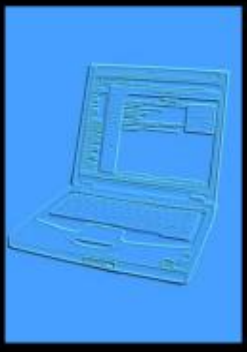
Advances in Network and Communications Engineering 2

Proceedings of the MSc/MRes Network Systems Engineering and MSc/MRes Communications Engineering & Signal Processing

2003 - 2004



Edited by

Dr Steven M Furnell
Dr Paul S Dowland

Advances in Network and Communications Engineering 2

Proceedings of the MSc/MRes Network Systems Engineering and MSc/MRes Communications Engineering & Signal Processing

2003 - 2004

Editors

Dr Steven M Furnell
Dr Paul S Dowland

School of Computing, Communications & Electronics
University of Plymouth

ISBN 1-84102-140-7

Preface

This book presents a series of research papers arising from *MSc/MRes Network Systems Engineering* and *MSc/MRes Communications Engineering & Signal Processing* research projects undertaken at the University of Plymouth. These one year masters courses include a significant period of full-time project activity, and students are assessed on the basis of an MSc or MRes thesis, plus an accompanying research paper.

The publications in this volume are based upon research projects that were undertaken during the 2003/04 academic year. A total of 22 papers are presented, covering many aspects of modern networking and communication technology, including security, mobility, coding schemes and quality measurement.

The authorship of the papers is credited to the MSc/MRes student in each case (appearing as the first named author), with other authors being the academic supervisors that had significant input into the projects. Indeed, the projects were conducted in collaboration with supervisors from the internationally recognised research groups within the School, and the underlying research projects are typically related to wider research initiatives with which these groups are involved. Readers interested in further details of the related research areas are therefore encouraged to make contact with the academic supervisors, using the contact details provided elsewhere in this publication.

Each of the papers presented here is also supported by a full MSc or MRes thesis, which contains more comprehensive details of the work undertaken and the results obtained. Copies of these documents are also in the public domain, and can generally be obtained upon request via inter-library loan.

We believe that these papers have value to the academic community, and we therefore hope that their publication in this volume will be of interest to you.

Dr Steven Furnell and Dr Paul Dowland

School of Computing, Communications and Electronics
University of Plymouth, May 2005

About the School of Computing, Communications and Electronics

This new School was formed from a merger between the School of Computing and the Department of Communication and Electronic Engineering in August 2003. It is a large multifaceted School with interests spanning across the interface between computing and art, through software, networks, and communications to electronic engineering. The School contains 61 academic staff and has 1100 students enrolled on its portfolio of taught courses, 104 of which are at MSc level. In addition there are 78 postgraduate research students enrolled on a variety of research programmes, most of which enjoy sponsorship from external sources.

The bulk of the staff in the School are housed in the Portland Square building, a purpose built state of the art building costing over £25million and situated near the centre of the historic city of Plymouth on the University campus. The laboratories are located in the newly refurbished Smeaton Building, and the Clean room for nano-technology also recently refurbished courtesy of a Wolfson Foundation grant is situated in the nearby Brunel Building. All buildings are a short walk from each other, enabling a close collaboration within our research community.

This School sits alongside two other Schools in the Faculty of Technology, the School of Engineering (the merged School of Civil and Structural Engineering and Department of Mechanical and Marine Engineering), and the School of Mathematics and Statistics. There are research and teaching links across all three schools as well as with the rest of the University. The closest links are with the Faculty of Science, principally the Centre for Computational and Theoretical Neuroscience which started in Computing, and Psychology through Artificial Intelligence and Human Computer Interaction research.

Prof. P. Dyke
Head of School

Contributing Research Groups

Fixed and Mobile Communications

Head: Professor M Tomlinson BSc, PhD, CEng, MIEE
E-mail: mtomlinson@plymouth.ac.uk

Research interests:

1) Satellite communications
2) Wireless communications
3) Broadcasting
4) Watermarking
5) Source coding and data compression

www.tech.plym.ac.uk/see/research/satcen/sat.htm
www.tech.plym.ac.uk/see/research/cdma/

Network Research Group

Head: Dr S M Furnell BSc (Hons), PhD, CEng, FBCS, CITP, SMIEEE
E-mail info@network-research-group.org

Research interests:

1) Information systems security
2) Internet and Web technologies and applications
3) Mobile applications and services
4) Network management

www.network-research-group.org

Signal Processing and Multimedia Communications

Head: Professor E Ifeachor BSc, MSc, PhD, DIC, CEng, MIEE
E-mail eifeachor@plymouth.ac.uk

Research interests:

1) Multimedia communications
2) Audio and bio-signal processing
3) Bioinformatics

www.tech.plymouth.ac.uk/spmc/

Contents

SECTION 1 Network Systems Engineering

SECTION 2 Communications Engineering and Signal Processing

Section 1

Network Systems Engineering

Passimages : An Alternative Method of User Authentication

D. Charruau, S.M. Furnell and P.S. Dowland

Network Research Group, University of Plymouth, Plymouth, UK
e-mail: info@network-research-group.org

Abstract

This paper presents an assessment of an alternative to the predominant password and PIN-based methods of user authentication. Although these approaches are in widespread use, there are many recognised problems in terms of their usage and the consequent protection that they actually provide. Therefore a graphical method using PassImages has been created in which users are authenticated from the selection of six images, chosen from a set of one hundred. A trial of the technique has been conducted via a prototype implementation of a web-based authentication process. This assessment shows that the PassImage approach provides a high level of effectiveness, with 29 trial users achieving 95% successful authentication.

Keywords

Security, Authentication, Passwords, Graphical Passwords

1. Introduction

In modern society it is not unusual to have to authenticate ourselves on several IT systems. Most of the time, these systems require a password or a PIN, but faced with the requirement to remember such information, many users encounter difficulties, which tends to result in poor choices or other bad practices. For example, passwords are often based upon dictionary words or personal information, resulting in vulnerability to attack by brute force cracking tools or social engineering (VeriSign, 2000). By contrast, enforcing better selection practices may simply lead to compromise in other ways, such as passwords being written down and left nearby the computer (often in plain sight) for the legitimate user's reference. All the while, of course, they are equally visible to potential impostors. In view of such problems, alternative methods are desirable, and common recommendations include the use of token-based or biometric approaches (Smith, 2002). However, one of the inherently attractive characteristics of a password is its low cost, and the aforementioned alternatives will typically incur additional expense. In addition, if a web-based service operator wished to authenticate users on the basis of such techniques, there would be no guarantee that the users possessed appropriate hardware. As such, the use of alternative secret-knowledge approaches may remain preferable in many contexts. Therefore, an experiment has been conducted in an attempt to evaluate an alternative method based upon selection of images rather than the recall of text sequences. This method is based on the conclusions of two previous studies conducted by Irakleous *et al.* (2001) and Furnell *et al.* (2004).

The paper begins by presenting an outline of the problems with existing password-based approaches, as well as previous attempts to utilise image-based methods as an alternative. It then proceeds to discuss the design and implementation of an alternative approach, and the results observed from a practical user trial. The implications of these results are then

discussed, along with opinion-based feedback from the trial participants, leading to the suggestion of future research directions in the concluding section of the paper.

2. Background

The vast majority of user authentication methods in operating systems, applications and websites involve the use of passwords. Indeed, passwords remain the method of choice in spite of recognised vulnerabilities, many of which arise from the behaviour of users. Passwords have been the way to authenticate on IT systems since the first computers were created in the early 1960's (Morris and Thompson, 1979). In the last two decades, other aspects of computer interfaces have changed significantly (e.g. the arrival of Graphical User Interface (GUI) environments), but people are reluctant to change their security systems for something new (Bensinger, 1998). As a result, an authentication method inherited from the command line age is still in use. Studies have shown that the end users' behaviour introduced the majority of the password weaknesses, by sharing their password or by choosing passwords that are easy to remember. For an intruder these passwords became easy to guess (Boroditsky, 1998). For example, a previous study has shown that on a sample of 15,000 passwords 21% of them have been cracked in less than a week and 2.7% in less than 15 minutes (Klein 1990). This suggests that allowing end users to choose their passwords effectively introduces weaknesses in the security system. In order to increase the security, administrators tend to provide passwords to the users, but then other problems arise: because the password is no longer simple to remember, people start to write it down, and the effect is even worse (Boroditsky, 1998). By the early 1990's an Internet Engineering Task Force (IETF) request for comments (RFC) was already taking the matter as a serious security threat, and proposing the minimum requirements that a password must comply with – namely being at least 6 characters long, and composed of characters drawn from mixed case alphabetic, punctuation symbols and digits (Holbrook and Reynolds,1991).

In many ways GUI-based authentication methods using images are considered better than passwords. The reason is that images are easier to remember than a string of letters. This is due to the fact that the human brain has difficulty remembering information when it is not part of a context. On the other hand, an image can easily provide a context by itself (Bensinger, 1998). According to psychology researchers, the human brain is good at recognising images. Two studies are used as references to explore this ability. In the first test, 2,560 photos were presented to an audience, with each image shown for a few seconds. The users then had to examine a set of images composed of new and already seen images. During the test, participants had to indicate the images seen before. The result of this experiment was a 90% recognition rate (Standing *et al.*, 1970). Another study was carried out and followed a similar principle. The audience saw 10,000 pictures in two days and performed a recognition rate of 60% (Standing, 1973).

In addition to this ability to easily recognize images, a study has shown that image pin based methods were easy to use: in a study involving 27 participants, 63% were successfully able to authenticate. In parallel to this experiment, password authentication was studied and gave approximately the same rate of success (Irakleous *et al*., 2002a). Other methods have also been developed such as the "Déjà Vu" authentication method created by Dhamija and Perrig (2000). This method is based on the memory of images, but does not require a precise

sequence. Two types of images have been used to evaluate this method, namely photos and random art images. The photos are complex sceneries and the random art images are images drawn by a computer using random parameters inserted into a mathematical formula. The probability to be able to masquerade as another person with this method is unlikely to occur since a sequence of 5 images on a matrix composed of 25 images is required to authenticate. However, as described below, this still yields notably fewer combinations than a (correctly used) six character password.

The aim of the research at this stage was to devise a potential replacement for password-based authentication, while retaining a secret-knowledge based approach and providing a comparable level of protection to a password selected on the basis of the recommendations in RFC 1244 (Holbrook and Reynolds, 1991). A new method (hereafter referred to as the PassImage method) was devised that attempted to provide a user-friendly authentication approach based upon the selection of on-screen images.

3. Methodology

The guidelines of the aforementioned RFC 1244 indicate a total of 95^6 possible password combinations. Therefore it has been decided to create a method that allows the user to choose six images from a total of 100 images. In order to prevent "shoulder surfing" the images are displayed randomly on four different grids each time the process is launched. The images themselves all depict objects from everyday life, as illustrated in Figure 1. It is therefore hoped that users will be able to recognise the objects, and select six that they are most comfortable with.

Figure 1 : PassImage example

In order to create a method that can be used by the largest number of potential users it was decided to design a web-based authentication procedure. The use of a web interface had many advantages compared to other means of assessment, since there was no need to distribute software to potential participants, and most operating systems can support the method as it was written in JavaScript.

The authentication process was made as simple and secure as possible. Therefore in order to select an image, the user is only required to do a simple click on the image that they want to choose. For security purposes, images chosen are not displayed since it would be easy for prying eyes to catch the selection. Therefore a system of 'traffic lights' was implemented. Each selection from the user switches on an amber light (see figure 2). Once all selections have been made, and if the user achieves authentication, all the traffic lights become green (see figure 3), on the other hand if the user fails, the traffic lights become red. Users can change the grid and cancel the last selection either with the button provided or with keyboard keys. Shortcut keys are useful to accelerate the choice of the images by reducing the need for moving the mouse pointer from the grid to the buttons and back to the grid.

In order to simplify the authentication procedure it was decided that the system will assist the legitimate user in recalling the correct sequence of images. To achieve this, the six pictures that comprise the PassImage are always displayed back in the right order on the login grids. For example if a user chose the PassImage shown in Figure 1, then in all subsequent authentication sessions these would appear in the grids in this order (i.e. the image of the chair will always be the first one that the user will encounter when looking through the choices available). This enables the user to scan each grid from left to right, top to bottom, with no requirement to hunt back and forth between the grids (i.e. unless the user inadvertently misses one of their images, they should only need to advance forward to the next grid, rather than back to a previous one).

The concept is illustrated in figures 2 and 3, which depict the PassImage login in operation. The required selections are indicated by the shaded images, along with a number indicating the sequence in which each image has to be chosen (note: the shading and numbering do not appear in the live operation of the system, and have been added to the screenshots to help clarify the process involved). It should be noted that although the example depicts the user's images being spread over two different grids, this will not always be the case. Apart from ensuring that the images appear in order, their placement is done randomly; so on different occasions they may be spread over up to four different grids.

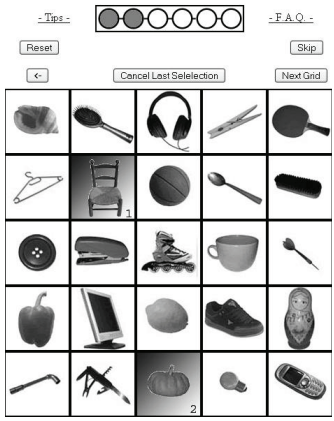


Figure 2 : Authentication Process

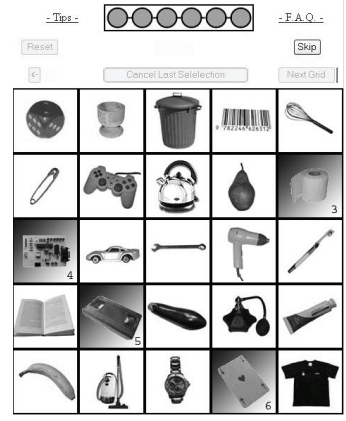


Figure 3 : Authentication achieved

In addition to measuring the success or failure of authentication attempts, the system used for the trial was also able to log the time that users took to choose the images composing the PassImage, the number of access attempts they made during the trial period, and the time taken for each authentication attempt. At the end of the trial period, a small survey of the participants was conducted via an online questionnaire, in order to collect user opinions regarding the PassImage method.

4. Experimental results

Twenty-nine users were involved in assessing the method, during a total period of 90 days. During this time, the PassImage website was set as the new homepage for each participant's web browser. As such, each time they loaded the browser, they were prompted to provide their user identity and then authenticate themselves via PassImage. Successful authentication then initiated automatic redirection to their original browser homepage. In order to foster goodwill amongst the trialists, they were not obliged to use the method each and every time they loaded the browser, and a 'skip' button was offered as a quick route to their normal homepage. For participants who had forgotten their PassImage, the system offered an option for it to be recovered.

None of the trialists used the authentication method for the full 90 days of the study period. The average period of usage was 38 days, with users having performed an average of 31 trials. The numbers of trials varied considerably from one user to another. For example, the maximum of trials was 213, while the minimum was six.

The result shows that the users achieved a high rate of authentication. From a total 911 trials, the users were able to authenticate on 867 occasions. This gives an authentication rate of 95%, and a rejection rate of only 5%. However, in addition to this result it should be noted that users had to retrieve their PassImage on only three occasions. Therefore if only the retrievals are taken into account to calculate the number of authentication failures, the authentication success becomes 99.6%. Another interesting point is that there was only one occasion during the trial in which a user made three errors in a row. This suggests that a standard security policy of blocking the account after three consecutive rejections is likely to have low impact upon the activities of legitimate users.

A measurement, which is very important for such techniques, is the time spent by the users to set up an account and then to authenticate during subsequent logins. The selection of the PassImage was a relatively long process, and on average, users spent two minutes to perform this task. This is, however, justifiable in the sense that users should consider their choices carefully in order to ensure that they remember them later. Figure 4 illustrates that as users made more use of the system, the time taken to authenticate steadily decreased. The results of the measurements made on the time spent to authenticate, show that after a short usage period, users are, on average, able to authenticate in around twenty seconds. This is still somewhat longer than the typical time taken for password-based authentication, but this could arguably be set against the potential security benefits of the new approach.

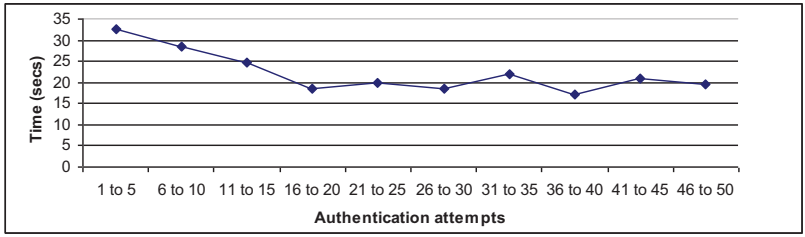


Figure 4 : Average time taken to authenticate

Since security was one of the main objectives of the experimentation, the choice of PassImage has been scrutinised in order to find weaknesses implied from a poor set of image selections. Although the selection process did impose some restrictions upon the users' choices (e.g. they were not permitted to choose multiple instances of the same image), there were other ways in which potentially weak choices could be made. For example, users could conceivably choose all six of their images from the same category of pictured object (e.g. food and drink, clothing and footwear, etc.). Therefore all the images were categorised and all the PassImages were parsed to see if they met such criteria. This revealed that, out of 37 PassImages chosen by users during the trial, 5 were classed as weak choices because the constituent images all belonged to the same category. Also related to the issue of image choices is that the post-trial comments from one participant suggested that two of his relatives had nearly found his PassImage. This suggests that the method may face some potential for attack through social engineering.

5. Discussion

In addition to the analysis performed, the users were asked to provide their opinions about the method. All users felt that the implementation using a web-based interface was appropriate, 89% of the users felt that the method operated quickly, and 8% of the users found that the images were hard to recognise. Despite the high rate of successful authentication, almost a quarter (23%) considered it hard to use (against 27% who found it 'very easy' and 39% who classed it 'easy'). From the users' perspective, the perceived ease of use of the method was closely tied to their ability to remember the necessary images - 27% thought that remembering the six images was very easy, 39% thought it was easy and 23% thought it was hard. The theoretical chance for an impostor to guess the correct PassImage is one in 858,277,728,000 (based upon six images chosen from 100, without duplicates). To determine how the security was perceived by users, they were asked to rate the chances of a person remembering their PassImage witnessed during the authentication process, 42% thought that it would be very hard, 35% that it would be hard and only 4% of the users thought that it would be easy. A question on how many images users would have chosen was also asked; it showed that 54% of the users would choose six images. This result is not surprising since the experiment was based on the same choice. However 34% of the users preferred to choose fewer images and only 11% of the users would choose more than six images. Even though the analysis of the authentication time showed that users, on average, were able to authenticate in around twenty seconds, 39% of the users thought that the method was too time consuming. The last question asked if the users thought that this alternative method could

replace the present means of authentication, 73% believed that it could, while 27% considered that it would not be feasible. The main reason expressed in the latter case was the time taken for authentication when compared with typing a password. A secondary factor was the difficultly in remembering the images.

When comparing the results to earlier studies, some further positive observations can be made. In the study conducted by Irakleous, a similar technique only achieved 63% success. The present method has also achieved a better effectiveness than the study carried out by Furnell *et al.* (2004), which had an effectiveness of 84% from 378 attempts. The measurements resulting from this analysis can also be compared with results from the "déjà vu" research performed by Dhamija and Perrig (2000). In the "déjà vu" research, users only spent a minute to choose their images whereas in this method results show that the choice of the PassImage is quite a long process since users, on average, took more than two minutes to make their selections. The "déjà vu" research also showed that the users were able to authenticate in an average of twenty-seven seconds, whereas with the current experimentation users spent an average time of twenty seconds after a short usage period, and that the average time taken over the whole experiment was twenty three seconds. Furthermore it should be noted that the requirement for the authentication was not the same. In the "déjà vu" experiment, the users were asked to select five photos from amongst twenty other photos on the same screen. Therefore it can be concluded that the choice of displaying simple objects rather than complex images may simplify the user's choice.

6. Conclusion

The practical study revealed a good approval rate of the PassImage method, and a high level of effectiveness (albeit amongst a relatively small user population). However, some issues need to be addressed. Even though users considered the web-based interface to be appropriate, some imperfections have to be addressed, such as more attention to the production of a better set of images. It is believed that if better images can be produced, the difficulties remembering them may decrease, as well as avoiding obvious categorisation issues. In order to prevent social engineering, a possible way would be to create a larger image database and to filter the images that the user can choose in accordance with a questionnaire about his/her work and hobbies.

In terms of the implementation, a better way to assess such a method would be to integrate it into a system in which users traditionally expect to login, rather than as a voluntary additional layer within an application that normally proceeds without authenticating the user. Another necessary evaluation would be to perform the assessment with trialists using several accounts. Therefore the effect of having to remember multiple PassImages could be studied, revealing whether it is possible to use the PassImage as intensively as the PIN and passwords that are currently used across many different systems. Other techniques to reduce the time spent in order to authenticate have to be found since it will lead to a better acceptance of the method.

Once these issues have been addressed, the method would benefit from a larger scale trial (without the option for the users to skip the authentication process – which would help to yield a more accurate impression of their acceptance of the technique).

7. References

Bensinger, D. (1998) “Human Memory and the Graphical Password”, p.2. www.PassLogix.com (accessed 10/12/03).

Boroditsky, M. (1998) “Passwords Security Weaknesses & User Limitations”, p.2. www.PassLogix.com (accessed 10/12/03).

Dhamija, R. and Perrig, A. (2000) “Déjà Vu: A User Study Using Images for Authentication”, In *Proceedings of the 9th USENIX Security Symposium*, August 2000.

Furnell, S.M., Papadopoulos, I. and Dowland, P. (2004) “A long-term trial of alternative user authentication technologies”, *Information Management & Computer Security*, vol. 12, no. 2: pp178-190.

Holbrook, P and Reynolds, J. (1991) “RFC 1244 Site Security Policy Handbook Working Group”, p.58, www.ietf.org (accessed 10/12/03).

Irakleous I., Furnell S.M., Dowland P.S. and Papadaki M. (2002) “An experimental comparison of secret-based user authentication technologies”, *Information Management & Computer Security*, vol. 10, no. 3, pp100-108

Klein, D. (1990) “Foiling the Cracker: A Survey of, and Improvements to, Password Security”, in *Proceedings of the Second USENIX Security Workshop*, Portland, Oregon, August 1990, pp5-14.

Morris, R. and Thompson, K. (1979) “Password Security: A Case History”, *Communications of the ACM*, vol.22, no.11, pp594-597

Smith, R.E. (2002) *Authentication. From Passwords to Public Keys*. Addison Wesley.

Standing, L., Conezio J. and Haber R., (1970) “Perception and memory for pictures: Single-trial learning of 2500 visual stimuli”, *Psychonomic Science*, 19(2):73-74, 1970.

Standing, L. (1973) “Learning 10,000 pictures”, *Quarterly journal of Experimental Psychology*, 25:207-222, 1973.

VeriSign. (2000) *The Security Risks of Using Passwords*, VeriSign White Paper, available online: itpapers.news.com (accessed 1/10/2004).

Reliability of Commercial Biometric Authentication Solutions

C. Mustiere and N.L. Clarke

Network Research Group, School of Computing, Communications & Electronics,
University of Plymouth, UK
e-mail: info@network-research-group.org

Abstract

Since informatics exists, security measures are evolving as fast as the IT (Information Technology) world nevertheless; the threat of stolen data is still present and remains capital for company assets. Furthermore, the biometric technique is growing fast and hence, some devices using that technology are commercially available. Biometric technology provides a higher security than passwords thus; vendors use that aspect to claim more money. In addition, vendors of such devices are conducting their tests in perfect laboratories conditions and therefore, tend to over estimate the performance rates of their devices. In addition, a brief background will be provided for people unfamiliar with biometrics then experiments and results will be described in details. Hence, this study is going to investigate the reliability of biometric solutions under various test conditions. Moreover, the experiments are carried out on five devices tested in normal conditions and then in unexpected conditions in order to fool the system. Besides, in order to provide a short overview what can be already said is that the rank in term of reliability from the most secure one to the less is in the following order: iris, fingerprint, signature, face and voice regarding the study's results.

Keywords

Biometrics, reliability, fingerprint, face recognition, voice recognition, signature verification, iris recognition.

1. Introduction

Since IT birth, security is one of the most well known problems for companies, investing a huge amount of money because they obviously know that they can loose much more. Security begins with the unauthorized access to information clearly because most companies have sensitive, confidential or classified information. The most widely and traditional developed control access to an informatics' system is the classical alpha numeric password, technically poor compared with biometrics using human body characteristics. Nowadays, how many people have a password that you can easily guess or obtain by the owner himself or someone else? Even some users write down their passwords instead of learning them. Those following points represent the three categories of authentication that can be used separately or combine together (Despina, 1997):

- What you know (e.g. password or PIN)
- What you have (e.g. ATM card, Smart card, SecureID tag or PKI certificate)
- What you are (biometrics)

In the IT world, security is widely spread and improves fast. That is why for many years now, some engineers thought about using human characteristics body instead or in addition to

passwords. They have started to call that access control “biometric”, using two different methods to authorise the access of a user in a system: identification or authentication.

Identification is the method called “one to many”, matching one identity with many others. This technique is more successful than authentication because in that case users try to found their matching file in a database including several other users. Therefore, identification should be distinguished from authentication.

Authentication is the method called “one to one” search. Here, the method consists in knowing if the person exists and compared the current template only with a registered pattern of the person. Compared to the first technique of identification, it is less accurate and it is rarely a 100% matching. Moreover, an enrolment phase is needed before accessing the system, where people have to be registered for the first time usually in a database.

This project is dealing with the reliability and the acceptance of a secure system depending on how the system is protected against threats and its effectiveness to identify system’s abuses. Different biometric technologies are applied and increased their uses in the following sectors: Pubic Services, Law Enforcement, Banking, Physical Access Control and Computer & Networks.

The aim of this project is to test the reliability of commercial biometrics devices. A number of the most widespread commercial biometrics will be tested in order to have significant statistics and avoid looking at unused biometrics which is not relevant enough. The tests have been done collecting results in order to make calculations and be able to build conclusions. In order to make relevant statistics with a relatively large collection of information the need to have several users using the biometric devices in normal condition is necessary, for the project thirty six users were available.

2. Background

This project is looking at a number of key commercial biometrics already implemented and available on the market at that time. Biometrics was waited like the next biggest event in term of security in the IT world, is it justified? Since the growth of biometric techniques and the fast evolution of the high technology market prices of such devices tend to be a bit cheaper. Hence, security is brought to the single user even if it remains expensive and stays a high skilled field.

In term of history the word "Biometrics" is derived from the Greek words bio (life) and metric (to measure). Basically, in the field of technology, the definition given by Zimmerman (2001) is identification and authentication of individuals based on physical attributes.

Furthermore, the need to identify people is as old as humankind. For every people our own body uses different techniques to recognize someone else instinctively, even without noticing it. How does the biometrics work to differentiate people without mistakes? Human are recognizing himself naturally instead of biometric devices that need physical and scientific proofs of changes such as distance, temperature, speed etc. Biometrics are using two main different categories to analyse the human body (Maltoni et al, 2003):

- Physiological characteristics such as fingerprints, hand geometry, face recognition, vein checking, voice recognition, iris scanning or biological characteristics such as DNA, blood, saliva, odour, urine.
- Behavioural characteristics such as keystroke analysis, signature verification

In addition, the first obvious question coming into everybody's mind is: is it possible to differentiate twins? Fortunately, it has been recognized by the British government that every fingerprints of any people are unique even twins (Maltoni et al, 2003). In order to evaluate the performance of biometric techniques, a number of measures are used. Referred to as the Far, FRR, EER (Clusif, 2003):

- FAR: False Acceptance Rate: The errors where impostors are falsely believed to be the legitimate users.
- FRR: False Rejection Rate: The errors where the system falsely identifies the legitimate user as an impostor.
- EER: Equal Error Rate. It is the point where FAR and FRR meet. It is often used as a comparative measure of performance between biometrics. The lower this rate the better the overall accuracy is considered to be.

Figure 1 illustrates the relationship between both FAR and FRR, showing where the EER should be. On one hand, if the FAR is high, the security is low but the device is considered more users friendly. On the other hand, every intruder will gain access to the system because of the poor security. The other curve FRR is the complete opposite of FAR. Thus, the most important point is to be able to define the trade-off between security and convenience to the user.

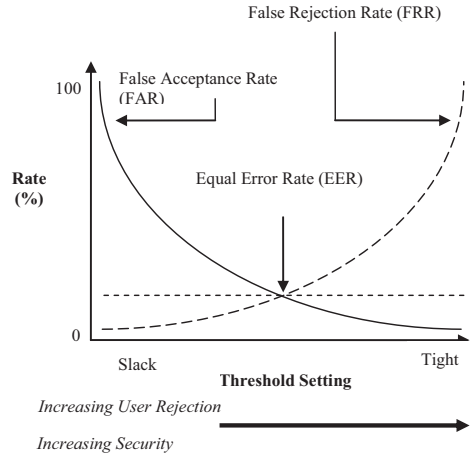


Figure 1 : FAR and FRR graph explaining the EER as a threshold

3. Iris recognition

The enrolment under normal conditions was done with both eyes of each user. The total number of users was fifteen with thirty attempts of each eye and one attempt per eye for the other users account. Moreover the not enrolled users trying to access by one time each eye. The device was allowing only five users hence; in order to have more users the need to install three OS was necessary. The problem was to make the enrolment and authentication as fast as possible that the users are not waiting therefore; the use of too many OS would have been a waste of time to reload them.

The results have shown a FRR = 203/900 = 22.55% which is not so good over 900 attempts. According to Mansfield et al. (2001) the FRR should be around 2%. Obviously the results obtained are much higher. This poor FRR can be explained by the very few users involved and adding to that some errors. Moreover, an error made over fifteen users increase drastically the statistics compare to an error made over thirty six like with the other devices. The problem was that during the enrolment phase some user could not open their eye so much and therefore, did not make a good template. Hence, the authentication phase was always rejecting them and the statistics drops down. Moreover, the FAR = 0/870 = 0% is significant because nobody over 870 attempts could access other users account which has not been the case for the following devices. According to Mansfield et al. (2001) the FAR should be 0.0001% and the results obtained during the test are similar. However, that statistic of 0.0001% will never be obtained because of the so small number of users.

4. Signature recognition

The enrolment phase is simple and has been done via a webpage. The link was on the CIC Company's webpage, with an enrolment of three signatures in order to build the template and which could be rejected by the application because the pattern was too. Thus, even the easiest signature has to go thought that security.

To conclude on the experiment and because some users were not present the second time for the authentication phase the FRR = 121/960 = 12.60% which is not so good. According to Qu et al. (2004) the FRR should be around 3.33% much better results than the experiment. The explanation can come from the use of the graphic tablet and also because most of the signatures are written fast therefore the pen could not build the template and thus the users were rejected during the authentication phase. The problem was for the people most of the time to get use to the device. Another point to bear in mind is after so many signatures the users get bored and the pattern is no so accurate anymore. Compared to the FRR the FAR = 12/1107 = 1.08% is high because some easy signatures were forged by several other users. Most of the time it happens to the signature made slowly and that the name can be clearly read. According to Qu et al. (2004) the FAR is 6.67% which is really high compare to the project's result. This difference can be explained easily because most of the users could not access to their account and therefore other users either make the small percentage instead of the expected one.

5. Fingerprint recognition

This device was easy of use and thirty six people were testing it with the same finger and thirty six attempts were made in order to compare it with the authentication phase of one user against thirty five other users

To conclude the thirty six users have made thirty six attempts in order to obtain that result for the FRR = 387/1152 = 33.59% which is really bad. According to Huseyin (2003) the FRR should be around 0.2% which gives an EER of one over five hundred. The main reasons was, even when the template was good the user has to get use to the device and takes almost ten attempts to remember the right position of the finger and then get access every time. Therefore, the results tend to rise rapidly adding to them all the users who could not access at all the system because of a poor enrolled template. Instead of, that FRR result the FAR = 1/1120 = 0.09% is due to one user who can get access to another user account in normal conditions. According to Huseyin (2003) the FAR should be around 0.2% thus in the experiment case the results are higher. This result can be explained because of the numerous users with a poor template compared to the total number of users therefore, any other users could not access the account when even the legitimate user was rejected.

6. Face recognition

The thirty six users were asked to make ten pictures of a normal front face such as an ID picture but without any smiles or grimaces. The software is sensitive to any aspect of the face hence; the more straight faces give better results. The attempts for the users were thirty times for his own account and because it was made via the software seventy times for the other users' account.

To conclude the face is an element really changeable with glasses, beard or even different hair style therefore, the results obtained for that biometric are bad. The FRR = 322/1080 = 29.81% is bad but can be explained by the system which sometimes did not capture the face at all and hence, the recognition with someone else could never be done and raised the statistics dramatically. According to Huseyin (2003) the FRR should be of 10% so the results are worst but not so far away regarding the number of total users in the study. In spite of, those results for the FRR the FAR = 136/88200 = 0.15% is good and could be explained by a high threshold of fifty percents. According to Huseyin (2003) the FAR is around 1.0% higher than the result obtained in the project. This percentage can also be explained by the fact that some users had bad captured templates and therefore could never match with anybody else.

7. Voice recognition

The enrolment was simple and fast with thirty six users having thirty attempts detailed in three times ten sentences. The first sentence, said ten times, is the one proposed by the software itself: "One swallow doesn't make a summer", the second sentence, is the one proposed by the Word application in order to get all the characters in a font: "The big brown fox jumps over the lazy dog" and finally the same sentence coming from Microsoft Word®

but in French: “Servez ce Whisky au petit juge blond qui fume”. The enrolment was made simply by recording the voice via a microphone.

To conclude the result is a FRR = 985/7420 = 13.27% which is good but the explanation can come from the quality of some recorded files. According to Huseyin (2003) the FRR must be between 10% and 20% thus the project result is closer to the 10%. These results can be explained by some voice templates rejecting or allowing every user because of the background noise or even the too low speech almost non audible. Furthermore, the result for the FAR = 159/88200 = 0.18% is good and can be explained by the difference in user voices recorded files, with some user having a better quality record than others. According to Huseyin (2003) the FAR should be between 2% and 5% and the result obtained in the test is lower. The explanation is coming from the little number of users in the first time then in the second time because several templates were inconsistent and reject every user even the legitimate one.

8. Discussion

Now to show all the results and be able to build relevant conclusion here is a summary presented in the Table 1 of all the biometrics tested in the project and the FRR and FAR associated in percentage:

Devices/Rates	FRR (%)	FAR (%)
IRIS	22.55	0
SIGNATURE	12.60	1.08
FINGERPRINT	33.59	0.09
FACE	29.81	0.15
VOICE	13.27	0.18

Table 1 : Results of FRR and FAR of the biometric tested

This table above is a summary of all the results obtained during the project. Compared to companies results, most of the time the FRR is much higher except for the voice recognition with better results due to the quality of the templates. Moreover, the FRR are poor because of the few users involved in the project compared to the number of users used in companies most of the time around two hundred instead of thirty six like here. Furthermore, the use of the device was not obvious and the time that user get use to it was increasing the errors. In addition, one error over thirty six users has more impact than one error over two hundreds. Unlike, the FAR are all the time better than the expected ones. This can be explained because there is a relation between FRR and FAR as previously mentioned to calculate the EER therefore if the FRR is poor that have an influence on the FAR making it better. The fact that a user can never access another account is mainly due in this experiment to the fact that even the legitimate user can not access it. Thus the FAR was good because of technical problems and not because of the device itself.

9. Conclusion

In order to conclude on the experiments, the results concerning the reliability of selected commercial biometrics are much higher than the company's results. First of all the reliability of the five devices as iris, fingerprint, signature, face and voice recognition system, can be ranked from the most secure to the less one as previously enumerated. Furthermore, several parameters are involved in the study leading to some different FRR and FAR. One of the first noticeable element is that behavioural biometrics are much more accurate than physiological. Moreover, much more efforts were needed to fool them because it uses the way the legitimate user behaves and it often changes from one individual to another. Secondly, the number of users influence the statistics. Thus, a rejection of a user because of a bad enrolment gives a zero percent of success and therefore; the statistics are increased by a higher percentage compared to a larger number of users. To sum up an error gives important changes in the results instead of companies using more users for their experiments and hence errors have less influence.

We also have not to forget that there are not any devices 100% accurate. The aim is to make the system as secure as possible by having so complex and expensive way to fool the system that anybody thought the reward do not justify this amount of effort. Security is all about providing adequate protection for the user but should remain commercially accessible. Finally, when you combine all the definitions to access a system what you know, what you have and what you are, you will achieved the highest level of security expected for applications or systems.

10. References

Clusif, (2003) "*Techniques de controle d'acces par biometrie*", Club de la securite des systemes d'information Francais, France
[Online] http://www.clusif.asso.fr/fr/infos/event/pdf/ControlesAccesBiometrie.pdf [8th September 2004]

Huseyin, A. (2003) "Digitized and digital signatures for identification", IEEE ICASSP2003 Tutorial –Hong Kong, [Online] http://akhisar.sdsu.edu/abut/DL/DLLecture-June_2003.pdf [8th September 2004].

Maltoni, D., Maio, D., Jain, A.K. and Prabhakar, S. (2003) "*Handbook of Fingerprint Recognition*", Springer, ISBN: 0387954317

Mansfield, T., Kelly, G., Chandler, D. and Kane, J. (2001) "Biometric Product Testing Final Report", Technical Report, Centre for Mathematics and Scientific Computing, National Physical Laboratory, [Online] http://www.securimetrics.com/articles/gfx/cesg-trials-report.pdf [8th September 2004]

Polemi, D. (1997) "*Biometric Techniques: Review And Evaluation Of Biometric Techniques For Identification And Authentication, Including An Appraisal Of The Areas Where They Are Most Applicable*", Institute Of Communication And Computer Systems National Technical University Of Athens, [Online] http://www.cordis.lu/infosec/src/stud5fr.htm [8th September 2004]

Qu, T., El Saddick, A. and Adler, A. (2004) "*A Stroke Based Algorithm for Dynamic Signature Verification*", Proceedings of the Canadian Conference on Electrical and Computer Engineering (CCECE), Niagara Falls, Canada, 2-5 May, pp461-464.

Zimmerman, M. (2001) "*Biometrics and User Authentication*", SANS Institute Reading Room, [Online] http://www.sans.org/rr/whitepapers/authentication/122.php [8th September 2004]

User Authentication by Service Utilisation Profiling

A.M. Aupy and N.L. Clarke

Network Research Group, University of Plymouth, Plymouth, United Kingdom
e-mail: info@network-research-group.org

Abstract

In the age of the information society, computer security is becoming a vital aspect of the information systems involved. Many organizations are enforcing new IT policies calling for better security in an attempt to safeguard one of their most crucial resources: data. New concepts, technologies and products, such as biometrics, will become commonplace in tomorrow's IT architectures. The work presented here aims to define a new behavioural biometric based on human-computer interaction and service utilisation. The gathered patterns (e.g. keystrokes and applications launched) are classified using a neural network approach in order to discriminate between legitimate users and impostors. This paper includes a description of the test procedure (pattern gathering, classification, and evaluation) and a discussion of the performance obtained, with an overall EER of 7%, proving that this tool shows the required quality as a potential new behavioural biometric.

Keywords

Behavioural Biometrics, Security, Neural Networks, Pattern Classification, Human-Computer Interaction

1. Introduction

In the last few years, biometrics systems such as fingerprint or iris recognition have evolved from science-fiction features to trustable security solutions. These solutions have been pushed forward by industries as traditional approaches (usually passwords or tokens) have shown their limits when trying to efficiently secure critical systems. However, biometrics solutions are sometimes criticized as their forgery can be easier than that of other approaches. When designing a biometric solution, extensive benchmarks must be achieved so that the end user understands the service level they might expect.

This paper describes the construction of a new kind of behavioural biometric which provides a non-intrusive way to authenticate the current user of a computer. A piece of software monitors human-computer interaction (HCI) and, using intelligent pattern classification techniques, attempts to recognize the user based on their activities. The efficiency and trustworthiness of this approach need to be discussed. The concept suggests that Service Utilisation Profiling, or SUP (i.e. identifying recurrent patterns in user activities), could achieve user authentication (i.e. being at least able to separate between legitimate and illegitimate users). The solution would ideally allow users to bypass the burdensome password-based OS login in a corporate network environment: knowing that each PC would be assigned to one legitimate user, a few minutes of illegitimate user activities would report a theft of identity (along with raising any chosen countermeasures such as station locking, intrusion detection system (IDS) actions etc.).

After a preliminary study, this research had been split up into six main tasks: define all HCI that would be relevant to SUP, define a database (DB) format to store the patterns, develop and deploy a program able to transparently capture those HCI and store them in the DB, develop software to extract data from the DB and inject it into the intelligent system, fine-tune and validate the system, and finally compute biometric statistics (BEM, 2002; Bolle, 1999) to confirm the level of authentication performance.

2. Background

User authentication has been using two simple techniques for decades: passwords and tokens (e.g. magnetic cards). With globalisation as it is witnessed today, the distribution of businesses worldwide creates a need to distribute the computing environments. Thus, passwords are widely used for remote logins and therefore circulate over weak networks such as the Internet. In essence, passwords could be secure if they were long enough (i.e. dozens of characters), as a good secure system would only be penetrated by brute force attacks. However, long random passwords comprising alphanumerical sequences in upper- and lower-case are unlikely to be remembered by users. For this rationale, and based on the motive to enforce enhanced security policies, new authentication techniques are actively being researched. Passwords relied upon what a user knows (that can be forgotten), and tokens on what a user has (that can be lost). Biometrics now relies upon what a user is, and that is fundamentally the real meaning of user authentication.

Biometrics splits up into physical biometrics (e.g. fingerprint recognition, retinal scan etc.) and behavioural biometrics (e.g. gesture recognition, signature recognition etc.) (Cobb, 1996). Behavioural biometrics is a recent technology which is characterized by (and may suffer from) fuzzy decision patterns compared to, for instance, straightforward fingerprints (AfB, 1999). Biometrics requires at some point a decision process between possibly several million patterns. This implicates pattern classification techniques such as neural networks (NN) which have been around since the 1960's (Picton, 1994; Duda, 2001). Since then, exhaustive research produced numerous neural network designs. Today's computer power allows the efficient implementation of complex classification techniques (Looney, 1997; Bishop, 1995), with applications in biometrics. These techniques can be neural as mentioned, but also analytical and statistical.

Within the framework of this research, the point of convergence of the previous concepts is HCI. Indeed, HCI serves as a new authentication method to improve security; it is described in terms of a behavioural biometric and therefore evaluated as a biometric; and finally it operates as inputs to a pattern classification technique. HCI has been studied since the early days of computing, allowing computers to evolve overtime from dumb console-based user interfaces to rich and powerful graphical user interfaces (GUI), which simplified the interactions (Myers, 1996). There are approaches to converge HCI with NN dating back to the 1990's (Beale and Finlay, 1992). However, the subject itself of using HCI as a behavioural biometric trait has not been explored yet therefore calling for the feasibility study presented here.

3. Methodology

3.1 Data collection

The needed user interactions capturing software and DB extracting software were chosen to be developed in Borland® Delphi™. 20 HCI events, called "actions", were chosen to be captured in the first piece of software named "Logger". When later performing data mining, only the three most relevant actions were kept, and they are presented in Table 1.

Code	This action is raised each time…
KEY	A full word has been typed in. The word is recorded, along with the title of the window where it has been entered.
OPN	A window is opened. The name and class of the window are recorded.
CLO	A window is closed. The name and class of the window are recorded.

Table 1 : The three main user actions captured by Logger

Logger's basic *modus operandi* is the following:

- It silently starts in the background when Windows® starts, only showing an icon in the taskbar,
- It checks for the DB integrity and other errors,
- It starts its capture systems: OPN, CLO (and 4 other actions) are captured by *polling* every second; KEY (and 13 other actions) is captured by *interruption*.

Low-level Windows API programming was used to perform the capture of the various system events. These actions are then saved thanks to the development of a proprietary simple database management system, which was required to allow little memory/CPU resources consumption, and on-the-fly compression. The DB is saved on the hard disk as two separate files: Logger.db and Logger.dbx. The compression technique employed is named string indexing technique. It associates a cardinal index with each string (and stores this association in the dictionary Logger.dbx), and thus saves only the 4-byte index into Logger.db instead of the full 255-character string.

Once fully debugged, Logger was deployed in the UK and France to 29 testers, ranging from mainstream to power users. A careful examination of the privacy issue had been undertaken as many users were initially reluctant to let this software captures all of their activities. Informing the users was the main element (stating that to reconstruct their full behaviour from the sole DB would be tremendously hard considering the quantity of gathered noisy information), but concrete developments in Logger also permitted to reassure the users (e.g. no alphanumerical password is recorded). The test period lasted 50 days. Due to OS reinstallation issues and faulty hardware, only 22 users out of 29 had (working) databases.

3.2 Neural network design

Before defining precisely the architecture and parameters of the intelligent pattern classification system, it was important to examine the data through a manual stage, enabling

to search for patterns in users input data. The question of how to inject wide range analogic data (such as timestamps and strings) into the NN was also raised. A second piece of software was therefore developed, entitled "ReadDB", which could perform data mining from the DB files retrieved and export the results directly into Matlab® .M format. Four fields are exported per action, and appropriate quantization levels were defined. For example, timestamps are divided into quarters of hour (96 quarters per day), and the strings' numerical indices from the dictionary are directly injected. This association string–index is made possible by the aforementioned string indexing compression technique. ReadDB writes all the required headers, and the main part of the .M files is made of a matrix containing the numerical version of the data, i.e. a succession of four-field actions. These actions are grouped by samples of user activity. The size of each sample had to be decided. Indeed, a single 4-field action is not relevant of a user's habits. The NN is trained with samples made of 300 actions, corresponding in average to 10 minutes of user activity. This means it is supposed that around 10 minutes of activity are enough to discriminate between two users. It must be clear however that 300 actions are not necessarily collected in 10 minutes. Variable levels of activity can produce them within a few minutes or many hours. As each sample is made of 300 consecutive 4-field actions, the NN's input layer faces 1,200 inputs per sample. It must be noted that even more data (such as the day of the week) were initially exported. However, profiling a user based on, say, his/her Monday activities would require months of data which were not available.

All NN architectures are managed in Matlab's NN toolbox using the net object. This complex structure contains all the NN parameters. For this research, an optimum 50x1 FANN architecture was found, i.e. 1,200 inputs, 50 hidden neurons and 1 output neuron. The hyperbolic tangent sigmoid, tansig, was the choice of activation function for all neurons. The training phase used traingdx as the learning method (gradient descent with momentum and adaptive learning rate back-propagation) and was set up for 1,000 epochs. The performance goal was 10^{-5}.

As early Matlab tests gave bad results, it was understood that injecting the direct strings' indices into the NN was not the right way: the same very index from two different DBs would not correspond to the same string. Therefore, a new piece of software, entitled "MergeDic" (standing for Merge Dictionaries), was developed to merge the 21 bases altogether, and to re-index accordingly the associated Logger.db files. Many attempts and code optimizations (notably using *hash tables*) were required to shrink down the computation time from several hours to a few minutes. Testing MergeDic showed excellent results in Matlab. Among a total of 320,928 strings, only 80,308 were common. This meant the 22 users did quite different things over the testing period. It is understandable: even if everybody uses Internet Explorer, the visited websites are different, and so are the titles of Internet Explorer's windows, and so are the strings found in each user's dictionary. Similarly, even if everybody uses Microsoft® Word, the opened documents are different, and so are the titles of Word's windows recorded by Logger.

The collection period was split into 4 fortnights. The Matlab programs oppose all impostors in turn to the legitimate user, one at a time. A first program loads the .M data, a second one scales it down to -1..+1 (i.e. normalizes it), a third one creates the NN and trains it with the first fortnight of users' activities, a fourth one runs the NN simulation for all the fortnights

and plots the results and a fifth one calculates the FAR, false rejection rate (FRR), equal error rate (EER) and optimum threshold for these simulations.

