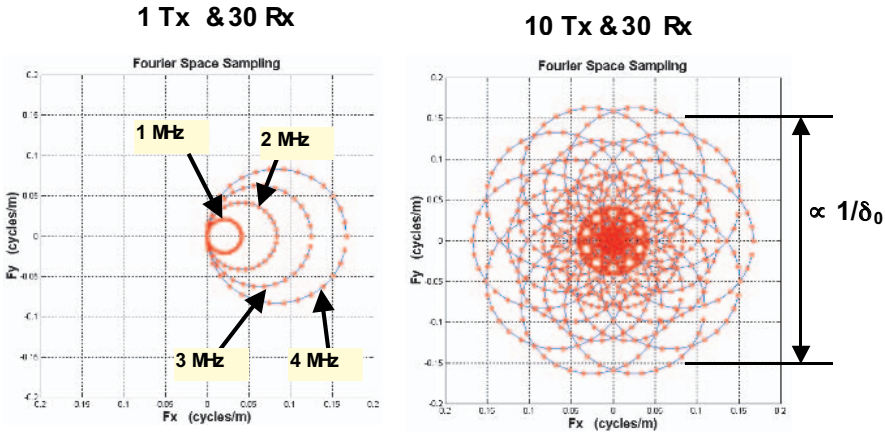


Figure 11. Fourier Space Sampling and Expected Resolution.

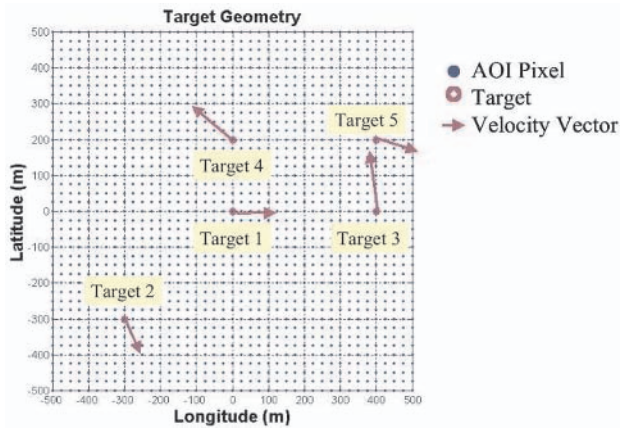


Figure 12. Scene Geometry — Zoom In.

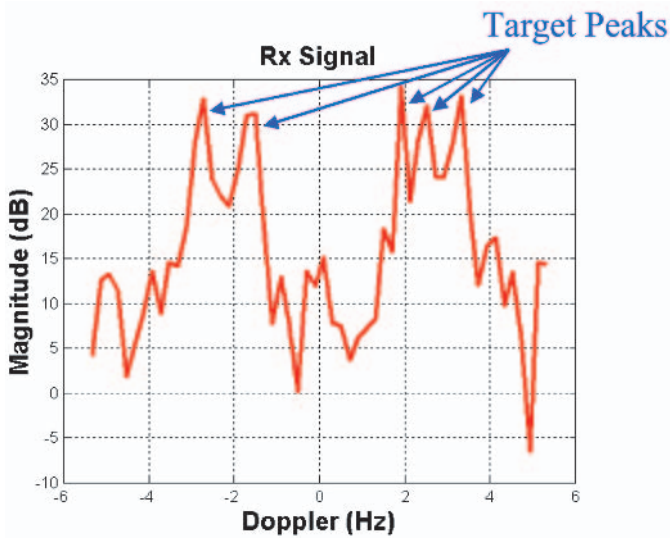


Figure 13. Received Signal Spectrum — Single Tx/Rx Pair.

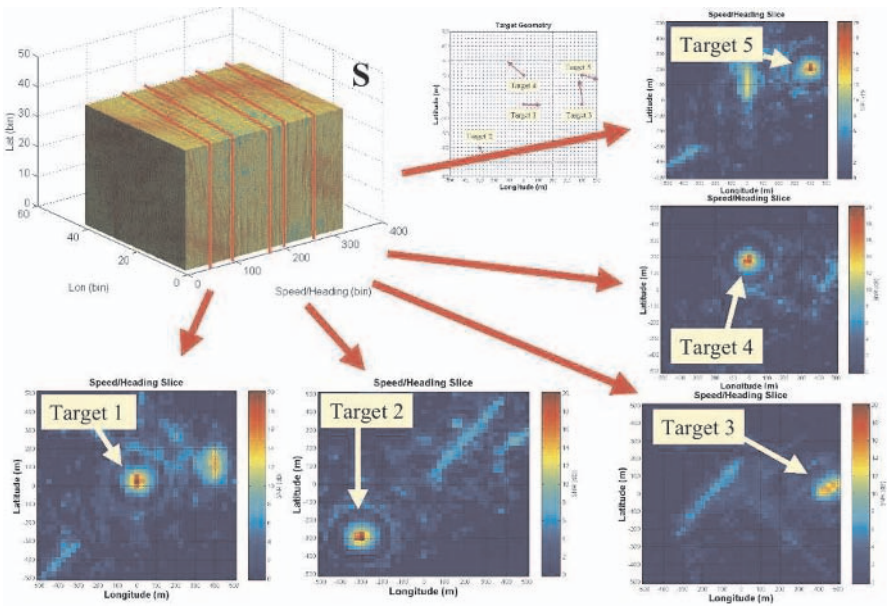


Figure 14. MFP Output Cube and Velocity Slices.

noise ratio (SNR) of at least 20 dB. Note that ground clutter was not modeled and is assumed to be removed by employing the MTI filter.

Target #	Target Truth Position and Velocity					Target Estimates				
	Lon (m)	Lat (m)	Alt (m)	Speed (m/s)	Head (°T)	Lon (m)	Lat (m)	Alt (m)	Speed (m/s)	Head (°T)
1	0	0	0	124	89. 1	0	25	0	123	282
2	-300	-300	0	104	156. 9	-300	-300	0	146	47
3	400	0	0	193	351. 9	450	25	0	200	313
4	0	200	0	151	309. 6	0	175	0	192	172
5	400	200	0	130	108. 2	400	200	0	131	156

Figure 15. Target Detections.

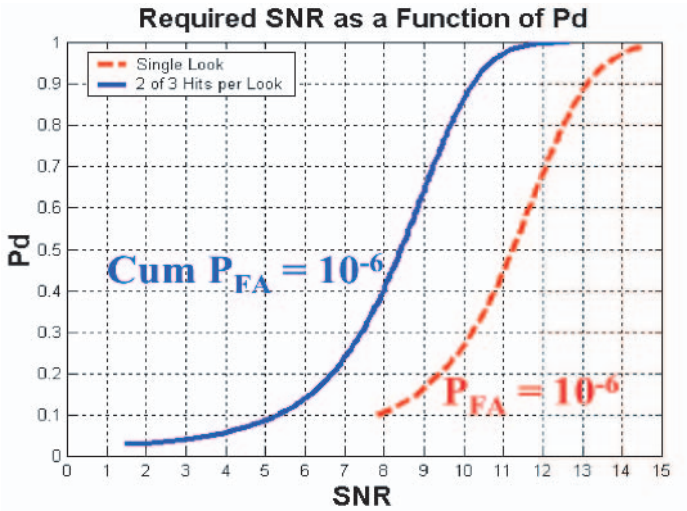


Figure 16. Detection Performance.

The MFP results are shown in figure 14. The output cube **S** is in the upper left. Slices of latitude and longitude, for a particular velocity (speed/heading), are shown. These slices show the likely targets that a 3-D CFAR process would detect. The targets clearly stand out from the background. The targets are also resolved spatially, demonstrating that geometrically diverse UNB systems can provide high spatial resolution.

The MFP output SNR was good (> 15 dB) for 4 of 5 targets. The table in figure 15 shows the target truth and the MFP estimates. The target location estimates were quite encouraging. The estimated velocities were moderately acceptable, this aspect will require further analysis to find the appropriate spatial and frequency diversity to improve the velocity estimate.

7. Detection Perfomance

The probability of detection (Pd) performance versus SNR, of a non-fluctuating target with a probability of false alarm (P_{FA}) of 10^{-6}, for a single look or CPI is shown by the red (dashed) curve in figure 16. For an SNR of 12 dB, the single look Pd is 0.7. The cumulative Pd for an M hits of N looks scheme, with M=2 and N=3 and a cumulative P_{FA} of 10^{-6}, is shown by the blue (solid) curve. This shows that the same 12 dB SNR provides a cumulative Pd of 0.99. The TMT detection processing will likely employ such schemes to improve detection performance.

8. Summary

The TMT concept shows promise for providing high resolution surveillance of ground and airborne moving targets with geometrically diverse UNB transmissions. The UNB signals provide relief when faced with the consequence of ongoing spectrum erosion. The simulation activities have begun to probe fundamental issues of imaging quality and required diversity in frequency and space. Future work will bring more reality into consideration. Issues such as clutter and MTI filters, target cross section fluctuations, and netted radar architectures will be explored.

9. Acknowledgements

The authors would like to acknowledge the management of the Air Force Research Laboratory, Sensors Directorate, for supporting this research.

References

[1] H. Bascom, J. Clancy. HF Imaging Techniques for Ground Penetration, Phase II Final Report, Decision-Science Applications, Inc., Report No. 147/1762, Aug 1998.

[2] H. Bascom, J. Clancy. Ogrodnik HF/VHF Bistatic SAR for Buried Target Detection Experimental Results. *44th Annual Tri-Service Radar Symposium* , June 1998.

[3] H. Bascom, J. Clancy. Ogrodnik Bistatic Approaches to Below Ground Target Imaging, *43rd Annual Tri-Service Radar Symposium Record*, June 1997.

[4] R. McMillan, et al. Bistatic Surveillance of Surface Targets - Concept Analysis, *42th Annual Tri-Service Radar Symposium Record*, June 1996.

[5] C. Jakowatz, D. Wahl, P. Eichel, D. Ghiglia, P. Thompson. *Spotlight-Mode Synthetic Aperture Radar: A Signal Processing Approach*, Kluwer Academic Publishers, 1996.

[6] H. Li, F. Lin, Y. Shen, N. Farhat. A Generalized Interpretation and Prediction in Microwave Imaging Involving Frequency and Angular Diversity, *Journal of Electromagnetic Waves and Applications*, vol. 4, no. 5, pp. 415–430, 1990.

[7] N. Willis. *Bistatic Radar*, Artech House, 1991.

[8] D. Mensa. *High Resolution Radar Cross-Section Imaging*, Artech House, 1991.

NEAR INFRARED IMAGING AND SPECTROSCOPY FOR BRAIN ACTIVITY MONITORING

Il-Young Son, Birsen Yazici
Rensselaer Polytechnic Institute
Troy, NY 12180, USA

Abstract The first demonstration that near infrared (NIR) light can be used to monitor the state of cortical tissues noninvasively through the skull was presented by Jobsis in 1977 [53]. About a decade later, researchers started looking at the potential use of NIR spectroscopy for functional brain activity monitoring. Early studies began with simple motor and sensory tasks demonstrating the feasibility of the technology for noninvasively assessing the state of cerebral activity in a localized area. More recent studies have attempted to monitor more complex cognitive tasks such as warfare management [48] and aircraft landing simulations [102]. In this chapter, the research surrounding the application of NIR imaging and spectroscopy to noninvasive monitoring of functional brain activity is reviewed. A comprehensive review of equipment technologies, mathematical models, and past studies is given with some emphasis on the technology's potential in security and defense applications.