4. Biometrics results

The typical NN simulation results shown on Figure 1 are excellent for the first month. For the 2nd month, the impostor is still very easily identified, but the legitimate user output becomes noisy. With a very low threshold, the distinction is still achievable, which is enough to give good biometric results. The outputs become noisy overtime because the legitimate user must have changed its activities near the end of the test period. Typically, he/she became working on different documents (e.g. Word documents with different titles). The method logically calls for re-training the NN every month or so, as it is well understandable that user's activities evolve overtime.

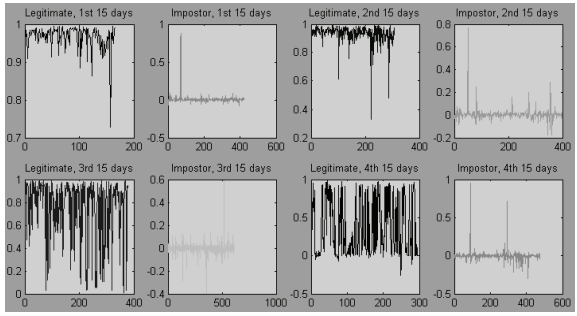


Figure 1 : Typical Y outputs plot

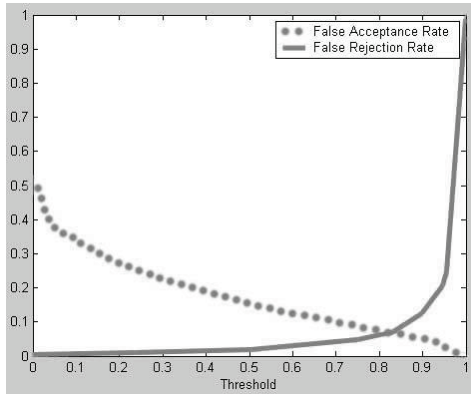
Eventually, twenty-one users (out of a theoretical 22) were tested as legitimate users against all 21 others, treated as impostors. The total number of tests carried-out was therefore 441 (21 legitimate users tested against 21 impostors). This represented 14 GB of generated data. The total computational time is evaluated to more than 300 hours. Most of the tests have been computed on one to six Pentium4 2.6 GHz Hyper-threading (emulating bi-processors systems), 512 MB RAM, Windows XP. Eventually, the test procedure was automated (i.e. MergeDic, ReadDB and Matlab were run automatically, loading the right files, self-clicking on buttons etc.) so that 357 tests could be carried out within one week only (instead of 84 when doing it manually).

The average means of Y outputs and their standard deviation over the 60 days for the legitimate user and the impostor are given in Table 2. The table also gives the optimum threshold and EER for each user. It can be concluded that if the overall biometric quality was to be evaluated based on averages only, it would be very impressive. Indeed, the average output for the legitimate user is always very close to 1, and those of the impostors close to 0. Similarly, the standard deviation is very low for legitimate users, meaning that the NN does not fail to recognize the legitimate user it has been trained with. This is very good from a security point of view. The standard deviation is near 0.3 in average for the impostors, which

is still quite good. However, these means are obtained over long periods of time, and the biometric quality must be also assessed on its responsiveness (i.e. the biometric's answer after several patterns only). The final curves for the 441 tests give a final threshold of 0.8208 and a final EER of 7.09%.

User	Average mean legitimate	Average standard deviation legitimate	Avg mean impostor	Average standard deviation impostor	Optimum threshold	EER (%)
1	0.96	0.04	0.02	0.28	0.8533	3.76
2	0.85	0.17	0.01	0.25	0.477	9.35
3	0.98	0.03	0.1	0.31	0.951	7.6
4	0.98	0.03	0.04	0.31	0.9218	3.83
5	0.87	0.03	0.01	0.28	0.5392	8.49
6	0.97	0.04	0.05	0.31	0.916	7.6
7	0.96	0.05	0.05	0.3	0.876	6
8	0.96	0.04	0.02	0.28	0.866	6.13
9	0.98	0.03	0.04	0.29	0.9155	5.35
10	0.86	0.09	0.02	0.28	0.6151	10.47
11	0.95	0.05	0.02	0.29	0.8201	6.72
12	0.96	0.05	0.03	0.31	0.85	6.9
13	0.97	0.04	0.01	0.3	0.8707	5
14	0.96	0.05	0.01	0.3	0.8594	5.09
15	0.87	0.12	0.01	0.28	0.6653	10.65
16	0.99	0.03	0.02	0.3	0.9492	3.21
17	0.94	0.07	0.01	0.27	0.8182	6.16
18	0.96	0.05	0.03	0.3	0.9016	6.04
19	0.93	0.06	0.01	0.29	0.7406	7.436
20	0.92	0.08	0.02	0.28	0.7715	6.84
21	0.97	0.04	0.03	0.31	0.878	6.18

Table 2 : Main statistics and final FAR/FRR curves



Optimum threshold: 0.8208
Associated EER: 7.09%

Figure 2 : Overall System Performance

5. Discussion

The high biometric quality achieved (EER 7.09% or 1:14) is akin to that of other behavioural biometrics such as keystroke analysis (Furnell, 2000; Clarke, 2003). However, it cannot compare to physical biometrics such as iris recognition or fingerprint recognition. This means that SUP could be a trustworthy biometric in terms of one-to-one identity verification, but not in terms of one-to-many identification/recognition. The user would still need to initially claim an identity via a traditional password-based login.

Despite the good EER figure, several issues, inherent to the concept of SUP, must be investigated. For example, at least 10 minutes of user activity after login would be required before trying a first "guess" about the user verification. If the optimum threshold of 0.82 is set up from the beginning, it can also be questioned if the EER would stay near 7%. From Table 2, it can be seen that most of the time the required threshold is indeed near 0.8; when it is not the case, it is justified by the unrealistic small sizes of the associated training sets.

Another point is regarding the learning phase. If not modified, it currently supposes to record 15 days of continuous legitimate user activity. It sounds quite long and no real impostor activity must occur during it. On the other hand, sample impostor activity must be injected into the NN for it to recognize illegitimate users, and not just produce white noise as by default. Hence, various impostor samples should be supplied with the prototype.

Retraining the system every month or so would turn off the system during two weeks, therefore counterbalancing the security benefits brought by this new biometric. It has been shown that the NN bases user recognition on their current trends, e.g. a document they may be working on daily, or websites they are visiting repeatedly during a period of time. Overtime, the NN does not really achieve user profiling that could stay truly authentic for months or

years. This is also why real identification cannot be accomplished. At the end of the day, rather than user profiling, the NN precisely does service utilisation profiling, in its true meaning which fundamentally implies that it can evolve rapidly overtime. Thus, service utilisation, as many behavioural biometric traits, suffers from its lack of *permanence*.

A final necessary remark when developing a biometric is regarding *circumvention* (Matyas, 2000), i.e. is it possible to fool the system. If a malicious user could watch a legitimate user's activities (e.g. by spyware means), he/she could reproduce them quite easily to stay identified on this user's station. It would be interesting, as future work, to investigate how malicious activity could be dissimulated among recreated legitimate activity. For example, one minute of malicious impostor activity would likely be ignored by the NN if it was found in the middle of 9 minutes of (simulated) legitimate activity.

To continue this research, the following could also be undertaken:

- Develop a prototype integrating the NN into Delphi (using Blum, 1992).

- Improve the NN performance and its ability to profile service utilisation. In fact, this profiling should be more guided than it is today, by defining, for example, special flags that describe very discriminating actions (e.g. DOS commands typed in a command prompt are likely malicious if the PC pertains to a mainstream user). A complete expert system approach could even be investigated to complement the NN: a list of pertinent questions would be compiled and an associated scoring system would help profiling the user.

- Enhance data mining. This means using the rest of the collected data (notably mouse actions), identifying recurrent patterns in it and filtering it efficiently.

- Try bigger and/or better and/or different NN architectures such as self-organizing NN. A good idea would be to use two or more simultaneous NN and merging their outputs using linear combination or even another NN (Chakraborty, 2003; Maren 1991). The principle of using multiple NN should allow each one of them to be forced to concentrate on a particular aspect of SUP: for example, morning, midday and afternoon activities profiling (3 NN).

- Further investigate the privacy issue. This means the acceptability factor of this biometric should be assessed; but also, solutions should be proposed such as cancellable biometrics, data encryption and so on.

6. Conclusion

As early as the 90s, it had been identified that everyone behaves differently when facing a computer so that useful information could be extracted from the user's habits. It could help enhance the GUI and HCI, but, in the present times, new demand calls for new supply: the computer now takes a crucial part in business assets whilst monolithic security techniques are still in use. Therefore, new security developments are actively being researched, and now well-developed biometrics can provide credible security alternative solutions.

Technology often tries to mimic nature. Therefore, the idea to recognize a user based on several of their personal traits, as it is done in real life, logically comes to mind. In this case, the motive in researching a new behavioural biometric has been the opportunity to gather several domains such as data mining, security, biometrics and pattern classification in order to develop a new solution from scratch, which could eventually be applied to concrete industrial issues. The good results of this feasibility study have demonstrated that user authentication, or rather, user verification could indeed be achieved via service utilisation profiling. Thus, this latter could act as a promising new non-intrusive authentication approach, taking a valuable part in information systems' security chain.

7. References

Association for Biometrics and International Computer Security Association. (1999) *Glossary of Biometric terms*. Self-publication, Northampton, UK

Beale, R. and Finlay, J. (1992) "*Neural Networks and Pattern Recognition in Human-Computer Interaction"*. Ellis Horwood Workshops, England

BEM Workgroup (2002) *Common Criteria for IT Security Evaluation*. CESG, UK. www.cesg.gov.uk

Bishop, C. M. (1995) *Neural Networks for Pattern Classification*. Oxford University Press, New York, USA

Blum, A. (1992) *Neural Networks in C++: an Object-Oriented Framework for Building Connectionist Systems*. Wiley, New York, USA

Bolle, R., Pankanti, S., and Ratha, N.K. (1999) "*Evaluating authentication systems using bootstrap confidence intervals"*. In Proceedings of IEEE AutoID '99, pages 9-13, New Jersey, USA

Chakraborty, D. and Pal, N.R. (2003) "*A novel training scheme for multilayered perceptrons to realize proper generalization and incremental learning"*. IEEE Transactions on Neural Networks, Jan 2003 Vol.14, pp. 1-14

Clarke, N., Furnell, S., Lines, B., and Reynolds, P. (2003) "*Using keystroke analysis as a mechanism for subscriber authentication on mobile handsets"*. Proceedings of the IFIP SEC 2003 Conference, Athens, Greece. pp. 97-108.

Cobb, S. (1996) International Computer Security Institute

Duda, R.O., Hart, P.E. and Stork, D.G. (2001) *Pattern Classification*. Wiley, New York, USA

Furnell, S. and Ord, T. (2000) "*User authentication for keypad-based devices using keystroke analysis"*. Network Research Group, University of Plymouth, UK

Looney, C. (1997) *Pattern Recognition using Neural Networks*. Oxford University Press, New York, USA

Maren, A. (1991) "*Neural Networks for enhanced Human-Computer Interaction"*. IEEE Control Systems, Aug 1991, Vol. 11, Number 5, pp. 34-35

Matyas, V. and Riha, Z. (2000) *Biometric Authentication Systems*. Universitas Masarykiana, Czech Republic

Myers, B. (1996) "*A brief history of Human Computer Interaction technology"*. Human Computer Interaction Institute, School of Computer Science, Carnegie Mellon University, USA

Picton, P. (1994) *Introduction to Neural Networks*. MacMillan, London, UK

IT Risk Analysis for Small and Medium Enterprises

I. Kritharas, V. Dimopoulos and S.M. Furnell

Network Research Group, School of Computing, Communications and Electronics,
University of Plymouth, UK
e-mail: info@network-research-group.org

Abstract

IT risk analysis is considered imperative during the planning of security strategies providing the basis for establishing a cost-effective security program against the potential threats. However, currently the acceptance of risk analysis within SMEs is very limited. The reasons stem from the distinctiveness of these business environments. Characteristics such as lack of expertise, budget constraints, lack of awareness and lack of time prevent SMEs from performing IT risk analysis. This project aims to closely approach the issue by investigating three commercial risk analysis software tools, namely CRAMM, BUDDY SYSTEM and COBRA. It is an attempt to identify the main weaknesses of these tools which make them unsuitable for SMEs. The evaluation revealed that CRAMM and BUDDY SYSTEM although they are effective they require significant experience and expertise. CRAMM process needs also considerable time to be accomplished whereas both CRAMM and BUDDY SYSTEM are very expensive. COBRA even though being less expensive and simpler to use does not produce meaningful results. Once the reasons behind the problem has being realized, the project proceeds to the identification of the main security requirements for an effective risk analysis approach and it proposes a potential simplified model.

Keywords

Risk analysis, SMEs, tool evaluation

1. Introduction

A considerable percentage of small and medium sized enterprises (SMEs) have already realized the new business opportunities that Information Technology (IT) can provide and have started deploying it in many functions and areas of their business activities. The essential computing and networking infrastructure, as well as the emergence of new Internet-related applications such as e-commerce and e-business, are strategic enablers of their activities. However, these new business opportunities expose these organizations to a considerable number of important threats which can affect seriously their financial stability. In case of a disruption for even a few days could cause severe financial loss. According to the Information Security Magazine (ISM) Survey 2002 (Briney and Prince 2002), 49 percent of the small and 70 percent of medium sized companies reported loss resulting from security incidents. Additionally, the Department of Trade and Industry (DTI) 2002 survey (DTI 2002) shows that although smaller organisations encounter fewer threats, they are more vulnerable than larger companies.

A basic requirement in order for a security program to protect efficiently a company is a comprehensive identification and assessment of risks, and the subsequent implementation of mitigation techniques which result in the selection of the most relevant and cost-effective security safeguards. This procedure which is known as IT risk analysis (Pfleeger and Pfleeger

2003) is becoming an increasingly important business issue for SMEs. Nevertheless, this solution is not so easy to implement, as distinctive obstacles prevent SMEs from performing IT risk analysis.

2. Current attitude of SMEs towards IT Risk Analysis

To date, the security practices that SMEs deployed are quite inadequate compared to large organisations. DTI 2002 survey (DTI 2002) shows that very few businesses are concerned about the possible security threats to their organization over the next year and those that are concerned tend to be large businesses. As a consequence large organisations appeal to implement more security controls than smaller enterprises. For instance, large businesses are twice as likely as small ones to have a security policy. This lack of security practices within SME environments includes also the adoption of IT risk analysis which is very limited due to numerous obstacles. The magnitude of this problem is highlighted in the Business Information Security Survey-BISS 2000 (NCC 2000) conducted by the National Computing Center (NCC). According to this survey 62% of businesses with 0-9 employees and 31% of companies with 10-99 employees questioned have never performed an information risk assessment. These findings are also confirmed by the responses of a survey conducted by the authors. Specifically, only 37% of the overall SMEs (1-250 employees) participated in that survey carry out IT risk analysis and most of them are medium sized businesses (21-250 employees).

The above survey findings reveal considerable variations regarding the percentage of SMEs that perform IT risk analysis compared to large organisations. The reasons that justify this attitude of SMEs can be derived from a number of characteristics which are distinctive in these business environments. These attributes are identified below:

- **Lack of a standardised risk analysis methodology -** A major obstacle that prevents SMEs from performing risk analysis is the fact that there is not currently a standardized, integrated and widely adoptable methodology which can be replicated or modeled in a repeatable manner (Jones and Sutherland 2003). The solutions in this field that are currently available are based on the subjective nature of risk analysis theory and they do not assure accurate and detailed results. The large variety of proposed risk analysis methodologies depends heavily on the critical ability of the evaluator who should be able to modify the specific methodology according to the unique needs of the company. As such it is possible for even an experienced risk analyst to reach misleading conclusions.

- **Lack of security expertise** – The process of risk analysis is very demanding as the current tools and methodologies require high level of expertise and experience. However, SMEs do not tend to employ specialized technical staff in information security. This lack of security expertise can be confirmed by the findings of the ISM 2002 survey Briney and Prince 2002) that show 49 percent of small and 51 percent of medium organizations do not employ any staff with a specific IT security training. These enterprises tend to use a general IT administrator who is responsible for every task that should be implemented in the network. Apparently, the security administration is inadequate and most of the times it focuses only on the protection by common threats such as viruses and hackers (Briney and Prince 2002).

- **Lack of financial resources -** In addition to the above obstacles, the lack of budget is another significant factor which affects the SMEs from not performing risk assessment. This is confirmed by the findings of 2004 DTI survey (DTI 2004) which shows that only 27% of small businesses spend more than 1% of their IT budget on information security, compared to 39% of large businesses that do so. Usually, SMEs do not have a dedicated proportion of their available funds for their information security. They tend to include the required cost for their security within their overall IT budgets. However, this is not true for large enterprises which have a more organized and strategic plan regarding their expenditures in information security. Additionally, the high costs of risk analysis software tools prohibit SMEs from deploy them.

- **Lack of awareness -** Many SMEs believe that because they do not have a well-known brand name, they are not exposed to serious threats (Heikkila 2000), and therefore they consider that risk analysis is not necessary or critical for them. However, as it is has already mentioned previously, SMEs are more exposed to threats than large businesses due to their light security practices. These businesses are not aware of the value of IT risk analysis and the benefits which may be derived from it. They focus directly on the technical aspects of security without take into account the required level of information security and the related funds that should be invested. Most of the times companies deploy only the basic and well-known security controls such as antivirus software. According to DTI 2004 survey (DTI 2004) the major concern of administrators in SMEs are the virus attacks and hackers, while they pay less attention or even disregard other significant threats. As such, by enabling antivirus software and a firewall a company expects that its sensitive assets are secured. This case can be characterized as a false sense of security because the organisation can not realize its real security needs (Dimopoulos *et al.* 2004)

3. Risk analysis tools evaluation

Based on the findings so far, it is concluded that the problem of the adoption of IT risk analysis is focused especially on the SME environments. A main part of this research is concerned with the evaluation of three risk analysis software tools which are currently available in the market. It is an attempt to identify practically the strengths and weaknesses of these tools and understand the reasons that made them unsuitable for SME environments. The tools that are investigated are CRAMM, Buddy System and COBRA. The criteria for selecting these software tools were that all of them are designed and specified for the IT field. Thus, they provide more features and specialized functions particularly related to information security. Also, COBRA was selected for another reason which is the fact that its vendor asserts that this tool is designed primarily for SMEs due to its simplicity and relative low cost. As far as CRAMM is concerned, it is a well-known and respective tool designed for large organisations. However, the latest update (CRAMM Expert version 5) of this tool released in 2003 includes a new module, namely CRAMM Express, which is also available as a separate product to the full version of CRAMM. Express retains many features of Expert but has been designed to conduct a high level risk assessment in less than two hours without the need of security expertise. According its vendor, CRAMM Express is ideally suitable for SMEs due to its low cost and simple approach. A summary of the evaluation of the above tools follows.

CRAMM

CRAMM (CCTA Risk Analysis and Management Method) is a risk analysis and management tool developed by UK government's Central Computer and Telecommunications Agency in 1985 to provide government departments with a method for information systems security reviews. Since then, CRAMM has undergone major revisions and currently it is distributed commercially by Insight Consulting. Except from the IT sector, CRAMM is also deployed in the fields of finance, insurance and government (CRAMM 2004). Generally, although CRAMM Expert is quite comprehensive and effective risk analysis tool, it is not designed for SMEs due to its complexity and high cost.

As far as CRAMM Express is concerned, although it is a light version of CRAMM Expert and it provides a simpler, quicker and less costly assessment, it cannot be considered suitable for SME environments. Even though it is less complex than CRAMM Expert, it still requires significant experience and expertise as the valuation of assets is made by assigning a value from 1 to 10 to the potential impact of unavailability, destruction, disclosure and modification. Also, the user should possess technical knowledge in order to select the relevant threats. Furthermore, CRAMM Express does not address threats and countermeasures to the same level of detail as CRAMM Expert and it does not also support compliance for BS 7799 certification. In order for the risk assessment to be more detailed, the user should upgrade to CRAMM Expert, but with additional cost.

BUDDY SYSTEM

Buddy System is an alternative, less complex than CRAMM automated risk analysis tool. It was first introduced in 1987 by *Countermeasures, Inc*. and since then the tool has undergone various changes and updates. It consists of three integrated software modules: 1) *Survey Module*, 2) *Analysis Module* and 3) *Maintenance Module*.

The tool assesses the level of vulnerability based on safeguards already in place. The information about these safeguards is collected using the survey module. This module generates appropriate surveys which one or more individuals have to answer. These surveys are automated and they can reconfigure themselves to fit the respondent type or environment identified by the respondent. The completed surveys are imported into the analysis module for vulnerability and risk assessment by an evaluator. The latter can then interactively bring the vulnerability levels down to the acceptable level performing "what-if" modeling to determine the best course of action. Furthermore, appropriate cost benefit and Return on Investment (ROI) analysis is accomplished and the results can be displayed in comprehensive reports. Finally, the maintenance module allows the user to add or change countermeasures, threats, vulnerabilities, assets and other database elements without limitation. This feature permits users to update database contents as frequently as needed. Therefore, it is possible for the evaluator to customize the tool based on the special and unique needs of his/her company.

On the other side, Buddy System presents a number of weaknesses which make it unsuitable for SMEs. Similar to CRAMM, this tool requires the respondent to identify all the threats that apply to the system being surveyed as well as the likelihood of them occurring. However, these assumptions may lead to inaccurate and misleading results. Furthermore, the capability of the Maintenance module for customization of the tool according to the unique needs of

each organisation requires expertise due to the complexity of the procedure. However, even if the tool is tailored to fit the business environment, some of the proposed countermeasures are irrelevant or vague and may confuse the user. Finally, the cost of the Buddy System software package is prohibited for SMEs.

COBRA

COBRA (Consultative Objective & Bi-functional Risk Analysis) is an even more simple questionnaire-based risk analysis tool than CRAMM and Buddy System, and it is used exclusively in IT systems. It consists of two software products: 1) *The Risk Consultant* which is the main part of COBRA and is used for security risk analysis and 2) *ISO 17799 Consultant* which measures compliance with ISO 17799.

The conclusions that were derived from this evaluation were not expected at some extent. According to its vendor statements, it was expected for this tool to present features tailored especially for SMEs. Instead, COBRA seems a quite inadequate tool for an important procedure like risk analysis. It is based on very lengthy questionnaires, which the user should answer in order to generate a final report which is very superficial, without the recommended countermeasures offering substantial solutions including cost-benefit analysis. Also, even though its price is lower in relation to the other tools, it cannot justify these limited capabilities. Table 1 summarises the strengths and weaknesses of the evaluated tools in relation to SME environments.

	Advantages	**Disadvantages**
CRAMM Express	♦ Quick process ♦ Low cost	♦ High risk assessment. ♦ The asset valuation and threat identification require expertise and experience. ♦ It does not support compliance for BS 7799 certification
BUDDY SYSTEM	♦ Easy-to-use GUI. ♦ "What-if" modeling lowers the analysis effort. ♦ Cost-benefit analysis. ♦ Customization according to the unique needs of the organisation. ♦ Multi-language capability.	♦ The customization and configuration of the tool according to the unique needs of a company is complex and requires expertise. ♦ Even if the tool is tailored to fit the business environment, some of the proposed countermeasures are irrelevant or vague and may confuse the user. ♦ Some stages of the Buddy System methodology are based heavily on assumptions and the subjectivity of the user. ♦ Expensive.
COBRA	♦ Low cost. ♦ It measures compliance with ISO 17799.	♦ Very lengthy questionnaires. ♦ The recommended countermeasures are at high level. ♦ It does not include cost-benefit analysis.

Table 1 : Characteristics of risk analysis tools

4. Requirements for a simplified risk analysis methodology

The previous research reveals that the problem of lack of IT risk analysis focuses especially in SME environments. The evaluation of the three risk analysis software tools revealed the magnitude of this gap by identifying the limitations and weaknesses of these tools. IT risk analysis is considered imperative during the planning of security strategies providing the basis for establishing a cost-effective security program against the potential threats. Therefore, a new approach in this field should be made based on the specific elements that characterize the SME environments. In order for the new approach to eliminate the aforementioned constraints, the following requirements should be fulfilled:

- **Applicable to different types of SMEs.** The approach should be applied to a wide range of different business fields and have the capability to be customized to meet site-specific requirements.

- **Ease-of-use.** It is necessary for the method to be simplified so no special training required. The approach should not rely on the subjective expertise of the evaluator. It should be as much automated as possible providing the user adequate guidance throughout the risk analysis process and expecting minimal effort from the user perspective.

- **Quick.** The procedure should be implemented in a timely manner avoiding for instance the conduction of time-consuming and lengthy questionnaires.

- **Cost-benefit analysis.** The practitioner should know whether his/her investment in security controls will be economical viable. This feature can be realized through a cost-benefit analysis which will result in useful conclusions related to recommendation of the best cost-saver and effective solutions.

- **Provide substantial countermeasures.** The results of the assessment should include meaningful countermeasures with appropriate guidance for implementing them.

- **Comprehensive reporting.** The final results should be included in a variety of graphically supported reports providing a manager who lacks technical knowledge adequate guidance for the deployment of the selected countermeasures.

5. Methodology approach

According to the ISO/IEC 17799:2000 (BSI 2001), in order to select the most effective set of security controls, it is essential the information security requirements of an organization to be first determined. These security requirements dictate the level of protection required. As such, the main initial objective of a new risk analysis approach is the identification of these requirements. Such an approach would be based upon the assumption made by Haar and Solms (2003) that similar organisations may be exposed to similar risks and therefore have similar information security needs. The basic concept of this model is to use the character and properties of the organisation as a basis in the identification of the most common security requirements which can be realized through the implementation of common protection controls as well. The character of a company can be defined according to the type of the

industry which belongs to, its size in terms of the number of staff it employs, its geographical location as well as the type of building in which the assets of the company are located (Furnell *et al.* 1997). Although there is a large diversity of organisations, these properties allow the division of companies into distinct categories and each of them is characterised by specific security needs. As such, this approach can result in the deployment of the method by a wide range of business types

Once these organisational properties have been determined, the identification of the main critical business functions can be identified. Each company which belongs to a specific category defined by its industry sector and size can be assumed that it performs a number of common business activities which are supported by the IT infrastructure and therefore need protection. The next stage determines initially the most common assets that compose each of the identified business processes. For example, e-commerce is a critical business function that is composed by a number of assets which support this process. One of these assets can be customers' data such as credit card details which are considered confidential information and therefore need tight security countermeasures. Continuing the process, each asset can be associated with one or more threats that are automatically available to the practitioner by the tool. Consequently, the user does not have to guess the potential threats to each asset as it was seen in the evaluation of the tools.

A new simplified approach should also include a cost-benefit analysis in order to enable SMEs to focus their restricted financial resources on the critical assets of the company and establish a priority and required level of protection. The selection of cost-effective protective measures is based on the assumption that the cost of controlling any risk should not exceed the maximum loss associated with the risk (Jenkins 1998).

After the recommendation of the most common security controls based on generic organisational characteristics, a second phase should follow which includes a more detailed risk analysis. This procedure takes into consideration the unique business needs of each organisation which require specific countermeasures. However, this stage may need the involvement of the user in order to identify and evaluate the particular assets. This stage requires further and attentive investigation in order for the method to not rely on the evaluator's subjectivity and experience.

6. Conclusions

This project revealed the multidimensional problems of SMEs in adopting IT risk analysis. Inherent characteristics such as budget constraints, lack of technical expertise, lack of awareness and lack of time remain major obstacles for SME business environments. The evaluation of three risk analysis tools showed that are unsuitable for this type of businesses due to the above distinctive characteristics. The approach which was proposed in order to realize the aforementioned requirements indicated the fundamental and necessary functions that should be implemented for a more simplified and quicker risk analysis process. However, the research reveals several constraints in developing such a method due to the subjective nature of risk analysis. It seems impossible to develop a simplified risk analysis methodology which does not rely on the subjectivity, judgment and experience of the evaluator. The subjective nature of risk analysis theory does not guarantee accurate results. Valuation of

assets, likelihood of an event occurring, potential impacts are some fundamental elements of risk analysis that need to be estimated with some way in order to reach a result. It is very risky to rely on doubtful statistical and historical data in order to predict the future.

7. References

Briney A. and Prince F. (2002) *Information Security Magazine Survey, does size matter?* Information Security Magazine [Online].Available:http://infosecuritymag.techtarget.com/2002/sep/2002survey.pdf (10 March, 2004)

CRAMM. (2004) CRAMM risk analysis tool, website: http://www.cramm.com

Dimopoulos V., Furnell S., Barlow I. and Lines B. (2004) *Factors affecting the adoption of IT risk analysis*, 3rd European Conference on Information Warfare and Security Royal Holloway, University of London, UK, 28-29 June. [Online].Available:http://ted.see.plymouth.ac.uk/nrg/papers/ NRG%20Paper%20150.pdf (10 July, 2004)

DTI. (2002) *Information Security Breaches Survey 2002,* Department of Trade and Industry [Online].Available:http://www.btglobalservices.com/en/products/trustservices/docs/security_breaches_2002.pdf (22 April, 2004)

DTI. (2004) *Information Security Breaches Survey 2004,* Department of Trade and Industry [Online].Available:http://www.pwc.com/images/gx/eng/about/svcs/grms/2004Technical_Report.pdf (25 May, 2004)

Furnell S., Warren M. and Sanders P. (1997) *ODESSA: A new approach to healthcare risk analysis,* University of Plymouth [Online]. Available:http://ted.see.plymouth.ac.uk/nrg/ papers/ NRG%20Paper%2033.pdf (5 June, 2004)

Haar H. and Solms R. (2003) *A model for deriving information security control attribute profiles,* Computers & Security, Vol.22, No. 3, pp 233-244.

Heikkila P. (2000) *SMEs are the weak link in supply chain security,* Silicom.com [Online]. Available: http://software.silicon.com/security/0,39024655,11019992,00.htm (10 June, 2004)

Jenkins, D. B. (1998) *Security Risk Analysis and Management*, White Paper, Countermeasures, Inc. [Online].Available: http://www.cs.kau.se/~albin/Documents/RA_by%20Jenkins.pdf (30 January, 2004)

Jones A. and Sutherland I. (2003) *Information Security Bulletin*, Vol.8, issue 4 [Online]. Available: www.chipubishing.com/portal/backissues/pdfs/ISB_2003/ISB0804/ISB0804.pdf (26 June, 2004)

NCC. (2000) *Business Information Security Survey 2000,* National Computing Centre (NCC) [Online].Available: http://www.ncc.co.uk/ncc/biss2000.pdf (21 March, 2004).

Pfleeger, P.C. and Pfleeger, L.S. (2003), *Security in Computing* , 3rd Edition.

IT Security: A Human Computer Interaction Perspective

D.Katsabas, S.M.Furnell and A.D.Phippen

Network Research Group, University of Plymouth, Plymouth, United Kingdom
e-mail: info@network-research-group.org

Abstract

Faced with an increasing range of attacks, the appropriate use of available security features in computer systems and applications is becoming ever more necessary. However, although many applications provide ways in which users can protect themselves against threats, the design and implementation of these features can often be criticized from a Human Computer Interaction (HCI) perspective. This results in usability problems for novices and other non-technical users, which may compromise the level of protection that they can achieve. In this research, some standard principles of HCI have been used to devise guidelines to support the inclusion of security features within applications. Ten guidelines were created in total, and a number of existing applications have been assessed to determine their compliance. The results showed varying levels of adherence to the recommended practice, suggesting that current applications have some significant scope for improvement in their presentation of security functionality. To support this view, revised versions of user interfaces were designed for applications that achieved low scores, and the paper presents an example of the outcome to illustrate the approach.

Keywords

Human Computer Interaction, HCI Guidelines, Security, HCI-S

1. Introduction

There are many computer applications that provide some security functionality. This is particularly common in applications that require a connection to the Internet, where a great number of security threats emerge (Paller, 2002). An application may be able to provide significant protection from Internet threats. However if users do not know how to use it, their systems will still be vulnerable (Whitten and Tygar, 1999). In order to improve the usability of an application, Human Computer Interaction (HCI) principles should be carefully considered. There are many aspects in HCI that need attention, including the design of the user interface, the level of online help that can be provided, and the ease of use. If these aspects are not afforded sufficient attention, people may find it hard to understand a program or they may be put off by its complexity (Furnell, 2004). For example applications should not make it difficult to perform a specific task in an application, require too much time for it, or some level of technical experience. Even if some users overcome the difficulties and learn how to use a complex application, they will be likely to forget how to use it afterwards, as it will be infrequently used. Unfortunately, some applications have not applied HCI principles to the user interface and as a result, the security features are often overlooked.

Another considerable matter is that many users, and especially those that are not experienced enough with computers, are not able to customize the applications they use and simply use the default settings. They may not know that security options for the application exist, or how to modify them according to their needs. The reason for this is that the software applications that provide security options have been designed by technical people having a technical audience in mind. As a result, their complexity is high.

In this paper an attempt has been made to make the use of security features in applications easier. A number of guidelines have been used and applications were evaluated according to the level of attention that they afforded to the key issues. Moreover, new interfaces were created for the applications that were perceived to have bad HCI aspects, and a survey has been made in order to test the effectiveness of the new interfaces in making the use of security easier.

2. Background

In general, Human-computer interaction is "a discipline concerned with the design, evaluation and implementation of interactive computing systems for human use and with the study of major phenomena surrounding them" (Hewett *et al.* 1996). In order to examine and analyse the principles of HCI someone with knowledge only in computers is not enough. Skills from several different sciences are needed in order to study this subject. For that reason assistance has to be provided by people from computer science, psychology, sociology, anthropology, industrial design and other fields (Preece *et al.* 1994).

In the computer science world, Human-Computer Interaction is not explored enough in order to eliminate problems. HCI aspects help to make computer systems friendlier and easier to use by finding methods and processes for designing interfaces (Carroll, 2003). A suitable user interface that will be easy to learn and efficient to use is desirable in all computer applications. Moreover ways to implement an interface are found like algorithms that work efficiently, software toolkits and libraries. HCI is also concerned with the development of interaction techniques, new interfaces and methods for evaluating and comparing them (Mandel, 1997).

The greatest objective of HCI is to increase human creativity and improve the communication and cooperation between humans and computers. This can be achieved by designing computers and computer applications in such a way that people can fully utilize all the advanced features offered (Baecker, 2004). However, the developers may not consider the use of the functionality from the perspective of their end users and this causes difficulties in the way programs are utilized.

The extent to which a system is friendly may be minimized when security measures have to be taken (Swartz, 2004). For example, suppose that passwords have to be used in order for users to gain authenticated access. The more complex and longer the passwords, the more secure the system will be. Furthermore, the security will be increased if passwords are not the same on multiple systems, and are changed on a regular basis. On the other hand, human memory is limited, and cannot remember complex and long passwords (Krause, 2004). For the same reason, it will be hard for the users to memorize new passwords every time they

have to change them. This example clearly shows that usability and security can sometimes be contrasting objectives.

As mentioned by Johnston *et al.* (2003), HCI-S is defined as: "the part of a user interface which is responsible for establishing the common ground between a user and the security features of a system. HCI-S is human computer interaction applied in the area of security". Establishing such a common ground is vital in the sense that, without it, users will fail to relate to the options available to them. For example, they often do not use features that they perceive to be advanced or hard to use, and indeed from the presentation of security options in many applications, they may be perceived to be the preserve of experienced or technical users. Therefore, if guidelines can be created that improve the HCI-S aspects of an application, and if those guidelines can be applied correctly, the use of security options may be easier to apply. The purpose of HCI-S is to make a computer system more robust, reliable and secure by enhancing the application's interface.

3. HCI-S Guidelines

There are several HCI guidelines that an application should follow in order to have correct HCI aspects. Most of the guidelines used were drawn from those proposed by Johnston *et al.* (2003). Further guidelines were created by modifying the 10 usability heuristics proposed by Nielsen (1994). To further refine the guidelines the first principles of interaction design (Norman, 2003) were studied and a number of them were used to improve HCI-S. Ten equally significant guidelines were created and the applications were evaluated against each one of them.

1. **Visible system state and security functions:** Applications should not expect that users will search in order to find the security tools or have hidden features inside the application. Furthermore the use of status mechanisms can keep users aware and informed about the state of the system. Status information should be periodically updated automatically and should be easily accessible.
2. **Security should be easily used:** The interface should be carefully designed and require minimal effort in order to make use of security features. Additionally the security settings should not be placed in several different locations inside the application, because it will be hard for the user to locate each one of them. (Johnston *et al.* 2003)
3. **Suitable for advanced as well as first time users.** Show enough information for an experienced user while not too much information for a first time user. Provide shortcuts or other ways to enable advanced users to control the software more easily and quickly.
4. **Avoid heavy use of technical vocabulary or advanced terms:** Beginners will find it hard to use the security features in their application if technical vocabulary and advanced terms are used.
5. **Handle errors appropriately:** Plan the application carefully so that errors caused by the use of security features could be prevented and minimized as much as possible. However when errors occur, the messages have to be meaningful and responsive to the problem.
6. **Allow customization without risk to be trapped:** Exit paths should be provided in case some functions are chosen by mistake and the default values should be easily restored.
7. **Easy to setup security settings:** This way the user will feel more confident with changing and configuring the application according to their needs

8. **Suitable Help and documentation for the available security:** Suitable help and documentation should be provided that would assist the users in the difficulties they may face.
9. **Make the user feel protected:** Assure the user's work is protected by the application. Recovery from unexpected errors must be taken into account and the application should ensure that users will not lose their data. Applications should provide the user with the latest security features in order to feel protected. Furthermore some form of notification would be useful in case a security update is available.
10. **Security should not reduce performance:** By designing the application carefully and using efficient algorithms it should be possible to use the security features with minimum impact on the efficiency of the application.

4. Assessment of existing applications

Ten applications were used and assessed against the HCI-S guidelines designed in the previous section. In order to make comparisons on a like-by-like bases only well established software products were evaluated. Three antivirus applications were used (Norton Antivirus, Panda Antivirus and McAfee VirusScan). There were also two firewall applications, namely Agnitum's Outpost Firewall and Zone Alarm Firewall, as well as Opera and Mozilla Firefox web-browsers, Qualcomm's Eudora and Incredimail email client software, and finally Microsoft Word. This gave an overall mix of both security-specific tools, and more general applications that nonetheless included security functionality. Each application was tested according to the level of compliance with each of the 10 guidelines. A maximum mark of 5 could be achieved for each guideline so that the total mark obtained will be out of 50 (10 * 5 = 50). The same grading method was used for all the applications and the grades were from 0 to 5, as listed in Table 1.

Grade	Reason
0	Application diverges completely from the guideline
1	Application significantly diverges from the guideline.
2	Application has paid some attention to the guideline but still has major problems
3	Application has paid some attention to the guideline but still has minor problems
4	Application follows the guideline in some sections
5	Application completely follows the guideline in all possible sections

Table 1 : Grading method for HCI-S guideline compliance

The evaluation version of each application was installed and a number of tests were performed in order to assess the performance of the application for each guideline. For

example, the settings of the application were examined to check if they were easy to setup, if the security options could be modified easily, if the default settings were provided etc. Table 2 shows a summary of the score that each application achieved for each of the 10 guidelines. It can be noted that there are no guidelines that seem to score uniformly well or uniformly badly across all applications. As such, no consistent pattern can be observed in terms of where applications are failing to present security appropriately.

	Firefox	Outpost	Mc Afee	Eudora	Zone	Norton	Ms Word	**Incredimail**	Panda	Opera
Visible system state and security functions	2	3	4	3	5	2	3	4	3	3
Security should be easily used	4	3	3	3	4	4	3	5	3	3
Suitable for advanced as well as first time users	5	2	2	5	3	4	4	4	3	2
Avoid technical vocabulary or advanced terms.	2	0	4	0	2	2	1	2	4	3
Handle errors appropriately	3	2	3	2	4	2	4	3	2	4
Allow customization without risk to be trapped	2	2	0	2	1	2	1	2	1	2
Easy to setup security settings	2	5	5	2	2	2	3	5	5	5
Suitable security help and documentation	0	1	1	5	2	5	4	2	5	5
Make the user feel protected	3	4	4	5	3	3	4	2	4	3
Security should not reduce performance	3	4	1	0	1	3	4	4	4	4
TOTAL (/50)	**26**	**26**	**27**	**27**	**27**	**29**	**31**	**33**	**34**	**34**

Table 2 : Score summary for assessed applications

5. Applying the guidelines

In order to demonstrate the relative ease with which HCI-S can be improved, the user interface of a subset of the applications tested in Table 2 were modified in order to follow the HCI-S guidelines (Katsabas, 2004). Presentation of the full set of the modifications made is beyond the scope of this paper, and so a specific example is presented. The software tool Mozilla Firefox obtained a relatively low score because it did not conform to most of the HCI-S criteria (Table 2). Even though the privacy options had a separate tab, the security options were among options presented in an 'advanced' tab. Studies in HCI have shown that options classified as 'advanced' scare many users, especially beginners. Therefore, grouping the security settings in an advanced tab may result in a number of users never accessing them. In order to improve the usability of the security settings a separate tab was added named "security" that contained all the options relating to security (Figure 1).

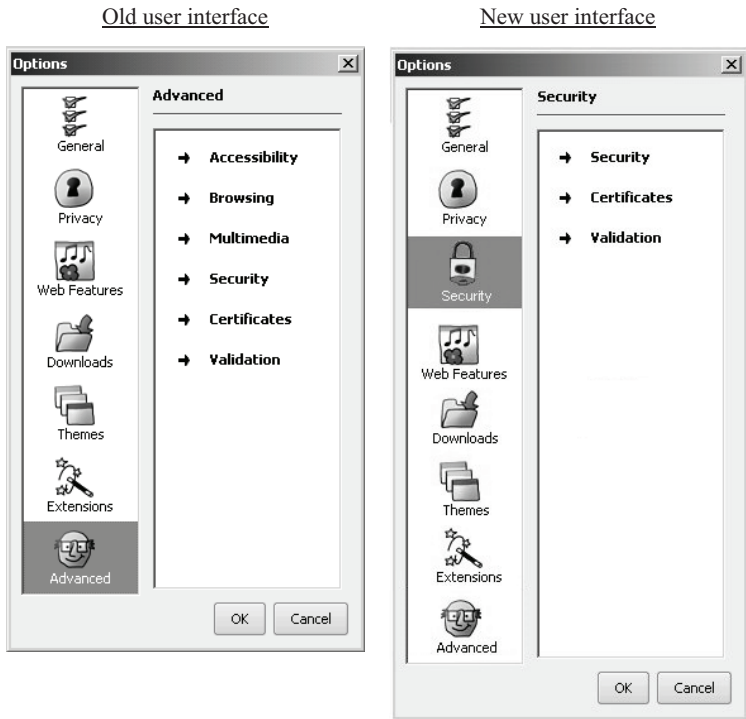


Figure 1 : A new options tab was created to store the security options

Having performed such modifications for a number of interfaces, a questionnaire was used in order to perform an initial evaluation of the new user interfaces compared with the old ones. The participants in this phase were employees in the IT department of a shipping company, and they were asked to indicate whether they felt the revised interfaces would contribute to improving the usability of the security features in the associated application (five usability grading options were available, from 'much less' to 'much more'). Figure 2 illustrates the average results observed, with almost three quarters of the participants considering that the modified interfaces would make it easier to understand and use the security.

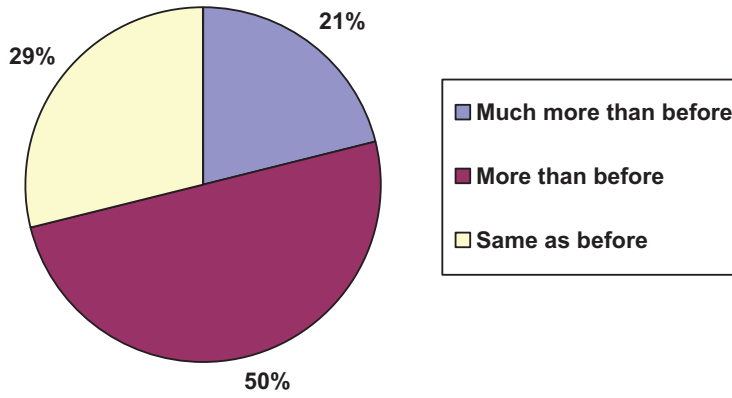


Figure 2 : Perceived level of usability improvement resulting from interface modifications

6. Conclusions

The score that most applications achieved was average, this means applications followed some HCI aspects, but there are still improvements to be made in order to reach a satisfactory mark. The average mark of the applications is 29/50. This score is 21 points away from 50/50 that can be achieved by applying simple guidelines when designing an application. Additionally from the scores in Table 2 it can be observed that the values achieved in each guideline vary, with none seen to perform consistently well or consistently poorly.

All of the proposed HCI-S guidelines are considered to be achievable. To demonstrate this, improvements were made to the interfaces of the applications that obtained the lower marks. These improvements intended to redesign the graphical user interface in such a way that users would find it easier to use. Furthermore additional attention to the HCI-S guidelines was paid in the new interfaces so that use of security would be improved. Some errors were minimized, additional functionality was added using buttons and options, more information and help about security was offered, and explanations were given for specific words, abbreviations and in sentences that could be easily misunderstood by new users.

Although the research to date has provided interesting results, it has only achieved a surface level assessment of how the proposed guidelines would persuade users to use the available security. The guidelines were applied mainly in the interface of the applications, and users could only have a look at the appearance of the new interfaces, rather than actually use them. A more useful assessment of the guidelines, and the effectiveness of the new user interfaces, would be obtained if the improved applications could be used in practice. This issue will be considered as part of the authors' ongoing research.

7. References

Baecker, R.M. (2004) “Goals and Aspects of HCI”, Wikipedia, http://en.wikipedia.org/wiki/Human-computer_interaction, [Accessed: 31 August 2004].

Carroll, J. (2003) *HCI Models, Theories, and Frameworks: Toward a Multidisciplinary Science*, Morgan Kaufmann, ISBN: 1-55860-808-7.

Furnell, S.M. (2004) “Using security: easier said than done?”, *Computer Fraud & Security*, April 2004, pp6-10.

Hewett, T.T, Baecker, R.M., Card, S., Carrey, T., Gasen, J., Mantei, M., Perlman, G., Strong, G. and Verplank, W. (1996) "*Curricula for Human-Computer Interaction*", ACM Special Interest Group on Computer-Human Interaction, http://sigchi.org/cdg/cdg2.html [Accessed: 5 June 2004].

Johnston, J., Eloff, J.H.P. and Labuschagne, L. (2003) “Security and human computer interfaces”, *Computers & Security*, vol. 22, no. 8, pp 675-684.

Katsabas, D. (2004) *IT Security: A human computer interaction perspective*. MSc thesis, University of Plymouth, UK.

Krause, B.R. (2004) "Security and Usability - Basic Principles, single sign-On" http://www.encentuate.com/resources/usability.htm [Accessed: 6 June 2004].

Mandel, T. (1997) *The elements of user interface design*, John Wiley and Sons, United States, ISBN: 0-47116-267-1.

Nielsen, J. (1994) “Ten Usability heuristics”, http://useit.com/papers/heuristic/heuristic_list.html, [Accessed: 5 June 2004].

Norman, N. (2003) “The first principles of Human Computer Interaction”, http://www.asktog.com/basics/firstPrinciples.html, [Accessed: 4 June 2004].

Paller, A. (2002) “*Why is computer security so important?*”, The Ohio State University, http://www.chemistry.ohio-state.edu/compsupp/Security, [Accessed: 9 January 2005].

Preece, J., Rogers, Y., Sharp, H., Benyon, D., Holland, S. and Carrey, T. (1994) *Human-Computer Interaction*, Addison-Wesley, Great Britain, ISBN: 0-20162-769-8.

Swartz, A. (2004) "Usability in the Real World: The Paradox of Usable Security", http://www.usabilitynews.com/news/article1875.asp

Whitten, A. and Tygar, J.D. (1999) “Why Johnny can’t Encrypt: A usability Evaluation of PGP 5.0”, *Proceedings of the 8th USENIX Security Symposium,* Washington, D.C., USA, August 23–26, pp169-184.

Approaches to Establishing IT Security Culture

C.Langue, S.M.Furnell and P.S.Dowland

Network Research Group, University of Plymouth, Plymouth, United Kingdom
e-mail: info@network-research-group.org

Abstract

People are widely considered as the weakest link in the Information Technology framework. Although companies have security tools such as anti-virus software or firewalls, the behavior of employees may cause huge incidents. People are not aware of security issues, and they hardly consult documents for advice and guidance. This problem is particularly relevant to SME's (Small and Medium Enterprises), which rarely have a dedicated IT security staff. Many guidelines and standards have been developed, but their complexity is not adapted to the needs of most companies. The aim of this research is to identify realistic means by which IT security culture can be enhanced and propose appropriate methods to train and improve awareness of staff. The resulting paper is a guideline for the development of a relevant IT security awareness and training programme, defining all the aspects to be considered: roles and responsibilities, budget, policy, topics to be covered, tools to develop for awareness and training and assessment of existing tools. By reading the paper, medium-skilled person should understand the issues of IT security culture and understand how to enhance it within a company.

Keywords

IT security culture, security policy, security tools

1. Introduction

The number of people connected to the Internet is expanding very fast and the purpose of its utilisation is changing. The Internet was first used by companies as a method of communication. Promoting products and contacting clients or employees was its main utility. Today, it is also used to exchange critical data and make financial transaction, which require high standards of security. Unfortunately, its development was so fast that people did not learn the basic rules of safety. Moreover, for economic reasons, companies developing software do not spend enough time to ensure of the security of their products. As a consequence, the number of weaknesses which may cause disaster on the business of companies increase at a huge rate - from 1,090 in 2000 to 3,780 in 2004 according to CERT/CC (2005). Although many businesses have anti-virus software, firewalls, policies or automatic backup, it is no longer sufficient to protect data.

The best way to protect a company against security incidents (consequences of direct/indirect attacks or accidental failures) is to train employees to adopt responsible and safe behaviour. People are the weakest link of the IT society, because they have never been taught and learned the basic rules of security. Improving this knowledge is now crucial for the protection of companies and for the protection of the Internet.

A variety of guidelines and recommendations have been developed to help companies enhance their IT security (NIST, 2003 and OMB, 1996), but most SMEs do not have

resources (material, financial and human) to apply them. Until now, the focus has largely been placed upon large enterprises, but 53% of the European population work in a company employing less than 50 persons (DTI, 2004). The intent of this paper is to help IT security managers to develop a relevant and cost-effective IT security awareness and training programme, based on a realistic step-by-step approach, an analysis and implementation of products existing, a description of products to be developed and of the topics to be covered while awareness and training.

2. The need for security

What are the resources used by SME? Figure 1, based on the results from the UK Department of Trade & Industry (DTI) shows that in 2004 in the UK, over 90% of employees can access the Internet. Its main goal is still to communicate via e-mails, but the amount of transactional purposes jumped from 13% in 2002 up to 73% in 2004. Utilisation of wireless networks has increased from 2% in 2002 to 34% in 2004. Financial operations require high security level, but wireless networks are extremely vulnerable and unprotected. Indeed, other survey findings suggest that 53% of companies having a wireless network do not protect it (Gordon *et al.* 2004). Moreover, e-mails are the main carrier of viruses, Trojans and worms attacks; they affected almost 80% of US business and cost over $55 millions (Gordon *et al*. 2004).

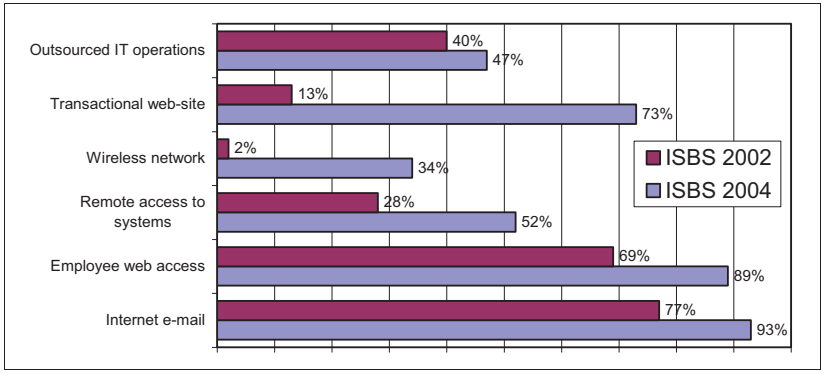


Figure 1 : Resources/services used by organisations (DTI, 2004)

Considering that 51% of information in UK is highly confidential, 52% would cause significant business disruption if corrupted and 44% would cause significant business disruption if not available (DTI, 2004), the issue of keeping the network secured is obvious.

Recognising that employees are often a cause of (or contributor towards) incidents, it is relevant to consider how that are made aware of their responsibilities. Further results from the DTI (Figure 2) show that employees are mainly trained during the induction or with handbooks. However, while these may set things in a sound direction, there is a need to ensure that the issue remains in the employees' minds. In this respect, it can be observed that only 23% of the surveyed companies regularly train their staff.

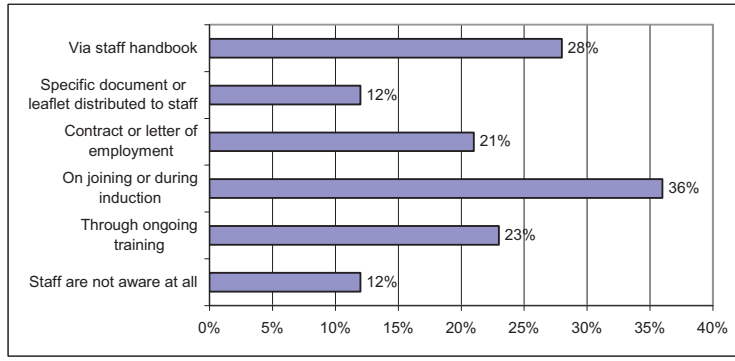


Figure 2 : How staff are made aware of IT security policies (DTI, 2004)

3. Building relevant IT security guidelines

The intent of the research is to deploy a guideline that security managers can easily use to build a relevant IT security awareness and training programme. Choosing appropriate awareness and training material, pertinent topics to develop, building a security policy, conducting needs assessment, defining roles and responsibilities are described in the guideline.

3.1 Inform the management

Before developing an IT security programme, security managers have to convince their organisation's management about its necessity. In this respect, published sources can be drawn upon to provide figures and arguments that managers are likely to be receptive to. For example:

- 25% of companies had a serious incident involving accidental system failure or data corruption (DTI, 2004).
- Within SMEs, the loss generated by a serious incident averages £10,000, rising to £120,000 in bigger companies (DTI, 2004).
- According to DTI experts, 5% of the IT budget should be spent on security.
- 99% of companies have anti-viruses, but last year 79% contracted viruses, Trojans or worms (DTI, 2004).
- Number of viruses increased 400% in the first term of 2004 compared with the same period in 2003 (Symantec, 2004)
- With specialised tools (passwords crackers/sniffers, port/network scanners, spyware etc.), hackers can break into a weak network within minutes (CERT, 1997).

Security managers may also exploit statistics issued from the network (number of incidents, malicious attacks, accidental failures, amount of financial losses, working days lost, etc.).

3.2 The IT Security Awareness and Training programme

Improving an IT security culture is a long process. Buying security tools and awareness products is not enough; a strategy has to be designed and a precise programme must be deployed. Many steps have to be reached before establishing the IT security awareness and training programme. Figure 3 shows processes and elements which must be considered to develop it, and the following description provides a brief overview of each part of the task.

The first step is to define the security action plan, which defines the strategy of the programme. The plan is built on the result of pre-evaluation (which identifies the state of the IT system and the current knowledge of the staff), and needs assessment (which establishes the level of education needed and the critical security gaps). The security action plan defines responsibilities, the schedule of the programme (long and short term), the security policy, and topics to be covered by the IT security awareness and training programme. These areas are addressed in published guidelines (NIST, 1998), but they are not always fully relevant to SMEs.

Studying resources needed is a critical point of the programme and managers must ensure that companies:

- have logistical resources to apply the program (computers, rooms, equipment).
- have skilled personnel to develop the security and awareness tools.
- have personnel to dispense the training.

Companies may lack resources to apply the programme and ask for the service of a third party. An accurate analysis of the human and material resources needed for the programme is hence necessary: if nobody can apply the programme, it is useless to develop it.

Budget must consider the cost of developing or acquiring material, but also the cost of post-implementing the programme (e.g. update of awareness and training tools, incident reporting, assessment of employees, etc.). The budget will probably not permit a security manager to outsource the entire programme (from the audit to the implementation and reporting), so a precise budget model (depending on each company) must be developed and respected.

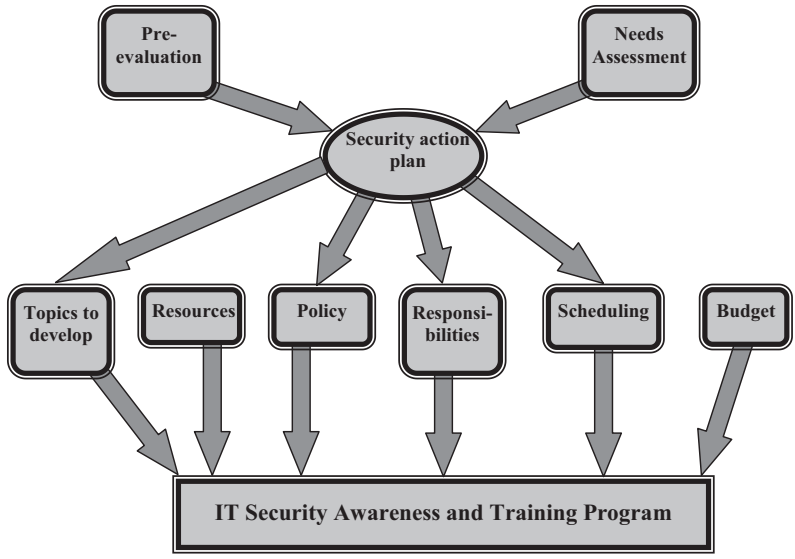


Figure 3 : IT security awareness and training program process

When all of these stages have been reached, the material to be bought or to be developed can be chosen and implemented. Many products have been developed to fill the gap of security knowledge. The basic solutions proposed are (most of them are awareness solutions):

- Provision of periodicals (these can provide a good source of information, however it may be difficult to persuade staff to read it and this does not guarantee that staff will apply or even understand the material).
- Screen-savers with security reminders (these can be used to ensure the security message is repeated to staff on a regular basis, however this may quickly become irritating to staff and will inevitably lead to staff attempting to bypass the screen saver or replacing with an alternative version).
- Posters (these can be a good solution for short messages, however they are likely to be ignored over time – the position of these will determine their effectiveness).
- Computer-based training (this requires considerable resources, in particular, an investment in time but can potentially be an effective solution).
- Lectures and seminars (these can prove very effective if supported by the majority of the staff and led by skilled presenters).
- Videos (these can be very good if sufficient relevant material/content can be found, or produced, to suit the company's requirements).

To enhance the effectiveness of the programme, efforts should be made to ensure that it is not annoying for employees, and users should be rewarded for their good behaviour. A help desk must be created to provide support for the employees.

Post-implementing the programme is a big issue: security tools, training products and topics must be up-to-date. Feedback of the programme and implementation of the strategy allow a good continuum of the learning. Several solutions are possible for reporting and assessment: personal interviews, audits, statistics, questionnaire, and external benchmarking.

4. Conclusion

This paper is a step toward the establishment of an IT security culture. Many training and awareness tools are available, but it is important to appreciate that such a culture is unlikely to be developed by a single step. It is then important to develop new guidelines which can be used by non-experts, and the paper has outlined a framework that can be used to guide the process. In terms of the incentive to adopt and follow it, managers must realise that it will often be less expensive to train staff than to repair their mistakes and suffer the consequences of a breach.

5. References

CERT/CC. (1997) “Security of the Internet”, URL: http://www.us-cert.gov/reading_room/tocencyc.html (accessed December 2003)

CERT/CC. (2005) “CERT/CC Statistics 1988-2004”, URL: http://www.cert.org/stats/cert_stats.html (accessed March 2005)

Gordon, L.A., Loeb, M.P., Lucyshyn, W. and Richardson, R. (2004) *Ninth Annual CSI/FBI Computer Crime and Security Survey*. Computer Security Institute.

DTI. (2004) *Information Security Breaches Survey 2004*. Department of Trade & Industry, April 2004. URN 04/617.

NIST. (1998) “Information Technology Security Training Requirement: A Role- and Performance-Based Model” URL: http://csrc.nist.gov/publications/nistpubs/800-16/800-16.pdf (accessed February 2004)

NIST. (2003) **“**Building an Information Technology Security Awareness and Training Program”, URL: http://csrc.nist.gov/publications/nistpubs/800-50/NIST-SP800-50.pdf (accessed February 2004)

OMB. (1996) Office of Management and Budget (OMB) Circular No. A-130, “Management of Federal Information Resources”, URL: http://www.whitehouse.gov/omb/circulars/a130/a130.html (accessed February 2004)

Symantec. (2004) Internet Security Threat Report volume VI, URL: http://enterprisesecurity.symantec.com/content.cfm?articleid=1539 (accessed September 2004)

ISEduT: An Educational Tool for Information Security

S.Sharfaei and S.M.Furnell

Network Research Group, University of Plymouth, Plymouth, United Kingdom
e-mail: info@network-research-group.org

Abstract

Lack of employee awareness is a recognised reason for the occurrence of many security incidents. While there are many approaches to assist in raising security awareness, small to medium size organizations face limitations in selecting appropriate methods due to their lack of budget. The result of this research is a computer-based tool that aims to improve security awareness efficiently and with low cost. The tool uses multimedia, graphics and texts to represent real-life scenarios existing in organizations. Users' security awareness is assessed through a set of questions involved in the different types of scenarios and their understanding is raised using the security countermeasures provided by the application.

Keywords

Computer Based Training, Security Awareness, ISO 17799, Education

1. Introduction

Survey findings have established a number of significant obstacles to the achievability of security within organisational environments, including inadequate end user awareness, budget constraints, and skills constraints (Ernst & Young, 2004). While the latter issues can often be addressed more easily in larger businesses (i.e. with more in terms of available resources), the first issue can represent a particular problem for organisations at all levels. Indeed, employees that are not trained to protect their data security are most of the time not even aware of the threats. Employees think that the IT department in the company can hold the security responsibilities that the employees are not aware of. This raises the necessity of a method that solves the information security problem by raising end user awareness with low budget investments. The paper investigates end users' awareness, different methods of raising security awareness, and the design of a computer-based tool for security awareness training. The outcome of the research is a solution for raising security awareness, by using a computer-based training (CBT) tool that does not require users to have a high degree of specialized knowledge.

2. Importance of Security Training

Problems related to information security arise as a consequence of computer-based technologies used in organizations. These problems are extremely dangerous because they can compromise a company's activities. This has become a major concern for institutions that possess sensitive data. Sabotage has always been present in each industry or market field and can produce economical disaster for a company. It is known that many times companies do

not report such intrusions fearing bad publicity that can make their situation even worse. Information security can be threatened both intentionally and unintentionally. It is well known that a system that is used in improper conditions can permit intruders to take it over.

Many organisations are failing to secure their technology and this is because their employees are not aware of security issues. Human mistakes can cause more security breaches than technology flaws. Figure 1 shows the common types of security breaches in UK businesses. It is clear that increasing employees' awareness in security related issues could reduce most of these security breaches such as theft of computers, staff misuse of information system, virus infections and even unauthorised access.

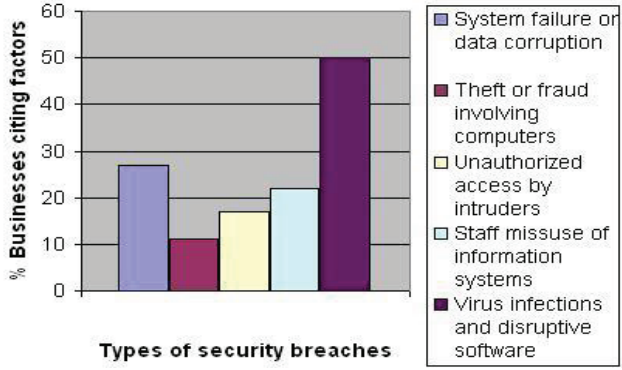


Figure 1 : Common type of security breaches (DTI, 2004)

A strong argument can be made that organizations can only have security if they have a security policy, and then educate their employees to act in accordance with it. Fortunately, the number of businesses with security policies in 2004 has increased comparing to 2002 (DTI, 2004). However, it is still disquieting because only 34% of small businesses and 45% of medium-size businesses have security policies. This means that more than half of the small to medium size organizations still do not have any security policy in place. In addition, findings from the Computer Security Institute (Gordon *et al.* 2004) state that while the majority of their 480+ respondents rated security awareness training as being important, many also considered that their organizations were not investing enough money to support it. Such findings suggest a definite need for training the employees and increasing their security awareness.

3. Promoting Security Awareness using CBT

Understanding the importance of security awareness and considering the widespread lack of employee awareness in information security can bring about the motivation for various organizations to enhance employee and user education and training in security matters. Some of the methods that can be considered to increase information security awareness are:

corporate endorsement, sending personnel to training courses or hiring security specialists, clauses in employment contracts, written materials, demonstration of security breaches to employee, threatening disciplinary actions and security awareness software (Furnell *et al.* 2003).

Currently many researchers responsible for integrating education with technology are trying to improve learning methods by making better use of computers in education. This method is better than corporate endorsement, clauses in employee contracts or written materials methods mentioned above. CBT gets the user involved and prevents them from forgetting the problem; they will manifest interest in a process that will improve their skills rather than reading some pamphlet that is usually forgotten in a drawer. Threatening disciplinary action and demonstration methods can easily produce irritation and tension amongst employees and it is known that these two actions will both affect their efficiency and autonomy. This can be avoided with CBT because employees will know that when they have a security issue all they have to do is to run the program and find out the answer.

Sending personnel to training courses or hiring security specialists are highly rewarding methods in the quality of the employee security awareness, but are also extremely expensive. The company must afford sending personnel to courses (sometimes at a rate of £500 per day, including accommodation and transport) or to hire security specialists. Compared to these costs computer based training may be even free. The economical aspect can be the most important factor of the whole discussion. Because everything relates to it: the need of security or the process of building security. CBT can be used on unlimited numbers of computers. This makes its costs lower than the other methods. The users can run the application whenever they like. CBT can be customized to newcomers into a company, for testing and training before endowing them with access to critical information assets.

4. The ISEduT system

The Information Security Educational Tool (ISEduT) tool has been developed to provide a CBT environment that promotes security awareness. The application is implemented using Visual Basic and Microsoft Access database. The database contains both the security countermeasures and information about the users (name, date, score and the elapsed time to complete the test). The exercises are fully compliant with the ISO 17799 standards (BSI, 2000). These standards are the guidelines for all the security issues presented by this software. The ISEduT is aimed to:

- Train the staff members of a small to medium size company.
- Provide effective training for information security program by defining the functions and responsibilities of staff members.

In order to meet the objectives of the application, extensive researches were carried out and some existing computer based awareness tools were studied. The appropriate security guidelines were extracted from ISO 17799 to make staff members aware of the companies'

security issues and policies in order to protect the organizational information assets from security breaches.

To signify the gathered information about security awareness in the best manner, the top technologies available in the computer are used to attain the best result in raising security awareness. Providing the information using one approach (e.g. text only) will make the test tedious and it reduces the effectiveness of the software. In order to make the tool more interesting for the users, the security information is represented using different techniques.

The application implements multimedia (audio and video), texts and graphics to simulate security problems into four types of pre-defined case study scenarios and situations. Some of the scenarios contain only one security problem (Single Problem Scenarios - SPS), whereas others contain a range of issues (Multiple Problem Scenarios - MPS). The test is in a way that makes the user think and find the security problems represented by each type of techniques used, and then select the security problem through multiple selection answers provided by the software. Also in some stages the application asks the user to find the appropriate solution for a particular security problem. The software is supplied with an optional feature through the test that provides additional information about specific security problems and useful tips. The application tracks the time taken to complete the whole test by each user.

The scoring mechanism in the software is done accurately. The application scores the users according to the number of questions answered. Each question has 8 marks. ISEduT allows two attempts for each question. In both SPS and MPS scenarios the user obtains the full mark when he/she selects the correct answer(s) in the first try. In SPS, the application marks the user 4 out of 8 when the user finds the correct answer in the second trial and no mark is given when the user does not succeed to find the answer in two attempts. In MPS, the scoring method is different if the user does not find the correct answer in the first trial. As the user selects the answers in the second attempt, the application checks if both answers are correct, then it scores the user 4 out of 8 (because it is the second attempt) and if one of the answers is correct and the other is incorrect it only marks 2 out of 8. No mark is given when both answers are wrong.

The tool is a multi-user system since it allows multiple users to perform information security test. Each user will own an account after registering with the application and their history will be saved in the database. The ISEduT provides a report containing the history of each user; accordingly employees' performance and awareness level can be checked through these sets of reports. To maintain privacy, each user can only view his/her history report when using the application. The report contains information such as number of questions answered during each test; the user's score, time to complete the test and the date on which the test was performed. These information are useful in the sense that the security awareness level of each user can be examined based on the employee's score and time taken to complete the whole test.

4.1 Case Study Scenarios

The ISEduT contains a total of 24 case study scenarios, presented using different types of audio, video, picture and textual information to convey the scenario environments. The scenarios contain measures that explain a real-life situation which intend to help the learners

understand the materials. The test models (except video scenarios) were built upon risk analysis, security policy and the ISO 17799 standards. Major risks that organizations can fall victim to were identified and the topics which employees are expected to be aware of, were selected from the identified risks. For the purposes of this prototype implementation, the contents of video scenarios were selected from existing cyber security awareness training videos produced by the Arizona Department of Administration (ADOA, 2003).

The scenarios generally involve 1-3 actors (except for the video scenarios) and the required information to represent the scenarios is provided in short dialogues. The advantage of using short dialogues with fewer actors is that the users spend less time in understating the scenario therefore it maintains users' concentration during the test. The tool contains three video case study scenarios that show short video clips of real-life cases that can occur in companies.

Figure 2 (a) shows one of the picture scenarios involved in the tool. The graphics involved in the application are created using different designing tools. Poser is a graphic designing tool that has environmental objects (systems, mouse, and keyboards), data objects (text and colour). The 3D studio provides the third missing dimension and it is based on three-dimensional images. It also has a set of already designed objects such as humans. These objects are ready to use and can be changed and modified as needed.

4.2 Situations

Situations present security problems using graphics that contain active pieces of image. One part of each graphic contains a security problem and the user identifies the problem by clicking on the image. The ISEduT prototype contains five different situations. The user has two chances to identify the security problems by pressing on the clickable parts of the graphics. If the user misses the chance of finding the security problem on the second trial, then the software will give no mark and the user will have to proceed to the next situation. Figure 2 (b) shows the ISEduT interface for one of the situations.

Figure 2(a) : Picture scenarios interface

Figure 2(b) : Situations interface

The application allows the users to select the type of scenarios and answer the number of questions that they like which means ISEduT has the capability to allow users to follow their own custom learning path. The software has a great benefit for companies as everything is saved in its database and the database is designed in way to can be easily upgraded.

5. Discussion

The ISEduT database covers a number of scenarios and case studies directly related to security awareness. One of the aspects that give value to the tool is diversity. The scenarios are presented using different methods. This is due to the fact that CBT must be entertaining and incentive for the users to keep the users concentrated during the test. But this is only one phase, because each different scenario trains the user into recognizing specific aspects that are better promoted when using a precise presentation. For example it might not be clear to someone that keeping passwords on notes on the monitor is insecure until he/she observes (in a picture or video) a situation where everyone behind him are writing down those passwords.

In order for the software to increase the learning efficiency the database material focus on a number of specific situations and subjects considered relevant to the majority of employees working in IT environments. This enables the user to immediately apply what he/she had learned from using the software. The ISEduT can be installed on the computers and the users would run it when they have unoccupied time. This provides another advantage: each user will use the ISEduT in conformity with his/her own learning cycle characteristics.

The efficiency of the software must be tested by the software developer and also by others. The tests were done using the latest version of the software and the scope of the testing was to see how people react to it. The objective behind the testing is to see if the educational goals of the software are achieved. Four persons were available for the testing. Each of them has different computer knowledge skills and backgrounds. The results of the testing were evaluated using history reports provided by the application. While the results of these testing are not concluding for every situation to be encountered, they still give valuable insight on how ISEduT software affects the staff. The test showed that each user ran the tool at least twice. This is a sign that the CBT instigates attention and motivation in the people using it. The general impression the users communicated was that the software was easy to learn and fun to use. All of the users believed that they learned something new and they would have used the software to expand their knowledge.

Based on the findings it is expected that with proper management the budget allocated to security decrease without affecting the security awareness level. This is possible due the reuse of inside trainers (if any) and due to inexpensive in-house training (the ISEduT). Security awareness is not to be done once and gets beyond; it is a process that never ends. It is important to understand that the ISEduT is not effective in an environment where no one promotes it or at least provides the proper equipment and access to use it. Although it may be possible that sometimes self trained users pick it up and start using it effectively. This is not probable to happen often therefore the greater support from the managerial line the greater the good results of the software will be.

To continue this research the following points can be considered:

- Case study scenarios could evolve into having animated images in order to offer more detail on the situations.
- A section in the software can be created that will train the user on a specific area. After that case study and scenarios tests could be run and depending on the results the users get a specific part of the training should be revised. This ongoing manner of training/testing would ensure that all information was assimilated by the user and that he/she is ready to go to another subject.
- The reporting tool can be modified so that it sends testing results over the network/internet to the managers. They will be able to evaluate the progress of each user and accordingly to estimate the effects of the software.
- The tool can be tested in several organizations and the results of the users' performance reports can be collected in order to assure its efficiency and positive impacts on security awareness.
- The ISEduT currently does not support automatic updates but this could be added in the future. This provides the tool a mechanism that will keep the database fresh all the time.

6. Conclusions

The ISEduT software proposed to solve the problem of security awareness in an efficient way with the least costs possible. The findings and research were concentrated on computer based training tools and security awareness enhancing mechanisms. The software now succeeds in raising interest over its content and it allows a small to medium size company a good start on the process of raising security awareness. The user-friendly environment and structured case study scenarios and situations attract the user into using the tool.

The ISO 17799 standard can be used as a single reference point for identifying the number of controls needed for most of the companies that use information system. Because the current research is not exhaustive, and the ISEduT tool can still be improved substantially, future work and research can be done. This research is important to the community because it provides a good way to raise the security awareness level without much economical and logical hassle. The research has been very useful to discover more subtle aspects of the computer based training. The result of this research (ISEduT software) is a ready to run application that when properly introduced and managed in a company is expected to raise the security awareness level significantly.

7. References

ADOA. (2003) *Cyber Security Awareness Training Videos*. Information Security Services, Arizona Department of Administration. http://www.security.state.az.us/Security-Awareness/securityawareness-training%20video.htm [Accessed 11 August 2004].

BSI. (2000) *Information technology. Code of practice for information security management*. BS ISO/IEC 17799:2000. British Standards Institution 15 February 2001. ISBN 0 580 36958 7

DTI. (2004) *Information Security Breaches Survey 2004*. Department of Trade & Industry, April 2004. URN 04/617.

Ernst & Young. (2004) *Global Information Security Survey 2004*. Assurance and Advisory Business Services. Ernst & Young. EYG No. FF0231.

Furnell, S.M, Warren, A.G. and Dowland, P.S. (2003) "Improving Security Awareness through Computer Based Training", in *Security Education and Critical Infrastructures*, C.Irvine and H.Armstrong (eds): 287-301.

Gordon, L.A., Loeb, M.P., Lucyshyn, W. and Richardson, R. (2004) *Ninth Annual CSI/FBI Computer Crime and Security Survey*. Computer Security Institute.

Security Technologies for a Virtual University

V-C.Ruiz, S.M.Furnell and A.D.Phippen

Network Research Group, University of Plymouth, Plymouth, United Kingdom
e-mail: info@network-research-group-org

Abstract

Virtual universities (VUs) are meant to provide a new learning model for students and a way to study away from traditional facilities. Security is now regarded as a high priority within computer systems, and has been increasingly introduced to the public over the years. Many new technologies have been presented as means to secure computer systems, especially when connected on large scale networks such as the Internet. As in every IT system, virtual universities are faced with potential attacks from outside or inside its network. This paper proposes a security model for such universities, in order for the students, staff, and the administrators to work in a safe environment, and provide the same degree of confidence as in a traditional learning environment.

Keywords

Security, Online Distance Learning, E-learning, Virtual University.

1. Introduction

In recent years, national studies concerning security within organisations (among them educational organisations) have been published to the public. The Department of Trade & Industry from United Kingdom and the Computer Security Institute in the United States have both published a survey concerning security problems encountered by companies during the year 2003 (DTI, 2004; CSI, 2004).

Many aspects of these survey results show an increasing number of incidents detected by companies during the past year, compared to previous years. This increase in the number of reported incidents may be viewed either as an increased awareness about what is really happening in an organisation's network, but could also represent an actual increase in the number of incidents. Nevertheless, results regarding these surveys can explain part of what an institution such as a virtual university may be faced with when connected to the Internet.

In the Symantec Internet Threat Security Report (Symantec, 2004), educational organisations are identified as one of the primary targets for attackers. In this report, Symantec noted that the education industry is the 8th most targeted on the Internet. From another perspective, recent years have also witnessed a number of attacks being launched onto the Internet from *within* universities. Since university networks are usually composed of a high number of machines connected to the Internet by a huge amount of bandwidth, it makes them prone to host attackers (students or external hackers that gained access to a machine on the network) (Borland, 2000; Roberts, 2004).

Adding to the threats of security breaches coming from the Internet and users, virtual universities are keeping highly sensitive information regarding its users. This information, along with copyrighted material, makes VUs highly sensitive organisations.

This paper intends to describe potential security issues that may occur in online distance learning environments such as virtual universities, and proposes a security model in the perspective of future integration in applications.

2. Security issues facing Virtual University environments

E-learning platforms (and therefore VU), rely on a client-server approach. Requirements regarding their security can be decomposed into three sub-categories: the server, the client, and the network. Issues may be addressed by implementing security measure in one or more categories. Traditional universities secure their network from improper outside use by using the latest anti-virus software, installing firewalls, preventing unauthorised software installations, etc. They usually also establish an IT policy usage that the users must agree with, in order for them to be able to use the network (Kvavik *et al*., 2003). A traditional university network is a private network which can be monitored, and if a student is spotted performing an unauthorised action, their account can be revoked immediately. Virtual universities, due to their distance approach and use of large scale networks, are more complex to monitor and maintain. Depending on the user's rights and actions on the server, issues may vary.

2.1 Authentication, authorisation

Authentication is a mechanism whose goal is that only registered and legitimate users get access to the system. In VUs, authentication is a crucial part of the security processes as it is used to grant access to sensitive material. Once the authentication has been established successfully, authorisation of access to certain parts of the server may be granted.

2.2 Confidentiality

Confidentiality within a university is not only a requirement for the university itself but also for the students. Indeed, communications between students and instructors are confidential, and therefore should not be spied on. In terms of data confidentiality, the virtual university stores different information regarding the users; students' grades are required to be kept secret, personal information about each individual must not be released on the Internet.

2.3 Copyright

Every piece of material on the server (e.g. instructors' lecture notes, books in the library, research paper, etc.) is copyrighted. Each individual who writes a paper would rather not see their work published under another name somewhere else in the world.

2.4 Other issues

Depending on its implementation, virtual universities may be faced with other issues. There are various ways to set up an online distance learning platform, one of the most common is by using a traditional web server (WebCT, Archimed Campus Virtuel, etc.). This approach has the advantage of being easy to set up, but lacks flexibility for implementing the best technologies available. Additionally it has the disadvantage of the incompatibility of different web browsers, and is prone to the different bugs and vulnerabilities that they may contain.

The second approach to creating an e-learning platform is by creating custom software (Wang, 2001). This approach requires the users to have the software installed on the computer from which they want to access the university. This latter approach therefore decreases the flexibility regarding mobility. Bespoke software developed in-house may also be prone to cracking from badly intentioned users if it has not been exposed to sufficient prior testing.

3. Security model and available technologies

In order for the VU to address and resolve issues it is faced with, a security model needs to be implemented. Various technologies can be used to fulfil the goal of securing the university, and some of the principal aspects are discussed in this section.

3.1 Server

The server is one of the most important parts of the virtual university. VUs usually use a web server which provides services to the user (Dean, 2002). It must therefore be able to handle at least authentication, integrity and privacy. The virtual university may rely on multiple servers (for example one for each service or for balancing the users to one another in order to increase the availability), which need to be administrated. A set of multiple servers is harder to maintain, but generally provide better availability. In this scenario, the whole server cluster needs to be secured, on software as well as on a physical basis.

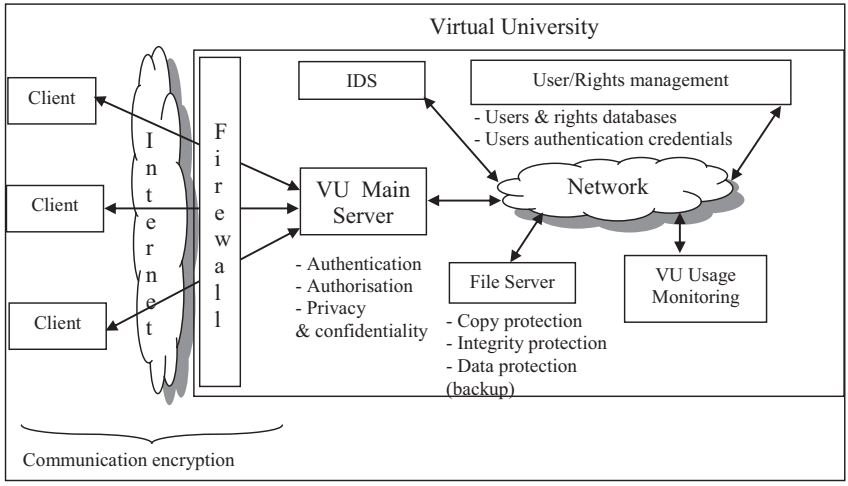


Figure 1 : Architectural diagram of a VU implementing the security model

Figure 1 shows how a virtual university can apply some protection measures, as part of the security model presented below.

3.2 Access control

3.2.1 User Authentication

One of the primary goals of the server is to be able to authenticate and identify the users reliably. Since the virtual university contains sensitive information, regarding the users as well as the taught material, this authentication process must be as thorough as possible and provide certainty concerning the users attempting to log in.

Three main types of user can be defined in a virtual university: the students, the instructors and the administrators. Since each one of these users have different rights within the university, their authentication method should be set accordingly. As an example, students could be required to only provide a user name and password combination (restriction may be applied on the chosen password length, form, etc.). A secure password authentication scheme should be preferred for this scenario (Lin *et al.*, 2003). Instructors would use a token, provided to them by the university, to answer a challenge from the server (Federal Financial Institutions Examination Council, 2001). Finally, the administrators would have to use a combination of digital signature (using public/private key pairs delivered to them by the university) and challenge-response authentication to prove their identity. A single sign on (SSO) approach is desirable since different services are used in a virtual university.

When someone connects to a server on the Internet, it should be sure that it is not connecting to an impostor. For this reason, the server should send a digital certificate (validated by a certificate authority) to its connecting clients, proving its identity.

3.2.2 User management

In order to authenticate each user, a user management policy must be put in place. An easy way to add, change, and revoke users from the system must be introduced. This user management system is a critical part of the virtual university functionality and should only be accessed by individuals having the right to do so, i.e. the university administrators. The implementation of the user management should support directory services such as LDAP.

3.2.3 Rights management

Part of the user management system may be used for rights management. In a virtual university, three different types of individual have been defined. Each type of users can be given a set of rights to different parts of the university.

Students, who can be considered as the end-users, may be able to read the material given for their lectures, post and read messages on forums, manage a calendar, access a virtual library and send or receive emails. They must not be able to change in any way the configuration of any part of the virtual university, except regarding their personal preferences.

Instructors are given more rights than students. Since they provide the knowledge, they should be able to post new material, change their lecture content, retrieve students' work, mark them, manage a students list and their marks, post new quizzes and exams.

Administrators are power users regarding the VU. They can be divided into many different categories, such as technical, program, accounting administrators, etc. Each one of these sub-categories may change one aspect of the university's configuration, and may access the university's sensitive data.

3.2.4 Usage monitoring

Universities usually put in place usage policies concerning their computer systems and networks. This kind of policy must also be put in place in a virtual university to prevent unauthorised actions within the university's computer system. On servers, every user's actions should be monitored and recorded. For extra protection, a dedicated server can be put in place for the purpose of collecting and storing logged data.

3.3 Protection from external threats

Virtual universities using the Internet as a communication medium should be especially careful regarding the data coming into the servers. Users, without knowing it, could easily bring down the network or server and stop the university from functioning correctly.

3.3.1 Malware protection

Viruses and worms are the most common threats for online servers (CSI, 2004; DTI, 2004; Symantec, 2004). Whereas viruses usually arrive through emails, worms propagate from one server to another by exploiting flaws in different services. The most effective method to

protect the servers would be to put anti virus software in place to scan every document arriving on the server (via email, or directly uploaded).

3.3.2 Firewall protection

To be protected against attacks, firewalls have proven to be very useful. The VU servers must integrate one as a basic security measure.

3.3.3 Intrusion detection system

An Intrusion Detection System (IDS) is a tool used to detect attempted of attacks or intrusions. Such a tool can detect incoming viruses, worms, etc. on a system. They can use heuristics in order to be able to detect unknown malware. Using such a tool would help the administrators detecting hostile activity within the university's network and react accordingly.

3.4 Data protection and storage

A sensitive part concerning virtual universities is related to data storage. Data is a broad term but in organisations such as VUs, everything should be considered as sensitive information.

Different types of data may be considered in a virtual university. Depending on the user, different information is kept by the university. Whereas the university keeps records of its students' performances (e.g. grades over the years), their payments, etc., it also keeps other information regarding the instructors and administrators, such as their salary, their taught courses, etc. All this information needs to be stored in a secure place and be made inaccessible from unauthorised users. For this reason, each item of sensitive information must be kept in a dedicated format (e.g. passwords should be kept encrypted) and in a dedicated space.

In order for the VU to face every possible event, this information needs to be backed up regularly, by using specific features of each storage and operating system the server uses (e.g RAID). Sensitive data should be evaluated to get the best available technology for protecting and storing them.

3.5 Services management

Virtual universities provide services to the students, such as email, chat, calendar, etc. These services may depend on the users that are using it and therefore need a services management policy.

3.6 Communications confidentiality

On the Internet, communications can be eavesdropped by badly intentioned people, providing them with sensitive information they could use wrongly. The only way to prevent this kind of attack is to encrypt data passed on the network between the server and its clients, therefore preventing anyone who is not part of client-server "conversation" to understand what is being said. SSL/TLS protocols have been developed for that purpose.

3.7 Document Copyrights & Integrity

Even if outside intruders can be stopped from getting sensitive data passed between the different parties, users who have access to the provided material can still make inappropriate use of it. Lecturers notes, virtual libraries books, research papers, and magazines are required to include either copy protection mechanisms or copyright notes (visible or invisible). Depending on the type of document (images, videos, text, etc.), different measures may be applied. Whereas text can use software specific features (e.g. Adobe Acrobat PDF protections) as well as electronic marking techniques (Brassil *et al.*, 1995), images and videos can be marked by watermarking or steganography. These two technologies hide information invisible to the naked eye, but can help to trace wrong uses. Ensuring that documents have not been tampered with should be an added feature to the environment. Lecturer's notes, as well as students' homework uploaded to the server are documents that require an integrity check. When uploading a file, the users should be required to provide a signature concerning the file (e.g. using a program to sign the resulting data with a private key) (Van Vlerken, 2000). This process could be automated, and students or lecturers would be sure that what they are reading is the original handed-in work.

3.8 Server upgrades/updates

To guarantee that the server does not have outdated software which could have security flaws, the technical administrators have to be very careful about announcements made by their software providers. Depending on their budget, and the expected services from the manufacturer, organisations have a lot of different software to choose from. The choice of software to use should be considered thoroughly in terms of services and support associated to it.

4. Summary of requirements

As previously indicated, each user is assigned a set of actions that can be performed on the platform. As a result, each action introduces associated security issues, as summarised in Table 1. For example, when a student hands in a coursework, the instructor marking it should be sure that it is the original one, and that it was genuinely uploaded by the student. The student would also rather not let anybody else see his work and copy it. As a result, the main issues concerning this particular action are copyright, privacy, integrity and non-repudiation.

Action by the user	Security Issue	Protection
Log in (all users)	Authentication & Authorisation	Strong authentication depending on the user. Must be immune to known attacks from outside hackers
Lectures study (students)	Authorisation Copyright	Rights management Copyright protection

Action by the user	Security Issue	Protection
Online Exam (students)	Authorisation Privacy Non repudiation	Rights management Monitoring
Online exams marking (instructors)	Privacy Authorisation	Rights management
Hand-in Work (students)	Copyright Privacy Integrity Non-repudiation	Copyright protection Document integrity measures
Coursework retrieval & marking (instructors)	Authorisation Integrity Non-repudiation	Rights management Document integrity measures
Communication (e-mail, chat) (all users)	Malware Privacy Integrity	Anti-virus/Firewall/IDS Communication encryptions and integrity protection measures
Virtual Library access (all users)	Authorisation Copyright	Rights management Copyright protection
Modify lecture & add new material, exam, quizzes(instructors)	Authorisation Copyright	Rights management Copyright protection

Table 1 : Usage of the security model depending on the users' actions

5. Conclusions

In computer systems, there is a trade-off between security and ease of use and access to information by the users. Most users are used to passwords and PINs as a mean to authenticate, thus making other available technologies harder to be accepted (Furnell *et al.*, 2004). Nevertheless, in order for the users to trust a VU as much as a traditional university, changes in peoples mind will have to occur.

This paper has proposed a general security model for the learning environment of a virtual university. Although the technological aspects may well become dated or prove ineffective in the future years, the model itself should remain a useful reference for future virtual universities.

Future work on the subject would require a thorough assessment of feasibility and test user acceptance within an actual virtual university. This model relies mostly upon operations done by the server, and therefore performances issues should be looked into; the security model

should not undermine the learners experience because of losing time online. The financial aspects for implementing the proposed protection measures should also be a concern.

6. References

Borland, J. (2000) "Universities likely to remain Net security risks", *CNET News.com*, February 2000. Available: http://news.com.com/2100-1023-236933.html.

Brassil, J. T, Low, S, Maxemchuk, N. F. and O'Gorman, L. (1995) "Electronic Marking and Identification Techniques to Discourage Document Copying", *IEEE Journal on Selected Areas in Communications*, Vol. 13, No. 8.

CSI. (2004) "2004 CSI/FBI Computer Crime and Security Survey". Available: http://i.cmpnet.com/gocsi/db_area/pdfs/fbi/FBI2004.pdf.

Dean, C. (2002) "Technology Based Training & On-line Learning – An overview of authoring systems and learning management systems available in the UK", December 2002. Available: http://www.baol.co.uk/PDF/authsys.pdf

DTI. (2004) *Information Security Breaches Survey 2004*. Department of Trade & Industry, April 2004. URN 04/617.

Federal Financial Institutions Examination Council. (2001) "Authentication in an Electronic Banking Environment", August 2001. Online: http://www.ffiec.gov/pdf/pr080801.pdf.

Furnell, S.M, Papadopoulos, I. and Dowland P. (2004) "A long-term trial of alternative user authentication technologies", *IMCS – Information Management & Computer Security*, Vol. 12, No. 2, pp178-190.

Kvavik, R. and Voloudakis, J. (2003) "Information technology security: governance, strategy, and practice in higher education", ECAR – Educause Center for Applied Research, October. 2003.

Lin, C-L. and Hwang, T. (2003) "A password authentication scheme with secure password updating", *Computers & Security*, Vol 22, No 1, pp. 68-72.

Roberts, P. (2004) "Attacks at Universities raise security concerns", *Infoworld*, April 2004. Available: http://www.infoworld.com/article/04/04/14/HNuniattacks_1.html.

Symantec. (2004) Symantec Internet Security Threat Report, Volume V, March 2004.

Van Vlerken, P. (2000) "Message Authentication, Integrity, and Non-repudiation from Paper to PKI". Available: http://www.imforumgi.gc.ca/new_docs/authentic_e.pdf.

Wang, Y. (2001) "Security framework for Online Distance Learning", *Master's thesis*, University of Plymouth, Plymouth, UK.

The Interaction Between Mobile IPv6 and Firewalls

L.Ghashash[1], S.M.Furnell[1], A.Akram[2] and B.V.Ghita[1]

[1]Network Research Group, University of Plymouth, Plymouth, UK
[2]Panasonic Mobile Communications Development of Europe Ltd., Thatcham, UK.
e-mail: info@network-research-group.org

Abstract

IPv6 protocol introduces new changes to the Internet infrastructure these include end-to- end IPSec, end-to-end connectivity and flexible extension headers. These new features are not supported by current commercial firewalls, which were built using different aims and objectives. Mobile IPv6 in its turn introduces new way of working by enabling users to keep their current session on the move regardless of their current location, but it also imposes critical changes by enabling untrusted users to be located inside the firewall-protected area. This feature along with IPv6 features changed the aims and objectives that traditional firewalls were built to fulfill.

This study highlights what those problems/challenges are. New modules for firewalling (Quarantine Module and Distributed Firewalls), are introduced which should tackle the problems in deploying Mobile IPv6. Companies wishing to deploy Mobile IPv6 in their network have one of the two options: to enhance the current infrastructure by isolating the Mobile Nodes, Correspondent Nodes and Home Agents in separate segments and protect each segment by a strict access list. Alternatively, they could install the new techniques to its infrastructure to handle mobility. The bottom line is that the success of Mobile IPv6 deployment depends on how much the security is a critical issue.

Keywords

Security, Mobile IPv6, Firewalls, Distributed Firewalls, Quarantine Module.

1. Introduction

Mobile IPv6 was designed to enable user mobility when used with IPv6 (Perkins et al, 2004). This protocol works by using two IP addresses, The Home Address (HoA) that is used to identify the Mobile Node (MN). The second address is the Care of Address (CoA) which is obtained in each location visited by the MN.

Using Mobile IPv6, users are not bounded to their home network, but are free to roam anywhere. This enhancement poses several challenges, as users are not identified by their location and may have any IP address or could be in any location, in addition to the new end-to-end IPSec. Firewalls have to cope with these changes in order to provide the required security, whilst remaining appropriately transparent for the mobile user.

2. The Problem of Mobile IPv6 and Firewalls

This section summarises a number of challenges that may be faced in terms of applying firewall protection to mobile IPv6 environments.

Changes in the aims and objectives of firewalls
Firewalls are not designed to handle IPv6 traffic or mobility; they only aim to protect fixed hosts inside the LAN based on by their IP addresses. No traffic is allowed to pass without inspection or logging (Vives and Palet, 2004). Normally firewalls trust internal nodes and do not trust external. In addition, protected nodes do not move outside the protected area. Any host is reachable from outside via the firewall only. Finally, only five elements are needed to make a decision: source and destination IP address, source and destination port, and protocol.

The above description aims fail to fulfill the requirements of Mobile IPv6 networks. Such environments introduce major changes in the Internet topology, such as visiting untrusted MN inside the organization and end-to-end IPSec.