Keywords: functional brain monitoring; near infrared spectroscopy; diffuse optical tomography.

1. Introduction

Near infrared imaging and spectroscopy is an emerging technology concerned with monitoring the changes in the state of biological tissues using light in the range of 600 to 900 nm. The optical properties of the major chromophores within this range makes NIR especially attractive for tissue imaging. The plausibility of using NIR to noninvasively monitor the state of cortical tissue was first demonstrated almost 30 years ago by F. F. Jobsis in his seminal paper, "Noninvasive, infrared monitoring of cerebral and myocardial oxygen sufficiency and circulatory parameter" [53]. He was the first to show that it was possible to penetrate the skull using near infrared light as the source. It would take another

Jim Byrnes and Gerald Ostheimer (eds.),
Advances in Sensing with Security Applications, 341–372.

decade before advances in technology allowed non-invasive monitoring of brain function using NIR methods. Some of the first demonstrations of brain activity monitoring were performed by Chance et. al. [12], Hoshi et. al. [46] and Villringer et. al. [109].

The research into NIR based brain activity monitoring was motivated by its potential as an alternative to older and more established imaging modalities such as functional magnetic resonance imaging (fMRI) and positron emission tomography (PET). There are several motivating factors for researching the potential of NIR based methods. For one, the NIR method provides information about physiological parameters not available in other modalities, such as oxygenation information. Secondly, NIR equipment has higher temporal resolution, in the order of milliseconds, compared to fMRI and PET [52]. This allows for, among other things, being able to model fast oscillatory noise related to normal physiological functions [34]. Thirdly, NIR equipment is relatively less restraining compared to fMRI or PET and generally safer than PET as it does not rely on ionizing radiation. Some types of NIR equipment, namely those using continuous wave signals, has also been made portable and in some instances, telemetric [11, 45].

With these motivations as a driving force, the research into NIR based brain activity monitoring have blossomed in the past 20 years with growing number of potential applications. So far, brain activity monitoring has seen applications in physiological studies of brain disfunction, preterm neonatal care, education and training, and cognitive workload assessment.

The list is not exhaustive but gives some idea of the variety of fields in which it finds utility. Workload assessment is particularly of interest to security and defense research. The DARPA's Augmented Cognition (AugCog) program, for example, have used EEG's, NIR, and fMRI technologies to monitor cerebral responses to a given complex cognitive task. The goal of such *operator monitoring* is to assess the cognitive state of the operator and adapt the system accordingly to mitigate the effect of information processing bottleneck in order to optimize his or her performance output. The assumption is that by noninvasively monitoring cerebral response to varying task conditions, it may be possible to extract patterns and infer the cognitive state of the task performer. Additionally, it has potential applications in emergency medicine in battle field, and operator fatigue assessment in stressful environments.

In this review, it is our goal to give an overview of the breadth of research conducted in the past 20 years. As with most review of this nature, we cover the breadth and do not intend to provide close in depth

survey of all the research available in this area. However, we provide a comprehensive review and point to motivations for further research.

With these goals in mind, we structured the chapter as follows: First, an overview of the types of imaging system used in NIR spectroscopy studies is given. This is followed by sections describing in some detail the two different kinds of signals that NIR method can measure. The two types of signals are characterized by their response time, namely fast response and slow response signals. The fast signal is associated with changes in neuronal tissue and slow signal is related to changes in the state of hemoglobin concentration and oxygenation (hemodynamics). The fast signal have latency of around 50 to 300 ms and slow signal have latency of about 10 seconds [32]. Lastly, an overview of human studies with an emphasis on security and defense applications is given. This is followed by some closing remarks on the future direction of research.

2. NIR Imaging and Spectroscopy Systems

Three types of light sources have been suggested for use in NIR imaging and spectroscopy systems [11, 7]. The simplest of these are continuous wave (CW) sources. As such most commercially available systems use CW light as their source [44]. CW-type instruments are able to assess regional cerebral blood flow by measuring light attenuation through the cortical tissue and calculating their hemodynamic responses, i.e. changes in hemoglobin concentration and oxygenation, using these attenuation measurements. With these absorption measurements, however, it is very difficult, if not impossible, to gauge the absolute concentration changes since the real path length of light photons are unknown and cannot be measured or inferred. CW-type instruments rely on simplified assumptions about the nature of the media being probed and the changes occurring inside the sampling volume. As such, only relative concentration changes from some baseline measurement can be assessed. The advantage of using a CW-type instrument is that they are inexpensive and can be made portable. An example of CW-type system used in our lab is pictured in figure 1. As can be seen from the figure the entire system is quite compact. The probe itself is highly flexible, thus relatively comfortable to wear and consists of array of photosensors and diodes. A close-up of the probe is pictured in figure 2.

In order to better quantify the absolute value of chromophore concentrations, time of flight (TOF) must be measured in addition to light attenuation. This may be achieved using time-resolved or frequency domain methods. Time-resolved spectroscopy (TRS) was first pioneered by Delpy et. al. [19], Patterson et. al. [85] and Chance et al. [12, 13].

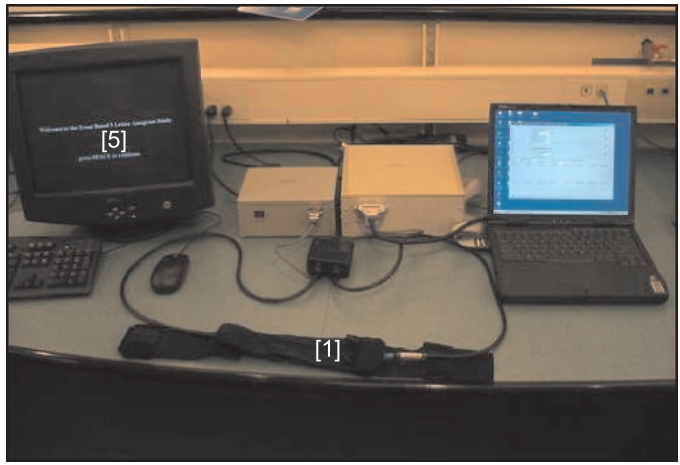


Figure 1. An example of CW-type system: 1. probe 2. data processing unit for pre-filtering and managing control signals 3. power supply 4. laptop computer for recording data and sending control signals to the probe 5. computer for hosting the task.

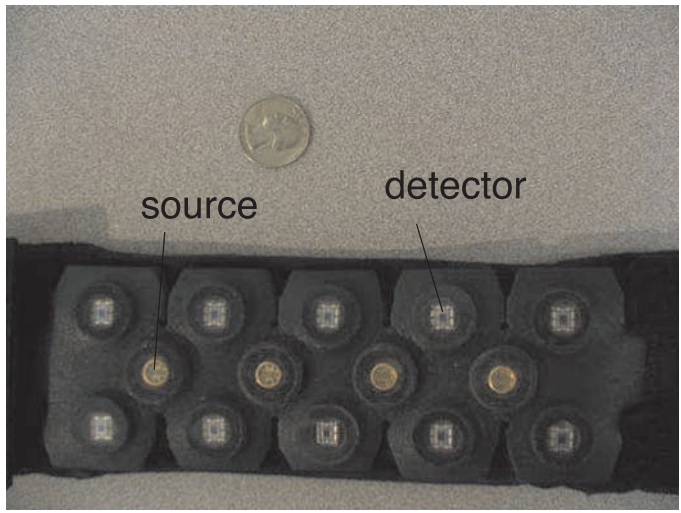


Figure 2. Close of the probe used in our laboratory. The quarter is for size reference.

Since then, number of researcher have studied and experimented with TRS including Chance and Oda [61] [63] [76] [80] [81] [73] [82] [113] [114]. TRS instruments rely on a picosecond pulsed laser with a detector that is designed to detect the time evolution of the light intensity [44]. With the time profile of light intensity through the medium, it is possible to measure both absorption and reduced scattering coefficients [32]. A ma-

jor draw back of TRS instruments is that they are relatively expensive. For this reason, these instruments have mostly been built for research purposes and are not readily available commercially.

Frequency domain method was first suggested by Gratton et. al. in [31]. Frequency domain approach uses radio frequency intensity modulated sources. In addition to the DC component, the light intensity attenuation, frequency domain systems can measure phase and modulated amplitudes which can all be related back to the input signal. Frequency domain method is mathematically related to TRS via Fourier transform. However, frequency domain systems have an advantage of being an inexpensive alternative to TRS systems. Also, in practice, frequency domain instruments display higher signal to noise ratio (SNR) and are generally faster than TRS instruments. The disadvantage of frequency domain systems is that they can only provide information at a finite number of modulation frequencies. An example of commercially available frequency domain system (ISS 96208) is shown in Figure 3.

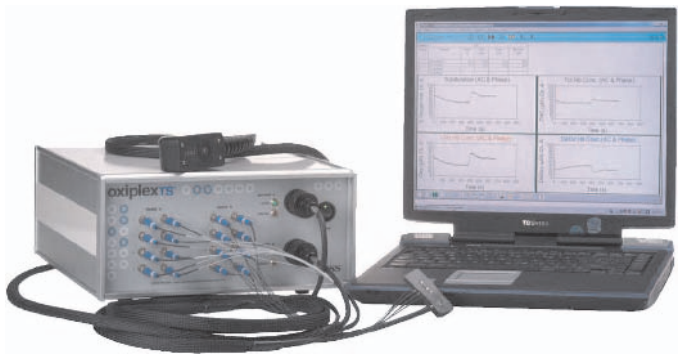


Figure 3. An example of frequency domain instrument. ISS, Inc.'s oximeter model 96208.

3. Hemodynamic Response

3.1 Modified Beer-Lambert Law

By far, the most widely used model in calculating hemodynamic response is based on the classic Beer-Lambert law. The Beer-Lambert law is derived from solution to radiation transport equation under several simplifying assumptions [91]. It describes a linear relationship between absorbance, A, of light through a medium and wavelength dependent extinction coefficient, $\epsilon(\lambda)$. This relationship is given by **Equation (1)**

below.

$$A = -\log \frac{I}{I_0} = \epsilon(\lambda)cL \tag{1}$$

where $I = I_0 \exp -\epsilon(\lambda)cL$ is the intensity of the transmitted light, I_0 is the intensity of the incident light, c is the concentration of the chromophore and L is the path length. In continuous wave NIR imaging, a modification of the Beer-Lambert law has been successfully applied, albeit with some caveats due to inaccurate but simplifying assumptions about the biological media. Delpy et. al. were the first to introduce the modified Beer-Lambert law in [19] [17] [87]. Modified Beer-Lamber law (hitherto referred to as MBLL) relies on several simplifying assumptions. These were identified by Obrig and Villringer and are paraphrased below [78].