IPv6 extension header and routing header
Due to the flexible nature of its specification, the IPv6 protocol allows a packet to contain one or more extension headers. In addition, skipping unknown extension headers by firewalls is dangerous (Savola, 2003). The routing header (one of the IPv6 extension header) poses an additional challenge because Mobile IPv6 specifications do not restrict the MN from forwarding the packet to an address different than his Home Address (HoA).

End to end IPSec
End-to-end IPSec is a compulsory part of IPv6 (Deering and Hinden, 1998). In Figure 1, host A communicates with host B using end-to-end IPSec. In this scenario, Firewall X (FWX) cannot read the packet contents, thus it cannot inspect the packet and log it. Some companies may find this behavior a possible source of information espionage because only host B can read the content of this packet.

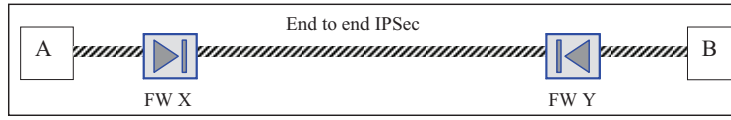


Figure 1 : End to end IPSec communication

Normally, FWX creates a state entry for the outgoing message, when a reply comes back, it compares it with that entry (Strebe and Perkins, 2002). By using IPSec, the state entry cannot be created properly thus, FWX cannot make a correct decision about the packet. Firewall Y does not trust host A, and it does not know what the content of this message is (since it cannot read it). Any content filtering or pattern analyzing is not possible in this case in order to detect any malicious activity. Many companies may desire to block this end-to-end connectivity due to these reasons.

Internal Security concerns
In Mobile IPv6, MNs (Mobile Nodes) are not internally protected anymore. There are several dangers when hosting unknown MN in the same area; attacks may be launched from a malicious visiting MN, as they do not have to traverse the firewall. In addition, any MN (local and visiting) may carry viruses/worms that can spread inside the network.

State Entry Problem

This is simply because the firewall does not have a state entry in his state table to match incoming message (Strebe and Perkins 2002), so it drops it. Each firewall protecting Home Agent (HA), Correspondent Node (CN), MN will make a problem for different messages

Firewalls protecting CN

The firewalls protecting CNs will make a problem for three messages: Home Test Init (HoTi) coming from HA, CoTi, and Binding Update (BU) coming from MN.

Prior to return routability test (Perkins et al. 2004) , all the traffic between Home network and CN was data, when a HoTi message arrives to the CN firewall, this message does not match any previous entry in its state table, and therefore it is dropped by the CN firewall
CoTi and Binding Update messages coming to the CN, the CN firewall does not have any entry in its state table for MN address (CoA), thus these messages are dropped by the firewall.

Firewalls protecting HA

Firewalls protecting HA will create a problem for BU, Dynamic Home agent Discovery (DHAAD), Mobile Prefix Solicitation (MPS) and HoTi messages. All of these messages are initiated from the MN toward HA, the HA firewall will not have any state entry for any of these messages therefore it drops them.

Firewalls protecting MN

a. Firewalls in the Foreign Network
   Since the MN is the node that initiates all the messages, there will not be any problem with that firewall, as each message has a state in the firewall entry table. However, there will be a problem for the Data coming from HA, as it does not any entry in that firewall. Any data initiated by the CN (and tunneled by HA to the MN) may be dropped, as it does not match any entry in the FN firewall (this is the case for peer-to-peer - p2p - applications).

b. Personal firewalls in the MN
   Since the MN is the initiator of all the traffic either toward the HA or the CN, thus the MN personal firewall should pass all Mobile IPv6 traffic to the MN.

Table 1 shows a summary of the problematic messages. We notice that most of the problems are related to Return Routability Test and route optimization. (Shima, 2003) suggests an improvement to distinguish if the MN and the CN are in the same network (Internet or intranet). If both are in the same location, route optimization may be performed (because no firewall separate them), elsewhere MN should continue communicate via it is HA by using tunnelling.

No.	Firewall	Message	Source	Destination
1	HA	BU	MN(CoA)	HA
2		HoTi	CoA	HA
3		DHAAD	MN (CoA)	HA
4		Mobile prefix solicitation	MN(CoA)	HA
5	MN	Data	HA	MN
6	CN	CoTi	MN (CoA)	CN
7		HoTi	HA	CN
8		BU	MN(CoA)	CN

Table 1 : State entry problems in MIPv6 traffic

3. Suggested improvements

3.1 Distributed Firewalls

Distributed firewalls are software installed on each peripheral. Such software processes permission/deny to all packets. The policy to control each firewall is distributed locally from central server and all traffic is sent to another server for logging and analyzing if needed. To achieve a specific policy language, the distribution language and entity authentication must be provided.

The distributed firewalls will coexist with current perimeter firewalls (current border firewalls that reside on the edge of the network). Perimeter firewalls provide first security checkpoint, their main purpose is to prevent attacks. The distributed firewall will be installed on each client computer to monitor and inspect other issue in traffic like IP Addresses and other factors.

3.1.1 Using distributed firewalls with Mobile IPv6

Distributed firewall architecture must be modified to satisfy the needs of Mobile IPv6. Policy should be divided into general and local policy. A user may roam to a location that forbids any MN users to use p2p application; however, he may want to use these applications when he gets back home. If a single policy is enforced on the user device, the p2p will be either enabled or disabled. Having a two-part policy will solve the problem. The first part controls how the user generally wants his connection to be (General policy), and the second policy control (overwrite) the user access based on that location policy (Local policy). If that location does not have any special restriction or policies, then the general policy on that computer applies. For more information refer to (Vives and Palet 2004).

Using distributed firewall in MIPv6 has the following advantages:

1. It is topology independent.
2. Monitoring and logging end-to-end IPSec is possible.
3. Host identification is done by cryptographically signed certificate instead of IP address
4. The traffic has to pass one firewall only: the destination host firewall (CN).

5. It solves the problem of hosting local and visited MN in the same location.
6. Devices that do not have enough power to apply the security policy can use other nodes as their security gateways (Vives and Palet 2004).
7. It protects the area that does not have security gateways.
8. It does not have single point of failure (Vives and Palet 2004).
9. It is software based; thus, it provide better update facilities for future needs.

3.1.2 Distributed firewall drawbacks

Distributing the security policy will enable the users to read it (since it is on their computer), a fact that may lead to identification of flaws, as all users can determine the boundaries of their privileges. Further, users may disable this security policy if they decide that, rather than protecting them, the policy appears to obstruct them from using certain network applications or performing specific operations. A proper method should be applied to ensure that users will not be able disable this module. In case that he disables it, no traffic may be exchanged with outside. This responsibility will be encompassed within the border firewall functionality.

In addition, the distributed approach has the following disadvantages:

1. It requires higher CPU usage processing at each node (it might be an issue on PDA and other peripherals that does not have powerful CPUs).
2. It is more complicated than border security method.
3. It requires more traffic on the network to exchange policies, logging and inspection.
4. Misusing the policy or vandalizing it may be possible (it may not be easy).

3.2 Quarantine Module

This module was introduced by (Kondo et al, 2004). It divides the network into zones (virtual segments), with each zone having its own security levels, and its own in/out passing blocking lists. There is one or more security enforcement point (a firewall or a dedicated server). The segmentation is done by layer 3 switch or by IP address.

When a node is attached to the network, tests are made to ensure its security level, antivirus version, OS patch, etc. Based on the result, it is placed in the trusted, untrusted or quarantine zone. Any node with unrecognized (unsupported) OS will be placed in the untrusted zone. If a node from untrusted zone whishes to move to trusted zone, it has to apply specific updates and patches.

3.2.1 Quarantine module for Mobile IPv6

Different access list are applied between zones. For example, we can enable trusted nodes to trusted nodes communications, as shown in Figure . Different zones can be adopted (not just trusted, quarantine and untrusted). A Company may wish to have several zone levels; each zone has its own pass/block policy. Kondo et al. (2004) stated that special paths and QoS could be assigned trusted zone communications. In Figure 2 a secure path is established to communicate between trusted zones, which bypass the firewall. This will eliminate the state entry problem in the firewall.

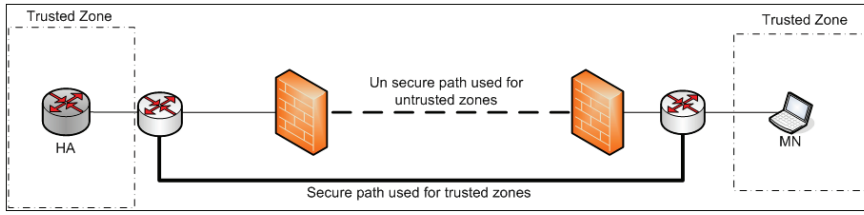


Figure 2 : Using quarantine module to bypass firewall

Figure 3 shows the deployment of Quarantine module in the Mobile IPv6 environment. In the figure: QS+PE is the quarantine server + policy enforcer, which is the server responsible for storing the security policy, OS patches, AV update and any other security requirements. It is also responsible for distributing and enforcing users to apply this policy.

Log and monitor server is responsible for monitoring all the activities in the network if it suspects of any malicious traffic from any internal node, it requires this node to pass the security policy again or it isolate it in the untrusted area. All the traffic from trusted zones bypasses the firewall by direct connection. However, Site A and site B must trust each other.

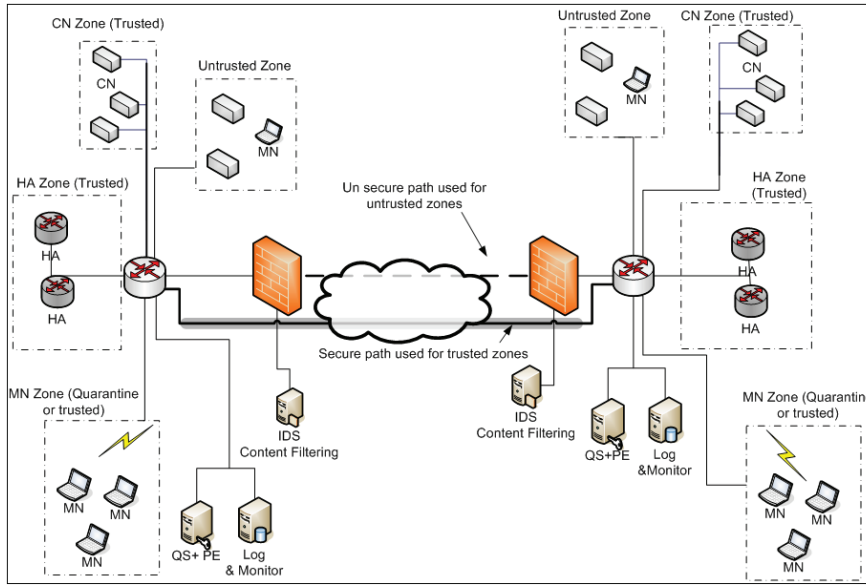


Figure 3 : Deploying quarantine module in MIPv6

When visiting a new location, the MN presents its current security certificate to the local QS; the QS (Quarantine server) will communicate with QS of the MN's home network (Home QS). It performs the security tests required in order to grant the MN current certificate. If it is

necessary, the QS will force the MN to issue a new certificate and bypass a different test (if the Home QS does not satisfy the security requirements of the Local QS or if the local QS was not able to contact Home QS)

It is essential to apply a strict policy on the trusted zone, because the traffic will bypass the firewall. In addition, strong mutual trust relation must be establish between sites in order to trust that trusted zone in site A may bypass firewall of site B and vice versa. The additional requirement is to encrypt the traffic passing through the secure path using strong encryption.

3.2.2 Quarantine Module drawback

The main drawback of the approach presented is that IPSec traffic will pass unmonitored. Another issue is to ensure that MN that is located in trusted zone is not a malicious node. Even if it passes all security tests, this does not guarantee that it will act as a legitimate user. In addition, this module protect the network itself from malicious users, however, it does not protect users from each other. Adopting this module does not guarantee that a Mobile users in the trusted zone is safe from other user or attacks in the same zone.

3.3 The complete solution

Combining quarantine module with the distributed firewall module and keeping the border firewall (refer to Figure 3) will create a complete solution.

The QS+PE server is the central entity for the proposed security solution. This server enforces distribute a personal firewall module to all the hosts in its network. Any host that did not operate this module (because he disables it or because his OS is not supported) will be either placed in the untrusted area or will be disconnected completely from the network. The distributed software contains in addition to the distributed firewall Security policy, Anti virus, content filtering, OS patching and checking

Once this module is distributed, any node will be challenged to pass security tests based on the company's policy, if it passes the entire tests, it will be placed in the trusted zone. If it failed to pass some tests, it will be placed in other zone (different levels of quarantine zone). If it disables this module, it will be placed in the untrusted zone. This procedure ensures that all hosts in the network are grouped in zones controlled by access list between them and with the border firewall.

Site B will follow the same steps as mentioned before. When site A and Site B initiate a connection, they first send an initiating session containing information about each site tests and policy. If both sites agree with the tests, the trusted zones in those sites will be able to communicate directly (bypassing the firewall).

This solution will introduce better security, as it ensures that all traffic even IPSec from trusted zone is monitored and logged. This will help if someone was able to bypass the security test and places himself in the trusted node, his traffic will still be monitored and logged in central location. Any user that disables the firewall module on his peripherals will automatically be placed in the untrusted area or disconnected from network(blocking his incoming/outgoing traffic).This is done by sending an update to all nodes in the network to

drop his traffic, and another update sent to the firewall is to block all his incoming/outgoing traffic.

4. Conclusions

This study proposed a security solution is only a framework it needs lots of research and deployment before it becomes an acceptable solution for Mobile IPv6. this will affect the cost and complexity of adopting it. Companies willing to tolerate IPSec traffic traversing their firewall, which accept the idea of untrusted users inside their network, should easily deploy Mobile IPv6. Such organisation would include ISPs, Mobile operators and other general access networks. However, mobile users are expected to provide their own security methods.

Other private companies that are critical toward their security will have to modify their infrastructure to handle the new problems issued by the Mobility users. This includes Isolating MN, HA and CN in a separate segments. In addition, Provide local MN with personal firewall to protect them from malicious visited MN and passing their incoming traffic. Additionally, apply strict access list to all segments make sure that there is no traffic leak. Finally passing IPSec traffic to pass from known hosts (local MN that are away) to their HA and enabling IPSec traffic to pass from and to MN (visited). We should not forget to block any traffic coming from untrusted (unknown) addresses to internal network

Other solutions like Quarantine module and distributed firewalls are still in their early stages, they need good amount of time before they become a reasonable solution. The method proposed (combining both methods) is only a framework and need deep inspection and deployment. However, it may solve of the security related challenges of Mobile IPv6.

5. References

Deering, S. and Hinden, R. (1998) *RFC 2460: Internet Protocol Version 6 (IPv6) Specifications*, IETF, http://www.ietf.org/rfc/rfc2460.txt

Kondo, S., Suzuki, S., and Inoue, A. (2004) *Draft: Quarantine Model Overview (draft-kondo-quarantine-overview-01)*, IETF, http://www.join.uni-muenster.de/Dokumente/drafts/draft-kondo-quarantine-overview-00.txt

Perkins, C., Johnson, D., and Arkko, J. (2004) *RFC 3775 Mobility Support in IPv6*, http://www.ietf.org/rfc/rfc3775.txt

Savola, P. (2003) *Draft: Firewall Consideration for IPv6 (draft-savola-v6ops-firewalling-02.txt)*, http://www.watersprings.org/pub/id/draft-savola-v6ops-firewalling-02.txt

Shima, K. (2003) *Draft: Route Optimization hint option (draft-shima-mip6-rohints-00)*, IETF, http://www.watersprings.org/pub/id/draft-shima-mip6-rohints-00.txt

Strebe, M. and Perkins, C. (2002) *Firewalls 24 Seven* SYBEX, California.

Vives, A. and Palet, J. (2004) *Draft: IPv6 Security Problem Statement (draft-vives-v6ops-ipv6-security-ps-00.txt)*, IETF, http://www.watersprings.org/pub/id/draft-vives-v6ops-ipv6-security-ps-00.txt

World Wide Web Content Study Based on Anonymised Network Traces

E. Salama, B.V. Ghita and S.M. Furnell

Network Research Group, University of Plymouth, UK
e-mail: info@network-research-group.org

Abstract

Current Internet packet traces, used to observe the characteristics of current network applications, must be anonymised when stored, due to legal reasons. This process reduces the application-level statistics that can be later performed on the traces collected. This study evaluates the amount of information that may be retrieved from packet traces that were anonymised, while retaining the HTTP header tags and proposes an anonymising method that supports current research of non-intrusive www characteristics without breaching user privacy. The second part of the study uses the technique proposed to provide detailed statistics about the characteristics of HTTP dialogues, as extracted from anonymised network traces. The results revealed possible sources of bias, such as large files for average object sizes, a relatively high of HTTP 1.0 servers, considering its limitations, and the majority of pages having an age of less than one year.

Keywords

HTTP, packet traces, privacy, web transfers

1. Introduction

Current Internet studies rely heavily upon offline analysis of packet traces, collected from aggregation points in the network and stored for later analysis. The packet collection process is firstly a demanding tasks in terms of the hardware and software involved. The process involves a typical network sniffer, such as TCPDump, which will collect the packets arriving on the network interface at the collection point. The collection and storage of the trace are closely related as, subject to the studied network, the packet trace file could grow by as much as several megabytes a minute. This is why the both the hardware involved and the media must be relatively high-spec, in order to support high data transfer speeds. Moreover, the studied network might have to be reconfigured in order to allow packet capture. For instance, a sniffer plugged into a network switch would capture packets sent from or to this interface only. Moreover, during the capture, the network card will have to be in promiscuous mode, which could impair the studied network. As a result, only a few packet traces repositories exist on the Internet, such as the WIDE project repository (Wide, 2005), the NLANR PMA (NLANR, 2005) or the Internet Traffic Archive (Danzig et al, 2005).

The second level of difficulty relating to packet traces is the technical and legal perspective. Network managers are rather reluctant to provide access to core points in their network, as any change, such as port mirroring, required to replicate the traffic between two interfaces of a switch, may affect the performance of the supervised environment. This reluctance will impact heavily on the availability of such collection points. Apart from these technical issues,

releasing and storing packet traces is, from a legal point of view, not an easy task either, because according the Data Protection Act (DPA, 1992), "No person [...] shall disclose any information which [...] relates to an identified or identifiable individual or business". Because they contain private data and network topology information about the studied network, traces collected by researchers can not be released, hence preventing these traces to be shared and compared for statistics purpose. Security is another issue that must be taken into account when releasing a trace. Indeed, with some knowledge, it may be possible, using network-related information present in the trace (such as IP addresses, sequence numbers, etc), to infer some characteristics of the network in which the capture has been made (such as running Operating Systems, network device brands, software versions...). These characteristics could then be used by an attacker and enable him to mount an attack against the studied network.

However, prior studies promoted network trace anonymising tools, used to sanitise the traces before storing or releasing them to the public. This process consists in removing or transforming private information (e.g. HTTP header, TCP options, original and remote IP address) contained in the trace, so as to not disclose it. This way, resulting traces may be released safely. The anonymisation process is hence a prerequisite before releasing a network trace. Several such network anonymisation programs, are currently available (Minshall, 2005), (Paxson, 2005). The typical behaviour of such tools is to randomise the IP address fields from the IP header and remove any information beyond the end of the TCP header.

While anonymisers are extremely effective, being able to remove any personal information that may link an individual with the collected trace, they also are limited in the sense that the process applied to the trace removes certain information that could be used for various types of analysis. Such information may include statistics regarding network performance, HTTP transactions outcome, HTTP implementation running on the remote server, content and object size for the downloaded web pages, and so on.

The aim of this study is to propose a framework which will allow some of the HTTP headers to be retained while maintaining user privacy. A proof-of-concept implementation will then be used to perform an analysis of the actual network traces. The goal of this study is also to propose to the research community a set of statistics from such anonymised traces that will provide further insight into the Internet performance and WWW structure.

After this introduction, the paper continues with an overview of the requirements of the anonymisation process in section two. Section 3 will then identify the sensitivity level for each of the fields of a typical HTTP header and discuss the method proposed in this study to anonymise the associated sensitive information; in the end, subsection 3.6 provides a brief overview of a proof-of-concept implementation. The software was applied on a traffic trace collected from the University of Plymouth backbone, as indicated in section 4; the analysis of the results is presented in section 5. Section 6 concludes the paper, highlighting the achievements of the study.

2. Anonymisation process requirements

As presented in section 1, prior research provides several alternative methods to anonymise a packet trace, depending on the amount of information to be anonymised. This study is based on the techniques offering the best balance between privacy, security, and efficiency.

The resulting traces must be anonymous so that no information about the studied network will be disclosed. This impacts on another main goal of the technique, which is to allow these traces to be released in the public domain, therefore any reference to a specific network should be removed. Such references would facilitate a breach of privacy of the involved users and also may lead to a network attack against the studied endpoints. The anonymisation process has to be efficient and fast enough to be able to process large amounts of data in real time.

3. Identifying sensitive information

After proposing the above guidelines, the study continued by identifying the sensitive information in the packet fields. The following subsections analyse each header included in packets. The description begins with the libpcap packet header, due to its usage as a packet capture library by tcpdump, and then follows the TCP/IP layers for typical WWW traffic. The corresponding studied headers are Ethernet, as the main type of connectivity for current LAN environments, IPv4, TCP, and HTTP. This information is used as a basis to determine which fields must be removed from the anonymised packet.

3.1 The libpcap packet header

The packet header does not include the actual content of the packets, but some information about each captured packet. The only private information in it is the timestamp that indicates when the packet was captured, which may reveal certain information regarding the timing of the capture. To ensure anonymity, timing information was considered for anonymisation.

Radical methods, such as replacing the timestamp with zeroes or randomising the values were excluded from the outset, because while such a modification would still allow overall statistics to be calculated, it would leave no information at all for performance evaluation. The preferred alternative, based upon anonymising all timestamps in a trace relatively to a reference timestamp, was chosen due to the fact that it provides the best balance between information and privacy. For simplicity, the timestamp recorded for the first packet in the trace was chosen as the reference timestamp; this timestamp is subsequently subtracted from all following packet timestamps. As a result, the first timestamp of the anonymised trace will be 00:00:00.000000 (the last 6 digits provide microsecond accuracy for the timestamp). The method maintains relative time differences between any two packets in the trace without disclosing any private information.

3.2 The Ethernet header

In Ethernet frames, the source and destination MAC addresses contain private information, as potentially they can be used in support when identifying one particular network device on the Internet. The considered method was to anonymise both address fields by replacing them with

zeros. This does not affect the resulting statistics, as the analysis will typically focus on end-to-end results rather than local network performance.

3.3 The IP header

Prior studies promoted two alternatives for this task: sequential anonymisation and random anonymisation. Both techniques have a similar aim: to provide a table of correspondence between the anonymised and the non-anonymised addresses; this is required so that request/response dialogues can be identified and analysed in the anonymised trace. The difference between the two comes from the way these two tables are constructed. Sequential anonymisation replaces the first IP found in the trace with the IP "10.0.0.1", the second one with "10.0.0.2", and so on. Random anonymisation uses an algorithm to calculate a corresponding value from the real value of each encountered IP address in the trace.

While it can be argued that randomising will provide further information to the analysis, as the resulting values are correlated with the real values, the scope of the statistics to be provided does not include localisation, fact that dismisses the advantages of this method. From the security perspective, randomising also represents a potential threat, as an attacker could be able to derive some information from these sequences of random numbers. Finally, randomisation requires further processing, as it requires an encryption or hashing algorithm (either reversible or not) to be applied to the IP addresses. Due to this combination of factors, sequential anonymisation was the preferred method for this study.

3.4 The TCP header

The TCP header contains several fields revealing or leading to private information: checksum, source and destination ports, sequence numbers, and a TCP timestamp (Jacobson et al, 1992). In order to eliminate any possible security threats, it was decided to anonymise all these fields. As a result, source and destination ports are anonymised sequentially, using the same technique as in the case of IP addresses. The checksum is zeroed, based on the assumptions made in the IP header. Finally, the timestamp and the sequence numbers are anonymised relatively, as seen with the Ethernet header.

3.5 The HTTP header

Apart from the headers, the most privacy-sensitive part of the packet is the actual payload. However, while prior tools chose to treat all data beyond the TCP header as payload, the proposed method parses the HTTP header as well in order to identify and retain relevant non-sensitive information.

HTTP anonymisation requires a different approach in comparison with the headers presented before. Ethernet, IP, and TCP headers are compulsory, have a very strict (binary) format and predefined size for each field, in order to reduce processing overheads. In comparison, HTTP headers are flexible, as they use text encoding and encompass a variable number of fields, named tags in the protocol terminology, all determined by endpoint factors, such as the server implementation. One of the reasons behind this flexibility is the purpose of the HTTP tags – while some of them are required as part of the HTTP functionality, several such tags, such as the Server Version tag, have only informative purpose.

From the outset it was decided that, for HTTP packets, the HTTP tags can be kept, but only after being anonymised; also, the payload (HTML code, images, and so on) must be removed from the packet. In order to reduce processing, text parsing was the chosen method for HTTP headers identification. To speed up the process, packets containing HTTP headers from GET request-response transactions are detected by looking only at the first four bytes in the packet. If the packet starts with "GET" or "HTTP", then the packet is further parsed for HTTP packets. This does reduce the functionality of the anonymising tool, as other methods such as POST are ignored; further methods will be added in future versions, but, for the purpose of this study, it was assumed that typical web traffic consists of GET requests

HTTP headers anonymisation requires much more processing than the others parts, because they are more complex, and each of them has to be anonymised following a specific method. They may be classified in the following categories: URLs (Host, Location, Referer…), IP addresses (Client-IP), Dates (Date, Last-Modified, Expires), Other (Cookies, Etags…)

3.5.1 URLs

URLs, as specified in the defining documents (Berners-Lee et al, 1998), are character strings which use the following format: protocol://web_address:port/path/file.extension?parameters . For statistics purposes, it is important to anonymise each unique URL the same way. For instance, if there are two requests for the URL http://www.google.com/ads/index.html, in both instances (and any following ones), the URL should be anonymised to the same result, e.g. http://xhjsauiopd.com/iii/kdzpw.html. With this method, it is still possible to count the number of times each URL was accessed, and the number of accessed web pages on each server.

The protocol, file extension, and the server suffix parts are kept, as they are too generic to reveal any private information and, more important, they are very useful in statistics. The whole host name is anonymised with random characters, including the dots composing it, because leaving the dots in clear would reveal too much information (for instance, in http://rty.mklkm.com, it can be assumed that 'rty' stands for 'www'). In order to reduce the overheads, the path is replaced with 'i's, while the filename is anonymised with random characters (similar to the host name process). Finally, additional parameters like '?param1=10' are replaced with '?zzzzzz=zz'. In order to maintain some of the information, the following characters are kept unchanged: & ? ; , % + = ! $ @ " ' .

3.5.2. Timing information

The timing information, as appearing in the Date: or Last-Modified: HTTP tags must be anonymised in order to eliminate any absolute temporal reference to a specific page. The process uses the first date as a reference which, when encountered in the parsed headers, is stored for later use. Every following date encountered is then anonymised relatively to the first stored date, which now corresponds to Tue, 01 Jan 1980 00:00:00 GMT. The concept is similar to the one used for anonymising timestamps.

For example, let us assume that a trace contains two requests: the first one asks for a page created on the 4th June 2004 at 14:21, and the second one for a page created on the 18th

November 2003 at 02:31. The first date will be anonymised to 1st January 1980, and the second to 16th June 1979, keeping the time period constant between the two requests.

3.5.3. Other headers

Every other header containing private information, such as Cookies, Entity Tags, Authentication Strings, etc is replaced with 'X's. Some headers are excluded from being anonymised on purpose, such as User-Agent and Server, which reveal the versions of web browsers and servers, and possibly Operating System versions, such as "User-Agent: Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1)". It could be a problem for some people or companies to release such information, but this choice was made because this information is used for the statistics and reflects the current state of the World Wide Web. Moreover, this information is not sufficient by itself to present a high risk.

3.6 Implementation

The observations presented in section three were integrated in a proof-of-concept software tool, programmed under Linux using the libpcap packet capturing library. The implementation had two main objectives: to respond to all anonymity issues identified in the previous section, and to introduce small overheads, so that the program would not lead to high numbers of dropped packets. The software is available on-line (Salama, 2005).

4. Data collection

The program was used to perform anonymisation and statistics on a set of traces collected from the University of Plymouth backbone network in February 2005.

The trace included approximately 1.5 million HTTP request-response transactions. Due to the large volume of traffic, the trace was collected in less than a day. The procedure involved data capture and anonymisation using the implementation described in section 3.6, followed by statistical analysis of the resulting anonymised traces, using the built-in scripts from the implementation.

The following section presents the resulting statistics, extracted from the anonymised trace. It has to be stressed that these results are obtained from a trace that may be stored without breaching the privacy of the users involved and cannot be extracted once available anonymisers, such as tcpdpriv, are used to capture and process a packet trace.

5. Statistics

The subheadings within this section present the statistics obtained from the annonymised traces by analyzing information from each of the encountered HTTP tags. It is important to highlight that not all transactions yielded data for all listed tags, either due to the web server not sending the specific tags to the client or due to other issues, such as malformed/non-standard tags.

5.1 Web server vendors distribution

The *Server:* field of the HTTP header indicates the implementation details of the web server that replied with the requested resource. The syntax of the field is *Server_implementation*/*Implementation_Version*. Even after filtering out the *Implementation_version* part of the header field, the set of traces still produced 246 different server implementations. Among these, the outstanding distributions are, as expected, Apache with 35.3% and Microsoft, with 46%. These figures do suggest a strong bias in the study, especially when compared with similar holistic studies, such as the "Web Server Survey" run by Netcraft (2005). The Netcraft survey, after sampling 60 million websites, concluded that Apache, powering 68.8% of the web servers, leads the web server market over Microsoft, which has a market share of only 20.8%. One of the possible causes of this bias may be the University of Plymouth website, which may have been accessed heavily during the survey, and was running Microsoft IIS at the time.

5.2 Types of requested resources

As expected, the analysis of the MIME type of the resources indicated that most requested resources on the web are text-based and image-based objects. Out of the 828,076 retrieved objects, 43.7% were identified as "image/" in the *Content-Type:* header, while 45.1% were text-based objects, based on the associated "text/" tab. Further analysis revealed that the text-based objects were virtually all html - 80% of all "text/" tags were "text/html"; the images were equally split between jpeg and gif images – 49.8% and, respectively, 48.7% of the "image/" tags.

Another significant proportion was recorded by the "application/" category – 8.9% of the resources, with "application/x-javascript" counting for almost half (4.85% of the total) of them. It is interesting to note that, in spite of the multimedia hype of the web, based on the tags carried, only 0.16% of the resources returned by the web servers were audio objects and 0.48% were video objects.

As abnormal behaviour, it is interesting to note the existence of two misspelled versions of the "application" tag, one of them named "aplication" and the other "applation". Due to the size of the actual traces, it was not possible to save the raw packet traces for further analysis, but a critical observation was made that such malformed strings, due possibly to customised versions of web servers, may lead to disclosure of the web server. It is therefore considered as a future improvement of the implementation to remove or modify such malformed strings from the resulting anonymised trace.

5.3 Content freshness

The anonymised trace does not include any absolute timing information, as both *Date:* and in the *Last-Modified:* header fields are anonymised. However, the difference of the two values may be calculated, as both fields are anonymised using relative to the same value - the first timestamp encountered in the trace. The comparison of the two values produced more than 400000 samples. A distribution of the encountered values is presented below in Figure 1.

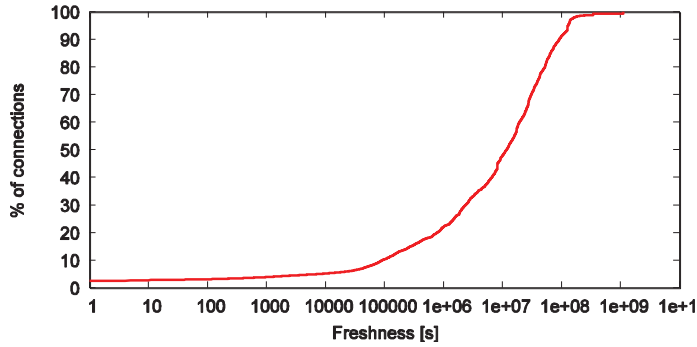


Figure 1 : Page freshness cumulative distribution

Based on the continuous distribution obtained, the figures indicate that 9.62% of the content is less than 1 day old, 30.7% is less than 1 month old, and 70.3% of the content is less than 1 year old. It is relevant to observe the possible inaccurate information due again to malformed dates: while 99.2% of the content was reported to be less than 10 years old, the top values indicated ages of 30+ years, figures which were likely to have been produced by incorrect dates.

5.4 Top level domains analysis

An analysis of the top level domains was made as part of this study, using the non-anonymised information from the packet trace. Such analysis does not implicitly provide information about where the web server is physically located, but can help to indicate the type of business or the service that the web server provides. While this information is far too generic to endanger the identity of individuals, it is one of the resources likely to be at least biased by the geographical location and interest focus of the monitored network.

Most of the captured traffic was focused around the .com domain, with 51.28%, followed by the .uk, .net, and .org domains with 31.45%, 7.8%, and, respectively, 2.3% of the total number of sites.

5.5 HTTP response codes analysis

This is another area where analysis of the anonymised traces may still reveal the performance of current web downloads. Within the analysed set of traces it was observed that most HTTP transactions were successful, with 80.9% returning a “200 OK” code. However, a significant proportion of the retrieved files produced “304 Not modified” (11.2%) and “302 Found” (2.8%) responses, while the “404 Page not found” and “500 Server error” errors accounted for 1.6% and, respectively 0.9% of the requests.

5.6 Requested objects size

The average value of requested object is 2.24MB, but it is accompanied by a significant standard deviation, 3.29e+07, due particularly to several large objects being downloaded. The cumulative distribution of values is presented below in Figure 2(a).

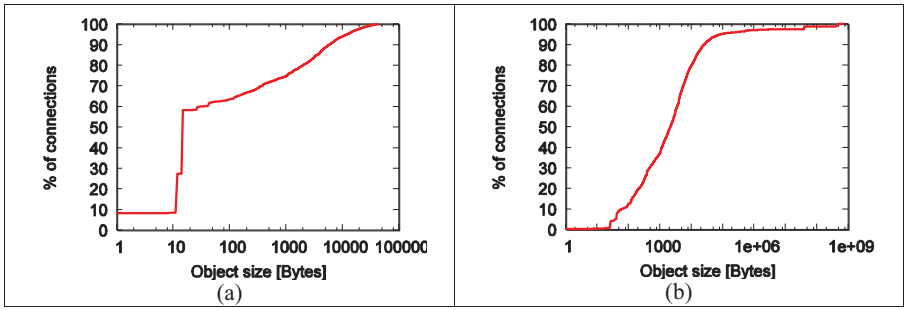


Figure 2 : Cumulative distribution of the requested object size (a) before and (b) after filtering the 12-byte and 15-byte objects

As can be observed from its shape, the distribution is significantly biased by two values. The two corresponding object sizes are 12 bytes, accounting for 19.1% of the total, and 15 bytes, accounting for 30.8% of objects. This is why the data was filtered to remove these values, as well as the zero size samples (9.8% of the total figure) and then the distribution was redrawn. The result, presented in Figure 2(b) provides a clearer image, with 40.2% of the objects having a size of less than 1200 bytes, therefore fitting in a full Ethernet frame, including a margin of 300 bytes for the IP, TCP, and HTTP headers.

5.7 HTTP version

Studies such as (Padmanabhan and Mogul, 1995) showed that multiple transfers of small files can degrade the performance of HTTP protocol due to the additional burden introduced by the three-way handshake TCP connection establishment and closing. To reduce the TCP impact, persistent connections were implemented, starting from HTTP 1.1 (Fielding et al, 1999), enabling multiple objects to be transferred within the same connection. In spite of this clear benefit, analysis of the traces indicated a rather balanced distribution, with a 47.1% / 53.9% split of the HTTP protocol versions 1.0 / 1.1. While the number of HTTP v1.1 servers is higher than the number of HTTP v1.0 servers, the ratio of the two alternatives is still far from a preferred clear supremacy of HTTP v1.1.

5.8 HTTP latency

From a performance perspective, the average response time between HTTP requests and responses can be analysed. This time includes the travel of the request to the server, the server processing time, and the travel back of the corresponding response. Due to nature and the functioning of TCP/IP, the paths followed by the request and the response may be different. Moreover, the delay expressed here is the time difference between the request and the first

packet of the response. As it may be observed in Figure 3, virtually all responses were ranging in the 1ms-1s interval, with 56.1% of requests experiencing delays in receiving a server reply after less than 100 ms.

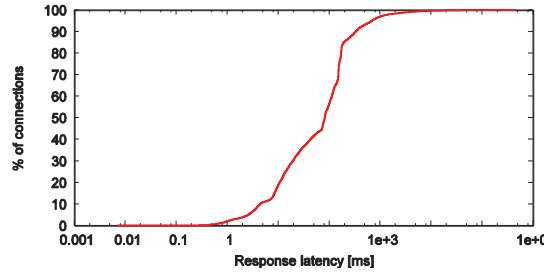


Figure 3 : Cumulative distribution of the response latency for HTTP requests

From a server activity perspective, the initial response is amongst the more demanding phases of the HTTP dialogue, therefore it is very likely that the actual data transfers led to higher results, fact that indicates that the connectivity between the studied network and the visited web sites induced low values for latency. It was considered beyond the scope of this study to establish a correlation between the response delays and the top level domains of the visited sites, but this relationship will be investigated as future work.

6. Conclusion and outcomes of the study

This study proves that HTTP statistical analysis of anonymised network traces is possible. A flexible network tool was developed, which has the ability to capture, anonymise and analyse packet traces. This tool is primarily targeted at HTTP packets, but could be easily used and extended to support other protocols.

The developed program was used to retrieve a large number of statistics from an anonymised packet trace, statistics that may be used to describe the state of the Web at the time of the capture, beyond the information provided by the TCP/IP headers. Statistics show precise figures about HTTP characteristics and performance, including web server distributions, most requested resources, content freshness and average size, and top-level domains distribution. HTTP performance was evaluated using server status codes and client-server dialogue latency.

The developed program will hopefully enable many researchers to release the traces they collected, and to make precise statistics from them. The released traces will be full anonymised traces, and will not be only a textual summary of a stripped trace like the ones processed with *Scripts for Sanitizing TCPDUMP Trace Files* (Paxson, 2005).

The presented study also presents a snapshot of current web transfers characteristics, as observed through the traffic generated by an end-network. The HTTP transactions have indicated that, for most connections, the 1.1 version of the HTTP protocol is used to transfer

data objects which are typically small. With regards to the timelines of the web, most transferred objects appear to be fairly recent, with approximately 30% of the web content requested being less than one month old.

7. References

Berners-Lee, T., Fielding, R., Masinter, L. (1998) “Uniform Resource Identifiers (URI): Generic Syntax”, RFC 2396, http://www.ietf.org/rfc/rfc2936.txt

Danzig P., Mogul J., Paxson V., Schwartz M. (2005) “The Internet Traffic Archive”, http://ita.ee.lbl.gov/

DPA. (1992) “*Data Protection Act*”, HMSO, http://www.hmso.gov.uk/acts/acts1998/19980029.htm

Fielding, R., Irvine, U.C., Gettys, J., Mogul, J. (1999) “Hypertext Transfer Protocol -- HTTP/1.1”, RFC 2616, http://www.ietf.org/rfc/rfc2616.txt

Jacobson, V., Braden, R., Borman, D. (1992) “TCP Extensions for High Performance”, Request For Comments 1323

Minshall, G. (2005) “tcpdpriv – Program for eliminating Confidential Information from Traces”, http://ita.ee.lbl.gov/html/contrib/tcpdpriv.html

Netcraft. (2005) “Netcraft Survey”, http://news.netcraft.com/archives/web_server_survey.html

NLANR. (2005) “Passive Measurement and Analysis Project”, within Measurement and Network Analysis Group, http://pma.nlanr.net/

Padmanabhan, V.N., Mogul, J. (1995) “Improving HTTP Latency”, "Improving HTTP latency", *Computer Networks and ISDN Systems*, vol. 28, pp. 25—35

Paxson, V. (2005) “Scripts for Sanitizing TCPDUMP Trace Files”, http://ita.ee.lbl.gov/html/contrib/sanitize.html

Salama, E. (2005) “Bernardo home page”, http://bernardo.sourceforge.net

WIDE Project. (2005) “Packet traces from WIDE backbone”, http://tracer.csl.sony.co.jp/mawi/

Survey of Wireless Access Point Security in Plymouth

M. Voisin, B.V. Ghita and S. M. Furnell

Network Research Group, University of Plymouth, UK
e-mail: info@network-research-group.org

Abstract

The development of wireless networks has brought a lot of security issues inherent to the communication medium and to the different security approaches available. While WEP encryption is not anymore considered secure, it remains one of the easiest to use and very widespread wireless security mechanisms, provided with most 802.11 wireless equipments. In order to assess wireless access points security (based on WEP encryption), a survey with a handheld device and the appropriate "wardriving" software has covered a part of the city of Plymouth. While the results were insufficient to link WLAN security with demographic information, they allow to draw a global picture of the spread of wireless technology mostly concentrated in the city centre and confirm the fact that many AP are used "out of the box" with default settings which are, or were until recently, not using any encryption.

Keywords

Wireless networks, Security survey

1. Introduction

With the democratisation of wireless equipments, numerous wireless local area networks (WLAN) have been set up, during the last years, both for private and commercial use. While security issues on these networks have been, and are still, denounced regularly by the Medias (Farrow, 2001), activities like wardriving have spread and become a kind of game for hacker's apprentices.

The study presented in this paper focuses on different aspects of these wireless networks. The discussion is focused on two areas. While the first part aims to analyze the spread of these "rather new" tools (e.g. wireless cards and access points), the second part is dedicated to the assessment of security that reflects the common user's knowledge.

The conclusions of a survey led in the city of Plymouth (United Kingdom) during spring 2004, have been used to assess those criterions. After presenting different security mechanisms usually found in wireless networks, this paper details a survey (with hardware and software used) investigating on WLANs security, more precisely, on the use of encryption within wireless networks, and its results. A final section is then dedicated to a discussion of these results.

2. Review of wireless security

In the context of the survey of wireless access points (AP) security, it is essential to distinguish basics of existing security protections. This section aims to give a first idea of

some of the most used mechanisms, the level of security they provide but also their limitations, in order to place the following work in the context of wireless security.

2.1 Frame level mechanisms

To address security issues, two basic mechanisms act in the low level frames. The first one consists of muting APs so that they do not broadcast the Service Set Identifier, needed for any client to establish a connection to the network. Indeed, default settings are usually set to broadcast it regularly in plaintext to facilitate clients' connections. While default SSID can usually facilitate finding the administrator's password on the Internet for a device used "out of the box", setting this identifier to the name of the company it belongs indicates to potential hackers what categories of resources may be connected to this network However, such measure implies the use of another process to supply the SSID to legitimate clients so that they can connect the network. The second mechanism relies on a filter based on a list of Medium Access Control (MAC) addresses of authorized devices. However, the level of security it brings is weak as MAC addresses are transmitted in plaintext in each 802.11 frame (whether the connection is encrypted or not) and can then be "spoofed" by almost any wireless card, using appropriate software.

2.2 Encryption

A higher level security mechanism can be reached by the use of encryption algorithms. Wire Equivalent Privacy (WEP) encryption is probably the most famous but also the most contradicted security standard for wireless networks. While it was the first wireless specific security method, its implementation suffers from different flaws, some of which (the most known may be the weakness in the key scheduling of Rivest Cipher 4 (RC4) encryption algorithm) have been used to provide WEP cracker software, now widely available on the Internet. According to security experts, WEP protocol suffers from different flaws (Khan and Khwaja, 2003) among which: lack of authentication key limited lifetime, vulnerability to "disassociation requests" injections, low security MAC level authentication and identification, lack of central security management and weakness of the cipher algorithm WEP is based on due to the Initialisation Vector (IV) generation method used.

With the breaches discovered in WEP based wireless security, the IEEE 802.11i workgroup dedicated to security was created in order to publish a standard on Robust Security Network (RSN) (IEEE, 2004). This method relies on Temporal Key Integrity Protocol (TKIP) and on the separation of user authentication and the message protection (preventing the possible decoding of data thanks to the observation of authentication processes). WiFi Protected Access (WPA) (WiFi Alliance, 2002) has been implemented on the Wireless Fidelity (WiFi) manufacturers' initiative to release secure replacement for WEP as fast as possible, without the need of major hardware changes. This method relies on TKIP and also includes mechanisms such as a Message Integrity Check (MIC) and extended Initialisation Vector (IV) with sequencing rules and re-keying mechanisms that address the previous breaches included in WEP implementation. However, WPA being based on a Pre-Shared Key (PSK), usually generated from a passphrase, it has recently been proven to be prone to different kinds of attacks (Moskowitz, 2003). WAP2, an evolution of WAP based on the Advanced Encryption Standard (AES) and on a new MIC implementation should replace the first version in a near future.

2.6 Authentication Server

The 802.1x standard (IEEE, 2004), based on interactions between a supplicant, an authenticator and an authentication server, can bring another security level to wireless networks. The service requested by the supplicant to the authenticator will be granted after the verification of its authorized services to the authentication server (usually RADIUS). In spite of the strength of this method, allowing regular and automated key changing within a connection, it presents weaknesses and has already been cracked because of "similar design flaws within 802.1x, EAP and 802.11… lack of message authenticity and lack of state machine synchronization" (Mishra and Arbaugh, 2002).

3. The survey

3.1 Tools

In order to lead a survey of wireless security, both software and hardware tools are needed. Most of the so called "wardriving" software have common features like channel hopping and detection of the network parameters: SSID, BSSID (MAC address format identifier of the network), received signal strength and noise, maximum data rate, channel used, type of network and whether encryption is enabled or not. The major difference between all is the platform they run on and the wireless cards supported. Other special "wardriving" software exist and take advantages of special hardware drivers (existing on Linux and some BSD OSes) to passively monitor WLAN. Unlike promiscuous mode, which needs the wireless device to establish a link with the AP for sniffing "every" frame, monitor mode (RFMon mode) enables a device to monitor any wireless packet (even frames with bad CRC) without emitting any signal. These software, such as Kismet (Kershaw, 2004) and AirSnort (Shmoo, 2004) are probably the most famous ones) allow then to find the cloaked AP when some legitimate clients are connected. Incidentally, they can give access to a WEP protected network, by computing WEP keys using information from those grabbed frames. The last step, in order to assess accurately APs security used, would consist in establishing a connection to the network, using the above tools to defeat cloaked APs, MAC filters and to compute WEP keys. At this stage, an impossible connection would reveal the use of further security mechanisms such as 802.1x, giving a full assessment of the AP security.

As the use of appropriate software is crucial for the accuracy of results, the choice of hardware devices, each of which having its own inherent advantages and limitations, is also an important factor in such survey. Different hardware solutions are available:

- A laptop computer equipped with a wireless peripheral (PCMCIA wireless cards, USB WiFi dongles or built-in wireless equipment like Centrino processors laptops).
- A Personal Data Assistant (PDA) with built-in wireless or with an expansion slot in which a Compact Flash wireless card can then be plugged in.
- WiFi detector devices like Kensington WiFi Finder or Smart ID WiFi Detector which are compact devices that indicate the strength of any received wireless signal, or the newer WiFisense wearable WiFi detector that can identify the signal strength but also whether any encryption method (WEP or WPA) is used.

- Another solution consists in developing a new device for this survey, through the use of WLAN "stand-alone" chips newly appeared on the market (Dallas 2004).

As for locating the AP, a Global Positioning System (GPS) device seems the best solution. These devices usually use a serial communication protocol and exist as stand alone, integrated in Compact Flash cards or use BlueTooth communication protocol.

While the development of a new device is time consuming and the new WiFi detectors seem to lack accuracy, the PDA approach has been chosen for its discretion. The equipment used in this survey was a PDA Toshiba e740 with integrated WiFi (Prism2 based card) and a Pretec Compact Flash GPS device. The software chosen was Kismet (Kershaw, 2004) for its passive monitoring mode allowing to detect cloaked APs and its GPS compatibility.

Due to legal issues, any survey can not list precisely the protection used in a particular network (implying illegal data eavesdropping to compute WEP key and illegitimate connection to the network so as to test the presence of authentication request). Therefore, the conclusions based out of investigations only take into account measures that could be observed without any connection to the network (e.g. encryption).

3.2 Preliminary work

After testing the different devices supplied in order to determine their limitations (range, battery lifetime, sensitivity to interferences) some effort have been made to run Linux on the e740 PDA to use passive monitoring software. As they do not lead to good results, the revision of advantages and disadvantages of each solution led to the choice of Microsoft Pocket PC and MiniStumbler (Milner, 2003) software, reducing then the security characteristic of each AP to the use of any encryption.

3.3 Data collection

The survey in the city of Plymouth was led during spring and summer 2004. The first areas covered were wards where the student population was the most present, according to 2001 census data (National Statistics, 2001), in order to observe student wireless activity before most of them leave for holidays. Then the city centre and neighbour wards were covered, walking through streets to gather wireless data and locate AP.

Along the survey, various problems have been encountered and all of them could not be solved. The poor reception of the GPS antenna, even worse while used during cloudy weather, for instance, has led to unusable data. The study of GPS theory helped in identifying the best conditions but could not solve totally this problem. Furthermore, the low lifetime of the PDA battery and few bugs that the software suffers from were source of troubles and limited the number of APs discovered.

4. Results

In order to work with all collected data, a database approach has been adopted, with creation of scripts to adapt data format, to feed the different tables and finally to process the maps and

statistics on which the conclusions of the survey are based. The results were then compared with the findings of a similar survey, led during winter 2003 (Wilks and Ghita, 2004).

4.1 Evolution of wireless cartography

Due to the difference of areas covered by the two surveys, the difference in the number of APs found (see Table 1) does not reflect the evolution of wireless spread between the two periods, but the proportions of protected and "unprotected" (in the WEP encryption sense) networks lead to the same conclusion: more than 60% of these APs are unsecured.

	Winter 2003	Summer 2004
Wireless AP found:	96	228
Protected networks:	34 (35.42%)	86 (37.72%)
Unprotected networks:	62 (64.58%)	142 (62.28%)

Table 1 : Comparison of two surveys results

As for the evolution of WLANs in the areas covered by both surveys, all of the previous APs have been found with the same configuration parameters but a third more wireless networks appeared between winter 2003 and summer 2004. Among these "new AP", about 33% are protected, getting close to the global results from Table 1.

4.2 Spread of wireless technology

If the first goal of the survey is to profile wireless security, a study of the use of wireless technology is necessary to give sense to further results. The graph represented on Figure 1 presents the repartition of the detected APs in function of their distance from the city centre. If half the detected APs were no more than 500 meters away from the centre, the graph reveals that this number decreases linearly up to 2km and seems to be really low further.

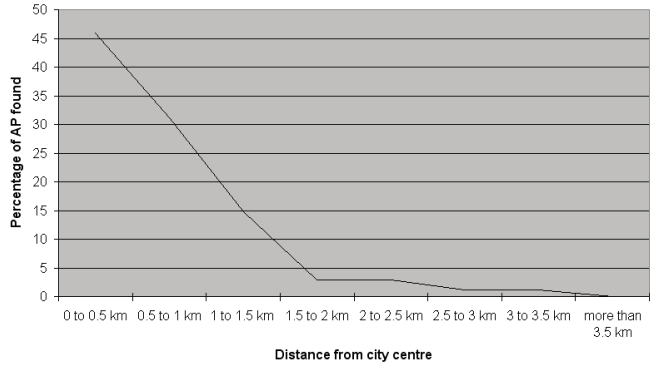


Figure 1 : Repartition of WLAN in Plymouth

4.3 Profiling wireless security

In order to correlate the results from this survey with demographic data from 2001 census, a repartition of the APs by ward has been processed. While any correlation is hard to observe, most of the covered wards seem to obey to the global percentages given in Table 1 (about 1/3 of WLAN use WEP based encryption). If the number of APs is not proportional to the population (103 APs found in Drake whose population is 8,831 compared to 18 APs found in Stoke with a population of 12,146), the Figure 2 reveals a strong correlation between the number of APs found per ward and the percentage of students in the population, except an unexplained exception for a ward.

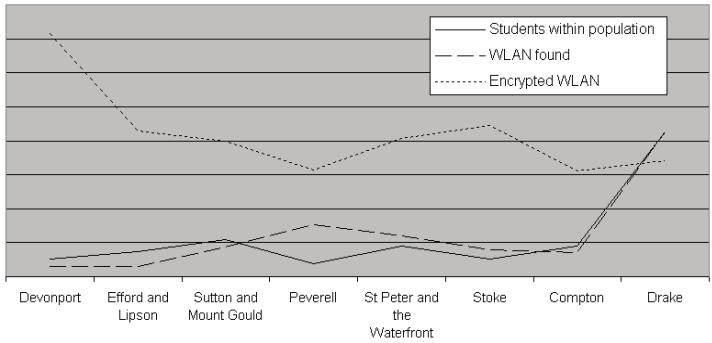


Figure 2 : Number of WLANs in function of student population

The representation of the percentage of encrypted WLANs by ward indicates that the proportion of secure networks does not follow the same law: while there appears to be more encrypted networks in wards with few students and globally few APs, this result can not be generalised to every ward.

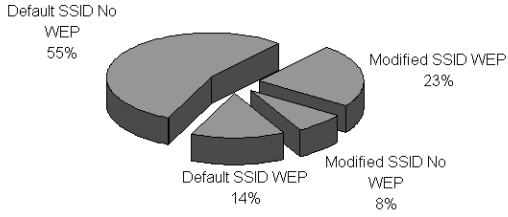


Figure 3 : Encryption and SSID

An approximation of the number of APs used with default settings is possible through the observation of SSID. Figure 3 shows clearly the predominance of these networks, with 79% of the APs found. This graph gives also an idea about the security policies adopted by wireless equipments vendors: only 14% of the networks found seem to have WEP as default security setting. These numbers are approximations due to the non exhaustive list of default SSID used.

This survey also confirms that wireless equipment can still be qualified as emergent technology for commercial use as the number of networks with multiple APs is limited (only 7 have been found). Therefore only few networks allow their users to roam the organization, keeping connected at least to one AP. Among the data collected during the survey, only one big network has been found using more than ten APs: the University of Plymouth network, which spreads over about a kilometre, relayed by wireless APs in most of the university buildings.

5. Discussion

Profiling wireless security with demographic data did not lead to consistent results. One of the causes of this inconsistency may be the fact that this survey is based only on WEP (and WPA) protection hence does not take into account any other security method mentioned in Section 2.

If no mathematical model was found to describe the repartition of protected and unprotected AP, other parameters collected during the survey may help such a modelling: among the different information observed, some parameters seem to be more likely found in unprotected WLAN, leading to the conclusion of the conditional probability of protection which illustrated by the difference between the two last columns of Table 2 (with $\mathbf{P_A}$: the probability of parameter **A**, $\mathbf{P_{WEP}}$ the probability of the network being protected and $\mathbf{P_{[A \cap WEP]}}$ the probability of a network having both the parameter **A** and the WEP protection).

A	$P_A * P_{WEP}$	$P_{[A \cap WEP]}$
Data rate: 54Mbps	0.0366	0.0284
Data rate: 22 Mbps	0.0427	0.0397
Vendor: Askey	0.0305	0.0114
Vendor: Belkin	0.0284	0.0056
Default SSID	0.2095	0.1079

Table 2 : Examples of unconditional and conditional probabilities

From these observations, the use of a Bayes probability approach, like Bayesian networks, may help to identify unprotected networks. Using an initialisation dataset and feeding it with the data collected for a particular AP, such network would then be able to compute the global probability of this AP using encryption. This probability is based on the observation of the number of protected and unprotected AP in which each parameter appears; each new probability computed brings a new sample and refining then the evolution of the network. The advantage of such “learning” Bayesian network based algorithm is the automated process of evolution for the different parameters which are subject to change: the more APs with a particular SSID and the same settings will be found, the higher will be the probability for any AP with the same SSID to have also the same encryption setting; allowing more accurate conclusions as they do not rely anymore on a static, non-exhaustive, list of default SSIDs.

However, due to a matter of time, no implementation could be produced to confirm or invalidate these results.

6. Conclusion

The survey confirms that while wireless networks seem to be used more for private than commercial use, they spread widely across the city but remain concentrated mostly around the city centre. This geographical repartition is likely influenced by the interest for the student population for all the "mobile" technologies. In spite of the Medias denunciating the lack of security on most of these networks and the efforts from manufacturers to facilitate the use of basic security methods, a large number of WLAN does not seem to be protected by any visible security mechanism, offering hackers an easy way to access personal data or even to use the network's resources. A relevant point revealed by the survey is the evident lack of awareness and interest of WLAN's users about security. In order to solve this issue, identifying the knowledge and needs of inexperienced users is necessary, implying a census of existing wireless networks and their security. However, this approach proves to be tedious, hence the interest of finding an empirical classification. On the other hand one should notice that such tool is double-edged as it would also allow hackers to identify easily their potential preys.

7. References

Dallas Semiconductors. (2004) "802.11b WLAN transceiver shrinks circuit board and bill of materials", *Maxim Engineering Journal* Vol 50, pp12-15.

Farrow, R. (2001). *Wireless Security: A Contradiction in terms?*, [Online]. http://www.networkmagazine.com/article/NMG20011203S0008 [Accessed 11 Feb 2004].

IEEE. (2004) *802.11 standards* [Online.] http://grouper.ieee.org/groups/802/11/ [Accessed 27 Nov 2003]

Kershaw, M. (2004) *Kismet*, [Online]. http://www.kismetwireless.net [Accessed 23 Nov 2003].

Khan, J. and Khwaja, A. (2003) *Building secure wireless networks with 802.11*, Wiley, Indianapolis.

Milner, M. (2003) *NetStumbler.com The New World of WiFi* [Online]. http://www.netstumbler.com [Accessed 23 Nov 2003].

Mishra, A. and Arbaugh, W.A. (2002) *An Initial Security Analysis of the 802.1X Standard*, [Online] http://www.cs.umd.edu/~waa/1x.pdf [Accessed 27 Nov 2003].

Moskowitz, R. (2003) "Weakness in Passphrase Choice in WPA Interface", *Wi-Fi Networking News*, [Online.]. http://wifinetnews.com/archives/002452.html [Accessed 23 Jan 2004].

National Statistics (2001) *Census 2001: The most comprehensive survey of the UK population*, [Online.]. http://www.statistics.gov.uk/census2001 [Accessed 5 Apr 2004]

The Shmoo group. (2004) *Airsnort*, [Online]. http://airsnort.shmoo.com [Accessed 23 Nov 2003].

WiFi Alliance. (2002) *Wi-Fi Alliance Announces Standards-Based Security Solution to Replace WEP*, [Online.]. http://www.wi-fi.org/OpenSection [Accessed 12 Feb 2004]

Wilks, A. and Ghita, B. (2004). "An analysis of wireless security implementations", poster presentation, *4th International Networking Conference*, July 2004, Plymouth, UK

Design of an Architecture for Wireless Community Networks

A. Perry and P.S. Dowland

Network Research Group, University of Plymouth, Plymouth, United Kingdom
e-mail: info@network-research-group.org

Abstract

With the mainstream use of Wi-Fi technologies, security concerns about them have consistently been raised. The level of security offered by the standards does not aim high, and is just good enough for home users. These technologies are used by a variety of organisations, ranging from companies that need secrecy, to communities that wish to create parallel wireless networks. It has emerged from this that the need for an appropriate level of security, including reliable authentication and strong cryptography, is required. Solutions exist for companies that can afford it but something has to be done for small non-profit organisations that need a standard architecture that is easy to deploy. This paper investigates existing *de jure* standards from the IETF, their weaknesses and their ease of use from an end-user point of view. Based on this, the aim is to define a possible architecture that could be easily used in most cases, and since it is based on open source technology, is free of charge.

Keywords

Networks, Wireless, Security, Authentication

1. Introduction

When the high-frequency radio technologies appeared, people willing to emit using them needed a licence that had to be bought to the government in most countries. However, people began using them anyway, and this gave birth to the radio-amateur trend. Following a short course of merely a few hours, and accepting some simple rules, radio-amateurs could obtain a licence. From 1999 onwards, a new radio technology has emerged aimed at computers. Code-named Wi-Fi, it was a set of standards that used microwaves frequencies to transmit data. This technology has almost immediately met a large public, since it is cheap enough and standard-based, unlike its predecessors. It was not long before some users thought about the days of packet radio (the use radio-amateurs had of their computer radio interface) and how this was simpler and cheaper to obtain comparable results on low distance. However, Wi-Fi standards suffer from security issues, in addition to those that any publicly accessible network would have.

Indeed, the biggest problems come from the fact that a community network is made to be accessible by people that the owner of an access point might not know personally. However, if somebody does something wrong from an access point, such as defacing a web server, the owner has to be able to identify the person who did it, or the responsibility might fall on him. When speaking specifically of Wi-Fi, the obvious security issue that has made the headlines is the privacy of the data that is transmitted over the air. Indeed, inherently to the way they work, radio communications are broadcasted. This means that anybody in the range of an emitter can receive all of the data that is transmitted. The standards provide a simple

cryptographic capability for Wi-Fi, with little overhead and need for processing power. However, this standard has not been designed to be secure, and indeed is not.

2. Existing solutions

While researching for the best solution, several existing methods were encountered and studied. The most obvious or interesting ones are detailed below.

2.1 Simple use of 802.11 standards

The 802.11 standards, specifically their a, b and g revisions, were designed with efficiency in mind, more than security. Indeed, using radio waves to transmit data is inherently insecure, since the antenna in standard transmitters broadcasts all data it emits. The designers of the protocol were aware of this issue, and provided two different means to encrypt the data transmitted with Wi-Fi. The first one, created with the b revision of the standard, is called *Wired Equivalent Privacy*, or WEP. The advocated goal of WEP is to bring approximately the same level of security to wireless that a wire would give (IEEE, 1999). Though WEP is known to be deeply flawed (Gast, 2002), it makes it necessary for an eavesdropper to make some kind of effort to be able to get anything but noise. However, WEP works with pre-shared keys, which makes it unsuitable for the purpose of this research. Indeed, if all participants in a wireless network use the same key, nothing is hidden from any of them, and since the network is public, anybody would have access to the key.

The second encryption scheme is called WPA for *Wi-Fi Protected Access*. WPA is a *de facto* standard, inspired from the work of the task group I of the 802.11 workgroup at IEEE. It uses *Temporary Key Integrity Protocol*, or TKIP, to resolve one of the flaws in WEP, and 802.1x and *Extensible Authentication Protocol*, or EAP, for user authentication. This means it allows individual keys and a sufficient level of security, which is pertinent to this research. However, hardware that works with WPA is still not widely deployed; it would then be unrealistic to design an architecture aimed at communities that would force them to buy new hardware when they already have things that work.

2.2 Commercial solutions

Several commercial solutions aimed at businesses and large organisations exist that allow them to secure their wireless networks and authenticate their users. The one closest to the goals of this research was aimed at *Internet Services Providers*. Indeed, in the context of the research, the community is acting as an ISP and tries to provide Internet services to users through users themselves.

The commercial solution reviewed here is made by Sputnik. With this architecture, one centralised server takes care of authenticating users and granting them authorisations on services they might use on the network. Users access the network through *access points*, referred to as AP in the following, disseminated in several places (Sputnik, 2004). When the user tries to access the network, the access point redirects him to a login page. The login information is sent to the authentication server, which sends back authorisation information to the access point. The access point then modifies its firewall rules accordingly to give the user

access to the allowed services. Any parameter can be modified and any statistic can be consulted from anywhere on the Internet by the administrator of the authentication server. This would be a suitable solution if it were not as centralised as it is, leaving the configuration of a node to its owner. All these properties are interesting and should be thought of while designing the architecture that will result from this research. However, the Sputnik solution fails to address some simple security issues: most importantly, nowhere are user communications encrypted, even with WEP. This means anybody, even if they are not a user of the ISP, can "hear" anything the user sends or receives in clear. This is not acceptable for a network that is concerned with security. Secondly, the authentication is done through a simple login/password scheme. Passwords are widely accepted as the most simple login method, and users are used to them. However, security analysts also agree that unless a strict policy is in place for changing them regularly, with accordance to some simple rules, password protection is impressively weak (Beverstock, 2003). There are now other means to authenticate users, and some of them will be reviewed later.

2.3 Community aimed (and developed) solutions

The best-known community solution at this time is the NoCat home grown software. NoCat is a community of wireless enthusiasts that decided they needed something to control the use people made of their freely accessible APs. Their main software, NoCatAuth, is a perl script that acts as a "captive portal" (NoCat, 2004). This means the user that tries to access the web through an access point on which he has not been identified will get redirected to a web page on which he has to either login, or accept an "*Acceptable Use Policy*". Once this is done, the user can use the allowed services seamlessly. This is similar to the Sputnik way, except the owner of a node keeps control over it. However, in addition to the aforementioned Sputnik issues, until recently there had been no effort to use a common user database for all willing communities. However, the NoCat developers seem to show good will on any issue that might be reported to them, and their project is getting more and more support.

3. Methodology

The research for a good architecture design has been made in several steps: determining interesting and important features used in other projects and issues to address, defining new features and requirements, looking for standard-based free software that could fill those needs, and finally testing their interoperability.

3.1 Features and issues

Some of the main features and issues have been reviewed previously. The following summary also takes into account new objectives for this research:

- The system should use a captive portal to ease login operations
- The system should not use a password as a means to authenticate a user
- The system should be easy to administrate, from the network if possible
- A node should be administrated by its owner
- Node usage, user logins, and other statistics, should be collected on a central server as well as being accessible to the node owner and the user

- The database for user authentication should be centralised, but scalable
- Communications between the user and the node should be encrypted with an individual key distribution
- A solution should be proposed to allow certification of users' identities

3.2 Software review and testing

Before looking at software that implemented standards, it was important to decide which standards to use, and how to make them interact together. After much reading, it had to be decided which implementation to use, based on its support for the standard important features, or based on its ability to add to the standard when needed. The basic needs appeared to be: an *Authentication, Authorisation and Accounting* architecture that would take care of allowing users to use the system with respect with their credentials. A *Virtual Private Network* system that would be as much as possible based on standards, with as low an overhead and processing power needed as possible, but still strong enough cryptography. Lastly, the final requirement is a solution to build some kind of "web of trust" that would help to certify users' identities without having a central authority that they should all visit. The reviews, tests and results undertaken are presented as part of the next section.

4. Findings

Several standards were reviewed and software packages tested while conducting this research. The detailed findings about the most important of them are presented below.

4.1 FreeRADIUS

Remote Access for Dial-In User Service, or RADIUS, is an *Internet Engineering Task Force* standard presented in *Requests For Comments* RFC2865, RFC2866 and RFC2869 among others. It defines a complete AAA architecture with many possibilities as to the way users get logged in, the type of connection that will be brought up, the parameters of it (e.g. the IP address the user obtains), and so on. One of the recent additions to the RADIUS protocol that was of interest for the project was the EAP, since it allowed users to log in through the mechanisms defined by the *Institute of Electrical and Electronics Engineers* in their 802.1x standard (also called Port-Based Network Access Control). RADIUS allows several tunnelling protocols to be used, but all of them use a password scheme as an authentication means. EAP presented the advantage of being extensible, meaning it was possible to write an additional module ("plug-in") for any software to allow any other kind of authentication (Hassel, 2002). 802.1x presents many advantages when used with WPA; however, as previously stated, only recent hardware can make use of it. Since this research is focusing on informal communities, it had to allow them to use their already acquired hardware, and thus WPA and 802.1x were not an option. The RADIUS protocol, without the EAP was not versatile enough to be used for the goal of this research. Since the project would have benefited from a robust and resilient AAA architecture, investigation on this issue was extensive, but with no viable outcome.

Since it seemed to be the best RADIUS implementation available at the time of this research in terms of resiliency, robustness and compatibility, the FreeRADIUS software was used. It

was both successfully operated against a *Lightweight Directory Access Protocol* directory containing user login information and queried with a RADIUS web client developed in PHP, for *PHP: Hypertext Pre-processor*. The community around FreeRADIUS was very reactive when concisely exposed to undocumented questions, and helped a lot in understanding the RADIUS shortcomings detailed above.

4.2 OpenLDAP

LDAP is comprehensive, lightweight protocol, originally designed to address the need for a simpler way to access X.509 directories. The IETF has standardised the LDAP in many RFCs, the most important and central one being RFC2251, which standardises version 3 of the protocol. Directories are close in functionalities to databases, but are different from them in some aspects. In particular, directories are mostly "read" and rarely "written", and this is reflected on the performances. Directories are often used to store user credentials and information, for example by Microsoft with its Active Directory (which is, in fact, a modified LDAP directory). LDAP directories define their schemas using an *Abstract Syntax Notation One*, or ASN1, syntax, which enables easy modification and addition of new object classes and new parameters (Howes *et al.*, 2003). In addition to this, LDAP has an impressive amount of support from the free software community, thus making it very practical to use. The biggest advantage in the context of this research is certainly its ability to easily distribute and replicate data.

OpenLDAP is the reference LDAP implementation in the free software world. It implements the LDAPv3 specifications and is considered to be robust. It is as interoperable as one could wish and is particularly easy to configure. It can use *Simple Authentication and Security Layer* to secure communications from one directory subset to the other, or from the master base to its replicate. OpenLDAP was successfully used to store user data.

4.3 NetSNMP

SNMP (*Simple Network Management Protocol*) is the Internet-Standard Network Management Framework as defined by the IETF (mainly in RFC3410 to RFC3418). It is based on the same underlying technologies as LDAP (the ASN1 syntax and the *Object Identifier* hierarchy), and is as resilient as its "cousin" (Mauro & Schmidt, 2001). The goal of SNMP however differs completely from the goal of LDAP: SNMP provides an easy and extensible way of monitoring and controlling a network. Since RADIUS was not a suitable architecture, no way to record and consult statistics about access points or logins is provided, if it were not for SNMP. By defining the appropriate *Management Information Base* it is possible to allow monitoring of such parameters that are of interest to the communities. This is easily done through ASN1, and only requires a freely obtainable *Internet Assigned Numbers Authority* number to be able to officially create such monitored objects. NetSNMP is considered a stable and complete implementation of SNMPv3 and contains many tools to ease statistics creation based on the SNMP collected data.

4.4 GnuPG

Gnu Privacy Guard (also known as GnuPG or GPG) is a free implementation of the *Pretty Good Privacy* IETF standard defined in RFC2440 that was first implemented in the OpenPGP

product. PGP is a standard for securing electronic messages using public key encryption. As opposed to many *Public Key Infrastructures*, PGP does not make use of a certification authority for certifying public keys. Instead, if a user wants to sign a key as "trusted", he has to verify in real life the identity of the owner and his public key fingerprint. Since the users have to meet in real-life before they can be sure of each other's keys, one still has to meet the owner of a node to be able to use it. Fortunately, there is more to PGP than the simple "ultimate" trust as it is called. Users can build a "web of trust", allowing them to trust the key of a user they have never met (Ashley, 2004). Indeed, every user acts as a certification authority, and if A trusts B and B certifies that C is trustable then it follows that A can trust C. This allows the building of a trusted community, and one can be confident the name associated to a key is indeed the name of the user of that key. Then, if somebody does something wrong, he can easily be identified using login statistics and his trusted key. The only drawback with this method is the time it can take to build such a web of trust. However, the PGP/GPG community being online for some time, the existing web of trust is therefore already well established, and can be used.

GnuPG is very good and has often been praised for its impressive compatibility features with other implementations of PGP, and is even considered the new reference implementation. It is also considered as being very robust from a security standpoint. Its use in the architecture would allow users to be identified reliably, and without the need for a central authority. Users new to the network or the PGP world could login with restricted privileges until they get their key signed by another member of the community.

4.5 Kame/racoon

Internet Protocol Security, or IPsec has its specification defined in several IETF RFCs, the most important being RFC2401: Security Architecture for the Internet Security. As for previously mentioned protocols, a complete description of IPsec would be beyond the scope of this paper. However, IPsec is not just a simple VPN technology, and that is what makes it so interesting. Indeed, if IPsec has a tunnel mode, which works just like a VPN by encrypting and then encapsulating a complete IP packet into another, it conveniently also offers a transport mode. As the name indicates, the transport mode works on the transport layer, thus not needing a second IP address, and reducing the overhead introduced. Unfortunately, this mode cannot be used in the situation of this research without the involvement of a complex routing mechanism that to the knowledge of the author has not yet been implemented anywhere. It also has an impressive choice of encryption and hashing algorithm to choose from, to allow the reduction of processing power needed to make it run on embedded devices. However, less processing power means a weaker algorithm, and in turn less security. The *Internet Security Association and Key Management Protocol*, or ISAKMP, and the *Internet Key Exchange*, or IKE, are here to compensate for this by providing an automatic way to change the symmetric keys as often as necessary. They also are responsible for the initial key exchange, and for verifying user certificates.

Kame is an implementation of the IPsec architecture in the sense that it can deal with adding *Authentication Header* and *Encapsulated Security Payload* according to the *Security Policy Database* and provided the *Security Association Database* contains the needed entries. Racoon is the associated daemon that deals with ISAKMP and IKE negotiations. Pre-shared keys can be defined, that will only be used for the negotiation of the communication keys, or

X.509 certificates can be used, so that the authentication of a user is unique, and the communication between their computer and the gateway stays protected from any eavesdropper. Kame and racoon were first developed for the NetBSD operating system, but were later ported to other operating systems such as the FreeBSD, Apple MacOS X or GNU/Linux. It has been shown to interoperate well with other implementations, including Microsoft Windows NT stack, or early GNU/Linux implementation FreeS/WAN.

5. Final design

From the previous results, it was possible to define an architecture for a wireless community network that would prove to be efficient, secure, and reliable. Identification is based on the web of trust of a community. If a user has had their key signed by a trusted member of the community, they become a trusted member and can then access the network through any node, with more than just *World Wide Web* access. Communications between the user and the node would be encrypted using IPsec, based on X.509 certificates generated for the users and sent to them by the node/authentication server encrypted with their PGP key. Several cryptographic schemes could be used with IPsec, depending on the available processing power. Since using algorithms such as the *Data Encryption Standard*, which has long been proven easily breakable, would lower security, it would be advisable that embedded devices set a short delay for changing keys. The authentication data, and optional user information, would be stored in a decentralised LDAP directory, and statistics about the node could be obtained from SNMP agents installed on them. This sets up the big picture, and offers a mostly complete design, as shown on Figure 1 : A possible design resulting from this research.

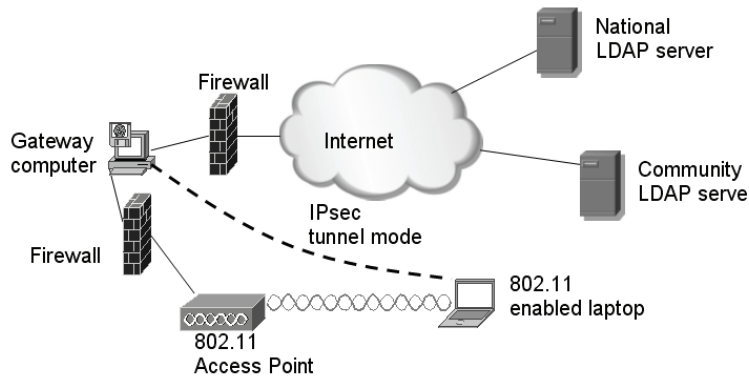


Figure 1 : A possible design resulting from this research

While this model should efficiently protect a user of the wireless network from eavesdropper, and protect the node owner's data from being accessible thanks to the firewalls, it is not free from issues. First, a user has to trust the owner of the node he is using, since the data is deciphered on it and the owner could easily dump any data that goes through it. However, websites that ask the user for sensible information without forcing a secure connection tend to

disappear, and the current trend in the web browser edition world is to consider security a major issue. Secondly, the owner of a node still has some issues to solve. Using the web of trust will help him ensure that he knows the identity of the owner of a given IP address at a given time. However, it is still not clear how the provision of this information would be accepted in a court. Currently, simply stating that a *cracker* took control of a computer is enough for the owner of that computer to be cleared (BBC News, 2003), but to the knowledge of the author, no law exists that guarantees such a judgement. A court might not trust a simple Wi-Fi enthusiast as it would trust an ISP when using logs to identify the origin of an attack. Laws and precedents might differ depending on the countries, but that is the role of the communities to check.

6. Conclusion

This research has shown it was possible for wireless communities to build secure and interoperable networks for just the price of hardware. Indeed, the free software community has made available such great tools, based on Internet standards, that allow designing a complete architecture, with high security prerequisites. Agreeing on such a design worldwide is still to be done, but it would allow a user from a town in the USA to use his existing account to roam in any other American city, and even in England. Once a consensus is reached among a great enough amount of communities, easy to use software still has to be developed, to allow willing people to easily deploy new nodes, without having to worry about legal or technical issues too much. Luckily, the free software world has recently seen a phenomenon develop around live CD distributions of operating systems such as Linux, and tools have been developed to easily create such CDs. A user would then just have to boot from it on the gateway computer, answer some simple questions, and optionally install it on the hard drive to get it running.

7. References

Ashley, J.M. (2004) *'The GNU Privacy Handbook'*, The Free Software Foundation. [Online] http://www.gnupg.org/gph/en/manual.pdf

BBC News (2003) *Teenager cleared of hacking* [Online] http://news.bbc.co.uk/1/hi/england/hampshire/dorset/3197446.stm

Beverstock, D. (2003) *Passwords are DEAD! (Long live passwords?)* [Online] http://www.sans.org/rr/catindex.php?cat_id=6

Gast, M.S. (2002) *802.11 Wireless Networks*, O'Reilly & Associates, Inc., Sebastopol.

Hassel, J. (2002) *RADIUS*, O'Reilly & Associates, Inc., Sebastopol.

Howes, T.A., Smith M.C. and Good G.S. (2003) *Understanding and deploying LDAP directory services*, Addison-Wesley Professional.

IEEE. (1999), *IEEE 802.11, 1999 Edition* [Online] http://standards.ieee.org/getieee802/802.11.html

Mauro D. & Schmidt K.J. (2001) *Essential SNMP*, O'Reilly & Associates, Inc., Sebastopol.

NoCat. (2004) *NoCatSoftware* [Online] http://nocat.net/moin/NoCatSoftware

Sputnik. (2004) *Sputnik Solutions for WISPs* [Online] http://www.sputnik.com/products/scenarios/casewisp.html

Development of a Linux-Based Management Service Using the Simple Network Management Protocol (SNMP)

M. Mochamet and B.V. Ghita

Network Research Group, University of Plymouth, Plymouth, United Kingdom.
e-mail: info@network-research-group.org

Abstract

This paper reviews the research and the development of a Linux based management service using the SNMP (Simple Network Management Protocol). The implementation took place in Intracom S.A Telecommunication industry and the management service concerned the POS (Point Of Sales) terminals that the company produces. More specifically, the project identifies the management needs of today's POS networks, while it implements a research trying to identify if SNMP can adequately cover those needs. Moreover, this paper brings forward a different dimension in the issue of managing POS terminals and introduces a new definition the *application* management, which corresponds to the terminals application operational status. Furthermore, it includes the implementation of two MIB (Management Information Base) files and an application which allows a network manager to retrieve event log files from a remote location by using the SNMP protocol.

Keywords

SNMP, POS Management, Agent, Management Information Base (MIB).

1. Introduction

In 1969, ARPANET, the first packet switching network, was developed by the U.S Department of Defense (DoD), in order to interconnect universities and government facilities. By the late 1970s the TCP/IP protocol suite had been standardized by the Internet Application Board (IAB), (Walworth, 2003), for use in military communications. By the mid-1980s the growth of the Internet was so rapid that the need for network monitoring and maintenance had become inevitable.

The initial idea of this project was the development of a Linux-based network management platform in a POS (Point of Sales) network using the SNMP (Simple Network Management Protocol). The project took place at Intracom S.A. Telecommunications Industry. More specifically the project aims included the extension of an SNMP agent by writing MIB (Management Information File) files which would meet a POS network's management needs. Moreover, the implementation concerned the installation of a NMS (Network Management Station) package in a TCP/IP network of terminals. The problem that the above implementation introduces, is the fact that a POS network's management needs are divided in two categories. The first category represents the networking which is combined with the network performance, configuration and fault management. The second category concerns the application that is running on the POS terminal. More specifically, the management service which is to be implemented must be able to handle any possible malfunctions of the terminal's application and to provide the network manager with information about its

operating status. The latter brings forward a different dimension in the problem and introduces a new definition the *application* management. The remained of this paper will analyse the management needs of a POS network, followed by a discussion about the implemented application

2. POS network's management needs

Before discussing about the management needs of the network devices, it would be sensible to analyze the general topology of a POS network. In more details a POS network consists of three different parts. These are, the Central System, the networking and the terminals part.

The initial idea was to implement a management service using the SNMP protocol, which would adequately cover the networking part but also and the terminals part. This way, it would be managed to have homogeneity in the management process of the whole network and it would require no extra staff to be trained for this purpose. The problem was that the terminals management needs, as it will be analyzed further on, are not concern only their network behavior, but also the application which is running on them. The latter brings forward a different dimension in the problem and introduces a new definition the *application* management. The question that the above entails and which the following paragraphs will try to answer is, if SNMP was the right choice for this type of management or if another solution should have been given.

First of all, as far as it concerns the networking part, all these devices are already having installed network management services using the SNMP protocol. Their management needs are based on the FCAPS (Fault, Configuration, Accounting, Performance, Security management), as these are defined by the International Standards Organization (ISO), using the first letter of each category (Alcatel, 2002). In more details, their needs have to do with the ability to gather information about dropped packets, transmission error rates and traffic, in order to predict possible link failures. Moreover, with the ability to change routing protocols or update routing tables. Finally, with the ability to boost the network's performance, by decreasing the delays, by implementing traffic distribution, and by replacing links which appears to be keen on failure. In general, the ultimate goal of the previous management needs is to provide the network manager with a service which would allow him to be proactive and not reactive to his network's operation malfunctions.

The second and the most complicated part, as far as it concerns the management needs, is the terminals. In more details, the terminals management needs, can be organized in three categories. First of all, includes the networking part which has to do with the network interfaces behavior. The management needs concerning this category are not far from those stated above. Secondly, they also require management, which has to do with the application, defined above as *application* management. Analyzing further the previous statement, the SNMP implementation should be able to handle any possible malfunctions of the terminal's software, to monitor interval variables from the terminal's NVRAM (Non Volatile RAM) and to retrieve event log files. The next section provides all the details of the practical implementation of these needs. Last but not least, the implementation concerns the hardware part of the terminal. As a matter of fact the terminals use several peripheral devices (e.g. printers and bar-code readers); these devices have a useful operating period, where after that,

problems may occur. According to the latter, the management service should be able to create alerts, informing this way the manager about any possible replacement or maintenance requirements of these I/O devices.

As it is referred in the literature SNMP is the dominant network management protocol in our days. However, its initially designed aims were not include the ability to provide *application* management into devices. The following paragraphs will present the problems encountered during the implementation by also stating the given solutions.

3. Agent Implementation

This section is referred to SNMP agent activation to the Coronis HE terminal which Intracom S.A designs and manufactures. However, the implementation appears to be totally robust and cross-platform.

3.1 UCD-SNMP tool

UCD-SNMP is an open source tool which was initially designed by University of California in 1995 (Schonwalder, 2002). Since then, several versions of UCD-SNMP project, has been implemented. In our days, the project has been renamed in NET-SNMP and it is freely available to download from the Sourceforge webpage in the following link *net-snmp.sourceforge.net*. For the needs of the management service implementation it has been used the 4.2.6 version of UCD-SNMP, as this is embedded to Linux distribution that Coronis terminals are using. The project provides the user with several tools concerning the SNMP protocol. More specifically the project includes:

- An extensible agent
- An SNMP library
- The SNMP set of commands for setting or requesting information from the SNMP agent
- Tools to create and handle SNMP traps
- A MIB browser

The main problem concerning the UCD-SNMP project is the lack of any documentation. The latter makes the tools installation and usage extremely complicated, especially for first-time users. In contrast, the fact that UCD-SNMP is available in any current Linux OS distribution renders this tool very popular inside the network management community.

3.2 Extending the agent

As it is referred in the previous part UCD-SNMP project provides an extensible agent. In other words by adding the proper files to the agent configuration the user is able to extent its functionality and the supported objects. In order UCD-SNMP to enable this option, includes the **mib2c** tool. The latter is designed to take a portion of the MIB tree (as defined in the MIB file) and to generate the necessary code skeleton in order to implement a new MIB module (Shield, 1999). Note that, mib2c tool helps the user to create a prototype of the required source code, but the files need further add-ins to become operational.

As it is referred in (Shield, 2004), in order to implement a new MIB module three files are necessary:

- a MIB definition file
- a C header file
- a C source file

The MIB file is the first file a user should create. It defines the objects that the new module will contain and the information, which they will provide to the network manager. Section 1.3.2, presented the structure of an MIB object, describing that it follows the ANS.1 syntax. When a new MIB module is to be created, the developer must assign to the module an OID (Object Identifier) number. In other words, the module has to be placed somewhere within the overall MIB tree. For this reason application forms are available in *www.iana.org* website, while the registration for an OID number is free of charge. The MIB file itself contains the following sections. Firstly, the imports part which include definitions of any other MIBs used. Secondly, the MIB file includes the Module Identity part, which gives a brief description and update information of the MIB (e.g. Last-Updated, Contact-info) and also defines the OID number which has been specified for the module. Finally, the Module Identity part is followed by the definitions of the MIB objects.

3.3 Coronis terminals MIB files

For the management needs of Coronis terminal, specified in section 2, two MIB files created. These were the Intracom-NVRAM-MIB file and the Intracom-LOGFILES-MIB. As far as it concerns the first one, this was created in order to cover the monitoring needs of the terminal's NVRAM (Non-Volatile RAM). More specifically inside the NVRAM there are several information concerning the terminal's operating status saved in struct variables using the C programming language. The MIB file created contains the definition of 153 objects which corresponds all the variables included in the NVRAM. All the defined objects are having read-only access rights, as modifying their values would result in the terminals unexpected behavior. The MIB objects accesses the terminals hard-drive were the NVRAM variables values are located in a binary form. In addition, the SNMP agent by defining a C programming struct variable, which simulates the original one, responds to the manager query with the selected values. The implementation of this file allows to the network manager to derive information about the terminal's operating mode, the transaction number, the installed wagering games and several other data.

The second MIB file which created was the Intracom-LOGFILES-MIB. This file defines two objects. The first one is read-write object which allocates the log-file path inside the terminal's hard drive (e.g. /etc/logfiles/). The second is responsible for retrieving the log file according to the path which is specified by the previous object. In other words, the network manager has the ability to retrieve any log file from the terminal by specifying its location using the *snmpset* command and then by query it using the *snmpget* command.

3.4 Terminal's log files retrieve process

The biggest research provocations of the project's practical implementation, was the terminals event log files recovery by the SNMP manager application. The following text illustrates a sample of the communication related log file that the Coronis terminals create.

```
Tue Jul 20 11:33:44 2004 - SERVER to LOTOS: After listen
Tue Jul 20 11:34:16 2004 - SERVER to LOTOS: After accept
Tue Jul 20 11:34:27 2004 - COMM_SERVER_SEND
Tue Jul 20 11:34:27 2004 - Sending coupon to HOST
Tue Jul 20 11:34:27 2004 - Trying to connect to HOST
Tue Jul 20 11:34:27 2004 - Connected to HOST
Tue Jul 20 11:34:28 2004 - Socket Closed---Transaction succeeded
Tue Jul 20 11:34:28 2004 - Coupon sent to HOST
Tue Jul 20 11:34:41 2004 - Timer Initialized in HOSTINIT
Tue Jul 20 11:43:16 2004 - ADSL/VSAT
Tue Jul 20 11:43:16 2004 - SERVER to LOTOS: After bind
Tue Jul 20 11:43:16 2004 - SERVER to LOTOS: After listen
Tue Jul 20 11:43:48 2004 - SERVER to LOTOS: After accept
Tue Jul 20 11:44:41 2004 - Timer Initialized in HOSTINIT
Tue Jul 20 11:49:27 2004 - ADSL/VSAT
```

As it can be seen the file contains a time-stamp of the occurring event and it describes the communicating status of the terminal. From the latter it is obvious that the manager can extract very useful information about the terminal's behavior at each time, while similar log files exist concerning the terminal's application's behavior and the hardware's operating status.

The main problem had to do with the fact that SNMP was not designed for processing such a query. In more details, the terminals log files are a cyclic buffer with size of 100Kbytes, while the maximum amount of information that a SNMP object is able to carry is 4000 bytes.

The first idea was to split the log files in parts of 4000 bytes each and carry them through the network. The process was using an SNMP object of string type which was reading 4000 bytes each time from the file. In order to transfer the whole file 25 packets required and as a result 25 snmpget commands. The main problem with the above procedure, except of course from the large number of packets which SNMP had to carry, was the fact that when SNMP prints a string in the screen it uses the isprint( ) function, replacing this way all the non-printable characters (e.g. \n, \r) with another one (e.g. $, #, .). The result that the above entailed was the fact that in order to make the idea functional, the source code of the SNMP agent had to be modified.

In order to confront the first problem, the need of 25 snmpget commands, the idea was to zip the log files by using the zlib compression library. The fact that the log files was text based combined with the fact that many messages in the log file was repeated very frequently, led to a compression rate of 93% and new file size minimized in about 7Kbytes. The result was that in this case, 2 only snmpget commands would be enough to transfer the 100Kbytes, file through the network. However, a zip file is nothing else but a binary file and as a result it contains several non-printable characters, which SNMP would replace during the transmission.

The solution in this problem came with the idea of converting the binary zip file into a hexadecimal file. However, as it is referred in the literature SMIv1 is not able to define hexadecimal object types and as a result the only solution was to transfer the zipped file as a string of characters and then reconstructing the original file in the NMS (Network Management Station) side. Of course, this would result in doubling the zip file size, as two ASCII characters would be needed to present one hexadecimal, but transferring about 16Kbytes instead of 100Kbytes was a good improvement. Moreover, in order to help the original file reconstruction in the manager side of the network and to make the application as general as it could be, the file parts which constructed with above procedure contained a footer with various information such as the total number of parts that the original part was divided, the part size, the part consecutive number and other.

Finally, for the needs of this process, a GUI (Graphical User Interface) created in order to simulate the snmpget commands loop which required according to the size of the file. The following figure illustrates the application's window which was designed using Microsoft Visual Studio 6.0.

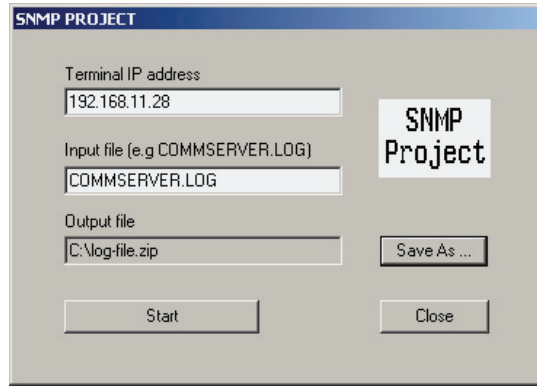


Figure 1 : Log-files retrieve process Graphical Interface

As it can be seen the manager is prompted to specify the terminal IP or hostname address, the name of the log file which he wants to retrieve and to specify the location where the file will be saved in the local hard-drive.

4. Conclusions

Network management is an evolving area of computer networks technology. As the network devices become more and more sophisticated the need for an effective management solution becomes a necessity. SNMP is the protocol which can effectively cover all the existing management needs of IT society in our days.

This paper presents the research and development of a Linux-based management service by using the SNMP protocol. More specifically, it described the POS terminals and networks

main characteristics and identified their management needs. As a result this paper defines a new dimension in POS networks management, the *application management*. Moreover, the practical implementation included the development of an application which allows the network manager to retrieve event log files from a managed workstation. The application is totally robust and is platform independent.

Overall, the project subject appeared to be very interesting. The fact that it was covering both a theoretical and a practical implementation of a network management service provided the author a very good background which will certainly form a base in his further working career. Furthermore, the project will be able to compose a base for further research in POS networks management. In more details, the SNMP service installation and testing took place in a small LAN network of terminals. A step forward would include the activation of the management application in the field, where the conditions of traffic and packet losses are realistic and the average number of terminals exceeds the three thousand. The latter would probably require the modification of the service in order to achieve more effective processing of the incoming management information.

5. References

Alcatel Internetworking. (2002) "*Network Management*", [online]. http://www.ind.alcatel.com/library/e-briefing/eBrief_NetworkManagement.pdf [Accessed 25/02/2004]

Shield, D.T. (2004) "*Extending the UCD-SNMP agent*", [online]. http://net-snmp.sourceforge.net/tutorial/agent/index.html [Accessed 13/08/2004]

Shield, D.T. (1999) "*How to write a MIB module*", [online]. http://www.csc.liv.ac.uk/~daves/Misc/UCD/01-intro.html [Accessed 13/08/2004]

Schonwalder, J. (2002) "*Evolution of Open Source SNMP Tools*", [online]. http://www.ibr.cs.tu-bs.de/users/schoenw/papers/sane-2002.pdf [Accessed 25/02/2004]

Walworth, S. (2003) "*SNMP*", [online]. http://www.nas.nasa.gov/Groups/Networks/Training/snmp [Accessed 27/02/2004]

A Knowledge Based System to Support Customer Service Agents Remotely Faulting Advanced Mobile Terminals

R. Ramchurn[1] and P. Reynolds[2]

[1]Network Research Group, University of Plymouth, Plymouth, United Kingdom
[2]Orange SA, Bristol, United Kingdom
e-mail: info@network-research-group.org

Abstract

Knowledge is a vital commodity for any organisation and its effective utilisation is essential for a company to survive in a competitive environment. In the telecommunications industry various knowledge sources are employed while diagnosing mobile phones for problems. These include qualified professionals, manuals and computer systems. This paper discusses how knowledge engineering techniques were used to structure the knowledge sources into a model to develop an expert system. This formed an innovative support tool that equipped the customer support agents with the necessary knowledge to tackle the mobile phone problems. The findings from the investigation implied that there is scope for development of new methods to make customer service more effective especially in the mobile phone industry.

Keywords

Knowledge, Knowledge engineering, Expert systems, Customer service agents

1. Introduction

The different methods used to determine faults on mobile phone handsets nowadays are limited because of the variety and depth of issues. The main reason being that so many new phone models appear on the market almost daily, each carrying with them new challenges for the customer service agents attending to the user's problems. One new way of addressing this area of concern is to utilise an expert system as a support tool to assist the agents in their daily tasks. The benefits of doing such an investigation were firstly to allow the capture of a wide area of knowledge in a specific domain and then to engineer this intelligent data into a diagnostic tool that solved mobile terminal problems. Many studies (Liebowitz, 1988) have revealed that expert systems have been useful in diagnostic applications. Experiments (Harmon *et al*, 1988) have shown that these systems have proved useful in diagnosing and offering treatment for certain blood infections. Other more recent developments in this area involve the "Automatic Control System" used to monitor and diagnose cellular network problems (Muñoz, 2003) and the "Bounced mail expert system" used by the White House to diagnose and correct internet traffic problems caused by bounced back mail (Nahabedian and Shrobe, 1996). This paper outlines a project conducted with Orange and shows how capturing maximum knowledge about a domain developed a diagnostic support tool and the way in which this was applied to solve featured mobile phone faults. The study describes the

knowledge acquisition methods, which enabled gathering of knowledge from various categories used to build an expert system. During the investigation a new approach was utilised to structure the knowledge captured so that the expert system could be realised. The work included capturing information from technical staff and system experts through questionnaires. The rest of the paper describes the system development methods applied followed by a discussion of the results observed.