1 *High, but constant scattering in the media.*

2 *Homogeneous medium.*

3 *Homogeneous change of parameters of interest within each volume sampled.*

For non-scattering media, following the classical Beer-Lambert law, L is equal to the distance between source and detector, denoted as d. For scattering media **Equation (1)** can be rewritten as

$$A = \epsilon(\lambda)c \cdot d \cdot DPF(\lambda) + G(\lambda) \tag{2}$$

$$= \epsilon(\lambda)c \cdot \langle L \rangle + G(\lambda) \tag{3}$$

where G is the contribution of the attenuated light due to scattering and $DPF = \langle L \rangle / d$ is called the differential path length factor. The $\langle L \rangle$ is the mean path length of the detected photons [91]. The differential path length factor describes the increase in path length due to tissue scattering. The DPF for various tissue samples have been both experimentally [21, 58] and numerically studied [41] and can easily be looked up on a table (for example, see [24]). Since the value of wavelength dependent G is not generally known, it is not possible to assess the absolute value of A. Under the first assumption given above, DPF and G can be assumed to be constant. This allows for assessment of *changes* in chromophore concentration by subtracting out G. This is generally considered a plausible assumption when measuring hemodynamics since changes in oxygenation and concentration of hemoglobin affects the absorption coefficient more significantly than the scattering coefficient [78]. The **Equation (2)** can then be rewritten as

$$\Delta A = A_t - A_{t_0} = \epsilon(\lambda) \cdot \Delta c \cdot d \cdot DPF(\lambda) \tag{4}$$

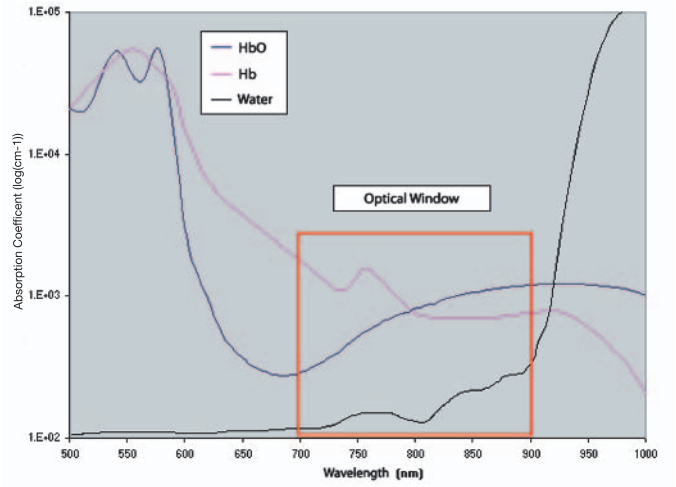


Figure 4. Graph of absorption coefficient versus wavelength of light. The optical window is denoted by the red box.

where A_t is the absorbance at some time t after the change in concentration of chromophores and A_{t_0} is the initial absorbance.

Each chromophores have a distinguishable extinction spectrum in the visible to the near-infrared range. This allows for the measurement of concentration changes in several chromophores simultaneously by taking optical measurements at multiple wavelengths. The main chromophores in the optical window[1] of 600 to 900 nm, are oxy- and deoxy-hemoglobin (denoted HbO and Hb respectively), water, lipids and cytochrome-c-oxidase. Figure 4 shows the graph of the absorption coefficients of HbO, HbO and water versus wavelength. The NIR range is denoted by the rectangular window. The main chromophores of interest in nearly all studies of NIR spectroscopy based brain imaging are HbO and Hb, as the other chromophores' changes are orders of magnitude smaller than that of HbO and Hb [9, 79]. It should be noted that there have been studies where transient increase in cytochrome-oxidase redox was observed for visual stimulations [39] [57] [77]. However, there are some questions as to the detectability of changes in cytochrome-oxidase redox. Uludag et. al. report that concurrent detection of change in cytochrom-oxidase redox state with those of hemoglobin concentration and oxidation might suffer from severe cross-talk error where the magnitude of the error for the redox state of cytochrome-c-oxidase may be in the order of those

[1]The term "optical window" denotes the spectral range where a low absorption by chromophores allow for significant penetration of light.

detected experimentally [107]. Cross-talk error is a general problem for MMBL based NIR imaging and will be revisited in some detail later in the chapter.

Oxy- and deoxy-hemoglobins are mainly of interest because they are related to the regional cerebral blood flow (rCBF). The focal change in rCBF determines the *activation* state. The term activation usually refers to the focal increase in rCBF whereas a decrease is called deactivation [78]. With the dual wavelength approach, one can derive two simultaneous equations to be solved for each of the two chromophore concentration changes. To this end, **Equation (4)** is split into two parts, separating the contributions from HbO and Hb. **Equation (4)** is then, rewritten as

$$\Delta A(\lambda) = (\epsilon_{HbO}(\lambda)\Delta[HbO] + \epsilon_{Hb}(\lambda)\Delta[Hb]) \cdot DPF(\lambda) \cdot d \qquad (5)$$

where $\Delta[HbO]$ is the change in HbO concentration and $\Delta[Hb]$ is the change in Hb concentration. Given **Equation (5)**, and assuming $\epsilon(\lambda)$ can easily be looked up from the extinction spectra for each chromophores, concentration changes ascertained by solving the two simultaneous equations at two distinct wavelengths. The generalization to more than two wavelength is straight forward and can be found in [16].

3.2 Some Issues Regarding MBLL

Figure 5. The assumed banana-shaped light path through tissue sample.

The simplifying assumptions of homogeneous medium and homogeneous change in differential volume are clearly inadequate for most biological media. The usual banana-shaped sampling volume [10, 108] from source to detector, as depicted in figure 5, is inadequate as the path of light is greatly affected by cerebrospinal fluid (CBF) [83] and the pial cerebral vessels on the surface of the brain [25]. This gives rise to a

layered effect on the path of light that is not accounted for under the MBLL assumptions. The greater part of the problem lies in what is termed *partial volume effect*. Under the homogeneous change assumption of MBLL, magnitude of concentration changes of chromophores are underestimated because the activated volume is usually smaller than the sampling volume [107]. Furthermore, the wavelength dependence of optical tissue properties means that this partial volume effect is wavelength dependent. As a result of this wavelength dependence, change in one chromophore concentration may mimic the effect of another, giving rise to crosstalk errors [84]. This is due to the fact that in the MBLL model, wavelength depedence is ignored by using a constant mean path length independent of wavelength. More precisely, focal hemodynamic change depends on the knowledge of the partial pathlength within the partial volume, which is unknown. Calculating the concentration changes of chromophores using measured absorbance changes at different wavelengths will give rise to distortions in the extinction spectra. There has been some progress in layered modeling based on Monte Carlo methods [96, 41]. Using Monte Carlo simulation, path length is estimated at each layer, which can be used to minimize the distortion and crosstalk.

There is some debate, as Obrig and Villringer point out [78], as to the significance of the crosstalk distortion. According to Obrig et. al.'s study, crosstalk between Hb and HbO elicits about 10% error. This is small enough that, although there is noticeable distortion on the ratio of magnitudes between chromophores, this will not significantly alter the quality of images.

Diffuse Optical Tomography. Boas et. al. in [9] suggested using a more sophisticated and accurate model for photon propagation in the brain to minimize the effect of crosstalk. In lieu of MBLL based image construction, Boas et. al. suggested a diffuse optical tomographic (DOT) method. The mathematical details of DOT methods are covered in [3, 4]. An excellent general review of DOT is given in [7]. Here, we limit ourselves to a very cursory overview of DOT method. The DOT is based on photon diffusion equation with Robin boundary conditions derived from radiation transport equation with less restrictive assumptions describing the propagation of light photons through a turbid medium. The diffusion equation is

$$-D(r)\nabla^2\Phi(\mathbf{r},t) + v\mu_a\Phi(\mathbf{r},t) + \frac{\partial\Phi(\mathbf{r},t)}{\partial t} = vS(\mathbf{r},t) \tag{6}$$

where $\Phi(\mathbf{r},t)$ is the photon fluence at position $\mathbf{r}$ and time t, $S(\mathbf{r},t)$ is the source, $D = v/(3\mu_s')$ is the diffusion coefficient, μ_s' is the reduced

scattering coefficient, $\mu_a = \epsilon(\lambda) \cdot c$ is the absorption coefficient and v is the speed of light through the medium. Assumption under which **Equation (6)** accurately models photon propagation is that of scatter dominated medium (same as the first assumption for MBLL model). The goal in DOT is to reconstruct the 3-D distribution of absorption and scattering coefficients of the medium given boundary data.

Under the first order Born approximation, the solution to **Equation (6)** is

$$\Phi_{sc}(\mathbf{r}_d) = -\int \Phi_{inc}(\mathbf{r})\frac{v}{D}(\epsilon_{HbO}(\lambda)\Delta[HbO] + \epsilon_{Hb}(\lambda)\Delta[Hb])G(\mathbf{r}, \mathbf{r}_d)d\mathbf{r} \tag{7}$$

where $G(\mathbf{r}, \mathbf{r}_d)$ is the Green's function, $\mathbf{r}_d$ is the location of the detector, Φ_{inc} stands for the incident photon fluence and $\Phi_{sc}(\mathbf{r}_d)$ denotes photon fluence at the detector. **Equation (7)** can be discretized by taking multiple measurements of $\Phi_{sc}(\mathbf{r}_d)$ at different source-detector positions. The integrand of **Equation (7)** describes the sensitivity of each measurement to the change in absorbance within each volume sample. For focal change in absorbance,

$$\Delta A = -\log\left(\frac{\Phi_{inc}(\mathbf{r}_s, \mathbf{r}_d) + \Phi_{sc}(\mathbf{r}_s, \mathbf{r}_d)}{\Phi_{inc}(\mathbf{r}_s, \mathbf{r}_d)}\right) \tag{8}$$

$$\approx \frac{\Phi_{inc}(\mathbf{r}_s, \mathbf{r})G(\mathbf{r}, \mathbf{r}_d)}{\Phi_{inc}(\mathbf{r}_s, \mathbf{r}_d)}\frac{vd\mathbf{r}}{D}\Delta\mu_a \tag{9}$$

where $d\mathbf{r}$ is the differential volume over which the absorbance change occurs.

In general Rytov approximation provides better reconstruction than Born approximation [8]. Here, MBLL is approximated locally at each sampling volume and the effective path length is estimated using the Rytov approximation as

$$L_j = \Phi_{inc}(\mathbf{r}_s, \mathbf{r}_j)\Phi_{inc}(\mathbf{r}_j, \mathbf{r}_d) \tag{10}$$

Change in the absorption coefficient is found by arranging the measurements and voxel combinations in vector-matrix form as $y = Ax$ where y is the change in absorbance detected at each source-detector pair, A is the so called system matrix derived from L_j and x is the optical parameters of interest, namely the absorption coefficients.

3.3 Physiological Interpretations

The oxygenation response over an activated area of the cortex can be described by a *decrease* in Hb along with a simultaneous *increase* in

HbO [78]. The accompanying increase in HbO is usually two to three times that of the decrease in Hb, hence the total volume of hemoglobin is expected to increase locally in the activated areas. Inversely, deactivation is typically characterized by decrease in HbO along with increase in Hb. If both HbO and Hb increases or decreases, this may indicate origin other than cortical activity. Therefore, it seems prudent to report both Hb and HbO responses along with the total hemoglobin displacement. Recording Hb response has an added benefit that it can be related to and compared with fMRI blood oxygenation level dependent (BOLD) response as increase in the BOLD contrast is highly correlated with a decrease in Hb.

Comparison with fMRI BOLD Response. A number of studies, comparing fMRI data with those gathered through NIR spectroscopy have been conducted [56] [72] [32] [105] [99] [106]. Such studies have addressed the problems of localization of detected signal and sensitivity of optical measurements to brain tissue [99]. Most of these studies have shown a high temporal negative correlation between BOLD and Hb responses [32] that confirms theoretical ideas about the nature of the BOLD response. The studies also show a good spatial collocation between BOLD and hemodynamic responses [56, 105]. Strangman et. al.'s study is of some interest since their results deviate from the expected results [99]. They observed that correlation between BOLD and Hb responses were highly variable between subjects. They've argued that this variability was due to the model error arising from simplifying assumptions described in the previous section. To account for these potential sources of errors, Strangman et. al. normalized each hemodynamic response for each subject separately with normalization factor equal to the ratio of the vascular response of interest and inverse BOLD response. Even so, their study found that correlation between BOLD and HbO were higher than between BOLD and Hb. This may be due to higher sensitivity of the optical measurement to changes in HbO.

Sources of Physiological Noise. Despite the apparent importance of Hb, in the literature, the change in HbO is most widely reported and sometimes is the sole response reported. HbO is sometimes preferred due to its higher sensitivity to NIR based measurements than Hb. However, HbO measurements are not without disadvantages. There have been some studies (see for example [77] [22] [23] [47]) which indicate the presence of low frequency modulation (~ 0.1 Hz) of both vascular (Hb, HbO) and metabolic (cytocrhome-c-oxidase) responses to visual stimuli with unknown origin. The cause of these low frequency oscillations are

not exactly known but are usually attributed to extracerebral activities such as respiration. Obrig et. al. have found that these modulations are most significantly expressed in the HbO responses and thus is most susceptible to extracerebral noise [77]. These low frequency oscillations are known collectively as Mayer wave and should be accounted for when analyzing data. Other sources of physiological noise are oscillations due to heartbeat and signal displacement due to motion artifacts. The oscillations due to heartbeat is faster than Mayer wave (~ 1 Hz) [98, 44].

Removing Physiological Noise. One method of removing the low frequency artifact is to convolve the response signal with a model of stimulus signal. Such methods have been used to increase the SNR in fMRI [54]. The stimulus signal is usually modeled as a pulse train with evenly spaced interstimulus interval as in **Equation (11)**

$$s(t) = \sum_{m=1}^{M} \delta(t - m\tau_{ISI}). \tag{11}$$

Here, M is the total number of stimulis and τ_{ISI} is the interstimulus interval. The convolved signal is used to fit a generalized regression model of the form

$$x(t) = \sum_{b=1}^{N} g_b(t \mod \tau_{ISI})\beta_b + e(t) \tag{12}$$

where the response signal $h(t) = \sum_{b=1}^{N} g_b(t)\beta_b$, is expressed as a linear combination of basis functions $g_b(t)$. **Equation (12)** can be rewritten in matrix form and generalized linear regression method can be used to fit β_b (see [59] for a review of linear regression).

Gratton et. al. in [34] describes a least squares regression method to model heartbeat artifacts and to filter it out adaptively. Since the heartbeat rate is approximately 1 Hz, it is necessary that data is sampled at a sufficiently high rate (e.g. above the Nyquist rate) so as to represent and filter the artifacts out correctly and have minimal impact on the signal of interest itself [98].

Izzetoglu et. al. have suggested an adaptive filtering technique for motion artifact cancellation using complementary sensors to sense motion [49]. Wiener filtering was proposed as a sensor-free alternative to adaptive filtering in [20]. Wiener filtering is a well established least squares method of reconstructing a signal degraded by motion artifacts. It is widely used in image processing and other signal processing fields. Suppose the response signal is modelled as the true signal contaminated by

additive stationary noise (the motion artifact), i.e. by $h(t) = x(t)+w(t)$. The idea is to construct a filter $g(t)$ or $G(\omega)$ in the frequency domain such that mean square error is minimized between the real signal $x(t)$ and the estimated signal $\hat{x}(t) = g(t) * y(t)$ where $*$ denotes the convolution operator. Without going into the details of the derivation, it turns out that the optimal filter $G(\omega)$ in the Fourier domain is

$$G(\omega) = \frac{P_x(\omega)}{P_x(\omega) + P_w(\omega)} \tag{13}$$

where $P_x(\omega)$ and $P_w(\omega)$ are the power spectral density functions of the signal and noise, repsectively. One disadvantage with this method of motion artifact removal is the estimation of P_x and P_w, which are not known a priori.

Zhang et. al. describes a new method based on eigenvalue decomposition [116]. This method differs fundamentally from others in that it uses spatial filtering instead of temporal filtering that previous methods employ. The justification for this line of thinking is that the physiological noise is usually systematic and global, whereas stimulus driven activation resides locally. The approach uses principle component analysis (PCA) to extract a set of basis functions, orthogonalize them, and use them to filter the signal. Let $\mathbf{H}_s$ and $\mathbf{H}_b$ denote the values of changes in hemodynamic response during the stimulation and rest (baseline) periods respectively, where each column corresponds to a channel or spatial location and each row corresponds to a time instance. Then, the matrix to be decomposed is the spatial correlation matrix

$$\mathbf{C} = (1/N)\mathbf{H}_b\mathbf{H}_b^T \tag{14}$$

where N is the total number of time samples. Then, the first M number of eigenvectors are chosen as the basis, $\mathbf{U}_{base} = [\mathbf{u}_1 \dots \mathbf{u}_M]$. The filtered stimulus invoked signal is then

$$\tilde{\mathbf{H}}_\mathbf{s} = (\mathbf{I} - \mathbf{U}_{base}\mathbf{U}_{base}^T)\mathbf{H}_s. \tag{15}$$

4. Neuronal Response

The acquisition of neuronal response using noninvasive NIR spectroscopy is less well-known than the slower hemodynamic response. However, there is a great deal of potential for its development and verification via comparison with other modalities such as event-related potentials (ERP) [32]. The origins of optically sensing this fast signal (usually known as intrinsic optical signal or IOS for short) can be traced back to Cohen [14]. Cohen showed that neuronal activity elicits a fast

change in light scattering of neural tissue. These properties were confirmed in invasive experiments where an isolated nerve and tissue slices electrically stimulated and optical properties measured simultaneously [14] [103] [104] [90]. Similar agreement between electrophysiological signal and light scattering change following stimulation is reported in [97]. Fast scattering changes were also observed on macroscopic structures. Frostig et. al. observed the IOS changes in hippocampal slices [29]. Experiments on exposed cortex of animals also confirmed the presence of scattering changes [66, 86].

Initial studies in detecting these scattering changes involved visual stimuli [33, 68] and tapping task [37] and used frequency domain methods [35]. Such a method is termed event-related optical signal (EROS). In EROS, frequency domain instrument is used at millisecond temporal resolution to detect phase shift of the photon density at the detectors [38]. With the phase information, it is possible to gauge the average time of flight of photons detected and thus infer the average path length. With this path length information, it is, in principle, possible to better localize the signal (compare to MBLL approach using CW equipment) by estimating roughly the depth that the photons travel [36]. Measuring phase information for detection of IOS seems, at first, principled since it has been shown that phase measurements are more sensitive to scattering changes than to absorption [30]. However, other studies have reported that the intensity data yield higher SNR [95] [26] [112]. Recently, Maclin et. al. examined fast signal response in somatosensory cortex to electrical stimulation of the median nerve recording both phase and intensity data simultaneously [62].

4.1 Data Analysis

Franceschini et. al. has employed a simple back-projection method described in [28] to create an optical image of the probed region. In this method the region of interest is discretized and each pixel is assigned a set of source contribution with weights. Each pixel is then a weighted combination of the signal from different sources. In this method, intensity signal is also first converted to a change in absorption coefficient using differential pathlength factor (DPF) approach described by Cope and Delpy in [17]. With EROS approach, the usual method of data analysis is to cross-correlate the signal, either phase or amplitude or both, with the stimulus signal. Also, in order to reduce erroneous reading due to equipment sensitivity to changes in superficial regions under source or detector fiber, cross-correlation between individual channels are assessed [32].

Removing physiological artifacts from fast signal is similar to removing artifacts from slow signal. For removing heartbeat artifact, the same approach described previously is used where a least-squares fit model of each pulse is constructed then subtracted out. For Mayer wave, since the fast signal dynamics is orders of magnitude faster than the oscillations of Mayer wave, it is possible to filter out the drift by simply employing a high-pass filter.

5. Human Subject Studies

Since the early 90s, noninvasive functional brain imaging of humans using NIR methods have been slowly gaining momentum despite existence of more established imaging modailties, such as PET, fMRI, and EEG. Part of the reason as stated previously, is because of its relatively high temporal resolution and its ability to monitor multiple tissue chromophores. The technique has been applied to adult as well as infant studies. NIR method is particularly suited for infant studies as the equipment, at least the CW kind, are minimally restraining, relatively safe, and portable [67]. Most neonatal studies focus on sensory stimulation such as visual, auditory and olfactory stimulations [69] [101] [89] [115] [6] [5], and cerebral disfunction [70, 71]. Our review will focus primarily on adult studies with some emphasis on defense and security applications.

Adult subject studies can be divided into two broad categories; response to basic sensory stimulation and response to more complex cognitive tasks. Recently, there have also been reports applying the NIR spectroscopy to brain computer interface research [18] and studyinh the correlation of hemodynamic response to computational cognitive models [94].

Motor and Sensorimotor Stimulation Studies. Maki et. al. has conducted studies on hemodynamic response to opposing finger movement stimuli in [64, 65]. They observed a significant increase in HbO and total blood volume and decrease in Hb as expected for cortical activation. Watanabe et. al. and Hirth et. al. conducted similar studies corroborating the results of Maki et. al. [111, 42]. Kleinschmidt et. al. also performed finger opposition task using NIR spectroscopy simultaneously with fMRI [56]. Colier et. al. studied response to coordinated movements of hands and feet both in phase and in anti-phase. Expected activation pattern was observed with decrease in Hb and increase in HbO with more significant changes seen in HbO. They found no significant difference between in phase and anti-phase stimulations [15]. Miyai et. al. conducted a more sophisticated study observing sensorimotor corti-

cal response to human gait. This study is noteworthy, in that, due to its constraining nature, it is not possible to study such a paradigm with fMRI. The subjects were measured during a 30 second treadmill walking. They report observing an increase in the level of HbO and total blood volume after 3 to 5 seconds while seeing slight decrease in Hb level [74]. Bilateral study was conducted by Franceschini et. al. using multiple stimulus. The volunteers were subjected to opposing finger, finger tapping, and medial nerve stimulations. They saw a consistent increase in HbO and decrease in Hb in the cortical region contralateral to the side stimulated [27].

Visual Stimulation Studies. Visual stimulations have also been studied in adult human subjects. Meek et. al. stimulated their subject by subjecting the volunteers to a 30s on/off cycles of computer graphics display [68]. They observed an increase in HbO and total blood volume during stimulation in the occipital cortex. They compared this with a measurement on frontal lobe which resulted in no significant change confirming the role of occipital cortex during visual stimulation. Heekeren et. al. conducted similar study with sustained two minute stimulation [40]. They observed an increase in HbO during the first 19 seconds of stimulation then plateauing during the entirety of the stimulation. The time-course of Hb was more dynamic. An initial decrease during the first 13 seconds was observed after which the subsequent 10 seconds saw an increase plateauing to a new level, near the baseline, after 40 seconds or so. After the stimulation was shut off Hb started to increase beyond the baseline.