2. Methodology

The system development process involved capturing knowledge about the different problem areas before encoding them into an expert system shell. Interview sheets were used to capture information about individual faulting sessions with customers. This investigation yielded a total of one hundred and eighteen (118) main problem areas structured in eight different categories. The exercise was split into five phases for better time management as follows:

2.1 Front End Analysis

This phase was concerned with familiarising with the problem domain. The work entailed identification of the problem with respect to customer needs and issues being tackled by the customer service agents. At the end of this exercise, a detailed description of the problem was produced together with the scope of the solution.

2.2 Task Analysis

This phase involved gathering information about how the problems were already being tackled. The various sources of knowledge used were then looked at for any patterns. The extensible mark-up language (XML) was used to capture the expertise into a knowledge model and covered all the potential mobile handset problem areas. This process was initially undertaken by listening to customer calls and then interviewing the customer agents to have a good understanding of how the customer's problem was solved. This included discussing the problem, the agent's job function, the problem solving techniques utilised and other sources of expertise. The categories of problems captured by the XML model are shown in figure 1.

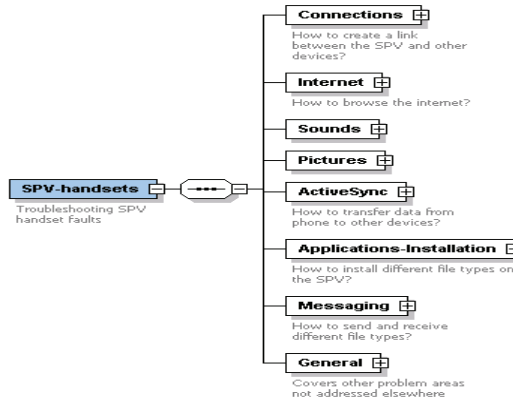


Figure 1 : Categories defined in XML model

2.3 Prototype development

This part of the process was concerned with the implementation of the captured data in a correct and efficient knowledge base. The development of the solution was undertaken using an expert system shell due to time constraints. An expert system shell is an efficient way of creating an expert system. It offers an inference mechanism, an explanation component and an empty knowledge base that can be filled in with knowledge from any domain. Figure 2 shows the architecture of an expert system that demonstrates the components of the expert system shell.

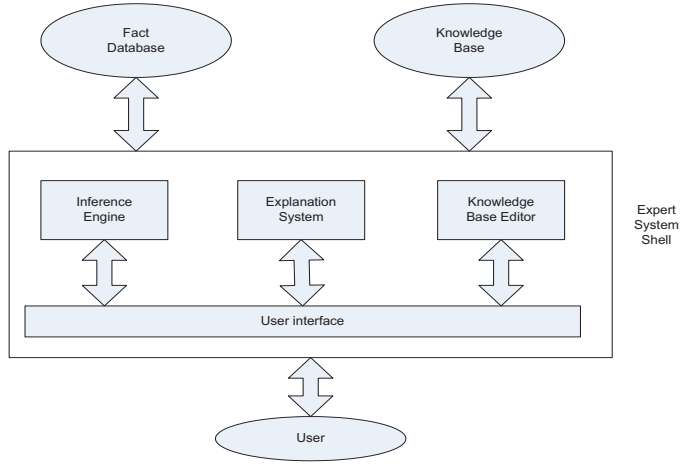


Figure 2 : Expert system architecture

In this exercise an evaluation of the various available expert shells was undertaken to assess their suitability for this study and the Xmaster system shell was chosen. In this phase only a few problems were used and tested into the shell. These were classified into problem areas and symptoms causing the problems. Once the inference mechanism was verified to perform according to the diagnostic process, the full system was implemented.

2.4 Solution development

This phase involved the encoding of all gathered knowledge into the expert shell to produce the final expert system. Initially, the core structure of the system was implemented by rearranging the problem areas and symptoms to allow smooth consultation. Then the design was adapted to the needs of the customers and to allow future expansion of the knowledge base. Finally the user interface was tailored using phrases and explanations to make it as easy and natural for the agent to use.

2.5 Testing

This step was used to test the consistency of the encoding process and identify any shortcomings in the expert knowledge base created. The different paths used by the system to reach a solution were analysed and modifications undertaken to allow completion of the expert system development.

3. Results

In each of the categories identified, the intention was to capture the maximum number of problems related to that area. The number of issues addressed depended on the service being dealt with. For instance problems pertaining to data support involved more symptoms to be gathered. The number of problems and symptoms captured in each category are shown in Table 1.

Category	Problem areas	Associated Symptoms
Connections	22	49
Internet	9	20
Sounds	16	24
Pictures	20	29
Active Synchronisation	12	23
Applications	7	13
Messaging	13	21
General	19	26

Table 1 : Breakdown of items captured per category

Once all these knowledge items were encoded into the expert system shell, the diagnostic tool was ready to address problems in real time. Whenever a customer called a service agent with a query, a search was done by the system to match the problem area. The expert system then ran a diagnosis on the handset and presented questions that were put to the user by the agent.

This allowed the system to identify the root of the problem in a short time, which then advised the agent how the fault could be resolved. The whole process took place as shown in Figure 3 below.

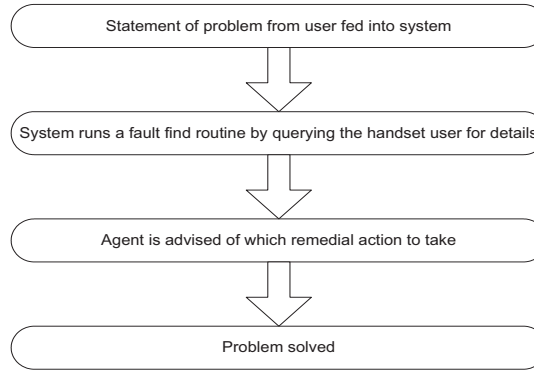


Figure 3 : Flowchart of expert system operation

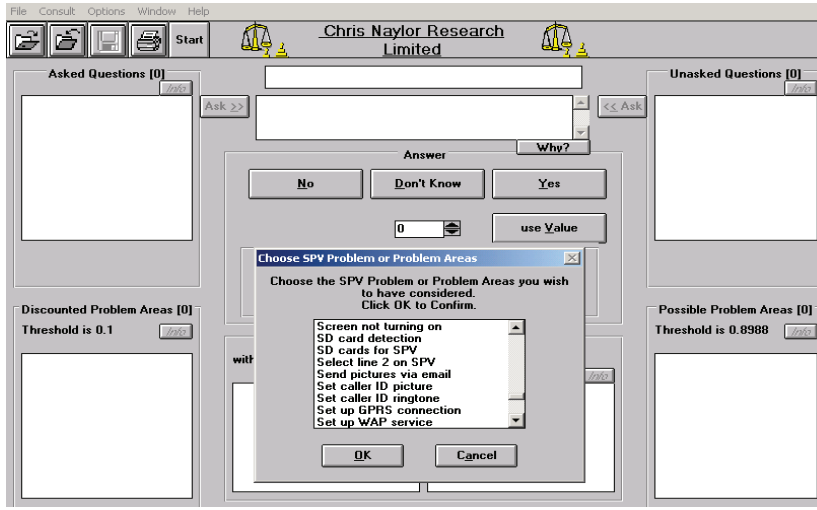


Figure 4 : Expert system user interface

Figure 4 shows the user interface of the developed expert system that equipped the customer service agents with the necessary knowledge to solve mobile phone problems. The categories encoded in the system were tested against real time problems and all were successfully

solved. Observations demonstrated that by using this system, the customer service representatives were able to improve the time taken to solve a query by 50% as they knew exactly what to do to tackle the problem. This was based on times taken during the testing experiments.

4. Discussion

In the first part of the investigation, the extensible mark-up language (XML) was used to acquire knowledge from various sources due to its hierarchical nature and versatility. The advantage of using XML was the fact that it could be converted into Java code that is widely used in programs on mobile devices. This novel approach in knowledge acquisition has added to the already existing techniques. This exercise required the use of good interpersonal skills, as people from various backgrounds were involved in the study.

The Xmaster shell was used to develop the expert system due to its ease of use that saved a lot of time. Initially the expert shell demonstrated some bugs, which were eventually fixed by the product supplier. Some limitations of the shell involved its fixed graphical user interface that could not be customised and the fact that there were no in-built advanced inference facilities.

The results show that once equipped with this solution, the customer service agents were able to solve problems in a consistent way. The time required to address a query was significantly reduced as the expert system quickly narrowed down the problem space and pointed to the root of the faults. The fact that knowledge from various sources were captured and preserved in the knowledge base meant that the agents could enrich their own knowledge by learning from the system. The developed tool has proved that the theoretical proposals for ‘a device diagnosis service’ by the research and innovation team of Orange UK could be further developed and implemented in practice.

5. Conclusion

This paper has considered the issue of developing an innovative support tool for customer service agents in the mobile phone industry. Even though the study has focussed on knowledge-based systems, there exist other systems such as intelligent systems capable of learning which are worthy of further investigation. The most important breakthrough by this tool is the ability to perform remote diagnostics on handsets, which can be miles away. This encourages work to be undertaken on data gathering over the air through a client program that can be downloaded to mobile handsets. The depth of the knowledge encoded in the expert system suggests that it can be tailor-made into a tutorial tool for mobile phone users. This is especially helpful when dealing with highly sophisticated phones. In this paper the power of knowledge engineering and expert systems has been demonstrated by description of an application in the mobile phone industry. This contributes to the understanding of knowledge acquisition and representation already available from current literature.

6. Acknowledgements

The authors wish to acknowledge the support of the Orange Research and Innovation department in the UK who sponsored this research project.

7. References

Harmon P., Maus R., Morrissey W. (1988) *Expert systems tools and applications*, John Wiley & Sons, Inc., USA

Liebowitz J. (1988) *Expert Systems Applications to Telecommunications*, Wiley Series in Telecommunications

Muñoz J. (2003) "Automatic Control of Mobile Networks" Article [online] available http://www.nokia.com [Accessed: 19th Feb 2004]

Nahabedian M. and Shrobe H. (1996) 'Diagnosing delivery problems in the white house information-distribution system' *The Artificial Intelligence magazine* [online] Available: http://www.aaai.org [Accessed 06th May 2004]

Decoding Schedules of Hybrid Concatenated Turbo Codes

I.F. Isnin and M.Z. Ahmed

School of Computing, Communications & Electronics, University of Plymouth,
Plymouth, United Kingdom
e-mail: mahmed@plymouth.ac.uk

Abstract

In this paper, performances of three decoding schedules of Hybrid Concatenated Turbo Codes were investigated. Every decoding schedule is differentiated among each other by having different MAP decoder component as their dominant decoder component. As the results, it was found that the performance of decoding schedule, with stronger dominant decoder component, is believed to have its error floor at lower level of bit error rate and experience less convergence problem, compared to other decoding schedules. And, the decoder of Hybrid Concatenated Turbo Codes system was found to achieve its maximum BER performance at little number of iterations before start to experience convergence problem.

Keywords

Hybrid Concatenation, Turbo Codes, Iterative Decoding, Decoding Schedule.

1. Introduction

In 1993, (Berrou et al. 1993) presented a new class of convolutional codes, called Turbo Codes. Berrou's original turbo codes encoder was built using parallel-concatenated scheme. It consists of two Recursive Systematic Codes, which were arranged in parallel and between of these two code components, there was one interleaver that permutes input bits sequence randomly. Then, outputs from these two code components were concatenated to become one single output of the encoder. In decoder, decoding process was done by implementing iterative decoding scheme. By using Turbo Codes, They achieved better performance in error correcting, where they proved that it is possible to get Bit Error Rate = 10^{-5} at Eb/N0 = 0.7 dB.

Because of the amazing performance that has been archived by Turbo Codes, after sometime it was introduced, many researches regarding this new convolutional code class, have actively been done. (Benedetto et al, 1996) have done Turbo Codes but this time it was in serially scheme. Compared to Parallel Concatenated Turbo Codes, Serially Concatenated Turbo Codes offer better performance at 10^{-6} to 10^{-7} bit error rates. Then in (Divsalar, 1997), a hybrid approach of concatenated turbo codes was taken. In term of its performance, hybrid turbo codes performed slightly better performance than Parallel Concatenated scheme since it could achieve BER = 10^{-5} at 15 iterations compared to parallel which achieved the same bit error rates at 18 iterations. When compared to Serially Concatenated scheme, generally, hybrid concatenated turbo codes performed similar performances as serial at high bit error rates. But at very low bit error rates, serially concatenated turbo codes are still remain better.

About the decoder of turbo codes in general, turbo decoder consists of several decoder components. During the decoding process of turbo decoder, the extrinsic information, which is one of the information that was generated by decoder component, is passed from one decoder component to other decoder components by following the sequence. The sequence of those decoder components is known as decoding schedule. The extrinsic information is passed through the decoding schedules iteratively, before hard decision processed is done at the end of decoding process.

The main idea of this paper is to investigate three decoding schedules of a hybrid concatenated turbo code system in term of their Bit Error Rate and Frame Error Rate performances. Therefore, Hybrid Concatenated Turbo Code simulation system that has been designed and developed are described in the beginning of this paper. Then it is followed by the explanation of simulation works that have been done to those designed decoding schedules. In the next section, result of simulation are mentioned and described generally. Finally, some discussion on the obtained results and conclusion regarding this work are included.

2. Hybrid Concatenated Turbo Codes System Model

The purpose of Hybrid Turbo Codes System Model is to simulate turbo-coding process of digital data transmission through Additive White Gaussian Noise Channel. Hybrid Turbo Code System Model is basically consists of an encoder with hybrid architecture, Binary Phase Shift Keying (BPSK) Mapping, Additive White Gaussian Noise (AWGN) Channel module and Decoder module.

2.1 Encoder Specification

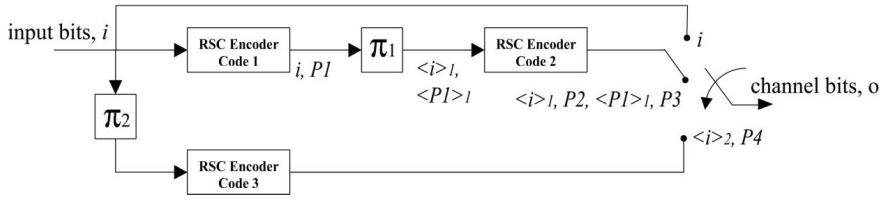


Figure 1 : Diagram of Hybrid Turbo Codes System Model Encoder

As shown in Figure 1 above, the encoder of Hybrid Turbo Codes System Model is designed with hybrid architecture with combination of Serial Concatenated Convolutional Code architecture and Parallel Concatenated Convolutional Code architecture. The Encoder consists of 3 identical Recursive Systematic Code (4,5/7) components (RSC Encoder Code 1, 2 and 3) and two random interleavers, which are Interleaver 1 (π_1) and Interleaver 2 (π_2). Size of Interleaver 2 is exactly the same as size of block length, while size of Interleaver 2 is double of the size of block length. For every single input, i, is inserted into encoder, then 7 bits concatenated codeword of the encoder is generated and denoted by,

$o = i, <i>_1, P2, <P1>_1, P3, <i>_2, P4$.[ϕ] However, bits that have been transmitted are $i, <P1>_1, P2, P3, P4$ bits. Therefore, the code rate of the encoder is $1/5$.

2.2 Decoder Specification

As shown in Figure 2 below, main components of the system decoder are three MAP decoder components, MAP1 MAP2 and MAP3. MAP1, MAP2 and MAP3 function to decode output bits of Encoder 1, Encoder 2 and Encoder 3 respectively. Every MAP decoder component implements BCJR Algorithm (Bahl et al, 1974). Besides MAP decoder component, decoder also consists of the same random interleavers as used in the encoder and its deinterleavers[♣]. In order to link all those MAP decoder components, interleaver and deinterleavers, several routes are designed as labelled by alphabetical characters. Arrows on routes mean possible direction of extrinsic information flow.

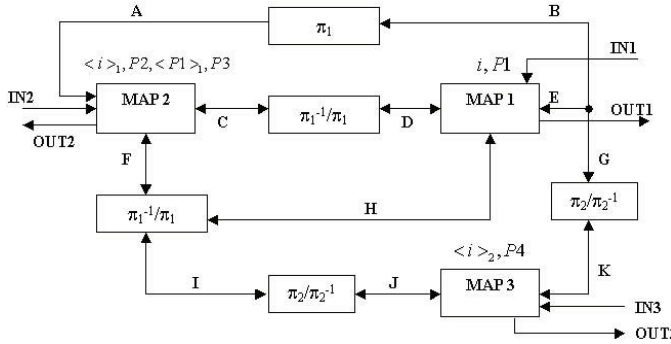


Figure 2 : The diagram of Hybrid Turbo Codes Decoder

2.2.1 Decoding Schedules of Decoder

Based on above architecture of system decoder, three decoding schedules (Schedule 1, Schedule 2 and Schedule 3) have been chosen. These decoding schedules are differentiated with number of different decoder component types that exist in one complete schedule. The justification of this selection is because every MAP decoder component has different decoding strength. MAP1 and MAP3 are expected to have same decoding strength. Differently with MAP2, which is expected to have higher decoding strength since it has higher number of input and higher number of decoded bits as labelled in Figure 2. Then, Figures 3 show diagrams of every chosen schedule with its decoder component sequence and routes.

[ϕ] $<i>_x$ means information bits interleaved by interleaver *x*, $<P1>_x$ means Parity 1 bits interleaved by interleaver *x*, *P1, P2, P3, P4* are parity bits from encoder 1, 2, 2, 3 respectively and i is information bits

[♣] Deinterleaver *x* is denoted by π_x^{-1}

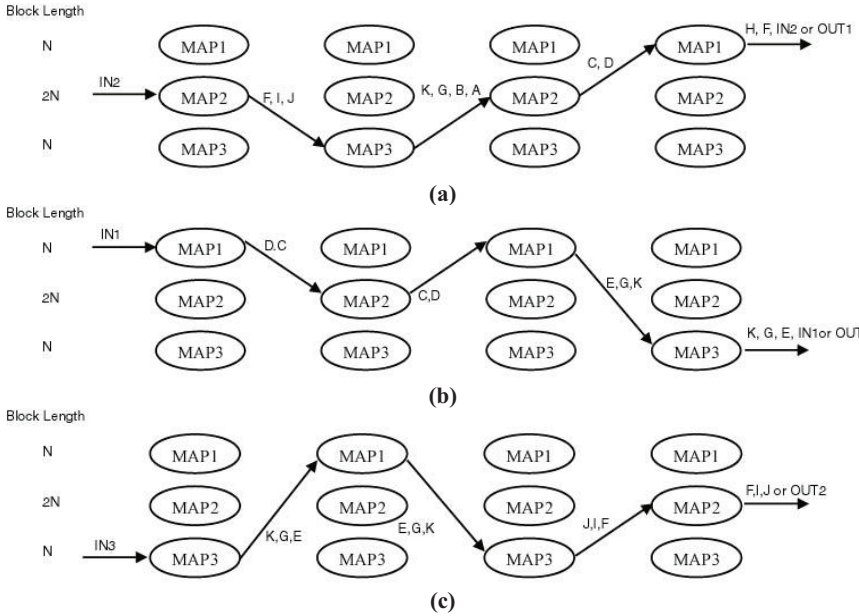


Figure 3 : Decoder component sequence and routes of (a) Schedule 1, (b) Schedule 2 and (c) Schedule 3

3. Simulation Works

In order to find out performance of those decoding scheduling for various conditions, every decoding schedule is simulated with different sizes of block length (500 bits, 1000 bits and 1784 bits) and a series of Eb/N0 values (from 0 dB to 6.5 dB). By using simulation system that has been designed as described above, every single simulation with a specific block length and specified Eb/N0 value, is executed 10000 times, with same random interleaver and deinterleaver but different sequences of input information bits. The new sequence of information bits is randomly generated for every single simulation. The random interleaver and deinterlever are only changed when size of block length is changed.

4. Simulation Results

Two error correction performance measurement units that are considered in this work are Bit Error Rate (BER) and Frame Error Rate (FER). By definition, BER is the rate of number of information bits in error over total number of information bits transmitted. And, FER is the rate of number of frames that contains error over total number of frames transmitted. Based on BER graphs that have been plotted, generally, decoder has reached its maximum performance at iteration 3. Therefore, iteration 3 is chosen to be the reference iteration for decoder scheduling performance comparison in Figures 4 and Figures 5 below.

4.1 Schedules BER Performance

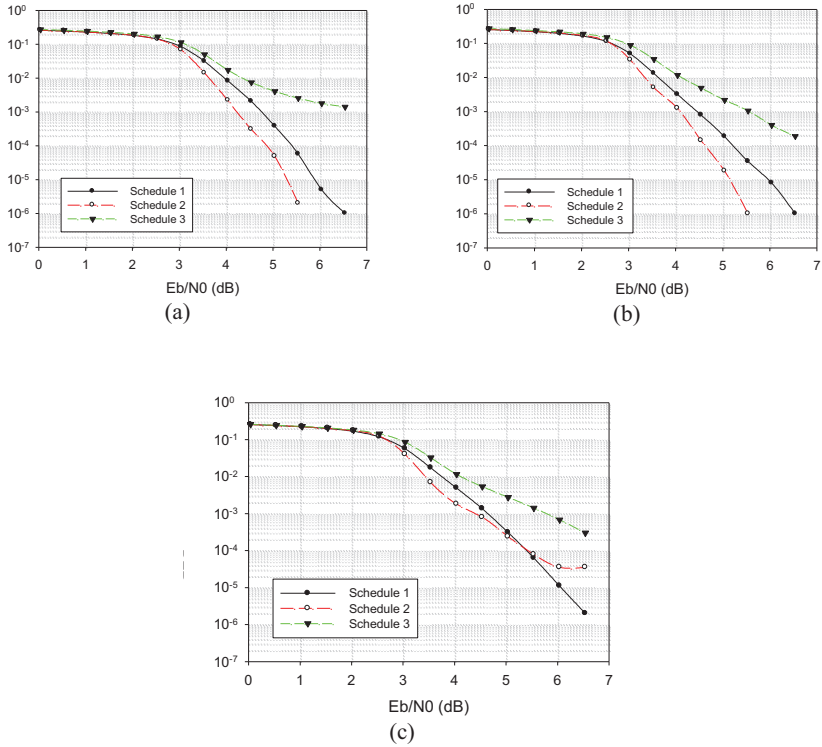


Figure 4 : Schedules Bit Error Rate (BER) Performance Comparison at iteration 3 with block length (a) 500 bits, (b) 1000 bits and (c) 1784 bits.

Figure 4 shows BER performances comparison of these three schedules with block length 500, 1000 and 1784 bits. It seems that Schedule 1 performed as the second position schedule after Schedule 2, which is considered as the best schedule that showed better performance with block length 500 and 1000 bits in term of coding gain performance. However, with block length 1784 bits, Schedule 2 performance was dropped because of error floor appearance at Eb/N0 higher than 4 dB at BER above 10^{-5}. Schedule 1 performance seems constant with no error floor observed for all block length used. And for Schedule 3, it seems to be the worse one because of its poor performance in both, coding gain and error floor appearance with all block length used

4.2 Schedules FER Performance

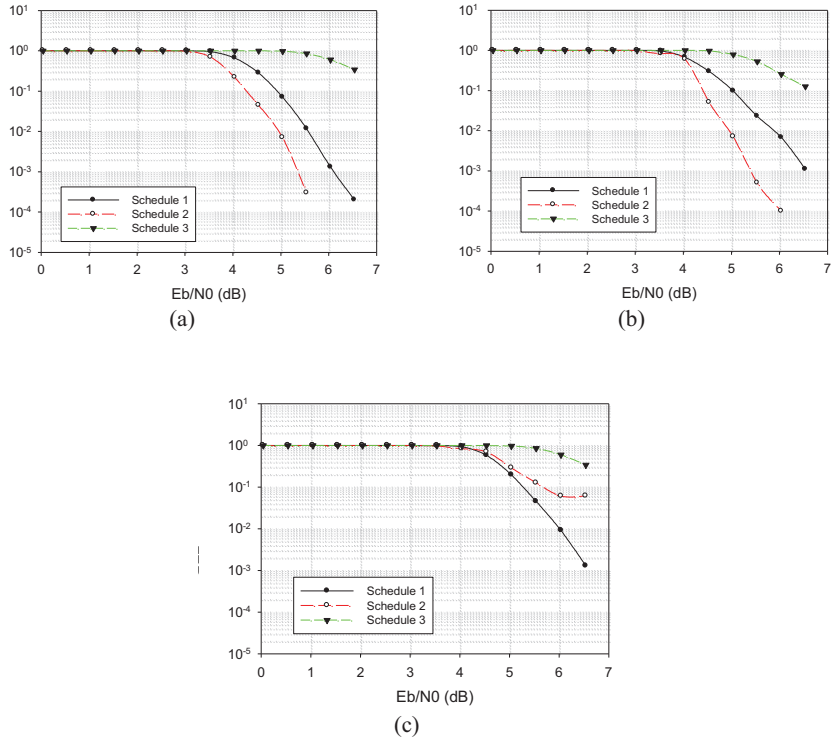


Figure 5 : Schedules Frame Error Rate (FER) Performance Comparison at iteration 3 with block length (a) 500 bits, (b) 1000 bits and (c) 1784 bits.

For comparison of FER performance as shown in Figure 5, Schedule 2 showed better performance in term of coding gain performance with block length 500 and 1000 bits. However, its performance seems to drop when block length 1784 is used, because there is an error floor appeared at Eb/N0 above than 5.5 dB at FER 10^{-1} region. And for Schedule1, its coding gain performances with block length 500 and 1000 are not as good as Schedule 2, however, Schedule 1 coding gain performance is better with block length 1784 bits and also there are no error floor observed for its performance lines for all block length used. And finally for Schedule 3, it seems to have worse performances among those three decoding schedules because of its poor coding gain performance and appearance for all block length that have been used.

5. Discussion

According to comparisons that have been done among those three decoding schedules, Schedule 2 seems to be the best schedule in term of coding gain performance. However, if performances at lower BER or FER are considered, Schedule 1 seems to be the schedule that would perform better. Because, there are no error floor observed on Schedule 1 performance as far that have been observed. It is believed that error floor of Schedule 1 is at lower BER and FER level compared to other schedules.

Besides that, decoder of this Hybrid Concatenated Turbo Codes system seems to achieve its maximum performance in BER measurement at a few iterations, which is at iteration 3 in most of the observation of this work. Afterward, it seems that the decoder starts to have convergence problem for next iterations. Based on performances graphs that have been plotted, it is believed that Schedule 3 experienced worse convergence problem and is followed by Schedule 2. Schedule 1 seems to have little convergence problem.

As mentioned previously, every decoder component has different decoding strength. In this system decoder, MAP2 has higher decoding strength compared to the others. And Schedule 1 is designed to have MAP2 decoder component as its dominant decoder. Instead of performed better in term of coding gain, Schedule 1 is believed to have its error floor at lower level and less convergence problem compared to other schedules. It seems like dominant decoder component could influence the level of error floor appearance and convergence problem of a schedule. It is believed that with higher decoding strength of dominant decoder, lower error floor is expected to appear and lower convergence problem is believed to happen.

6. Conclusion

In this work, a Hybrid Concatenated Turbo Codes simulation system has been designed and developed in software implementation. The encoder of the system is designed by combining serial and parallel schemes of concatenated convolutional codes. The system encoder consists of three identical RSC (4, 5/7) code components and two components of random interleavers. Noise channel that is used in this simulation system is Additive White Gaussian Noise (AWGN) Channel. The type of decoder components that are used in the system decoder is MAP decoder.

Based on the decoder that has been designed, three decoding schedules have been chosen, which every decoding schedule has different MAP decoder component as their dominant decoder. Every schedules are simulated with three different sizes of block length, 500, 1000 and 1784 of bits, for 0 dB to 6.5 dB of Eb/N0.

The decoder seems to find its maximum BER performance at very little number of iterations. Most of the result showed that decoder has achieved its maximum BER performance at iteration 3 before having convergence problem afterward. In term of FER, decoder performance lines seem to be steeper as the number of iteration is increased. Schedule with higher decoding strength is believed to have its error floor at lower level of BER and FER, and also believed to perform with less convergence problem compared to other schedules. It is

believed that with higher decoding strength of dominant decoder, lower error floor is expected to appear and lower convergence problem is believed to happen.

7. References

Bahl, L.R., Cocke, J., Jelinek, F. and Raviv, J. (1974) *"Optimal decoding of linear codes for minimizing symbol error rate"* IEEE Trans. Inform. Theory, vol. IT-20, pp. 284–287, Mar. 1974.

Bennedetto, S., Divsalar, D., Montorsi, G. and Pollara, F. (1996) "*Serial Concatenated of Interleaved Codes: Performance Analysis, Design, and Iterative Decoding"*, TDA Progress Report, Jet Propulsion Laboratory, August 1996, pp. 1-26.

Berrou, C., Glavieux, A. and Thitimajshima, P. (1993) *"Near Shannon limit error-correcting coding and decoding: Turbo-codes"* in Proc. ICC'93, Geneva, Switzerland, May 1993, pp.1064–1070.

Divsalar, D. and Pollara, F. (1997) *"Hybrid Concatenated Codes and Iterative Decoding"* TDA Prog. Report 42-130, Aug. 15, 1997.

Combined Data Compression and Error Correction

E. Venkatasubramanian and A.M. Ambroze

School of Computing, Communication and Electronics, University of Plymouth, UK
e-mail: M.Ambroze@plymouth.ac.uk

Abstract

Coding techniques are fundamental to digital communications systems and have today matured into complex compression and error control algorithms. In recent years there has been an increasing demand for efficient and reliable data transmission and storage systems. As a result of this there has been significant improvement in the way data is transmitted and stored.

This project investigates some of the contemporary techniques related to compressing, encoding and transmitting data over a noisy communication channel.

Source data in the form of text files are compressed using LZW compression algorithm. Different LDPC codes are used to add error correction. This is done by reducing Hamming matrices using Gaussian reduction method, and parity bits are computed and appended to the data. The resultant data is subjected to noise by introducing errors. The errored output is decoded and decompressed. The results and the processes are compared and plotted, and the various outcomes are studied and analysed. The performances of various compression ratios in conjunction with different coding rates are tested.

The entire software is written in 'C'. The software offers flexibility to test various parameters and offers extensibility and adaptability to allow addition and modification for future works.

Keywords

LDPC, Hamming, LZW, Bit Error Rate, Probability of Error, Coding Rate.

1. Introduction

1.1 Data Compression

Data compression is the representation of data by more efficient codes. Data compression results from the elimination of redundant fields of information while representing the data elements in the remaining fields with as few logical indicators as possible. The more we observe the actual data (both stored as well as transmitted), we can see that the same data can be represented in more efficient way by exploiting the structure of data and representing the redundancies with effective codes that would occupy lesser space and can be transmitted faster. The concept of redundancy is central to data compression. Data with redundancy can be compressed. Data without any redundancy cannot be compressed.

The theory of data compression was originally formulated by Shannon in his 1948 paper, "A Mathematical Theory of Communication". Data compression can be divided into two major families: lossy and lossless.

Lossy data compression concedes a certain loss of accuracy in exchange for greatly increased compression. It proves effective when applied to graphic images and digitized voice (Nelson and Gailly, 1996) as these are not sensitive to minor changes.

Lossless compression is applied to those types of data, which are sensitive to changes and cannot afford to loose information as this could be catastrophic. It would need an exact duplicate of the input data after compression and decompression. There are numerous methods by which we can compress a source without losing the actual information. Shannon, in his paper in 1948 established that there is a fundamental limit to lossless data compression.

1.2 LZW(Lempel-Ziv-Welch) Compression and Decompression

This project has implemented the most commonly used LZW compression technique also called the Lempel-Ziv-Welch compression, which is a renowned lossless compression technique and widely used in compressions like Winzip. The original Lempel-Ziv approach to data compression was first published in 1977. Terry Welch's refinements to the algorithm were published in 1984 and it became LZW compression.

The LZW compression algorithm is the basic of compression techniques like PKZIP and the Unix compress and is supported by the GIF and TIFF graphics formats. It adapts to the source statistics, so no prior knowledge of the source is required (Wade, 2000). It compresses by finding repetitions of strings of symbols in the source data. These strings are assigned a fixed length code which is usually considerably shorter than the length of the string. The codes corresponding to each string and the strings are stored in a lookup table. The codes are assigned to the string in such a way that the table does not need to be transmitted. The decompression algorithm can construct the table from the compressed data.

The code that the LZW algorithm outputs can be of any arbitrary length, but it must have more bits in it than a single character. The first 256 codes (when using eight bit characters) are by default assigned to the standard character set. The remaining codes are assigned to strings as the algorithm proceeds.

1.3 Error Correction

A major concern for the designers in communication systems is the control of errors so that reliable reproduction of data can be obtained. The pioneering work on reliable communication over noisy transmission channel was carried out by Claude E. Shannon in 1948. Shannon’s central theme was that if the signalling rate of the system is less than the channel capacity, reliable communication can be achieved if one chooses proper encoding and decoding techniques (Lin and Costello, 1983). The design of good codes and of efficient decoding methods was first initiated by Richard W. Hamming, a theorist at the Bell Telephone laboratories in the 1940s. From then on, the concept of error correction codes has evolved and many powerful codes are currently is use. Error correcting codes protects the data by adding redundant bytes to the original data in a systematic way, so that the actual data (and the information within) can be retrieved with minimum loss.

1.4 Hamming & LDPC Codes

Richard W. Hamming is best known for his work on error- detecting and error- correcting codes which is famously known as the Hamming codes. His fundamental paper on this topic appeared in 1950 and with this he started a new subject within information theory. Hamming codes are of fundamental importance in coding theory and are of practical use in computer design.

The throughput of a data link operating in a noisy channel can be increased by the ability of the receiving station to correct transmission error. This is called as Forward Error Correction (FEC). Forward-error correction is valued in communication links because it allows for virtually error free communications over a noisy channel. FEC improves system capacity by permitting high data rates within the communications link while providing improved transmission power efficiency.

Hamming codes provide for FEC using a "block parity" mechanism that can be inexpensively implemented. In general, their use allows the correction of single bit errors and detection of two bit errors per unit data, called a code word.

The fundamental principle embraced by Hamming codes is parity. Hamming codes, as mentioned before, are capable of correcting one error or detecting two errors but not capable of doing both simultaneously. We may choose to use Hamming codes as an error detection mechanism to catch both single and double bit errors or to correct single bit error. This is accomplished by using more than one parity bit, each computed on different combination of bits in the data.

For the past few years, turbo coding was discussed and touted as the key FEC technique for improving channel performance. Now, a new technique, called low-density parity check (LDPC), is emerging and could replace turbo coding as the FEC of choice by taking designers even closer to the Shannon Limit.

LDPC codes were first invented by Robert Gallager in his 1963 MIT PhD dissertation. Since then it was ignored for a long time until recently rediscovered (Sun, 2003) and now poised to become the standard in error-correction for applications from cell-phones to inter-planetary communication.

An LDPC code is a linear error-correcting code that has a parity check matrix H with a small number of nonzero elements in each row and column. Although LDPC codes can be defined over any finite field, the majority of research is focused on LDPC codes over GF(2), in which "1" is the only nonzero element. The code is the set of vectors V such that $HV^T = 0$.

To sum it up, the structure of a LDPC code is completely described by the parity check matrix H. The capacity of correcting errors in a codeword is determined by the minimum distance d_{min} where d_{min} is the least number of columns in H that sum up to 0.

The matrix H is called a sparse matrix as the number of 1's in each row and column are very few. There will be around 1% of 1's in the total matrix.

2. Results and Discussion

2.1 Different coding rates with same Compression ratio

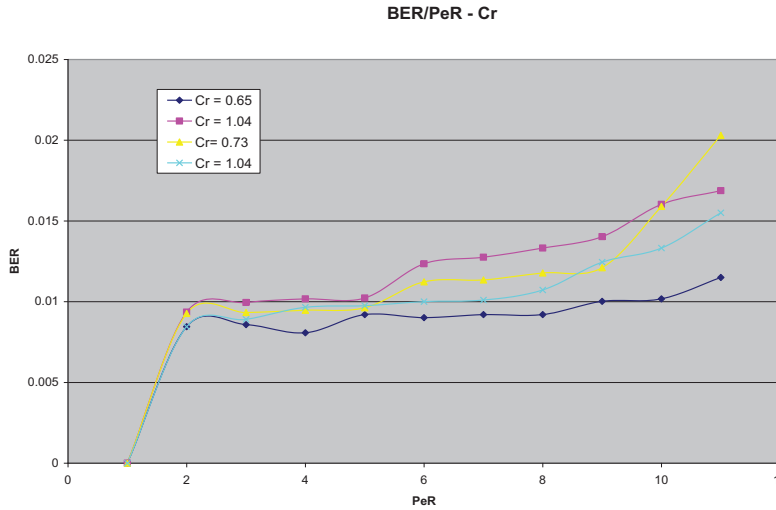


Figure 1 : Compression Ratio C=0.52 with different Coding rates R

The above graph depicts the performance of different C_R. C_R is nothing but compression rate C over code rate R given by,

$C_R = C/R$

For example, in the graph above, the $C_R = 0.65$ is obtained with compression ratio 0.52 and code rate 0.8, .i.e. 0.52/0.8 = 0.65.

R is the code rate of the matrix obtained by the equation R = k/n.

The value of C_R shows the combination of the best suited code rate along with the best compression. The best compression result obtained previously has been used, which is C=0.52 for all the iterations. The various matrix files considered are

1] 1024/1280 R = 0.8
2] 1024/2048 R = 0.5
3] 1024/1434 R= 0.71
4] 100/200 R= 0.5

where, for [1] 1024 is the information 'k', 1280 is the block length 'n' and this block will have parities n-k = 256.

We know that the compressed data is enlarged when error correction codes are added to it. Hence this should be maintained at a level where it does not compromise on the compression ratio.

From the graph above, we can decipher that the best performance with respect to the BER Vs PeR as well as compression ratio is obtained with C_R = 0.65 which is obtained with Index Size = 14, C = 0.52 and R = 0.8.

One more factor that needs to be looked here is that when we use matrices with lower rates R, the size of data that needs to be transmitted increases significantly. This is evident with matrices of coding rate 0.5. The compression ratio in this case exceeds 1 which is a severe contradiction on the purpose of compression. Investigations like the above would lead us into striking the right balance between the data size and the code efficiency.

2.2 Comparison between Encoding and No coding

The following graph has been plotted to show the comparison between not having encoding and transmitting just the compressed data and using an encoding before transmission.

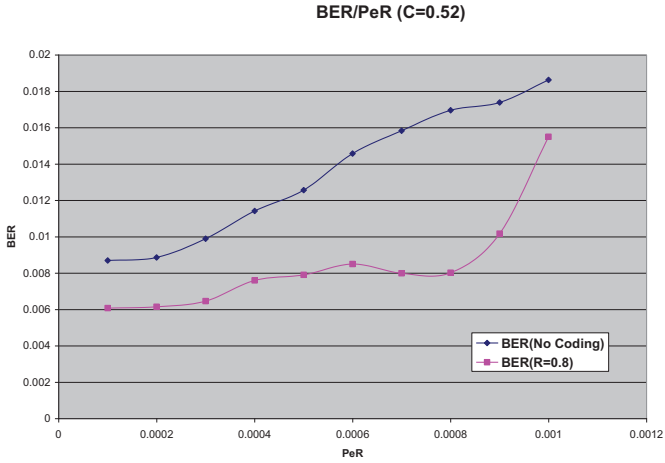


Figure 2 : Comparison between coding and no coding

In the figure above, the file was firstly compressed to compression ratio C = 0.52 and transmitted it (introduced noise). Again, to the same compressed file, I added encoding with R= 0.8, and transmitted the same with the same error probabilities. The result is phenomenal. Though using encoding and adding parities can compromise a bit of the compression, the reliability factor increases and compensates for this. Using an appropriate encoding scheme keeps the error level to a minimum, to a huge range of error probabilities. One obvious factor that can be seen in the graph is, after a certain level of error probability, when the PeR crosses a limit, the BER seems to shoot up. This is because the error gets compounded if the PeR crosses the threshold limit. This threshold limit depends upon the block length of the code and the size of the transmitted data. As we know already that Hamming codes are capable of

correcting 1 error per block, if the channel creates more noise than this, then the encoding/decoding process adds more error to it. This results in the BER level shooting up after the threshold level. So when we use an encoding scheme, the capacity and the behaviour of the channel must be studied well so that an appropriate code can be used.

3. Conclusion

What we have seen so far are a few pebbles from the river bed. The subject of information theory is vast, interesting and has undergone a significant amount of evolution. The main objective is to study the LZW compression and LDPC coding schemes, that are de facto in the industry, gain significant understanding of them and implement what we have learned using C programming.

In this section, the findings of the project and issues faced are summarised. Possibilities for further work on the topic are also discussed.

3.1 Findings

As we have seen so far, the BER Vs PeR graphs enables the selection of appropriate compression ratios and coding rates to be used for reliable transmission. A wise and cost effective approach can be arrived at using these analysis methods. The following factors can be inferred from the results arrived.

Best compression ratio can be obtained from sources with maximum repetitions and by using the appropriate Index size to pack the bits. Error correction should be able to not only correct errors, but also should not add too much of data to the source which would affect the overall compression ratio (including the encoding). BER/PeR performance depends on both the coding rate of the encoding and the compression ratio.

One of the main factor found was, the threshold limit set by the block size of LDPC codes. If the error limit crosses the threshold, the coding process would add even more errors to this and the BER would escalate rapidly. This is a limitation with respect to the LDPC codes.

Firstly, different Index sizes were experimented and the best compression ratio C = 0.52 was selected. Over this matrices with different coding rates were checked. The combined analysis of different coding rates with compression was found to be very useful in arriving at the best combined performance of compression and coding. From section 3.6, we were able to see that the appropriate matrix for the source file was the 1024/1280 with coding rate 0.8. This not only added fewer overloads to the data but also outperformed the other bulky matrices with best BER/PeR performances.

3.2 Further Work

For want of time, the study could not include the handling of input types other than text files. In future, further input type files should be tested. Currently for the tables and some arrays, fixed memory has been used. It would be desirable to change these to dynamic memory allocation. Further, testing a number of file types with different compression schemes and

coding rates so as to draw a broader conclusion about the performances would be appropriate. Finally, future work should test and analyse the results of a few other encoding methods as well. In particular the hard decision error correction BCH codes and other soft decision error correction codes and try to compare it with that of LDPC and be able to appreciate the difference.

4. References

Gallager, R.G. (1963) "*Low-Density Parity-Check Codes*" Doctoral Dissertation for MIT.

Lin, S. & Costello Jr, D.J. (1983) "*Error Control Coding: Fundamentals and Applications*". Prentice Hall Inc., New Jersey.

Nelson, M. & Gailly, J.L. (1996) "*The Data Compression Book*". M&T Books, New York.

Shannon, C.E. (1948) *"A Mathematical Theory of Communication"*, http://cm.bell-labs.com/cm/ms/what/shannonday/shannon1948.pdf , (Accessed December 17, 2004).

Sun, J. (2003) "*An Introduction to Low Density Parity Checks (LDPC) Codes*", http://www.csee.wvu.edu/wcrl/public/slideldpc.pdf , (Accessed December 17, 2004).

Wade G. (2000) "*Coding Techniques: An Introduction to Compression and Error Control*". Palgrave, New York.

Section 2

Communications Engineering & Signal Processing

Sonic Data Acquisition System

S.K. Annamalai[1], M.Z. Ahmed[1], M.A. Abu-Rgheff[1] and R. Bourne[2]

[1] University of Plymouth, Plymouth, United Kingdom
[2] J&S Marine, Barnstaple, United Kingdom
e-mail: mahmed@plymouth.ac.uk

Abstract

A duplex telemetry system, capable of carrying large amounts of data from an in-water array of hydro phones to an inboard processor was investigated and novel methods to improve the communication link between an array of sensors in the water and a submarine were explored. This paper aims at improving the various aspects of the technology of the previously developed data acquisition system. The digital communication link with reliability (of BER 10^{-5}) was simulated by Simulink. Digital IIR filters were designed to identify the node from which the data is coming. The use of OFDM was explored and time synchronisation was also explored. RBS was found to synchronise the sensor clocks upto a μs scale of precision.

Keywords

Coherent array processing, underwater acoustics, hydrophones, OFDM, time synchronisation, RBS, NTP.

1. Introduction

This research provides some unique and traditional solutions to the problem of designing an efficient communication link between the array of sensors (hydrophones) and the control unit inside the submarine. The operational configuration of the sensor array and the submarine is depicted in the Figure 1.

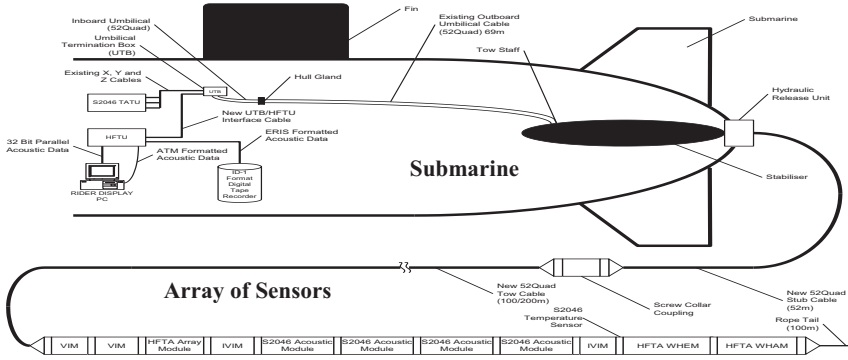


Figure 1 : TOWED ARRAY / S2046 Operational Configuration (Source: J&S Marines)

2. Background

Active research has been going on in this area with close collaboration between the University of Plymouth and J & S Marines. A duplex telemetry system which is capable of carrying large amounts of data from an in-water array of hydro phones to an inboard processor has been developed. This telemetry system is further capable of passing on the control signals from the inboard computer to the sensors to allow reconfiguration of the sampling rates, filtering and other built in monitoring.

2.1 CSTAR – Compact Submarine Towed Array

A brief introduction of the existing product is given in this section. CSTAR is a low cost, reelable, towed array system that offers the capability to deploy long towed arrays for blue water and littoral operations independently of shore of shore based support teams. Weir Strachan & Henenshaw, J & S Marine and QinetiQ are working together to develop the potential of the crustacean thin – line array originally developed as part of the UK Ministry of Defence Applied Research Programme (Source: J&S Marines). The system design incorporates a compact handling system that utilises advanced control and novel technologies.

The acoustic flow of the presently used system is shown in the Figure 2. This paper aims at improving the previously developed data acquisition system.

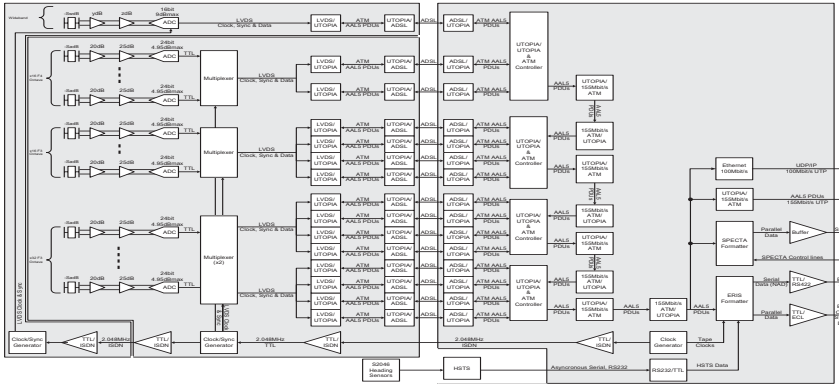


Figure 2 : TOWED ARRAY Acoustic Data Flow Diagram

3. Methodology

3.1 Simultaneous access

Multiple access schemes make it possible for many simultaneous users to use the same fixed bandwidth. Bandwidth is split into two to provide forward and reverse link. Major methods of sharing the available bandwidth to multiple users in a system are Frequency Division Multiple Access (FDMA), Time Division Multiple Access (TDMA) and Code Division Multiple

Access. Extensions of these major techniques are Orthogonal Frequency Division Multiplexing, hybrid TDMA and FDMA systems (Olofsson, M: 2002). The major techniques were analysed for their suitability in this application and OFDM was found to be more appropriate for this project.