Language Studies. Watanabe et. al. conducted language dominance study on 11 healthy subjects and 6 subjects with intractable epilepsy [110]. A word generation task was conducted, each lasting 17 seconds with 60 second rest period. The group saw increase in HbO, total blood volume, and Hb which seceded after the end of stimulation period for both types of subjects. This result is distinctly different from normal patterns of activation. The study unfortunately did not address this point. They corroborated their results of dominant hemisphere with Edinburgh questionnaire for health subjects and Wade test for epileptic subjects. Sakai et. al. examined speech processing using NIR spectroscopy [92, 88]. In their study, they used auditory stimulation where subjects were asked to track targets and press a button when the target shifted from one ear to another. They conducted two tasks; a repeat task where target was the repetition of a single sentence within a task block; and a story task where a target was a different successive sentences of

one continuous story. They found, consistent with fMRI results, that compared to the repeat task, story task was localized to the left superior temporal cortex. Also, story task saw a larger increase in HbO and decrease in Hb compared to white noise control task. Language lateralization was also studied by Kennan et. al. [55] and Noguchi et. al. [75]. Kennan's group used visual stimulation where subjects were asked to find syntactic and semantic errors in sentences presented. All subjects showed left hemispherical dominance to varying degree. The NIR spectroscopy results were compared with results garnered from fMRI modality where good consistency was found between the two. Similar error detection task but, using auditory stimuli instead, was conducted by Noguchi et. al. The error detection task was split into two, syntactic and semantic. Compared to the syntactic task, semantic task show very little activation in the left hemisphere. In the right hemisphere both task showed little activation, suggesting as with Kennan's group, that language processing is dominant in the left hemisphere.

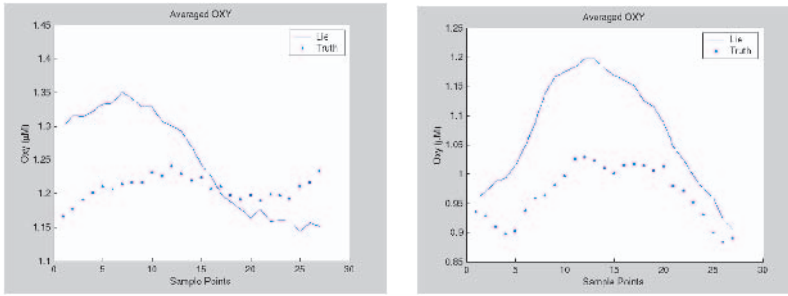
Mental Health Studies. Number of studies have been produced concerning subjects with brain disfunction, such as epilepsy [1], depression [100], and Alzheimers [43]. Adelson et. al. reports on the use of NIR spectroscopy on epileptic patients. They saw a preictal increase in oxygenation level 1 and 2 hours before and a decrease in oxygenation during seizure period [1]. A verbal fluency task was given to Alzheimer's patients and normal subjects by Hock et. al. They observed a decrease in HbO during the task in Alzheimer's patient in contrast to increase in HbO in normal subjects. Both groups saw slight decrease in Hb [43]. Suto et. al. conducted verbal fluency and finger tapping tasks on patients with depression, patients with schizophrenia and normal subjects. Compared with the normal subjects (control), depression patients saw smaller HbO increase during the first half of the verbal fluency task period whereas, schizophrenic patients were characterized by a small dip in the HbO during the start of the task and a re-increase in the post task period. The HbO increase in depression patients were generally larger than those of the control group. No significant difference was found between control and schizophrenic patients [100].

Cognitive Tests. Schroeter et. al. studied hemodynamic responses to classic Stroop task where mismatch in words spelling out various colors [93]. The task consisted of two rows and were asked to decide whether the color of the top row of letter corresponded with the bottom row spelling out a color. Three types of stimuli were presented, (1) neutral, where top row consisted of X's, (2) congruent, where top row

was a word spelling out some color and bottom row matched the color of the top row, and (3) incongruent where the bottom row mismatched with the color on the top row. Incongruent stimulus generated the largest rise in HbO and and largest decrease in Hb, congruent being second largest and neutral being the least activated. This result agrees with the intuition that the most difficult task would see the most activation.

Izzetoglu et. al. used target categorization and N-back tasks to assess hemodynamic responses in [50, 51]. In the oddball target categorization task the subject was given two stimuli, one more frequently occurring than the other. The objective was to press a button when the less frequent stimulus was given. They found that oxygenation change was higher when presented with less frequent stimulus. In the N-back task, 2-back saw the largest average oxygenation change with consistent rise from 0-back. However, 3-back had lower average oxygenation than 2-back. The group explained this result as a consequence of the difficulty exceeding the subject's ability to keep up with their working memory.

Recently, NIR spectroscopy also found application in analyzing more complex cognitive tasks and in workload assessment. These studies are of interest to defense and security researchers as they attempt to assess the cognitive state of the subject while the subject is given a series of tasks involving a cognitive function, such as problem solving or memory recall. Izzetoglu, et. al. for instance, studied oxygenation response to guilty knowledge task (GKT) which is commonly used to model deception [51]. An fMRI study using similar task setup was performed by Langelben et. al. [60]. These studies point to potential use of NIR based noninvasive brain monitoring as a lie detector. Both studies involved display of playing cards that are divided into four different categories, lie card, truth card, control card, and non-target cards. One card was shown for 3 seconds with 12 second interstimulus interval between each card. For truth, lie, and non-target cards, the subjects were asked, "do you have this card?" For the control cards, the subjects were asked, "is this the [control card]?" The subject was asked to lie about the lie card only. The goal was to see if the level of hemodynamic response differed significantly between lie and truth responses. Izzetoglu et. al. found that in 8 out of the 13 subjects, the level of the oxygenation change during the lie was higher than the level during the truth. Figure 6 shows the average oxygenation response of two subjects, averaged over all sixteen channels in the system used by Izzetolgu et. al. as reported in [51]. Correspondingly, Langelben's study found that lies elicited a rise in BOLD response in several areas including the anterior cingulate cortex, the superior frontal gyrus, and left premotor, motor, and anterior parietal cortex. The results of the two studies cannot be compared



(a) Sample Subject 1 (b) Sample Subject 2

Figure 6. Averaged oxygenation response for two subjects for GKT, ©2002 IEEE.

Table 1. t-test results for GKT with $\alpha = 0.05$, the numbers shown are the results of $1 - P$ where P is the probability calculated at the given α. ©2002 IEEE.

Ch1: 96.96%	Ch3: 88.14%	Ch5: 100%	Ch7: 100%	Ch9: 99.97%	Ch11: 100%	Ch13: 100%	Ch15: 100%
Ch2: 100%	Ch4: 100%	Ch6: 100%	Ch8: 100%	Ch10: 100%	Ch12: 100%	Ch14: 21.6%	Ch16: 100%

straightforwardly as Izzetoglu's system covered only the frontal lobe area and their results saw significant difference in almost all channels except one (see table 1). In contrast, the BOLD response were localized in several areas with perhaps the superior frontal gyrus as the region most likely corresponding to Izzetoglu's results.

Several of studies using NIR technolgy examined the potential for military applications by monitoring subject's hemodynamic response to simulations of complex military-related tasks with variable levels of difficulty. Takeuchi examined hemodynamic response to flight simulation tasks [102]. In it, an aircraft landing procedure was simulated. The aircraft was modeled after Japanese Air Self Defense Force (JASDF) T-2 jet trainer. The procedure for the task consisted of a minute of rest with eyes closed, then one minute of simulated level flight after which the subjects were instructed to descend under variable wind conditions. After the touchdown, the subjects were instructed to taxi the simulated aircraft along the runway center line and then brake to a full-stop as quickly as possible. Rise in HbO level in the left forehead was observed during the descent phase with peaks during the touchdown or the taxi phase (see figure 7). Takeuchi reports that the amount of increase in

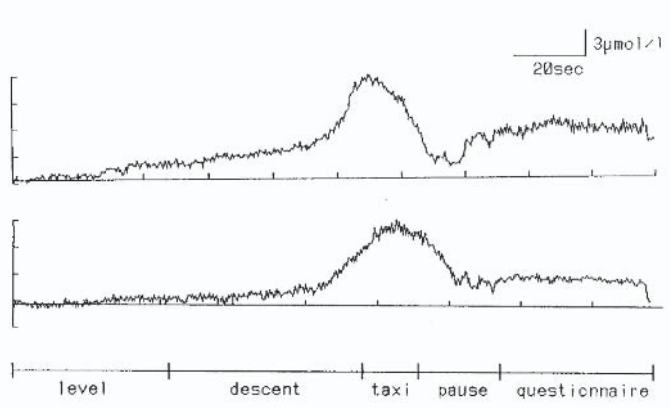


Figure 7. Change in HbO on the left forehead under 20 knot crosswind. Maximum amplitude is observed either during the touchdown (upper) or during the taxi phase (lower). ©2000 J. Occup. Health

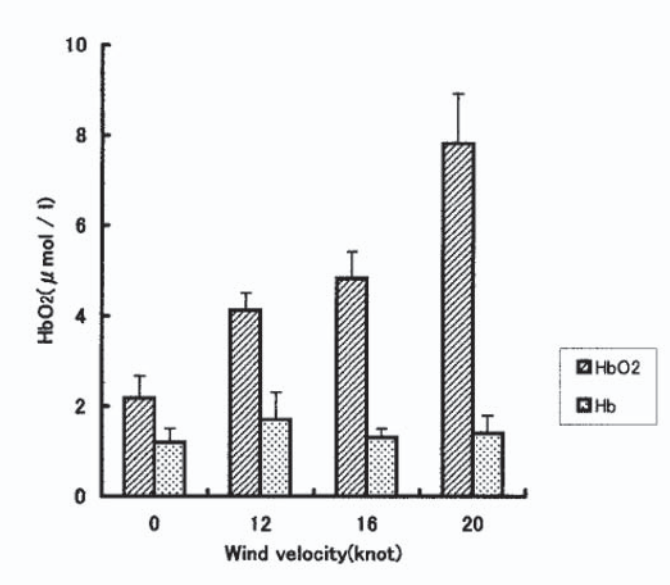


Figure 8. Averaged HbO and Hb on the left forehead under different wind conditions ©2000 J. Occup. Health

HbO varied directly with increase in wind velocity simulated. He also reports seeing little to no change in Hb level (see figure 8).

Izzetoglu et. al. conducted a study where air warfare management was simulated using Warship Commander (WC) task developed by Pacific

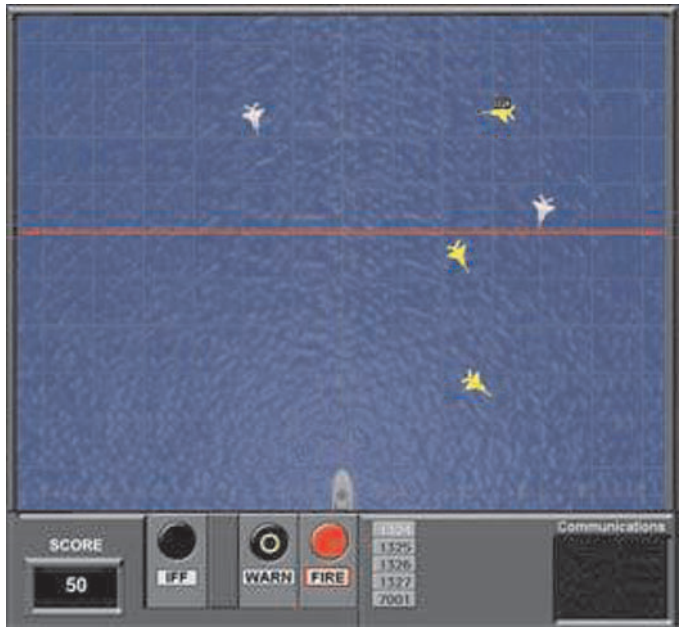
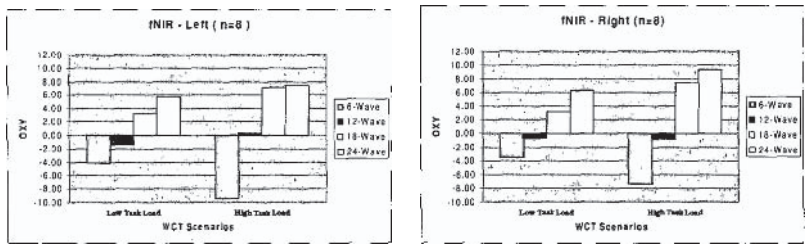


Figure 9. Warship Commander Simulator. The yellow colored tracks represent a more difficult track.

Science and Engineering Group under the guidance of Space and Naval Warfare Systems Center. Figure 9 shows a screen capture of the simulator. The objective of the task is to identify different aircrafts or tracks that appear on the screen as "friendly" or "hostile" and act accordingly while simultaneously monitoring ship status and communication. The difficulty was varied by varying the number of aircrafts and number of more difficult "yellow" tracks. They found that as number of aircrafts per wave increased, an increase in oxygenation level was observed with the lowest setting (6 aircrafts per wave) eliciting oxygenation decrease (see figure 10).

Our own work attempted to correlate the hemodynamic response to an auditory task with simulated response using ACT-R computational cognitive model. For reference on ACT-R model, see [2]. The subjects were given audio stimuli at varying interstimulus intervals (6, 4, 2, and 1 seconds). The task was divided into four 10 minute blocks with each block further divided into eight 72 second intervals, during which interstimulus interval remained constant, with at least 10 second rest period between each interval and a 2 minute rest between blocks. Our goal was to correctly identify the workload, as defined by the interstimulus intervals, from the hemodynamic responses. There were eight examples of



(a) Left forehead (b) Right forehead

Figure 10. Averaged oxygenation level for left and right forehead for varied number of aircrafts for high and low task load corresponding to the number of "yellow" tracks.

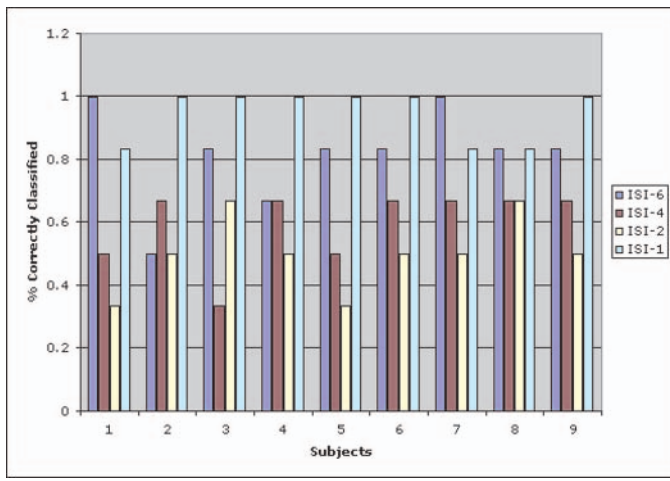


Figure 11. Percentage of correct classification for each workload for 9 subjects.

each interstimulus intervals. We trained using two examples of each interstimulus interval. The result of the classification is given in figure 11. The best result were obtained for the highest workload (interstimulus interval of 1 second) with the most confusion found in the middle two. These results, we found, were in agreement with ACT-R simulation in that ACT-R found the 1 second interval to be an "overloaded" condition and predicted that the response time needed was a little over 1 second. Thus, the response for 1 second interstimulus interval was expected to be more highly differentiable from the rest.

6. Concluding Remarks and Future Directions

To summarize, current state of NIR research consists of two major groups of instrumentation; (1) continuous wave and (2) time-resolved and frequency domain; and two major groups of parameters assessed (1) slow responding hemodynamic (HbO and Hb) parameters and (2) fast response neuronal parameter. Assessing the fast response parameter requires high temporal resolution provided by NIR equipment. Such temporal resolutions are currently not possible using fMRI modality. Thus, a natural complementary relationship exists between fMRI and NIR methods, where fMRI can provide better spatial localization and NIR better temporal resolution.

A preference of hemodynamic response is mainly due to better contrast-to-noise ratio and the ease of comparability with fMRI BOLD response which is inversely related to changes in Hb, over measurement of neuronal activity. However, measurement of the fast neuronal response can potentially benefit NIR methods by addressing the inherent pitfalls of MBLL assumptions in slow response measurements. Fast response, aside from being a direct measurement of neuronal activity, also has the potential to provide better spatial resolution.

Another promising alternative to the rather restrictive MBLL model is the layer model using diffusion equation. Diffuse optical imaging provides a better model for photon migration where approximation of path length is possible. This would provide a much improved focal change information and potentially eliminate crosstalk noise.

The major difficulty in any functional brain studies is inter-subject comparisons. Essentially, all NIR functional brain studies report high variability between subjects. No true comparison can be made due to relative nature of hemodynamic data. Since path length is unknown under MBLL, it is not possible to assess absolute levels of hemodynamic parameters. Instead, each subject is compared with their own baseline measurements. However, generally speaking, the shape of the time evolution of the signals are similar between subjects and thus can be analyzed using qualitative comparisons.

There are several future directions that NIR brain sensing and imaging research can take. In instrumentation, advances in time-resolve spectroscopic equipment may yield less expensive equipment and thus a more prolific use. This will allow for approximation of time of flight parameter providing a possible avenue for inferring path length. Theoretically, there is a need for better theoretical modeling to eliminate crosstalk noise. Possible improvements have already been introduced by Boas, et. al [8]. However, more human subject studies need to be conducted to

verify the feasibility of the model. Also, diffuse optical imaging is computationally more intensive than the MBLL based methods. Task specific variations in source-detector arrangements may also provide higher resolution in the areas of interest. A more dense arrangement would naturally provide better spatial resolution although there is a physical limit as the signal must penetrate through the skull. The authors acknowledge that the field is still young and results are preliminary. As a consequence the field provides abundance of research opportunities.

References

[1] P. D. Adelson, E. Nemoto, M. Scheuer, M. Painter, J. Morgan, and H. Yonas. Noninvasive continuous monitoring of cerebral oxygenation periictally using near-infrared spectroscopy: a preliminary report. *Epilepsia*, 17:89–99, 2002.

[2] J. R Anderson and C. Lebiere. *The Atomic Components of Thought*. Lawrence Erlbaum Associates, 1998.

[3] S. R. Arridge. Photon measurement density functions. part 1: Analytical forms. *Applied Optics*, 34:7395–7409, 1995.

[4] S. R. Arridge. Optical tomography in medical imaging. *Inverse Problems*, 15:R41–R93, 1999.

[5] M. Bartocci, J. Winberg, G. Papendieck, T. Mustica, G. Serra, and H. Lagercrantz. Activation of olfactory cortex in newborn infants after odour stimulation. *Pediatric Research*, 50:324–330, 2001.

[6] M. Bartocci, J. Winberg, C. Ruggiero, L. Begqvist, G. Serra, and H. Lagercrantz. Activation of olfactory cortex in newborn infants after odour stimulation. *Pediatric Research*, 48:18–23, 2000.

[7] D. A. Boas, D. H. Brooks, E. L. Miller, C. A. DiMarzio, M. Kilmer, R. J. Gaudette, and Q. Zhang. Imaging the body with diffuse optical tomography. *IEEE Signal Processing Magazine*, 18(6):57–75, 2001.

[8] D. A. Boas, A. M. Dale, and M. A. Franceschini. Diffuse optical imaging of brain activation: approaches to optimizing image sensitivity, resolution, and accuracy. *NeuroImage*, 23:S275–S288, 2004.

[9] D. A. Boas, T. Gaudette, G. Strangman, X. Cheng, J. J. Marota, and J. B. Mandeville. The accuracy of near infrared spectroscopy and imaging during focal changes in cerebral hemodynamics. *NeuroImage*, 13:76–90, 2001.

[10] R. F. Bonner, R. Nossal, S. Havlin, and G. H. Weiss. Model for photon migration in turbid biological media. *Journal of the Optical Society of America A*, 4(3):423–432, 1987.

[11] B. Chance, E. Anday, S. Nioka, S. Zhou, L. Hong, K. Worden, C. Li, T. Murray, Y. Ovetsky, D. Pidikiti, and R. Thomas. A novel method for fast imaging of brain function, non-invasively, with light. *Optical Express*, 2(10):411–423, 1998.

[12] B. Chance, J. S. Leigh, H. Miyake, D. S. Smith, S. Nioka, R. Greenfield, M. Finander, K. Kaufman, W. Lery, M. Yong, P. Cohn, H. Yoshioka, and R. Boretsky. Comparison of time-resolved and unresolved measurement of deoxy hemoglobin in brain. In *Proceedings of National Academy of Science*, pages 4971–4975, 1988.

[13] B. Chance, S. Nioka, J. Kent, K. McCully, M. Fountain, R. Greenfeld, and G. Holtom. Time-resolved spectroscopy of hemoglobin and myoglobin in resting and ischemic muscle. *Anal. Biochem.*, 174:698–707, 1988.

[14] L. B. Cohen. Changes in neuron structure during action potential propagation and synaptic transmission. *Physiological Review*, 53:373–413, 1973.

[15] W. N. Colier, V. Quaresima, B. Oeseburg, and M. Ferrari. Human motor-cortex oxygenation changes induced by cyclic coupled movements of hand and foot. *Experimental Brain Research*, 129:457–461, 1999.

[16] M. Cope. *The development of a near-infrared spectroscopy system and its application for noninvasive monitoring of cerebral blood and tissue oxygenation in the newborn infant.* PhD thesis, University College London, London, 1991.

[17] M. Cope and D.T. Delpy. System for long-term measurement of cerebral blood flow and tissue oxygenation on newborn infants by infra-red transillumination. *Medical and Biological Engineering and Computing*, 28:289–294, 1988.

[18] S. Coyle, T. Ward, C. Markham, and G. McDarby. On the suitability of near-infrared (nir) systems for next-generation brain-computer interfaces. *Physiological Measurement*, 25:815–822, 2004.

[19] D. T. Delpy, M. Cope, P. van der Zee, S. Arridge, S. Wray, and J. Wyatt. Estimation of optical path length through tissue from direct time of flight measurements. *Physics in Medicine and Biology*, 33:1433–1442, 2004.

[20] A. Devaraj, M. Izzetoglu, K. Izzetoglu, S. C. Bunce, C. Y. Li, and B. Onaral. Motion artifact removal in FNIR spectroscopy for real-world applications. In *Nondestructive Sensing for Food Safety, Quality, and Natural Resources. Edited by Chen, Yud-Ren; Tu, Shu-I. Proceedings of the SPIE, Volume 5588, pp. 224-229 (2004).*, pages 224–229, October 2004.

[21] A. Duncan, J. H. Meek, M. Clemence, C. E. Elwell, L. Tyszczuk, M. Cope, and D. Delpy. Optical path length measurements on adult head, calf and forearm and the head of newborn infants using phase resolved spectroscopy. *Physics in Medicine and Biology*, 40:295–304, 1995.

[22] C. E. Elwell, H. Owen-Reece, J. S. Wyatt, M. Cope, E. O. Reynolds, and D. T. Delpy. Influence of respiration and changes in expiratory pressure on cerebral hemoglobin concentration measured by near infrared spectroscopy. *Journal of Cerebral Blood Flow and Metabolism*, 16:353–357, 1996.

[23] C. E. Elwell, R. Springett, E. Hillman, and D. T. Delpy. Oscillations in cerebral hemodynamics. implications for functional activation studies. *Advances in Experimental and Medical Biology*, 471:57–65, 1999.