3.1.1 OFDM

OFDM scheme is a modulation technique proposed for high data rate wireless communications. Over the past several years, orthogonal frequency division multiplexing (OFDM) has received considerable attention from the general wireless community and in particular from the wireless LAN (WLAN) standards groups. OFDM has been selected as the best waveform providing reliable high data rates by IEEE802.11a and ETSI BRAN. This popularity is further highlighted by the recent selection of OFDM by the IEEE 802.11g committee as the modulation for extending the data rates of the very successful IEEE 802.11b. This project applies this technique to the underwater communications in a unique way. It applies OFDM to the sensor network.

Suitability of OFDM

- Time division multiplexing in its several forms lends itself to the handling of digital data, but the low cost and high quality of available FDM equipment, make it a reasonable choice for many purposes.
- The peak power in TDM(A) is considerably larger than in FDM(A).
- In FDMA each user is allocated a separate unique forward and a reverse channel with different frequency bands. Hence FDMA systems are preferred over TDMA systems for the current application.
- However in FDMA the bandwidth is wasted due to the distance between two successive carriers. OFDM is an improvement of FDMA in which the carriers are orthogonal and the distance between two successive carriers is reduced.

OFDM

- OFDM provides very high data rate and is robust (Olofsson, M : 2002)
- divides bandwidth into many narrow band channels (10 to 8000)
- Orthogonal carriers. Each carrier has an integer number of cycles over a symbol period.
- Therefore spectrum of each carrier has a null at centre. Freq of other carriers in system. Hence carriers could be placed as close as possible.
- Narrow bandwidth of 1 kHz is typical. Hence overhead of FDMA is solved here.

The system performance for the different number of constellations and QAMS were investigated and the results are presented in the following sections.

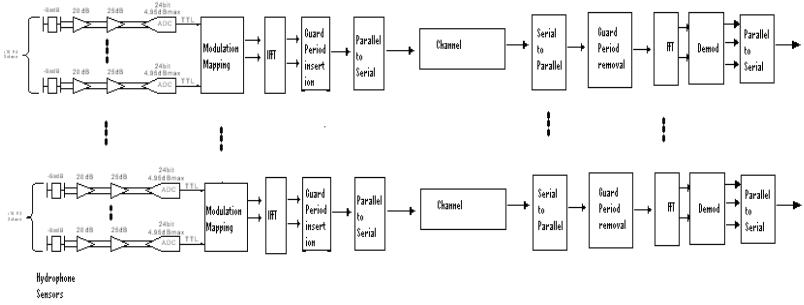


Figure 3 : OFDM block diagram

3.2 Digital Filter design

Any node from the array generates 24 bit value at a fixed frequency. The frequency varies from node to node but they are constant. Having a detector to identify the particular frequency will enable us to identify the node from which the data is coming. This is also important for J & S Marines. So digital filters were designed to meet this requirement and the results are presented in the following sections

Filters are used to alter or change some of the signals characteristics like wave shape, amplitude-frequency and phase frequency to suit our requirements. Improvement of quality of signal is one of the common objectives to do filtering. It reduces noise to improve the quality of a signal.

The functional block diagram of a typical IIR filter is shown in the following figure. It illustrates the dependence of the present output on past output samples.

3.3 Time stamp

The time difference between a pressure wave hitting different transducers is used to calculate the direction of the source, speed etc. It is important for the sensors to be sampled at right time and for the telemetry system to have a relationship between the data that is being passed up. Conventional solutions were analysed and the method based on transition of binary bit to pass on information was implemented in Matlab. This effectively improves the efficiency of the system.

In this method a single binary bit (SBB) is used to convey the time information. At t = 0 the binary ‘0’ is inserted in front of the data. During the next time instant at t=1 its complement (i.e a ‘1’) is inserted in front of the data. At t=2, SBB = 0 and for every successive time instant this procedure is repeated. This is depicted in the Figure 4.

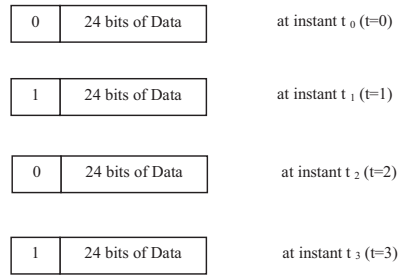


Figure 4 : Illustration of the variation of the proposed scheme to carry useful time information

3.4 Time Synchronisation

This section analyses the suitability of the various available algorithms to be applied to solve the problem of achieving time synchronization in an array of sensors in underwater environment. The time synchronization problem has already been investigated thoroughly in Internet and LANs. Several technologies such as GPS, radio ranging etc are currently used to provide global synchronization in networks. Complex protocols such as NTP have been developed that have kept the Internet's clocks synchronised. The time synchronization requirements differ drastically in the context of underwater sensor networks. The infrastructure in case of sensor networks is not as good as computer networks. Moreover GPS dose not work under water. Hence NTP / GPS are not sufficient to provide a solution to the existing problem. The sensor networks have diverse applications with diverse requirements.

RBS uses the receiver-receiver relationship and achieves μs precision of synchronisation by removing the largest source of non-deterministic latency from the critical path. This paper recommends the usage of RBS for obtaining time synchronisation in underwater array of sensors. The procedure to calculate the phase offset:-

1. The transmitter broadcasts m reference packets
2. The n receivers record the time individually at which the reference signal is observed
3. The observations of the receivers are interchanged.
4. The offset of any receiver i to any other receiver j can be computed as the average of the phase offset implied by each pulse received by both nodes i and j.

4. Results and discussion

4.1 Communication system

The end-end communication system was designed using the simulink and the various values obtained are tabulated as follows:

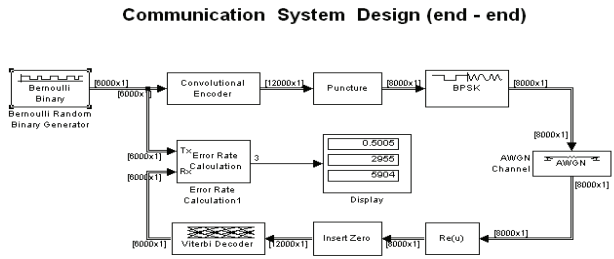


Figure 5 : Showing the BER rates for 6000 samples per frame

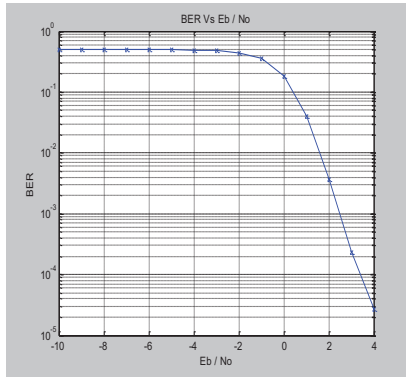


Figure 6 : BER Vs Eb / N

4.2 Digital IIR Filter design

Digital IIR design was implemented in matlab and the following frequency – magnitude response was obtained

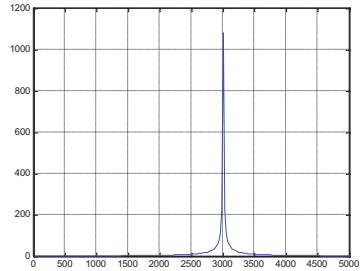


Figure 7 : Frequency – Magnitude response of a digital IIR filter for centre frequency of 3000Hz

For the centre frequency of 3000 Hz the transfer function was found to be

$$H(z) = \frac{1 - 1.4142\ z^{-1} - z^{-2}}{1 - 1.4124\ z^{-1} + 0.9974\ z^{-2}}$$

4.3 OFDM

The OFDM was implemented by using Matlab. The following results were obtained.

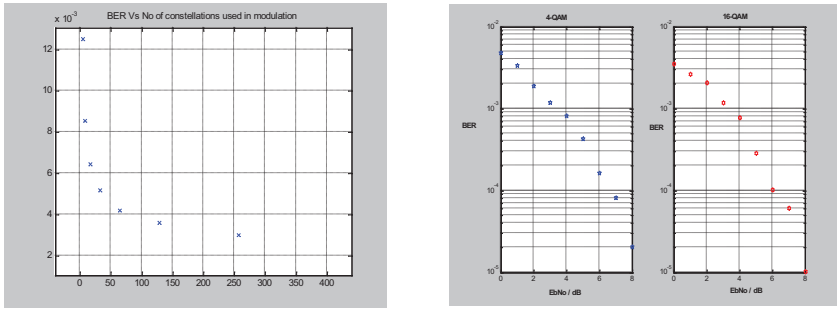


Figure 8 : Shows the BER Vs N_0 of constellations (for OFDM system) (right graph) and BER Vs Eb / N_0 for different QAM's (for OFDM system) (left graph)

Various multiple access schemes like TDM, FDM, WDM, CDMA and OFDM were analysed for a viable solution. OFDM has many advantages compared to its peers (as discussed above). It is not used currently in underwater communications and is proposed by the author as a method to increase the robustness of the communication link. This could lead to an increase of the achievable data rate.

4.4 Reference Broadcast Synchronisation

The precision of synchronisation can be calculated by the offset (Elson and et al: 2002) as follows

$$\forall i \in n, j \in n:\ \mathrm{Offset}[i,j] = \frac{1}{m}\sum_{k=1}^{m}(T_{j,k} - T_{i,k}).$$

where n : number of receivers , m : number of reference broadcast, $T_{r,b}$: r's clock when it received broadcast b.

The receiver group dispersion was calculated for various numbers of reference broadcasts and the following graphs are obtained from Matlab.

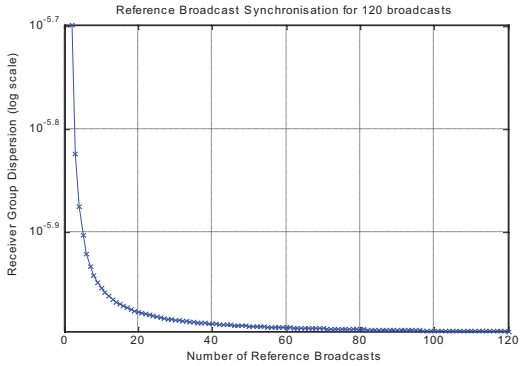


Figure 9 : The receiver group dispersion for different number of reference broadcasts

As the number of reference broadcast increases the precision of the synchronisation also increases. This is very evident from the simulated results.

5. Conclusion

This paper improves the technology of the previously developed data acquisition system of a submarine. The end-end communication system design was presented. OFDM has been used in a unique way. Digital IIR filters were designed to meet the specific requirements. Standard techniques like RBS, NTP and GPS were compared for the given channel and RBS was found to synchronise the sensor clocks up to a μs scale of precision. The other alternative technologies like ATM, Neutrinos and CAN bus topology were explored simultaneously to find a practical solution. Future work could be carried out to investigate the possibility of various multilevel access techniques and iterative Multistage decoding. New break through technologies like neutrinos and CAN bus could be explored for a data acquisition system of the future generation.

6. References

Elson, J., Girod, L, and Estrin, D. (2002) Proceedings of the Fifth Symposium on Operating Systems Design and Implementation (OSDI 2002), Report 020008.

Olofsson, M. (2002) 'Multiple access Methods for communication systems', Alba, Sweden.

Fuzzy Audio Signal Processing – Concept and Applications

S. Bawa, B. Hamadicharef and E.C. Ifeachor

School of Computing, Communications & Electronics, University of Plymouth, UK
e-mail: b.hamadicharef@plymouth.ac.uk

Abstract

A new audio signal-processing technique is proposed and investigated using Fuzzy Logic domain, to process audio signals from musical instruments tones. The new technique is intuitively understandable, simple to implement and is being developed into a user-friendly research tool implemented in MATLAB, and applied to various audio-processing tasks. The investigation considers audio signal processing defined by fuzzy rule based as would be defined by audio professionals. Applications include low pass filtering for noise reduction, adaptive filtering and features extraction, with particular interest for the noise reduction in pipe organ sounds during recordings.

Keywords

If-then based rules, fuzzy inference system, audio signal processing

1. Introduction

The paper is present the investigation of the new technique developed for processing audio signals from musical instruments. Audio data is being analysed in fuzzy logic domain based on a defined sets of rules (Russo, 1992a; 1992b; 1996) and (Russo and Ramponi, 1994). Low pass filtering is aimed at noise reduction. A model of *pipe organ sounds* is the instrument of interest for the experiment. In practice the impact of the noise affects the quality of the audio and the intelligibility of audio signal processing. This is especially true when you perform recording in large acoustic buildings like the cathedral and also for the listeners who have hearing disabilities (Umapathy and Parsa, 2003). The aim of the investigation is to assess the robust potential of using fuzzy logic for processing the audio signals. Particular interests are in development of tools for noise reduction (especially hiss noise removal), which is a problem during recordings of musical instruments such as pipe organ sounds. In this paper we show that fuzzy audio signal processing can be achieved good performance compared to classical techniques. Using Russo's definition of fuzzy signal processing, we exploit the potential of the concepts presented in (Russo, 1992a) and in (Russo, 1992b) and developed tools using MATLAB environment (Matlab, 1992) for processing the audio signals and music.

The remainder of the paper is organised as follows: Section 2 introduces the methods, which combines fuzzy logic within a signal-processing framework. Section 3 briefly describes fuzzy system and simulation. Section 4 presents illustrative examples. Results and discussion are held in Section 5 and Section 6 concludes the paper.

2. Fuzzy Signal Processing

The methods used in this paper consist of the following two: Fuzzy Logic and Audio Signal processing. The aim of this project is to use the new technique to process audio signals in a fuzzy domain, employing fuzzy logic and fuzzy sets techniques based on rules defined by human with special interest in noise reduction. The technique is evaluated and assessed on how much Root Mean-Square Error (RMSE) is been minimised

2.1 Basic Concept and Intelligent Fuzzy Systems

The basic concepts of fuzzy logic was first proposed by (Zadeh, 1965) to accommodate counting using linguistic approach that enables great flexibility in building and maintaining a workable rule base as defined by professional knowledge. The choice of a final action is based on the process of compromising of all the possible actions, often known as the defuzzification process. Over the last few decades, fuzzy logic has been applied successfully to many domains with examples being the a user friendly research tool for image processing (Russo 1992a), Electronic video camera images stabilizer (Egusa *et al*, 1995) and Generic optimisation of a fuzzy system for charging batteries (Surmann, 1996). In audio an application has been describe in which a real-time speech-music discriminator is describe making used of a fuzzy logic combiner. The recent research has moved the area of artificial intelligent systems in (Neural Networks, Genetics Algorithm and Fuzzy systems) to its boarders with professional knowledge couple with a bit of psychology. In this day and age fuzzy system has been used to model emotions, consciousness and awareness of individuals (Yanaru *et al.*, 1994). A lot of researches are being conducted world wide in different fields of science and technology using fuzzy domain.

2.2 Fuzzy Model and System

A typical fuzzy system shown in Figure 1 below has an input and output relationship and consist of the following main stages as.

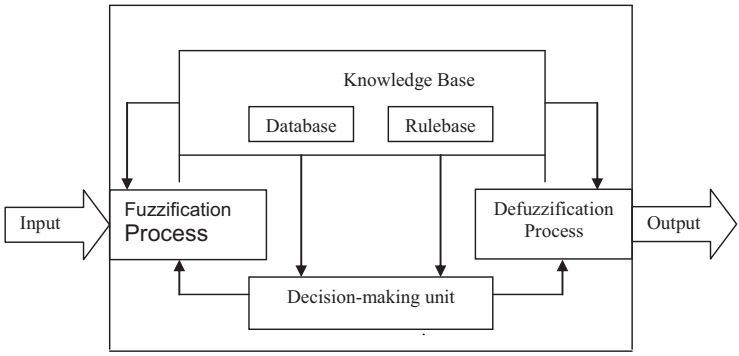


Figure 1 : Simple Illustration of a Fuzzy Model

- Fuzzification process: is the procedure of finding the membership degrees $\mu A_n(u_n)$ to which input data $u_1, u_2, \ldots u_n$ belong to a fuzzy set $A_1\ A_2 \ldots A_n$, which forms the antecedents part of the fuzzy rule.
- The fuzzy inference: is a working program that controls the functioning of the system. It contains an inference mechanism that maps up the input to the output
- Defuzzification process: this is a process where by a single-output value is calculated numerically for the fuzzy output variable(s) on the on the basis of the inferred resulting membership function for the variable(s). The commonly used methods are (i) the centre of the gravity method and (ii) the mean-of-max method
- Database unit**:** is the working memory in the set of fuzzy rules and contains current facts or past data. In some application the past data is used as additional source of knowledge to the knowledge base.
- Rule base unit: is the process natural to the human brains due to the underlying IF-THEN structure, which is forms, the basis of our daily language and natural logic. The rules are translated from linguistic into the calculus of basic inference algorithm.

2.3 Fuzzy Sets and Variables

Fuzzy logic receives its strength from its sets and variables. A fuzzy variable is made of two or more fuzzy sets. For convenient the fuzzy variables considered in this project is made of five sets with each having a linguistic terms “Very Low”, “Low” “Medium”, “High” and “Very High”. Between two sets there are some overlaps that allow the fuzzy logic to process the uncertainty or vagueness of the system. It is worth to note also that during processing there may be an occasion where a dead zone is created when the input hit an undefined region to produce some form of misleading output.

2.4 Fuzzy Rules and Inference

Under fuzzy logic domain any decision making process make use of linguistic rules this is normally exploited with the fuzzy inference. The rules structure involves many antecedents link by “AND” logic connectives and one (or more) consequent(s).

A basic example of a fuzzy rule including two antecedent clauses represented as follows:

```
IF (x is U) AND (y is V) THEN (z is W) . . .
```

Where U, V and W denote fuzzy sets associated to the quantities x, y and z, respectively. The antecedent clauses in the above equation are linked up by means of fuzzy AND operator to consequents. The new fuzzy filter developed is using following rule-base and one ELSE-rule that consider a situation where samples amplitude remains constant. The following statements or conditions describe three different scenarios.(Russo 1992B)

- *IF (value of the signal sample is high than those of the neighbours) THEN (decrease the amplitude of the signal.) (1)*
- *IF (value of the signal sample is low than those of the neighbours) THEN (increased the amplitude of the signal.) (1)*
- *ELSE do not change it (1)*

Since fuzzy reasoning represents a powerful framework for data processing that would usually require more than one rule, which resembles the mechanism for human decision-making, it allows high-level interface to complex problems. The overall rule-based structure for the propose project is given below with $U_{1,}\, U_{2,\ldots}U_n$ as the input samples and V represents the output samples:

IF (U_{11} is A_{11}) AND (U_{21} is A_{21}) AND (U_{n1} is An_1) THEN V_1'$_k$ is . . .

IF (U_{12} is A_{12}) AND (U_{22} is A_{22}) AND (U_{n2} isAn_2) THEN V_2'$_k$ is . . .

IF (U_{13} is A_{13}) AND (U_{23} is A_{23}) AND (U_{n3} isA_{n3}) THEN V_3'$_k$ is . . .

…….Else THEN V_n

The above rules are a function of sets of RuleList selected by MATLAB. The example of the RuleList matrix is given by.

```
RuleList = 5, 1 (1): 1
           4, 4 (1): 1
           3, 3 (1): 1
           2, 2 (1): 1
           1, 5 (1): 1
```

The aimed of the method, is to average the samples amplitude using the amplitude values from its neighbours, but simultaneously take care of the unwanted signals structures such as ripples, background noise etc in order to preserve the signal of interest.

2.5 Low-pass Filter

The low-pass filter developed is for filtering and suppressing the background noise during recording is the target of using the rule-base action aiming at reducing the samples whose values are "VeryHigh" or "VeryLow" than the sample in their neighbourhood.

Let consider Russo's definition for signal processing describe as follows: if we let s_0 be the input signal in a given range [0, N –1], $s_0 = s(t)$ is the signal to be processed at a given time (*t)* and $U = \{s_k\} = \{s_{1,}\, s_{2,}\, s_3, \ldots s_n\}$ be the set of neighbouring samples.

If we let the input variables to be defined as the difference of the samples variables given by the expression.

$$x_k = s_k - s_0 \qquad (1 \le s_k \le N\text{-}1)$$

Since $(0 \le s_k \le N\text{-}1)$, we have $(-N +1 \le s_k \le N\text{-}1)$. The output variables y is the processed term of the output, which must be added to obtain the new resulting samples z.

$$z_k = s_0 + y.$$

To demonstrate that the actual filtering has taken place we need to evaluate the fuzzy system that is the difference between the input data and the output data given below.

$$s_d = x_k - z_k.$$

To ensure that the technique works well different rules when applied for signal processing. The Root Mean-Square Error (RMSE) is evaluated over a given number of iterations and the results are compared with those obtained from the conventional methods. The least value of the RMSE obtained indicates that the set of rules used for that particular scenario provides the best results.

2.6 System Simulation

A simulation system is developed and performs in MATLAB 7.0 using fuzzy toolbox, which provides all the necessary functions to implement a fuzzy system and the additional I/O function to read and write the wave files, which allow the building of audio signal processing research tool.

3. Graphical User Interface

The graphical user interface have been developed is to provide a simple way of investigating of audio signal processing.

4. Results and Discussion

This section of the paper discusses the potentials and limitations of using Fuzzy Audio Signal Processing for processing audio signals. The main reason behind this study was to investigate and addresses the problems of background noise faced by audio professionals during recording musical instrument such as pipe organ tones (See Figure 2). The motivation was to design a fuzzy signal processing based filter to eliminate the noise. Set rules were selected using the algorithm and human intuition. Five sets of fuzzy membership function evenly distributed are used for fuzzification process of the input variables and for defuzzification of the output. The fuzzy output evaluation and signal difference (Diffsignal) were performed. Four types of signal were analysed. The performance of the 'FASP' was determine using Root Mean Square Error, (RMSE) by performing a given number of iterations under selected set of RuleList, various results were obtained and the least value of RMSE provides the best set of rules selected for the result. RMSE = 0.12. The limitation of the low pass filter is its inability to automatically extract the set of rules for the best result. A more advance adaptive and progressive method of filtering is required. A combination of Neural network, Genetic Algorithm and Fuzzy Logic algorithm using (Anfis method), where rules are extracted rather than selected, where training, testing and validation is performed over a given number of epochs would probably give a much lower valued of RMSE (See Figure 3). One other problem encounter was how to define a workable fuzzy membership function and set of rules for the application.

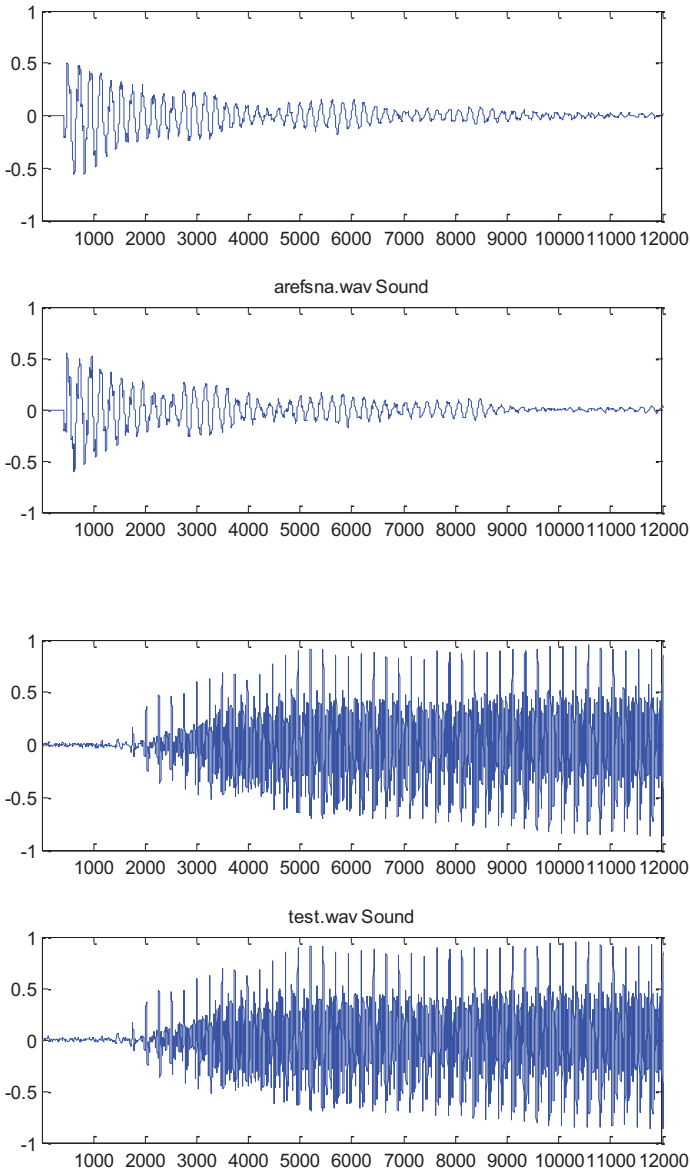


Figure 2: Results on pipe organ sounds

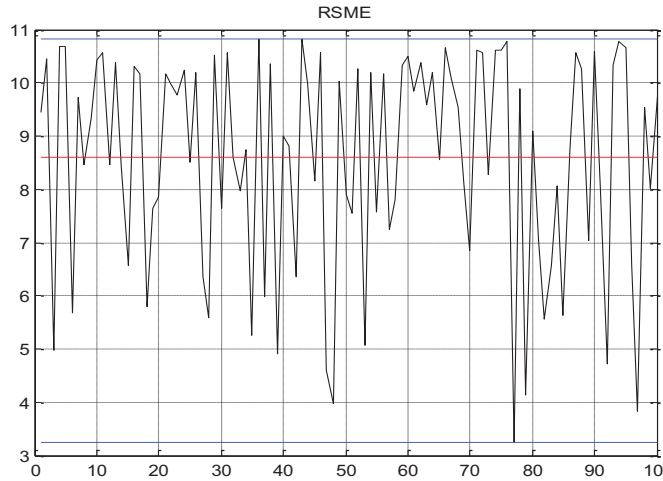


Figure 3 : RSME over 100 runs

5. Conclusion

In conclusion the new Fuzzy Audio Signal Processing technique is intended to process audio signals using fuzzy logic domain. The main aim is to develop a fuzzy-based low pass filter for noise reduction during recording of audio signals (pipe organ sound tones as examples). A choice of appropriate rules would mean a smooth filtering of noise without distortion of signal of interest. Result shows that fuzzy audio signal processing has potentials for processing many signal processing tasks. However more precise listening tests are necessary to decide whether the background noise reduction is pretty stationary after filtration is done.

Future work for the project is to consider more advanced optimisation technique that would exploit the fuzzy logic potentials, using a fuzzy-based adaptive (FASP) filter, to extract the best set of rules, train, test, fine tune and validate the audio signals. Complete the friendly graphical user interface research toolbox.

6. References

Aarts. R. M., and Dekkers. R. T. (1999) "A real time speech-music discriminator", *Journal of Audio Engineering Society*, Vol.47, No.9, pp. 720-725.

Breining. (2001) "A Robust Fuzzy Logic-Based Step-Gain Control for Adaptive Filters in Acoustic Echo Cancellation", *IEEE Transactions on Speech and Audio Processing*, Vol. 9, No. 2, pp. 162-167.

Cordon, Herrera and Villar. (2001) "Generating the knowledge base of a fuzzy rule-based system by the genetic learning of the data base", *IEEE Transactions on Fuzzy Systems,* Vol. 9, No4, pp. 667-674

Egusa, Y.; Akahori, H.; Morimura, A.; Wakami, N. (1995) “An application of the theory for an electronic video camera image stabilizer”, *IEEE Transactions on Fuzzy Systems*, Vol. 3, No. 3, pp. 351-356.

Goosh, Razouqi. Schumacher. J. H, and Celmins (1998) “A survey of recent advances in fuzzy logic in telecommnications networks and new challenges”, *IEEE Transactions on Fuzzy Systems*, Vol. 6, No.,3, pp. 443-447

Garibaldi J. and Ifeachor E. (1999) “Application of simulated annealing fuzzy model tuning to umbilical cord acid-base interpretation”, *IEEE Transactions on Fuzzy Systems*, Vol. 7, No. 1, pp. 72-78

Yinghua L., Cunningham, G.A., Coggeshall, S.V., (1997) “Using fuzzy partitions to create fuzzy systems from input-output data and set the initial weights in fuzzy neural network”, *IEEE Transactions on Fuzzy Systems*, Vol. 5, No. 4., pp. 614 - 621

Pham and Chem. (2002) “Article on some application of fuzzy logic in rule-based expert system”, *Expert system*, Vol. 19, No.4 pp. 208-223.

Russo, F. (1992a) “A user friendly research tool for image processing with fuzzy rule”, *Proceedings of the IEEE international conference on fuzzy system*, San Diego, USA, pp. 561-568

Russo, F. (1992b) “Fuzzy approach to digital signal processing: Concept and Applications”, *Proceedings of the IEEE 9th international conference on fuzzy system*, pp. 640-645

Russo, F. (1996) “Fuzzy Systems Instrumentation: Fuzzy Signal Processing”, *IEEE Transactions on Instrumentation and Measurement,* Vol. 45, No. 2, pp. 683-689

Russo, F., and Ramponi, G. (1994) “Fuzzy Methods for Multisensor Data Fusion”, *IEEE Transactions on Instrumentation and Measurement*, Vol. 43, No. 2, pp. 288-293

Surmann, H. (1996) “Genetic optimisation of a fuzzy system for charging batteries”, *IEEE Transaction on Industrial Electronics*, Vol. 43, No. 5, pp 541-548

Van De Ville, D., Nachtegael, M., Van der Weken, D., Kerre, E., Philip, W., and Lemaheiu, I. (2003) “Noise Reduction by Fuzzy Image Filtering”, *IEEE Transactions on Fuzzy Systems*, Vol. 11, No. 4, pp. 429-436.

Zadeh, L. (1996) “Fuzzy logic = counting with words”, *IEEE Transactions on Fuzzy Systems*, Vol. 4, No. 2 pp. 103-111.

Independent Component Analysis of Musical Instrument Sound

D. Chuckravanen, B. Hamadicharef and E.C. Ifeachor

School of Computing, Communications & Electronics, University of Plymouth, UK
e-mail: b.hamadicharef@plymouth.ac.uk

Abstract

In this research, a novel sound analysis technique method based on Independent Component Analysis is presented. ICA has the ability to extract signal components from sound blindly and that is to extract the harmonics or independent components from the Hammond musical sound. The Hammond organ sound is used because it is composed of nine harmonics and therefore, it will be observed how well can extract or separate these harmonics of this sound. Time domain or frequency domain representations are not enough to assess the quality of the extracted signals, quality measures are defined in this research and eventually implemented to compare quality of signal components produced by the ICA algorithms. Another part of the analysis assesses the robustness of the various ICA algorithms to noise. Moreover, preprocessing methods have been applied to achieve better extraction of the signal components. Some preprocessing techniques are promising.

Keywords

Independent Component Analysis, Audio Analysis

1. Introduction

The analysis and separation of audio signals into their original components is a crucial prerequisite to automatic transcription of music, extraction of metadata from audio, and speaker separation in video conferencing. (Uhle et al.,2003). Moreover, the well known cocktail party problem which is a good example of Blind Source Separation where by blind, it is meant that very little is known about the nature of the mixture or observed signals. In the Cocktail Party Problem, there are N people in a room speaking simultaneously and there are M microphones that are placed in different locations of the room. The problem is, given the recorded signals, it is required to determine the original speech signals, (Hyvarinen & Oja, 1999).

2. Theory of ICA

2.1 Introduction

The ICA algorithm is a technique that recovers a set of independent signals from a set of measured signals. It is assumed in this research that each measured signal is a linear combination of each of the independent signals and also, there are equal number of measured signals and independent signals. Each of the original independent signals is referred as s_i and each of the linearly combined mixed signals x_i where x is a column vector of n measured signals.

Each measured signal can be expressed as a linear combination of the original independent signals:

$x_i = a_1s_1 + a_2s_2 + ...a_ns_n$ …(1)

Therefore, the entire system of n measured signals can be expressed as:

$X = AS,$ ...(2)

Each row of X as shown in equation (2) is a set of readings for each signal x_i; each row of S is an original signal component and A is an N x N mixing matrix that generates X from S. In practice, from only the observed signals, it would be required to determine the matrix A and the original source signals S. The main goal of ICA is to search for specific features in S allowing a unique solution to be determined.

2.2 Constraints of ICA

[1] The real scale of S cannot be recovered

Any scale of the original signals can be normalized by the mixing matrix. Let S' be a set of signals with non-unit variances, and M be a diagonal scaling matrix such that A'S' results in a column whose elements have variance one, then it is shown in equation (1) that X = AS where S is the normalised signals.

$X = A' S' = (A'M^{-1})(MS') = AS$ ...(3)

Since the scale of the original signals cannot be recovered, signals with a variance of one by convention are found. The scaling matrix could be positive or negative and this is why ICA produces two answers for each independent component. However, any one of these two could be selected as sign (positive or negative) are not important in the extraction of these independent components.

[2] The order of the independent signals is not crucial:

A permutation matrix, that reorders the elements found in a vector, can show that permuted columns have equivalent solutions as shown in equation (4).

$X = A'S' = (A'P^{-1}) (PS') = AS$ …(4)

[3] Statistical Independence:

The main constraint that allows solution to be found is that the columns of S are statistically independent and the latter will be described in the next section. Each signal can be considered as a random variable and therefore, a matrix $W = A^{-1}$ (inverse of matrix A) is to be found as shown in equation (5) so that the signals in S are maximally independent.

$W\mathbf{X} = \mathbf{S},$ ...(5)

[4] Invertible matrix:

Lastly but not least, it is to be noted that A must be invertible for the equation (5) to be valid and the following equation is to be computed in the ICA algorithm:

$$A^{-1} = W \qquad ...(6)$$

2.3 Independence

Independence is a key issue for most ICA algorithms to work and without this feature, it is impossible to extract the original signals from the observed signals. Two random variables A and B are considered to define this property and these two variables A and B are independent if the conditional probability of A with respect to A is just the probability of A. In other words, it means that if a value of B is known, this value of B will tell nothing about A.

Moreover, another important feature that can be derived from statistical independence is that the mean of the product of any two functions f(A) and g(B) where $A \neq B$ is simply equal to the product of the mean of f(A) and the mean of g(B) and this is shown in equation (7).

$$\text{mean}\,(f(A)g(B)) = \text{mean}(f(A)) \times \text{mean}(g(B)) \qquad \ldots(7)$$

And also, the covariance between A and B is simply:

$$\text{cov}(A,B) = \text{mean}(A*B) - \text{mean}(A) \times \text{mean}(B) \qquad ...(8)$$

And now what happen when variables A and B are independent is that the covariance will be zero and this feature of zero covariance is used to find the columns of the demixing matrix W as used in equation (5).

3. Preprocessing techniques

3.1 Introduction

In section 2, the statistical principles underlying ICA methods have been explained. However, before applying an ICA algorithm on the data, that are the sound samples from the Hammond organ sound, it is important to do some preprocessing. The preprocessing of the data is done so that it is much easy to estimate the demixing matrix, W, if the measured signals have a zero mean, a variance of one and zero correlation. The data is preprocessed to meet these aforesaid criteria before the demixing matrix W is estimated. In this section, some preprocessing techniques are explained to make the problem of ICA estimation simpler.

3.2 Centering

The first preprocessing technique is centering. It is determined by subtracting the mean of a mixed signal (audio signal) vector **X**, as shown in equation (2), from each reading of that signal. This preprocessing is made to simplify the ICA algorithms. After estimating the mixing matrix A, with centered data, the estimation can be determined by adding the mean

vector of S, the original signals, back to the centered estimates of S. The mean vector of S is given by the product of the inverse of the mixing matrix and the mean **m** that was subtracted in the preprocessing is given by equation (9).

$$S = A^{-1}\mathbf{m} \quad \ldots (9)$$

where, **m** is the mean that was subtracted before preprocessing.

3.3 Whitening

Another important preprocessing technique that is to be considered is to whiten or "sphere" the observed variables. This means that before the application of the ICA algorithm and after centering, the observed vector **X** is transformed linearly so that a new vector $\tilde{\mathbf{x}}$ is obtained and this vector is considered as white where its components are uncorrelated and their variances are equal to unity. In other words, the covariance matrix of $\tilde{\mathbf{x}}$ equals the identity matrix and equation (10) illustrates it. (Hyvarinen, 1999)

$$E\{\tilde{\mathbf{x}}\tilde{\mathbf{x}}^T\} = \mathbf{I} \quad \ldots (10)$$

In equation (10), $\tilde{\mathbf{x}}^T$ is the transpose of the covariance matrix $\tilde{\mathbf{x}}$. The method which is used for whitening the data is the computation of the Eigen Value Decomposition (EVD) of the covariance matrix and the covariance matrix is given in equation (11).

$$E\{\mathbf{X}\mathbf{X}^T\} = \mathbf{E}\mathbf{D}\mathbf{E}^T \quad \ldots (11)$$

E is the orthogonal matrix of eigenvectors of $E\{\tilde{\mathbf{x}}\tilde{\mathbf{x}}^T\}$ and **D** is the diagonal matrix of its eigenvalues and it is represented as shown in equation (12).

$$\mathbf{D} = \text{diag}(d_1, d_2 \ldots d_n) \quad \ldots (12)$$

$E\{\mathbf{X}\mathbf{X}^T\}$ is determined from the mixture of signals. Now, whitening is done by this method defined by equation (13).

$$\tilde{\mathbf{x}} = \mathbf{E}\mathbf{D}^{-\frac{1}{2}}\mathbf{E}^T\mathbf{X} \quad \ldots (13)$$

The matrix $\mathbf{D}^{-\frac{1}{2}}$ is computed for each component and it is clearly explained by the following equation:

$$\mathbf{D}^{-1/2} = \text{diag}(d_1^{-1/2}, \ldots, d_n^{-1/2}) \quad \ldots (14)$$

Whitening transforms the mixing matrix into a new matrix Ã and by expanding equation (13), the relationship between the whitened observed signals, $\tilde{\mathbf{x}}$ and the new mixing matrix, Ã, is as follows:

$$\tilde{\mathbf{x}} = \mathbf{E}\mathbf{D}^{-\frac{1}{2}}\mathbf{E}^T\mathbf{A}\mathbf{S} = \tilde{\mathbf{A}}\mathbf{S} \quad \ldots (15)$$

The transformation done by the whitening process has produced a new matrix which is orthogonal and whitening is done to reduce the number of elements to be estimated from the original matrix A.

3.4 Illustration of whitening effect

To illustrate the effect of whitening, two random sources or signals A and B are mixed and at any point in time, the value of A is the value of a coordinate on the horizontal axis and the value of B is the value of a coordinate in the vertical axis. The joint distributions of the random sources A and B before mixing them, after mixing them and after the application of the whitening process are plotted in Figure 1. The length of data samples taken for the random sources is 2000.

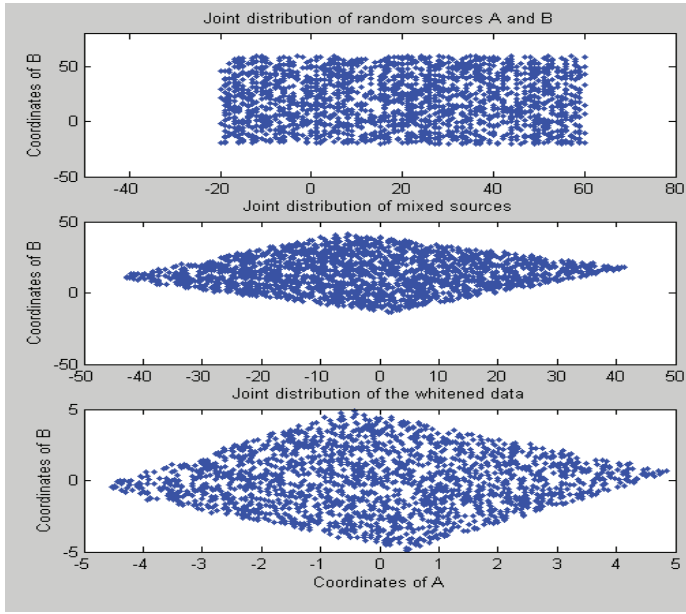


Figure 1 : Effect of whitening the mixed signals

The variance on both axes of Figure 1 is now equal and the correlation of the projection of the data on both axes is zero which means that the covariance matrix is diagonal and that all the diagonal elements in this matrix are equal. The job of ICA is to rotate this representation back to the original space of the joint distribution of A and B. Therefore, the whitening process is a linear change of coordinate of the mixed data. By minimizing the Gaussianity of the data on the projected on both axes, the independent components can be found.

3.5 Further Preprocessing

The success of ICA for the extraction of signal components from a mixture, depend also on some preprocessing steps such as filtering if the data consists of time-signals. Some experiments have been conducted and the results will be shown in section 5. Further preprocessing techniques have showed some promising results.

4. Sound Analysis

4.1 Tools

The toolboxes that have been used for this research are the ICALAB toolbox and the FastICA toolbox. The various stages that are available in the ICALAB toolbox are the preprocessing stage as the first one followed by the ICA stage where a particular algorithm is used to extract the signal components algorithm and then, there is the possibility at the final stage to do postprocessing on these extracted signal components as a means to do further analysis. The FastICA package for MATLAB is a program with parallel graphical user interface that implements the fixed-point algorithm for ICA. The graphical interface requires MATLAB 5 or 6 (Jarmo et al., 2004).

4.2 Data

The data used for analysis of the ICA algorithms consists of sound samples recorded from the Hammond organ. The Hammond sound was invented by Laurens Hammond (Vail, 2002). The Hammond organ uses a Tone Wheel generator to generate its characteristic sound. Each sound produced by the Hammond organ is composed of 9 harmonics and each harmonic is associated to a specific frequency. It should be noted that the proportions which the 9 harmonics are mixed are controlled by a set of nine harmonic drawbars. The drawbars have each one eight degrees of volume and the increasing number from 1 to 8 correspond to degrees of loudness where the number 1 is the softest and the number 8 is the loudest. In this research, the sound C2 which is the note number 13 on the 61 keys of the Hammond organ is used for analysis, (Nelson, 1999).

4.3 Measures of Quality

4.3.1 Introduction

To assess the quality of the signal components that are extracted from the observed signal, it is not enough to compare these real harmonics to those independent component components by visual representation on the time or frequency domain. Therefore, some quality measures are defined in this section.

4.3.2 Performance Index

The performance index is a good means to evaluate the quality of the extracted independent components from the Hammond organ sound. The performance index is defined by the following equation (Choi *et al*, 2001):

$$\mathrm{PI} = \frac{1}{n(n-1)} \sum_{i=1}^{n} \left\{ \left(\sum_{k=1}^{n} \frac{|g_{ik}|}{\max_j |g_{ij}|} - 1 \right) + \left(\sum_{k=1}^{n} \frac{|g_{ki}|}{\max_j |g_{ji}|} - 1 \right) \right\} \quad \text{…equation 4.3.2}$$

In equation 4.3.2, g_{ij} is (i, j) the element of the Global matrix G and $\mathbf{max_j}\ |\mathbf{g_{ij}}|$ represents the maximum value among the elements in the i^{th} column vector of the global matrix G. This performance index indicates how far the global matrix is from the permutation matrix. If the performance index is zero, this means that the extraction of the components is perfect. In practice, a performance index of 0.001 or less means the extraction is good. (Choi *et al*, 2001)

4.3.3 Measures of distortion

Several measures of distortion have been used in this research to evaluate the performance of the ICA algorithm. The total distortion includes interference from the other sources as well as noise and algorithmic artifacts. These measures of distortion of sources take into account the gain in amplitudes: they are the Source to Distortion Ratio (SDR), the Source to Interference Ratio (SIR), the Source to Noise Ratio (SNR) and the Source to Artifacts ratio. For a matter of simplicity, no noise has been added to the data samples of the Hammond sound.

5. Results

In this analysis, multistage filter is used as the preprocessing technique and values of the performance index for various ICA algorithms have been computed and the sound C3 with drawbars preset at 006876540 is used as sound data for analysis. Figure 2 shows the curves produced by the ICA algorithms. It is observed in this particular case that AMUSE algorithm has performed well the separation with performance index less than 0.001 followed by Fixed – Point ICA (FPICA) with values of PI ranging between 0.001 to 0.002- it is still good performance and Figure 3 shows the source to distortion ratio curves where each independent component is compared to its corresponding harmonic in terms of three aforesaid ratios.

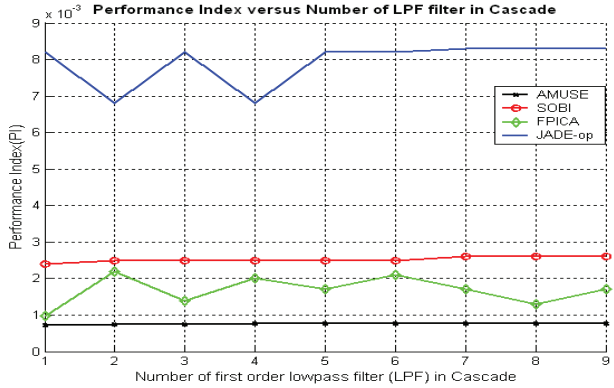


Figure 2 : PI versus number of LPF filter in Cascade

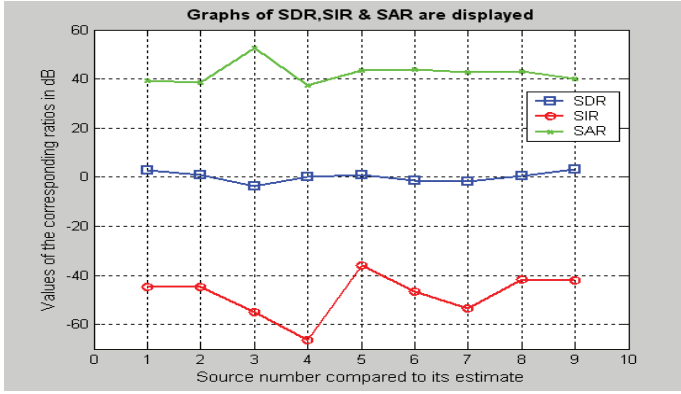


Figure 3 : Graphs of SDR, SIR and SAR are displayed

6. Conclusions

It is deduced from Figure 3 that distortion is mainly caused by the interference among the signals with SIR values negative, meaning there is great distortion in the signal components that are extracted and the distortion caused by interference is greater than that caused by the artifacts of the ICA algorithm and the algorithm used is FPICA. The SDR values have shown that low frequency component with high amplitude experiences more distortion than a high frequency component with low amplitude. Moreover, very low performance index values have been obtained when multistage low pass filter is used and thus very low performance index values indicate that very good quality signal components are extracted.

7. Future works

The postprocessing stage of the ICA analysis should be further extended to improve the quality of extracted signals. The use of PEAQ (Perceptual Evaluation of Audio Quality) is suggested since it has quality metrics to measure the quality of the signal components with great accuracy.

8. References

Choi, S., Cichocki, A. & Beloucharni, A. (2001) "Second Order Nonstationary Source Separation", *Journal of VLSI Signal Processing*, pp1-13.

Hyvarinen, A. & Oja, E. (1999) "*Independent Component Analysis: A Tutorial*", Helsinki University of Technology, Finland

Nelson, G. (1999) "*History of the Hammond B-3 organ*" http://theatreorgans.com/grounds/docs/history.html [10/07/2004]

Vail, M. (2002) "*The Hammond organ: Beauty in the B"*, pp. 5-19, 42-43.