[24] M. Essenpreis, C. E. Elwell, M. Cope, and D. T. Delpy. Spectral dependence of temporal point spread functions in human tissues. *Applied Optics*, 32:418–425, 1993.

[25] M. Firbank, E. Okada, and D. T. Delpy. A theoretical study of the signal contribution of regions of the adult head to near-infrared spectroscopy studies of visual evoked responses. *NeuroImage*, 8:69–78, 1998.

[26] M. A. Franceschini and D. A. Boas. Noninvasive measurement of neuronal activity with near-infrared optical imaging. *NeuroImage*, 21:372–386, 2004.

[27] M. A. Franceschini, S. Fantini, J. H. Thompson, J. P. Culver, and D. A. Boas. Hemodynamic evoked response of the sensorimotor cortex measured noninvasively with near-infrared optical imaging. *Psychophysiology*, 40:548–560, 2003.

[28] M. A. Franceschini, V. Toronov, M. E. Filiaci, E. Gratton, and S. Fantini. Online optical imaging of the human brain with 160-ms temporal resolution. *Optics Express*, 6(3):49–57, 2000.

[29] R. D. Frostig, E. E. Lieke, D. Y. Ts'o, and A. Grinvald. Cortical functional architecture and local coupling between neuronal activity and the microcirculation revealed by in vivo high-resolution optical imaging of intrinsic signals. *Proceedings of National Academy of Science USA*, 87:6082–6086, 1990.

[30] E. Gratton, S. Fantini, M. A. Franceschini, G. Gratton, and M. Fabiani. Measurements of scattering and absorption changes in muscle and brain. *Philosophical Transactions Royal Society of London*, 352:727–735, 1997.

[31] E. Gratton, W. W. Mantulin, M. J. vandeVen, J. B. Fishkin, M. B. Maris, and B. Chance. The possibility of a near-infrared optical imaging system using frequency-domain methods. In *Proceedings of of 3rd International Conference on Peace through Mind/Brain Science*, pages 183–189, 1990.

[32] E. Gratton, V. Toronov, U. Wolf, M. Wolf, and A. Webb. Measurement of brain activity by near-infrared light. *Journal of Biological Optics*, 10(1):011008–1–011008–13, 2005.

[33] G. Gratton, P. M. Corbaliis, E. Cho, M. Fabiani, and D. Hood. Shades of grey matter: non-invasive optical images of human brain responses during visual stimulation. *Psychophysiology*, 32:505–509, 1995.

[34] G. Gratton and P. M. Corballis. Removing the heart from the brain: Compensation for the pulsatile artifact in the photon migration signal. *Psychophysiology*, 32:292–299, 1995.

[35] G. Gratton and M. Fabiani. The event-related optical signal: a new tool for studying brain function. *International Journal of Psychophysiology*, 42:109–121, 2001.

[36] G. Gratton and M. Fabiani. Shedding light on brain function: the event-related optical signal. *Trends in Cognitive Sciences*, 5(8):357–363, 2001.

[37] G. Gratton, M. Fabiani, D. Friedman, M. A. Franceschini, S. Fantini, and E. Gratton. Photon migration correlates of rapid physiological changes in the brain during a tapping task. *Journal of Cognitive Neuroscience*, 7:446–456, 1995.

[38] G. Gratton, A. Sarno, E. Maclin, P. M. Corballis, and M. Fabiani. Toward noninvasive 3-d imaging of the time course of cortical activity: Investigation of the depth of the event-related optical signal. *NeuroImage*, 11:491–504, 2000.

[39] H. R. Heekeren, M. Kohl, H. Obrig, R. Wenzel, W. v. Pannwitz, S. Matcher, U. Dirnagl, C. E. Cooper, and A. Villranger. Noninvasive assessment of changes in cytochrom-c-oxidase oxidation in human subjects during visual stimulation. *Journal of Cerebral Blood Flow and Metabolism*, 19:592–603, 1999.

[40] H. R. Heekeren, H. Obrig, R. Wenzel, K. Eberle, J. Ruben, K. Villringer, R. Kurth, and A. Villringer. Cerebral haemoglobin oxygenation during sustained visual stimulation - a near-infrared spectroscopy study. *Philosophical Transactions: Biological Sciences*, 352:743–750, 1997.

[41] M. Hiraoka, M. Firbank, M. Essenpreis, M. Cope, S. R. Arridge, P. van der Zee, and D. T. Delpy. A monte carlo investigation of optical pathlength in inhomogeneous tissue and its application to near-infrared spectroscopy. *Physics in Medicine and Biology*, 38:1859–1876, 1993.

[42] C. Hirth, H. Obrig, K. Villringer, A. Thiel, J. Bernarding, W. Muhlnickel, H. Flor, U. Dirnagl, and A. Villringer. Non-invasive functional mapping of the human motor cortex using near-infrared spectroscopy. *NeuroReport*, 7:1977–1981, 1996.

[43] C. Hock, K. Villringer, F. Muller-Spahn, R. Wenzel, H. Heekeren, S. Schuh-Hofer, M. Hofmann, S. Minoshima, M. Schwaiger, U. Dirnagl, and A. Villringer. Decrease in parietal cerebral hemoglobin oxygenation during performance of a verbal fluency task in patients with alzheimer's disease monitored by means of near-infrared spectroscopy (nirs) – correlation with simultaneous rcbf-pet measurements. *Brain Research*, 755:293–303, 1997.

[44] Y. Hoshi. Functional near-infrared optical imaging: Utility and limitations in human brain mapping. *Psychophysiology*, 40:511–520, 2003.

[45] Y. Hoshi, S.-J. Chen, and M. Tamura. Spatiotemporal imaging of human brain activity by functional near-infrared spectroscopy. *American Laboratory*, pages 35–39, 2001.

[46] Y. Hoshi and M. Tamura. Detection of dynamic changes in cerebral oxygenation coupled to neuronal function during mental work in man. *Neuroscience Letters*, 150:5–8, 1993.

[47] Y. Hoshi and M. Tamura. Fluctuations in the cerebral oxygenation state during the resting period in functional mapping studies of the human brain. *Medical and Biological Engineering and Computing*, 35:328–330, 1997.

[48] K. Izzetoglu, S. Bunce, M. Izzetoglu, B. Onaral, and K. Pourrezaei. fnir spectroscopy as a measure of cognitive task load. In *Proceedings of the 25th Annual International Conference of the IEEE EMBS*, pages 3431–3434, 2003.

[49] K. Izzetoglu, S. Bunce, M. Izzetoglu, B. Onaral, and K. Pourrezaei. Functional near-infrared neuroimaging. In *Proceedings of the 26th Annual International Conference of the IEEE EMBS*, pages 5333–5336, 2004.

[50] K. Izzetoglu, G. Yurtsever, A. Bozkurt, and S. Bunce. Functional brain monitoring via nir based optical spectroscopy. In *Bioengineering Conference, 2003 IEEE 29th Annual, Proceedings of*, pages 335–336, 2003.

[51] K. Izzetoglu, G. Yurtsever, A. Bozkurt, B. Yazici, and S. Bunce. Nir spectroscopy measurements of cognitive load elicited by gkt and target categorization. In *Proceedings of 36th Hawaii International Conference on System Sciences*. IEEE, 2002.

[52] G. Jasdzewski, G. Strangman, J. Wagner, K. K. Kwong, R. A. Poldrack, and D. A. Boas. Differences in the hemodynamic response to event-related motor and visual paradigms as measured by near-infrared spectroscopy. *NeuroImage*, 20:479–488, 2003.

[53] F. F. Jobsis. Noninvasive, infrared monitoring of cerebral and myocardial oxygen sufficiency and circulatory parameters. *Science*, 198:1264–1267, 1977.

[54] O. Josephs, R. Turner, and K. Friston. Event related fmri. *Human Brain Mapping*, 5:243–248, 1997.

[55] R. Kennan, D. Kim, A. Maki, H. Koizumi, and R. T. Constable. Non-invasive assessment of language lateralization by transcranial near infrared optical topography and functional mri. *Human Brain Mapping*, 16:183–189, 2002.

[56] A. Kleinschmidt, H. Obrig, M. Requardt, K. D. Merboldt, U. Dirnagl, A. Villringer, and J. Frahm. Simultaneous recording of cerebral oxygenation changes

during human brain activation by magnetic resonance imaging and near-infrared spectroscopy. *Journal of Cerebral Blood Flow and Matabolism*, 16:817–826, 1996.

[57] M. Kohl, C. Nolte, H. R. Heekeren, S. Horst, U. Scholz, H. Obrig, and A. Villringer. Changes in cytochrome-oxidase oxidation in the occipital cortex during visual stimulation: Improvement in sensitivity by the determination of the wavelength dependence of the differential pathlength factor. In *Proceedings of SPIE*, volume 3194, pages 18–27, 1998.

[58] M. Kohl, C. Nolte, H. R. Heekeren, S. Horst, U. Scholz, H. Obrig, and A. Villringer. Determination of the wavelength dependence of the differential pathlength factor from near-infrared pulse signals. *Physics in Medicine and Biology*, 43:1771–1782, 1998.

[59] M. H. Kutner, C. J. Nachtsheim, and J. Neter. *Applied Linear Regression Models*. McGraw-Hill Irwin, 4 edition, 2004.

[60] D. D. Langleben, L. Schroeder, J. A. Maldjian, R. C. Gur, S. McDonald, J. D. Ragland, C. P. O'Brien, and A. R. Childress. Brain activity during simulated deception: An event-related functional magnetic resonance study. *NeurImage*, 15:727–732, 2002.

[61] H. Liu, M. Miwa, B. Beauvoit, N. G. Wang, and B. Chance. Characterization of small-volume biological sample using time-resolved spectroscopy. *Anal. Biochem.*, 213:378–385, 1993.

[62] E. L. Maclin, K. A. Low, J. J. Sable, M. Fabiani, and G. Gratton. The event-related optical signal to electrical stimulation of the median nerve. *NeuroImage*, 21:1798–1804, 2004.

[63] S. J. Madsen, B. C. Wilson, M. S. Patterson, Y. D. Park, S. L. Jacques, and Y. Hefetz. Experimental tests of a simple diffusion model for the estimation of scattering and absorption coefficients of turbid media from time-resolved diffuse reflectance measurements. *Applied Optics*, 31:3509–3517, 1992.

[64] A. Maki, Y. Yamashita, Y. Ito, E. Watanabe, Y. Mayanagi, and H. Koizumi. Spatial and temporal analysis of human motor activity using noninvasive nir topography. *Journal of Neurosicence*, 11:1458–1469, 1995.

[65] A. Maki, Y. Yamashita, E. Watanabe, and H. Koizumi. Visualizing human motor activity by using non-invasive optical topography. *Front Med Biol Eng*, 7:285–297, 1996.

[66] D. Malonek and A. Grinvald. Interactions between electrical activity and cortical microcirculation revealed by imaging spectroscopy: Implications for functional brain mapping. *Science*, 272:551–554, 1996.

[67] J. Meek. Basic principles of optical imaging and application to the study of infant development. *Developmental Science*, 5(3):371–380, 2002.

[68] J. H. Meek, C. E. Elwell, M. J. Khan, J. Romaya, J. D. Wyatt, D. T. Delpy, and S. Zeki. Regional changes in cerebral hemodynamics as a result of a visual stimulus measured by near infrared spectroscopy. *Proceedings of Royal Society of London*, 261:351–356, 1995.

[69] J. H. Meek, M. Firbank, C. E. Elwell, J. Atkinson, O. Braddick, and J. S. Wyatt. Regional hemodynamic responses to visual stimulation in awake infants. *Pediatric Research*, 43:840–843, 1998.

[70] J. H. Meek, L. Tyszczuk, C. E. Elwell, and J. S. Wyatt. Cerebral blood flow increases over the first three days of life in extremely preterm neonates. *Archives of Disease in Childhood*, 78:F33–F37, 1998.

[71] J. H. Meek, L. Tyszczuk, C. E. Elwell, and J. S. Wyatt. Low cerebral blood flow is a risk factor for sever intraventricular hemorrhage. *Archives of Disease in Childhood*, 81:F15–F18, 1999.

[72] D. J. Mehagnoul-Schipper, B. F. van der Kallen, W. N. Colier, M. C. van der Sluijs, L. J. van Erning, H. O. Thijssen, B. Oeseburg, W. H. Hoefnagels, and R. W. Jansen. Simultaneous measurements of cerebral oxygenation changes during brain activation by near-infrared spectroscopy and functional magnetic resonance imaging in healthy young and elderly subjects. *Human Brain Mapping*, 16:14–23, 2002.

[73] M. Miwa, Y. Ueda, and B. Chance. Development of time-resolved spectroscopy system for quantitative non-invasive tissue measurement. *SPIE*, 2389:142–149, 1995.

[74] I. Miyai, H. Tanabe, I. Sase, H. Eda, I. Oda, I. Konishi, Y. Tsunazawa, T. Suzuki, T. Yanagida, and K. Kubota. Cortical mapping of gait in humans: A near-infrared spectroscopic topography study. *NeuroImage*, 14:1186–1192, 2001.

[75] Y. Noguchi, T. Takeuchi, and K. Sakai. Lateralized activation in the inferior frontal cortex during syntactic processing: event-related optical topography study. *Human Brain Mapping*, 17:89–99, 2002.

[76] Y. Nomura and M. Tamura. Quantitative analysis of hemoglobin oxygenation state of rat brain in vivo by picosecond time-resolved spectrophotometry. *Journal of Biochemistry*, 109:455–461, 1991.

[77] H. Obrig, M. Neufang, R. Wenzel, M. Kohl, J. Steinbrink, K. Einhaupl, and A. Villringer. Spontaneous low frequency oscillations of cerebral hemodynamics and metabolism in human adults. *NeuroImage*, 12:623–639, 2000.

[78] H. Obrig and A. Villringer. Beyond the visible–imaging the human brain with light. *Journal of Cerebral Blood Flow and Metabolism*, 23:1–18, 2003.

[79] H. Obrig, R. Wenzel, M. Kohl, S. Horst, P. Wobst, J. Steinbrink, F. Thomas, and A. Villringer. Near-infrared spectroscopy: does it function in functional activation studies of the adult brain? *International Journal of Psychophysiology*, 35:125–142, 2000.

[80] M. Oda, Y. Yamashita, G. Nishimura, and M. Tamura. Quantitation of absolute concentration change in scattering media by the time-resolved microscopic beer-lambert law. *Advances in Experimental and Medical Biology*, 345:861–870, 1992.

[81] M. Oda, Y. Yamashita, G. Nishimura, and M. Tamura. Determination of absolute concentration of oxy- and deoxyhemoglobin in rat head by time-resolved beer-lambert law. *SPIE*, 2389:770–778, 1995.

[82] M. Oda, Y. Yamashita, G. Nishimura, and M. Tamura. A simple and novel algorithm for time-resolved multiwavelength oximetry. *Physics in Medicine and Biology*, 41:955–961, 1996.

[83] E. Okada, M. Firbank, M. Schweiger, S. R. Arridge, M. Cope, and D. T. Delpy. Theoretical and experimental investigation of near infrared light propagation in a model of the adult head. *Applied Optics*, 36:21–31, 1997.

[84] N. Okui and E. Okada. Wavelength dependence of crosstalk in dual-wavelength measurement of oxy- and deoxy-hemoglobin. *Journal of Biomedical Optics*, 10(1):011015–1–011015–8, 2005.

[85] M. S. Patterson, B. Chance, and B. C. Wilson. Time resolved reflectance and transmittance for the non-invasive measurement of tissue optical properties. *Applied Optics*, 28(12):2331–2336, 1989.

[86] D. M. Rector, R. F. Rogers, J. S. Sschwaber, R. M. Harper, and J. S. George. Scattered-light imaging in vivo tracks fast and slow processes of neurophysiological activation. *NeuroImage*, 14:977–994, 2001.

[87] E. O. Reynolds, J. S. Wyatt, D. Azzopardi, D. T. Delpy, E. B. Cady, M. Cope, and S. Wray. New non-invasive methods for assessing brain oxygenation and hemodynamics. *British Medical Bulletin*, 44:1052–1075, 2004.

[88] K. Sakai, R. Hashimoto, and F. Homae. Sentence processing in the cerebral cortex. *Neuroscience Research*, 39:1–10, 2001.

[89] K. Sakatani, S. Chen, W. Lichty, H. Zuo, and Y. P. Wang. Cerebral blood oxygenation changes induced by auditory stimulation in newborn infants measure by near infrared spectroscopy. *Early Human Development*, 55:229–236, 1999.

[90] B. M. Salzberg and A. L. Obaid. Optical studies of the secretory event at vertebrate nerve terminals. *Experimental Biology*, 139:195–231, 1988.

[91] A. Sassaroli and S. Fantini. Comment on the modified beer-lambert law for scattering media. *Physics in Medicine and Biology*, 49:N255–N257, 2004.

[92] H. Sato, T. Takeuchi, and K. Sakai. Temporal cortex activation during speech recognition: an optical topography study. *Cognition*, 40:548–560, 1999.

[93] M. L. Schroeter, S. Zysset, T. Kupka, F. Kruggel, and D. Y. von Cramon. Near-infrared spectroscopy can detect brain activity during a color-word matching stroop task in an event-related design. *Human Brain Mapping*, 17(61):61–71, 2002.

[94] I.-Y. Son, M. Guhe, W. Gray, B. Yazici, and M. J. Schoelles. Human performance assessment using fnir. In *Proceedings of SPIE*, 2005.

[95] J. Steinbrink, M. Kohl, H. Obrig, G. Curio, F. Syre, F. Thomas, H. Wabnitz, H. Rinneberg, and A. Villringer. Somatosensory evoked fast optical intensity changes detected non-invasively in the adult human head. *Neuroscience Letters*, 291:105–108, 2000.

[96] J. Steinbrink, H. Wabnitz, H. Obrig, A. Villringer, and H. Rinneberg. Determining changes in nir absorption using layered model of the human head. *Physics in Medicine and Biology*, 46:879–896, 2001.

[97] R. A. Stepnowski, J. A. LaPorta, F. Raccuia-Behling, G. E. Blonder, R. E. Slusher, and D. Kleinfeld. Noninvasive detection of changes in membrane potential in cultured neurons by light scattering. *Proceedings of National Academy of Science USA*, 88:9382–9386, 1991.

[98] G. Strangman, D. A. Boas, and J. P. Sutton. Non-invasive neuroimaging using near-infrared light. *Biological Psychiatry*, 52:679–693, 2002.

[99] G. Strangman, J. P. Culver, J. H. Thompson, and D. A. Boas. A quantitative comparison of simultaneous bold fmri and nirs recordings during functional brain activation. *NeuroImage*, 17:719–731, 2002.

[100] T. Suto, M. Fukuda, M. Ito, T. Uehara, and M. Mikuni. Multichannel near-infrared spectroscopy in depression and schizophrenia: Cognitive brain activation study. *Biological Psychiatry*, 55:501–511, 2004.

[101] G. Taga, K. Asakawa, A. Maki, Y. Konishi, and H. Koizumi. Brain imaging in awake infants by near-infrared optical topography. *PNAS*, 100(19):10722–10727, 2003.

[102] Y. Takeuchi. Change in blood volume in the brain during a simulated aircraft landing task. *Journal of Occupational Health*, 42:60–65, 2000.

[103] I. Tasaki and P. M. Byrne. Rapid structural changes in nerve fibers evoked by electric current pulses. *Biochemical and Biophysical Research Communications*, 188:559–564, 1992.

[104] I. Tasaki and P. M. Byrne. Optical changes during nerve excitation: interpretation on the basis of rapid structural changes in the superficial gel layer of nerve fiber. *Physiological Chemistry and Physics and Medical NMR*, 26:101–110, 1994.

[105] V. Toronov, A. Webb, and J. H. Choi. Investigation of human brain hemodynamics by simultaneous near-infrared spectroscopy and functional magnetic resonance imaging. *Medical Physics*, 28(4):521–527, 2001.

[106] V. Toronov, A. Webb, J. H. Choi, M. Wolf, L. Safonova, U. Wolf, and E. Gratton. Study of local cerebral hemodynamics by frequency-domain near-infrared spectroscopy and correlation with simultaneously acquired functional magnetic resonance imaging. *Optics Express*, 9:417–427, 2001.

[107] K. Uludag, M. Kohl, J. Steinbrink, H. Obrig, and A. Villringer. Cross talk in the lambert-beer calculation for near-infrared wavelengths estimated by monte carlo simulations. *Journal of Biomedical Optics*, 7(1):51–59, 2002.

[108] A. Villringer and B. Chance. Non-invasive optical spectroscopy and imaging of human brain function. *Trends In Neurosciences*, 20(10):435–442, 1997.

[109] A. Villringer, J. Planck, C. Hock, L. Schleinkofer, and U. Dirnagl. Near infrared spectroscopy (nirs): a new tool to study hemodynamic changes during activation of brain function in human adults. *Neuroscience Letters*, 154:101–104, 1993.

[110] E. Watanabe, A. Maki, F. Kawaguchi, K. Takashiro, Y. Yamashita, H. Koizumi, and Y. Mayanagi. Non-invasive assessment of language dominance with near-infrared spectroscopic mapping. *Neuroscience Letters*, 256:49–52, 1998.

[111] E. Watanabe, Y. Yamashita, A. Maki, Y. Ito, and H. Koizumi. Noninvasive functional mapping with multi-channel near infrared spectroscopic topography in humans. *Neuroscience Letters*, 205:41–44, 1996.

[112] M. Wolf, U. Wolf, J. H. Choi, R. Gupta, L. P. Safonova, and L. A. Paunescu. Functional frequency-domain near-infrared spectroscopy detects fast neuronal signal in the motor cortex. *NeuroImage*, 17:1868–1875, 2002.

[113] Y. Yamashita, M. Oda, H. Naruse, and M. Tamura. In vivo measurement of reduced scattering and absorption coefficients of living tissue using time-resolved spectroscopy. *OSA TOPS*, 2:387–390, 1996.

[114] Y. Yamashita, M. Oda, E. Ohmae, and M. Tamura. Continuous measurement of oxy- and deoxyhemoglobin of piglet brain by time-resolved spectroscopy. *OSA TOPS*, 22:205–207, 1998.

[115] P. Zaramella, F. Freato, A. Amigoni, s. Salvadori, P. Marangoni, A. Suppjei, B. Schiavo, and C. Lino. Brain auditory activation measured by near-infrared spectroscopy (nirs). *Pediatric Research*, 49:213–219, 2001.

[116] Y. Zhang, D. H. Brooks, M. A. Franceschini, and D. A. Boas. Eigenvector-based spatial filtering for reduction of physiological interference in diffuse optical imaging. *Journal of Biomedical Optics*, 10:011014–1–011014–11, 2005.

Topic Index