Application of Signal Processing Techniques to Detect Extrasolar Planets

O. Decugis and A.M. Ambroze

School of Computing, Communications& Electronics, University of Plymouth, UK
e-mail: mambroze@plymouth.ac.uk

Abstract

Discoveries of extrasolar planets are of interest of study in Signal Processing because they tend to reveal weak signals from a very noisy environment and improve the precision of the more common detectors. In our work we simulated a typical signal that could be measured in the method of radial velocity and compared different processing techniques to reveal the periodic wobbling signals. We show that autocorrelation and adaptive filtering can be combined to extract the main parameters of the hidden signal of the planet.

Keywords

Astronomy, Radial Velocities, Signal Processing, adaptive filtering.

1. Introduction

Extrasolar planets are planets that have been discovered revolving distant stars. The detection of these planets is a new area research since the end of the 1990s, when astronomers imagined several ways to detect these distant worlds. The success of the detection depends not only on the accuracy of the tools that are employed to detect but also on the capacity of the signal processing that is used with the measuring tool to reveal this signal. In the majority of the cases, the signal received from the star is noisy and can have more power than the desired one. The planet is expected to have a very weak influence on the star.

In this paper we first explain the different methods that are used to discover a planet around a distant star. Then we will show, while studying one of these methods that signal processing can help. The study will compare different techniques to cancel noise.

2. Methods of detection of extrasolar planets

The astrophysicians who study the light coming from the heavenly bodies have imagined several techniques to reveal from this light, the presence of a planet.

2.1 Radial Velocity and Astrometry

One way to discover the presence of the companion to a distant star is to study the gravitational influence it might have on the star. Each planet orbiting its star can be considered as a system of masses that attract each other. The two bodies will orbit around the barycentre of this system. Therefore the star could be seen wobbling around this point.

This phenomenon causes a Doppler shift on the stellar spectrum. The aim of the Radial velocity is to detect this effect. The Doppler shift for instance imparted by the Earth on the Sun was found at 0.1 m/s whilst Jupiter will cause it at 12.5 m/s. (Perrymann, 2000). The measure of the velocity shifts of that order can be done with a precise spectrometer. These spectrometers allow studying the light by wavelength. This method on this day contributed to the majority of the discovers made so far

The wobbling star might also be seen having a regular motion, induced by the planet. The method relying on astrometry will focus on measuring the position of the star relative to our solar system and the background of the sky. The astrometry relies on measure the parallax of stars over the year precisely enough for noticing its motions. This technique will be exploited in the future with satellites, because the precision required for the measurement cannot be achieved in the atmospheric environment. The mission GAIA of ESA that will be launched by the year 2010 will provide a precision enough to detect a planet of the size of Jupiter orbiting its star at 3 Astronomical Units, up to 700 light years.

2.2 Photometry

The photometry studies the variations in light of the star, in the magnitude. If the planet's orbit has favourable angle, the planet could pass in front of the star, causing a drop of its luminosity for a definite time. This is called a transit. In another circumstances, the planet on its orbit could have a certain separation that causes a gravitational lens phenomenon.

2.3 Imaging

Imaging consists on getting a light evidence of the planet, by reflection of the star light of the star by the planet. The Star and the planet must be enough separated to be resolved on a picture. Interferometers which are systems that can combine the light from several telescopes targeting the same object might have a resolving power sufficient, for this purpose. The image can be aided by image processing, while artificially removing the light coming from the star, to watch what remains from the supposed existing planets.

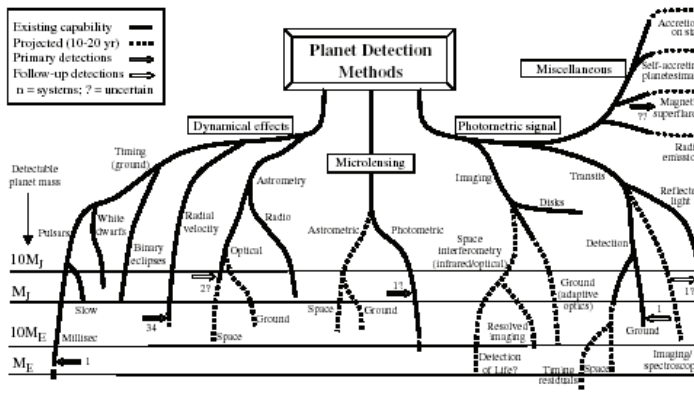


Figure 1 : Diagram of planet detection methods - Perryman(2000)

3. Simulation

3.1 Aims of the simulation

The simulation considered the evolution of one spectral ray of the star spectrum along time. The natural behaviour of the star without planet perturbation was simulated on the spectrum. For instance the star has a cyclic activity where it bursts energy with varying intensity. It was assumed that caused the main significant variation in the stellar behaviour. Other sources of noise coming from the nature of the light emitted the imperfections of the instrumentation, have various distributions. There were considered together as a global Gaussian source of noise. Also the level of noise emitted by the star can change according to its mass, size and surface temperature, and then a very brilliant blue star will be noisier than a yellow star like the Sun.

The Gaussian noise was then defined with a signal to noise ratio (SNR) and a standard deviation σ. Then adding the variation of the planet, it was compared when there is a planet and where is not. The methods studied were statistical and averaging, autocorrelation and then adaptive filters to cleanse the noise of the autocorrelation results. The simulation was made as if observations had been performed continuously in a long period of time of nearly 20 years; the aim of the simulation was essentially theoretical.

3.2 Equation used in the simulation

The radial velocity can be considered as a periodic variation that translates the velocity from an elliptic motion. Its amplitude is given by the radial velocity coefficient which is computed as follows, depending on the masses of the star and the planet, the period of their revolution around their system barycentre, and the eccentricity.

$$K = \left(\frac{2\pi G}{P}\right)^{1/3} \frac{Mp\sin(i)}{(Mp+Ms)^{2/3}} \frac{1}{(1-e^2)^{1/2}} \quad (1)$$ (Zeilik *et al.,* 1992)

The radial velocity coefficient K, here above depends also on the angle i at which the system is seen from Earth. G is the Universal Gravitational Constant.

The equation of the ellipse in phase domain was applied in time, as the angle that describes the planet around its star is performed during a certain time. The essential parameter that defines the ellipse is its semi major axis a, and its eccentricity e which is a measure of how stretched is the orbit.

$$r(\theta) = \frac{a}{1-e*\cos(\theta)} \Leftrightarrow r(t) = \frac{a}{1-e*\cos(t)} \quad (2)$$

To obtain the velocity equation in time of the ellipse motion function was derived.

$$v(t) = \frac{-a*e*\sin(t)}{(1-e*\cos(t))^2} \quad (3)$$

Finally we combined (1) and (3) to have the equation in time for radial velocity.

$$v_r(t) = -\left(\frac{2\pi G}{P}\right)^{1/3} \frac{Mp\sin(i)}{(Mp+Ms)^{2/3}} \frac{1}{(1-e^2)^{1/2}} * \frac{e\sin(t)}{(1-e*\cos(t))^2} \quad (4)$$

To simulate the natural behaviour of the star, we considered the period of its activity in a period comparable to the Sun's of 11 years the amplitude was chosen as common values reported by astronomers.

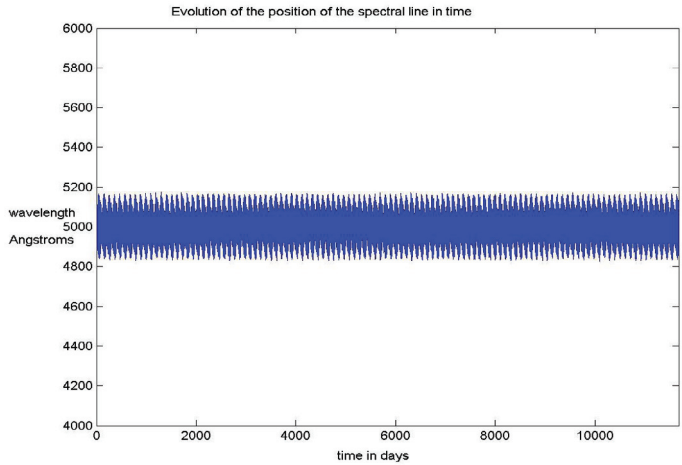


Figure 2 : Spectral ray evolving along time, example of a wide perturbation

4. Signal Processing Techniques

In this section we present the results and discussion of the results obtained in our programs.

4.1 Statistics and Averaging

One first signal processing method used consists in averaging the signal. The signal is scanned by bits of N samples and these are averaged by N, giving one value in stead.

Statistics were led in the cases of averaging and not averaging on the signal obtained by computing the probability density function (PDF) of the signal. It showed that averaging did not diminish the amount of noise from the width of the PDF lobe when there is noise was not reduced. It was explained by the big variation of the noise in our simulation that led to an invariant noise to averaging.

Although statistics proved that it was possible to detect a planet as we compared the ideal case when there is noise and when there is a planet and no noise. The order of size of the planets that can be discovered is beyond what we observe in our solar system and the planet has to

have a large eccentricity of nearly 1. Another condition is that the noise has to be low in power and variance to reveal a more common size planet like Jupiter or Saturn.

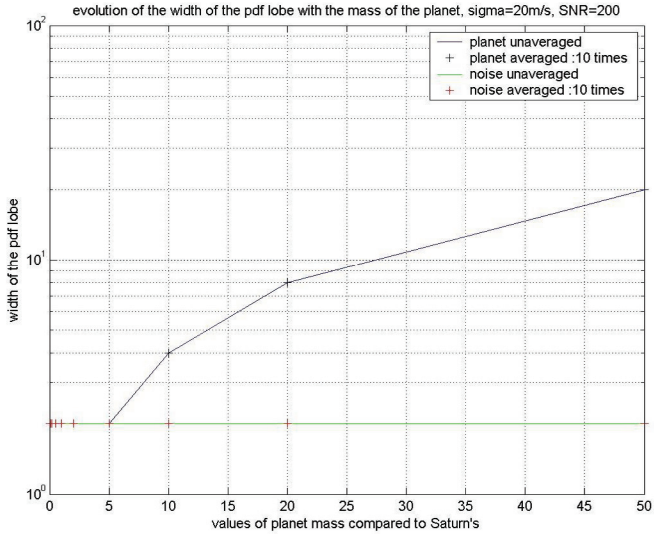


Figure 3 : Comparison of the widths of the PDF lobe when there is a planet and no noise and noise without planet around the star

4.2 Autocorrelation and adaptive filtering

As the astronomer would like to reveal the periodicity of the planet, we looked for a method that could do it. Correlation can reveal the periodic signal from the planet. This technique was used in spectrographs like ELODIE in southern France, in 1995 it allowed to realise a performance at the time of 15 m/s radial velocity and this spectrograph was the first that had a success in planetary discoveries. (Baranne *et al*, 1996)

In our work, the signal giving the evolution of the spectral line was auto correlated, and the resulting curve was afterwards baseline corrected, to reveal its variation compared to 0.
This method not only was able to reveal the wobbling signal but allowed to derive its periodicity.

The autocorrelation enabled the program to reveal planets with more common characteristics than the statistical method. However the signal can still be noisy and big planets could be found with just auto correlation. Astrophysicians and scientists are more interested in finding smaller planets than Jupiter or Saturn.

In the research, sometime are reported the discoveries of multiple planetary systems. But the three or four planets were not found at first. Several measures were performed with radial velocity and one first planet was found, and then later several others were found after updating new measures and revising the equations. The graphs obtained in research are like

group of points and astronomers have to find the best fitted curve. Such story can be read in the Extra Planet Encyclopaedia (Schneider, 2004), like a recent discovery of two Neptune like planets (September 2004), (McArthur *et al.*, 2004) and (Santos *et al.*, 2004)

This method is similar to adaptive filtering. Adaptive filters are structure in signal processing that can estimate the noise or error. It exist two structures the Least Mean Square, LMS and the Recursive Least Squares, RLS. The main difference is that RLS estimates the error with its variance, as LMS requires an ideal estimate of the curve. In our case the estimate could be a possible solution.

The results of both adaptive filter schemes are similar. But if noise or the estimate is not well approximated, the wobbling from the planet could by cleaned, as noise.

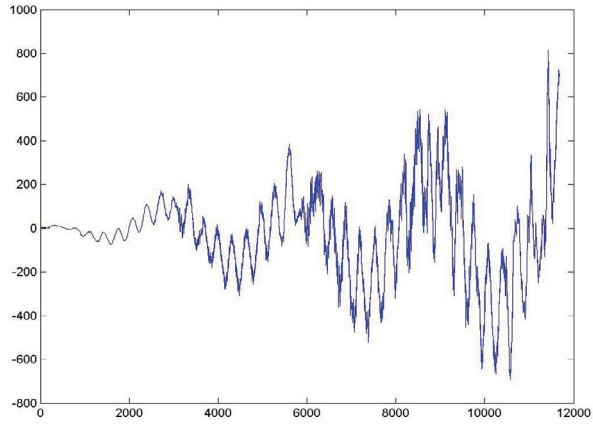


Figure 4 : Output of the adaptive filter after auto correlation. The curve reveals the wobbling of the planet superposed to the star one.

With this method we are able to localise planets of the size of Uranus, 10 times smaller than Saturn, as the one shown in Figure 4.

However the use of adaptive filtering has to be done with caution as it can lead to a false detection or errors if the parameters of the adaptive filter are not properly adjusted. Several estimate curves have to be tested, and compared to reveal the variations from the noise. The advantage of adaptive filtering is when the estimate is correct, the algorithm will make the curve tends to this estimate more rapidly than with a false one. The false one will still be noisy instead. Some of the variations then seen as noise, could then come from the planet's perturbations. Figure 5 shows the differences in these two algorithms and that the adaptive correction is more efficient to discover planet periodicity.

5. Conclusion

Comparing these techniques of noise cancellation that we studied averaging, autocorrelation, and then adaptive filtering showed us than in our simulation adaptive filtering was more efficient because it allows discoveries of smaller planets, when it is carefully used. Other methods might be explored in that sense to find the ways to enhance their results to find the evidence of faint distant worlds.

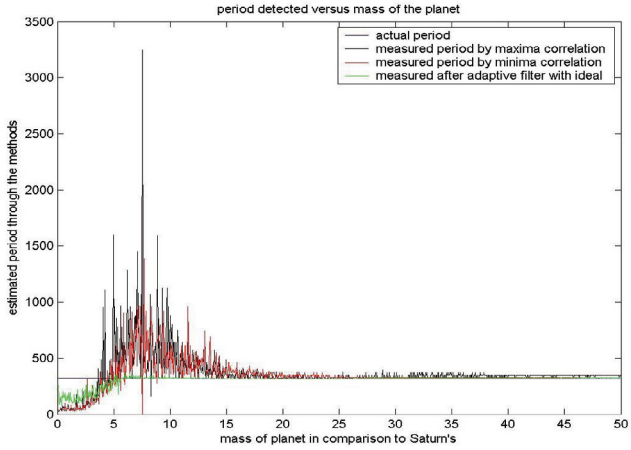


Figure 5 : Comparison in the estimation of the period found with the different algorithms with autocorrelation

6. References

Baranne A., Queloz D., Mayor M., Adrianzyk G., Knispel G., Kohler D., Lacroix D., Meunier J-P, Rimbaud G. and Vin A. (1996) "*ELODIE: A spectrograph for accurate radial velocity measurements*", Astronomy and Astrophysics supplement series 119, pp373-390.

McArthur B.E., Endl M., Cochran W.D., Benedict G.F., Fischer D.A., Marcy G.W., Butler R.P., Naef D., Mayor M., Queloz D., Udry S. and Harrison T. E.. (2004) "*Detection of a Neptune- mass planet in the ρ Cancri system usilng the Hobby Eberly Telescope*"; Mc Donald Observatory, University of Texas, Austin.

Perrymann M.A.C. (2000) "*Extra-solar planets*"; Leiden Observatory, University of Leiden, The Netherlands.

Schneider J. (2004) "*The Extrasolar Planets Encyclopaedia*" http://www.obspm.fr/encycl/encycl.html CNRS Paris- Observatory, last update: 01/09/2004, last visited 01/09/2004.

Santos N.C., Bouchy F., Mayor M., Pepe F., Queloz D., Udry S., Lovis C., Bazot M., Benz W., Bertaux J-L, Lo Curto G., Modasini C., Naef D., Sivan J-P, Vauclair S. (2004) *"The HARPS survey for southern extra-solar planets II. A 14 Earth-masses exoplanet around μ Arae*", Astronomy & Astrophysics manuscript, 25/08/2004.

Zeilik M., Gregory S.A. and Smith E.V.P. (1992) *Introductory Astronomy and Astrophysics*, 3rd edition, Saunders College Publishing, USA.

Fibre Optic Technology: A UK-Pakistan Comparison

A. Ghaffar and C.D. Reeve

School of Computing, Communications and Electronics,
University of Plymouth, United Kingdom
e-mail: creeve@plymouth.ac.uk

Abstract

In this paper fibre optic infrastructure is compared between a developed country such as United Kingdom and a developing country such as Pakistan so that it can be found that how much there is difference between these countries in this fascinating technology. To carry out this project, three sectors were used, Telecommunication companies, educational Institutions and government bodies in both countries. The research part of this study consists of an in-depth literature review, and explanatory research with questionnaires. For collecting data and information three different questionnaires were sent to relevant bodies in both countries. After analysing the returned questionnaires and other information, a significant difference is found between a developed country, UK and developing country, Pakistan. Fibre optics was invented in the UK many years ago and has been installed extensively, while in Pakistan it has only recently been introduced.

Keywords

Fibre optics, Optical fibre, Optical communications, Technology Comparison, Survey

1. Introduction

Optical fibre is a thin filament of glass or any other material through which light travels according to the principle of total internal reflection. In a fibre the signal propagates in the form of light rather than electronic. Typical dimensions for single-mode fibre are 8μm for the core and 125μm for the cladding. Fibre optics is being a backbone of today's modern telecommunications because of its high bandwidth, high speed and reliability. It is widely used in local Area Networks, Community Antenna Television, telephone exchanges and closed circuit cameras.

In this paper, the use of fibre optics is investigated in both countries. Optical Fibres are widely used in developed countries such as UK, USA, Germany, France, Japan, but some developing countries in the world are not so familiar with this high speed technology. For this project two countries are chosen for investigation. One is a developed country (UK) and other is developing country (Pakistan)

This paper describes a small-scale investigation that was conducted by the author in order to determine the fibre optic infrastructure in both countries. The investigation involved the distribution of questionnaires. The remainder of the paper describes the survey method that was employed, followed by a discussion of the results observed.

2. Methodology

Literature review will be the basis of the research project. First of all literature is collected related to the topic of project. Literature is searched from all available sources such as the library, internet, electronic and published journals, newspapers and books.

For this project, some data is collected from online sources and some is collected by contacting fibre optic companies, educational institutions, and government bodies in both countries by sending Questionnaires to them asking what have you done, what are you doing, and in future what you want to do with fibre optics. After collecting the literature, it is read viewed, analysed and then synthesized.

3. Data Collection

Data collection is a fundamental part of a research project. According to Kent, there are two main sources of data collection (Kent, 1999):

- Secondary Sources
- Primary sources

3.1 Secondary Data

Secondary data is information that already exists and that is being used for a second purpose within this category. We can include:

- Published articles in journals, books, newspapers or magazines.
- Data that has been published by various statistical sources, for example, government statistics or trade association statistics.

These kinds of data have been used in the literature review. Secondary data will also be need for completing the project. The authors have tried to get information from reports, registration books and other sources available in both countries.

3.2 Primary data

Primary data is the data specifically obtained for the research. There are several methods for obtaining this data. Two of the more common methods are:

- Questionnaire
- Personal Interviews

For the purpose of this paper, it was decided to design brief questionnaires with direct questions which were distributed to the telecom companies and universities in both countries. By doing this, the author hoped to obtain a higher number of answers about the main issues of the topic.

4. Designing the Questionnaires

It was selected a questionnaire approach to get information about fibre optics and its infrastructure in both countries that is not available by other sources such as internet, library, and books. It was planned to split a big questionnaire into three small questionnaires, one for telecom companies, second for educational institutions, and third for government bodies in both countries so that it will be easy to reply and excellent information can be received.

4.1 Designing the Questionnaire for Telecom/ Fibre-Optic Companies

The questionnaire for telecom and fibre optic companies was developed to obtain the following information:

- How many private or public fibre optic companies are in both countries?
- How many local and multinational companies
- What is max annual revenue of fibre optic companies?
- Future of fibre optic technology in both regions of the world.
- Where companies use their fibre optic products.
- Do companies provide any scholarships to their employees for higher education?
- What are the new fibre optic products?
- What are current fibre optic projects?
- How many companies provide fibre optic training?
- How many companies manufacture educational kits?
- How many MSc or PhD persons related to fibre-optic are working in companies?

4.2 Developing the Questionnaire for Universities

The questionnaire for universities was developed for the following purpose:

- How many universities are public and private?
- How many people are MSc or PhD in engineering faculties?
- How many universities teach fibre optic related modules and have fibre optic laboratory facilities in both areas of the world.
- Do universities provide short courses related to fibre optics?

4.3 Designing the Questionnaire for Government Bodies

This questionnaire was developed to obtain the following information:

- Current coverage of fibre optic infrastructure
- Future plans for fibre optic coverage
- Are fibre optic applications being encouraged by government?
- Are universities being connected through fibre cable?
- Main barriers to deployment and take up of fibre optics
- Essential elements for fibre optic deployment
- Current governmental fibre optic projects

5. Search for Companies, Universities and Government Organisations

Firstly, telecom and fibre-optic companies, Universities and government organisations were browsed by internet in both countries. By this research, 48 telecom companies in the UK and 25 in Pakistan were found. 54 universities related to telecom courses are found in the UK and 24 in Pakistan as shown in table below.

	UK	Pakistan	Total
Telecom/ Fibre-Optic Companies	48	20	68
Educational Institutions	54	24	78

Table 1 : Breakdown of Telecom Companies and Universities in UK and Pakistan

The email addresses of companies and universities are noted then the questionnaires were sent by email. Some emails were not delivered due to error. Later, questionnaires were sent to them by post. Reminder emails were also sent.

	UK	Pakistan	Total
Questionnaires sent by e-mail	43	18	61
E-mails not Delivered	5	2	7
Questionnaires sent by post	5	2	7

Table 2 : Breakdown of the Questionnaires sent by post and email in the UK and Pakistan

6. Extended Survey Method

A quick response was expected. Although a few companies and universities responded quickly and letter of thanks were sent to them, after three weeks the response looked pretty poor. The author, therefore, chose to send a reminder e-mail, which was sent approximately three weeks after the first one.

As a backup the author decided to send the survey by post that were not delivered by e-mail. These were 2 to Pakistan and 5 within UK.

7. Total Response

A breakdown of the total survey is shown in Figure 1.

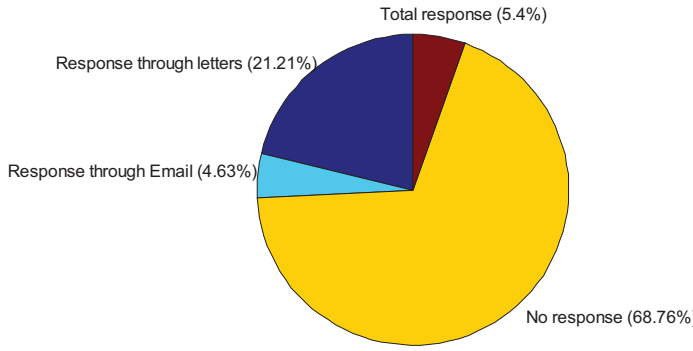


Figure 1 : Breakdown of Survey Response

The total response to the survey in the UK is 7 out of 102, which is a 6.68% response rate. The response rate for the survey sent to Pakistan is 8.33%, since 4 out of 48 were returned completed. The overall response is 11 out of 146, which results in a response rate of 7.53 %.

What can be concluded is that the response rate for the e-mail surveys in the United Kingdom is much higher. This may lead to the conclusion that the British companies, universities and their employees are more experienced with this way of communication than that in the Pakistan. The response rate for the survey by post is also higher in the UK.

The author was disappointed to receive only a few questionnaire replies, due to which the received information is less than the author's expectations.

8. Results

Basically, the aim of the project is to investigate the fibre optics infrastructure in both countries so, literature review forms the main results of the work conducted. Here are some more results that are found by getting information from questionnaires.

Below are the results of the questionnaires that are received by either post or by e-mail from companies, universities and government bodies from both countries. Because the questions of these questionnaires are open-ended due to which results are not drawn in table or chart format.

8.1 Results of Questionnaires Replies form British universities

From British universities, only 4 questionnaires out of 54 are received. These universities are:

- University of Warwick
- Lancaster University
- The University of Reading
- Queen's University

Out of these four universities, only university of Warwick has a lot of fibre optic facilities. In this university, school of engineering has five faculty members who are MSc or PhD in fibre optics. University has fibre optic laboratory, MSc and PhD degree and short courses facilities. This is an ideal university in the United Kingdom for a fibre optic professional.

8.2 Results of Questionnaires Replies form Pakistani Universities

From Pakistani universities, only one out of 24 questionnaire reply is received, may be because many universities is constructing their web-sites. This reply is from the private, Hamdard University. In this university no facility related to fibre has been found.

8.3 Results of Questionnaire Received from Telecom Minister, Pakistan

One questionnaire is replied by Minister for IT and Telecom from Pakistan.. According to this information, there is approximately 40% coverage of fibre optic infrastructure in Pakistan and further it is the government's plan to cover remaining non-fibre areas, at this time government has connected sixteen universities through fibre optic cable and further are being considered on demand. According to the minister's information, low teledensity is the main barrier in deployment of fibre optic cabling in the country. Bandwidth enhancement is the essential element of government's strategy for fibre optic deployment and a specific process is being organised to provide leadership in the design and implementation of the fibre optic strategy. The fibre optic deployment will be boosted in future.

Current fibre optic projects under the supervision of government are:

- 10,000 Km optical fibre to be laid in next three years.
- Access network in big cities.

8.4 Results of Questionnaire Received from Pakistani Telecom / Fibre optic Companies

2 out of 24 replies from telecom/fibre companies in Pakistan were received which are:

- Alcatel Pakistan Limited
- Siemens Pakistan Engineering Company Limited

Both companies are non-local, multinational and have been playing a vital role in telecom development in Pakistan for many years. According to information provided by both companies, fibre optics is still expanding in Pakistan and reaching more and more cities. Neither are providing any scholarships to their employees for higher study, but providing

fibre training facility in the country. No employee having MSc or PhD degree has been found in these companies in Pakistan. They are using their products in LAN, campus networks and linking switches in the cities and their annual revenue is excellent.

8.5 Results of Questionnaire Received from British Telecom/Fibre optic Companies

From British Telecom/fibre companies 3 out of 48 replies are received. These are:

- Halcyon Optical Services
- Fibreco Limited
- Ridgemout Technologies Ltd

All these are private and British-based companies with excellent annual revenue. All these companies have replied that Fibre Optic Technology is currently expanding. These companies are manufacturers of fibre products.
New products of these companies are:

- VF-20: Fibre Inspection Interferometer
- MPX: Mode Distribution Measurement
- Interferometers for end face symmetry measurements.
- Harsh Environment Expanded Beam products.

The products of these companies are being used in:

- Test equipment for fibre manufactures and installers.
- All communication applications; Telecom networks, military, Aerospace and transport
- Data Communication.

Current fibre optic projects and services of these companies are:

- Fibre measurements, optical design, consultancy
- Test and measurement services and equipment speciality cable assemblies and transmission equipment.
- One of these companies, Ridgemount Technologies Ltd provides fibre optic training as well. One MSC or PhD person is found working for this company.

9. Further Results

These are the results that are found during the literature search.

No	Comparison Points	UK	Pakistan
1	New products	Yes	No
2	New projects	Yes	No
3	Training provider companies	More than 30	2
4	Manufacturer of educational kits	Yes	No
5	Universities/ colleges with fibre optic modules	More than 30	2
6	MSc/PhD persons in fibre optic	5 in just one university	2 in all country
7	Member of ITU	Yes	Yes
8	PhD Opportunities	Yes	No
9	Scholarships by companies	No	No
10	Scholarships by Government	Yes	Yes
11	Data Rate of internet	High	Low
12	Fibre optic Vendors	Yes	No
13	Standard SDH/SONET	SDH	SDH
14	Fibre Optic Link between exchanges	All Optical	Now replacing copper to fibre
15	Public Companies		1
16	Private	More than 48	4
17	Local Companies	More than 30	2
18	Multinatioanl Companies	More than 15	4
19	Fibre optic installation standard	BS6701	None
20	Fibre optic association	Yes	No
21	Fibre optic magazines or journals	Yes	No
22	Fibre optic research laboratory	Yes	No
23	Budget for education	High	Low
24	Research in universities	Yes	No
25	Web site of companies and universities	Constructed	Under Construction
26	No of R&D Scientists	107,500	8428
27	No of R&D Scientists per Population	1886	85
28	Teledensity (year 2001)	58.8	2.4

Table 3 : Fibre optics Comparison between UK and Pakistan

10. Conclusions

10.1 Conclusion of Fibre-optic Technology

Over the last few years fibre optic technology has advanced at a tremendous rate in a rather quiet and reserved manner driven by the need for higher bandwidths on long distance backbone links.

This performance enhancement has gone hand-in-hand with the development of suitable transmission and access methodologies such as Synchronous Digital Hierarchy (SDH). The

higher rates defined by SDH would not be possible without the improvements that have taken place in fibre optics.

The fibre optics industry is a multibillion dollar business which is growing at almost 20% each year. (Jerry, 1997) In this industry, in which new technology dominates the headlines, the old is finding it still has a bright future.

10.2 Conclusion of Fibre-optic in the UK

The UK was one of the pioneers of fibre optics technology, and its science and industry base has continued to be at the leading edge of research and development world-wide. UK is in one of the top 50 international telecommunication routes in the world.

In particular, its academics have become much more adept at taking their work out of the laboratory and into the world of business

In United Kingdom, all telephone exchanges are linked via fibre optic cable. Further research is being done on its use in sensors and avionics, while in Pakistan, this technology is being introduced recently because of less knowledge and understanding.

As a result of government and EU funding over the past 15 years or so (JOERS, LINK etc) the UK is in a strong position to exploit the opportunities for nanotechnology in optical fibre communications systems (Baker, 1999).

10.3 Conclusion of Fibre optics in Pakistan

In Pakistan, IT and Telecom is being introduced and developed by both Government and private efforts. Many government and educational institutions are constructing their web-sites. Government has established many new institutions and universities in big cities that are running courses related to information technology. Many private sector universities are also playing a vital role in developing IT in the country, but if fibre-optic is discussed, it is still not so developed in the country. Fibre optics is secondary case, many parts of the country including the author's own rural area does not have the telephone facility, but major and main cities of the country are being linked through optical fibre cable, for high speed data and voice transmission.

There is an only one factory that is manufacturing optical fibre cable and other components, but many major fibre optic components are imported from developed countries to fulfil the demands. In some universities, modules related to optical communication systems is taught, but no fibre optic Laboratory facilities are found in the universities because of lack of knowledge of this technology and also because fibre-optic lab facilities are very expensive. Many web-sites of companies, educational institutions and government departments are found under construction. The main obstacle is lack of familiarity with this fascinating technology.

It can be concluded that there is a significant difference between the fibre optic infrastructure in the UK and Pakistan.

11. References

Benson J. (1997) *New Uses for Multimode Optical Fibres*, [online]. Available: http://www.azom.com/details.asp?ArticleID=622 [Accessed 12 September 2004]

Kent R. (1999) *Marketing Research: Measurement, Method and Application*, 1st Edition, p7

Baker, J. (1999) '14 Nanotechnology and Optical Fibre Communication Systems' in *Opportunities for Industry in the Application of Nanotechnology*, Draft report from Institute of Nanotechnology for the Foresight Materials Panel, March 1999.

Survey of Current Designs for Mobile Handset Phones and Future Trends Followed by Detailed Investigation / Design

A. Pistelas and C.F. Hamer

School of Computing, Communications and Electronics,
University of Plymouth, United Kingdom
e-mail: C.Hamer@plymouth.ac.uk

Abstract

The growing trend of decreasing mobile phone handset size has created the need for small antennas coupled with the need to fulfill growing demands of multiband operation. In this paper the various designs of mobile handset antennas are presented. The paper focused mainly on internal antennas due to the current manufacturing trend in mobile phones design. A low profile planar monopole antenna for multiband operation is designed and subsequently manufactured. A novel technique in planar monopole antennas, in which part of the ground plane operates as a parasitic element, was devised in order to overcome the planar antennas' problem of a narrow bandwidth and cover the three communications systems DCS (1710-1880), PCS (1850-1990), UMTS (1920-2170). Subsequently, a detailed description of antenna design is provided in order to describe the way the antenna operates and to demonstrate the role of the parasitic element in detail.

Keywords

External antennas, Internal antennas, Parasitic element

1. Introduction

In this paper, a survey regarding current antenna designs for mobile phones including design for future trends was presented. The limited power of mobile handsets and the growing trend for small devices over the last decade has led to significant improvement of antenna performance and different types of antennas being developed. The antenna designer's goal is to ultimately reduce the size of antennas without degradation of the bandwidth, gain, efficiency and ability to operate in multi band frequencies.

Mobile phone antennas are divided into two main categories, external and internal. The use of the internal antennas has become very common over the last few years, mainly because of their compact size.

Internal antennas are subsequently divided into six main types, namely the microstrip-patch antennas, the printed monopole, the planar inverted-F (PIFA) and the ceramic chip antennas. The most commonly used type is the planar inverted-F antenna due to its ability of combining small dimensions and flexibility in multiband operation. However, the constant demands for even smaller mobile phones in combination with the growing knowledge in dielectric antenna design have made the chip antenna the most promising type for the future mobile handsets.

In addition to the survey of mobile handsets antennas, the design and subsequent manufacture of a planar monopole internal antenna is also presented. The role of the ground plane as a

potential parameter in antenna design is discussed, tested and assessed based on simulations and real measurements.

2. External antennas

The dipole is the most common and efficient antenna. However, the use of a dipole antenna in a mobile handset is inappropriate because of its shape, so the λ/4 monopole antenna is preferred. Another type of external antennas that are used instead of the λ/4 monopoles is the Helical antennas (Kerry, 2002).

3. Internal antennas

Internal antennas are compact antennas built in the upper part of the handset. The ground plane contributes to the radiation of the internal antenna as an asymmetrical dipole.

3.1 Microstrip antennas

The light weight and low manufacturing cost of microstrip antennas makes them suitable as internal antennas in mobile handsets. Another advantage is that when the patch shape and mode are selected, the resonant frequency, polarization radiation pattern and impedance can be modified. On the other hand, the bandwidths of microstrip antennas represent their main drawback. A lot of techniques have been suggested to improve the bandwidth of microstrip antennas. The most effective is the etching of a small slot, known as U-slot (Guo et al 2002).

3.2 Planar monopole antennas

The planar monopole antennas create resonance by using metal plates. These plates are paths of one quarter wavelength. The planar monopoles are fed by a microstrip line and they are located in the top of the ground plane.

The planar monopole antenna has a large profile. It is essential to reduce its height in order to be implemented as internal antenna in mobile handsets (folded planar monopole antenna). The folded planar monopole antenna is achieved by wrapping the planar monopole into a compact box-like structure (Chang et al, 2002).

3.3 Meander Line Antennas

The Meander Line Antenna (MLA) is a type of 3D radiating element, but needs less height than PIFA. It is a combination of wire and planar strip lines which provides wide bandwidths in different frequency bands simultaneously. One of the basic characteristic of MLA is that its electrical length depends on the features of the meander line (Centurion Wireless Technologies, 2000).

3.4 Planar inverted-F Antennas

Planar Inverted Antenna (PIFA) is the most widely used antenna in applications of cellular communication because of its ability to be tuned in various frequencies, its small dimensions and wide bandwidth. PIFA typically consists of a rectangular planar element supported above the ground plane by a low-loss plastic (Centurion Wireless Technologies, 2000). Multiband operation is achieved by dividing the top patch into individual radiating elements. A multiband PIFA which also uses parasitic elements for bandwidth extension is illustrated in Figure 1.

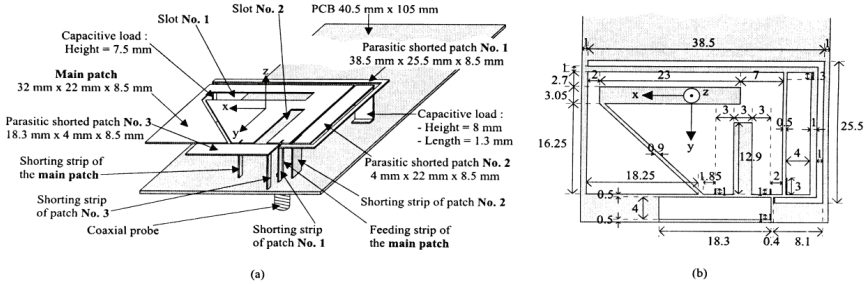


Figure 1 : Structure of a six band PIFA (Cias et al, 2004)

3.5 Chip antenna

In most of the chip antennas the radiation element is mounted on the chip which usually is either ceramic or plastic. The ceramic chip antennas have higher relative permittivity than the plastic ones. High relative permittivity reduces the size but also leads to narrower bandwidths. On the other hand the plastic chip antenna is lighter and less fragile (Wong et al, 2002).

4. Future trends

Although, current technology has found an answer to the market demands for small multiband internal antennas, it seems that future market demands can not be met by the same technology. The barrier for further size reduction of the most commonly used type of internal antennas (PIFA) is proven to be their physical height. Similar barriers can be found in other types of internal antennas where the minimum achievable size is limited by their physical length. Furthermore the size problem becomes even more intense due to the addition of parasitic elements within the antenna structure as a means of succeeding multiband operation.

All the today's design methods seem to have reached their fundamentals limits. Further reduction of mobile handsets size would reduce the available PCB area which acts as a ground in internal antennas. Dielectric materials seem to be the only way for further development of internal antennas. Chip antennas make use of the properties of the dielectric materials mainly for size reduction purposes. The most promising design method involves use of high dielectric materials as in the case of the High Dielectric Antennas (HAD) by Antennova.

The compact size of HAD antennas allows development of multi built-in antenna structures occupying less space than even the smallest currently used internal antennas. Such a breakthrough opens up the way for the use of Multiple Input Multiple Output (MIMO) systems. MIMO systems, combined with adaptive coding and modulation, could provide 30 times higher data rates than those of 3G systems. A drawback of MIMO is that until now bands bellow 1GHz cannot be covered. The current proposed solution is the use of parasitic elements for covering the AMPS and the GSM bands. (Antenova, 2004)

5. Design

The designed and manufactured antenna is based in a low profile planar monopole antenna proposed by Wong et al (2003), capable of covering four mobile communication systems.

One of the first problems that has been noticed was that even though it seemed the antenna could be tuned as Wong et al (2003) proposed (low frequency bands by the outside patch, and the three upper by the inside patch and the outside as half wavelength resonant mode) the obtained bands were too narrow.

The design has been divided in two parts. In the first part the antenna has been optimised based on Wong et al (2003) design and the effect of ground plane in the obtained bandwidth has been investigated.

The optimised design did not cover satisfactorily (-6dB return loss) any of the desired bands. Extending the length of the ground plane did not increase the bandwidth and the resulted shift towards the upper frequencies made the tuning of the antenna very difficult. The widest bandwidth was achieved with length of ground plane equal to the length of the feeding line. Figure 2 shows the designed antennas and the obtained bandwidths. Notice that there is difference between the two dB scales. The first is from 0 to -25dB and the second from 0 to -8dB.

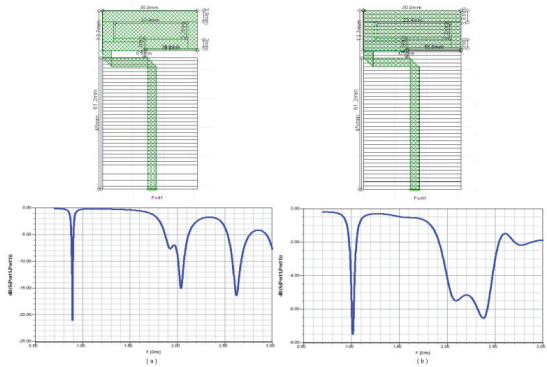


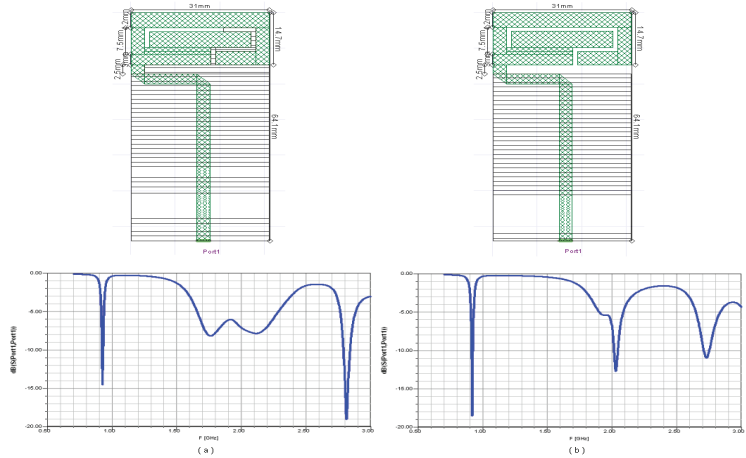
Figure 2 : Structures of the antennas and the obtain bandwidths in the first design part The optimized designed antenna and the obtained bandwidth (b) The same antenna with maximum ground length

In the second part, various methods have been investigated to overcome the problem of the narrow bandwidths.

The use of parasitic elements is a popular and effective method of extending narrow bandwidths. The disadvantage of this method is the increase of the antenna's volume. Inspired by Abedin and Ali (2003) who modifies the ground plane of a PIFA in order to reduce its height, the use of part of the ground plane as parasitic has been investigated. This method offers extended bandwidth without increasing the antenna volume, by enlarging the slit and fitting a patch of ground in the space between the two radiating patch elements. The length of the additional ground used as parasitic element depends on the frequencies under interest.

The parasitic element was tuned approximately at 1750MHz to ensure coverage of the lower limit of the three overlapping bands. The antenna has been modified (larger slit, dimension of patches, etc) so that the resonant modes of the two subpatches and the parasitic element are optimally coupled.

The modifications are limited in the width of the radiation elements, because any change in their lengths would disturb the tuning of the antenna. Further modifications have also been introduced into the patch which is attached to the feeding point of the antenna. The effect of the ground parasitic element has been further investigated by comparing the characteristics of this antenna versus a replica of the designed antenna with no extended grounds parts. Figure 3 illustrates the structures of the designed antenna and its replica and the obtained bandwidths.



(a) Structure and bandwidth of the antenna used the parasitic ground element

(b) Structure and bandwidth of the replica antenna without using parasitic ground element.

Figure 3 : Structures of the antennas and the obtain bandwidths in the first design part

6. Discussion

The dimensions of the first designed antenna, including the ground plane, are approximately the same of Wong et al (2003) antenna (30.9 x 61.2 mm^2), figure 2. The radiating rectangular patch dimensions are 13.7x30.9 mm^2. The length of the outside, longer, patch is 70.1mm and the length of the inside, shorter, patch is 29.65mm.

The outer patch operates as a quarter wavelength structure in frequencies at about 900MHz and as a half wavelength in frequencies around 1850MHz. The inner patch operates as a quarter wavelength in higher frequencies of about 2050MHz.

The disadvantage of this antenna is that although it is easily tuned, the bandwidths where the -6dB return loss limit is achieved, are too narrow to cover any of the desired communication standards. In frequencies close to GSM band, the bandwidth is only 20MHz and the next accomplished band resides between 1884 and 2103MHz. Although, Abedin and Ali (2003) state that larger ground plane results in wider bandwidths, it is observed that by using the largest possible ground plane (30.9x61.2mm^2) the bandwidth not only does not increase, but becomes narrower and shifts towards higher frequencies.

The size of the second designed antenna, including the ground plane, is larger (31 x 64.1 mm^2), figure 3. However, it is still small enough to fit in a mobile handset. The most critical modifications are the enlargement of the slit which separates the two subpatches and the addition of two parts of ground plane.

The first, rectangular, added part makes use of the observation by the first design. It has been added to slightly shift the band into higher frequencies so that the PCS and UMTS bands are covered by the two subpatches. The second part is a strip, which has been added in order to operate as a parasitic element and to extent the bandwidth. As an extension of the ground plane, it also contributes in the frequencies shift. It is tuned at about 1750MHz to cover part of the DCS band. Both of the added parts are located in areas where there is no layout on the other side of the PCB in order to avoid wider shift and destruction of the impedance (it would become capacitive).

Combining the bands that are provided by the two subpatches and the strip of the ground plane, a -6dB return loss bandwidth of 526MHz, from 1681MHZ to 2243MHz, has been achieved, figure 3. This bandwidth is almost two and half times larger than the one achieved by the first designed antenna. Comparison of the -6dB return loss bandwidth between this antenna and its replica without the two additional parts of ground would not give significant conclusions, because the extraction of the two ground plane parts affect the impedance of the antenna. However, figure 3 offers a picture of the shift, mainly caused by the addition of the rectangular ground plane. The enlargement of the bandwidth towards lower frequencies caused by the strip can be also observed. At frequencies around 900MHz even though the return losses are very low (-14.5dB in 927MHz) the bandwidth has remained too narrow. In all the designed antennas a band in the higher frequencies, above 2500MHz is observed. These bands are due to the operation of the inner subpatch as half wavelength in these frequencies.

Both of the orthogonal components are present. The antenna, including the whole ground plane, radiation is similar to a dipole antenna. Comparing the radiation pattern of the antenna with and without the additional parts of ground plane, it can be seen that the impact of a parasitic element in radiation pattern is small since this is tuned at lower frequencies. Moreover, an interesting observation is the extremely low radiation in 90, 90 and 0 degrees in Ex, Ey and Ez respectively. The lack of radiation in these directions is due to the dielectric of the PCB. The thicker the dielectric becomes the bigger the angle with low radiation. The obtained gain varies between 2 and 4dB. By comparing the two obtained gains, with and without parasitic element, it is observed that the parasitic also contributes in the increase of the antennas gain.

The real measurements of return loss exhibit significant divergences from the simulated results mainly due to difficulties in lining-up the ground plane and more specifically the parasitic strip with the slit which separates the two.

7. Conclusion

The design of a planar monopole antenna using the antenna proposed by Wong et al (2003) as a template resulted in severe narrow bandwidth problems. The use of the ground plane as a parasitic element has led to an enlargement of the bandwidth by approximately two and half times and subsequent covering of DCS, PCS, UMTS communication systems. Furthermore, this technique has increased antenna gain by almost 2dB in some frequencies.

The technique of using parasitic elements in order to extend the bandwidth is currently implemented in internal antennas. The difference in using the ground plane as parasitic element is that the increase of antenna size is extremely small. This technique has been applied in MLA, but more often than not results in an increase of antenna size.

Due to the flexibility of this technique, in terms of tuning the parasitic element, it is possible to implement this method in other types of internal antennas, such as multiband microstrip-patch antennas, where their patch is separated by a slit. The only tuning limitation of the parasitic element is the available space in antenna structure.

Finally, future modifications in order to assimilate Bluetooth operating frequencies in addition to GSM, DCS, PCS and UMTS bands in this designed antenna could be investigated. This may be possible by thorough choice of material characteristics in order for the four bands to be covered by the antenna elements as Wong et al (2003) proposed. The author of this paper proposes coverage of the Bluetooth frequency band using the band provided by the inner subpatch which operates as a $\lambda/2$ resonant mode at about 2500MHz in addition with the parasitic element tuned in at 2400MHz for achievement of the above. This modification is feasible both on a technical and commercial standard in the current climate of mobile phone antenna design.

8. References

Abedin, M. F. and Ali, M. (2003) "Modifying the Ground Plane and Its Effects on Planar Inverted-F Antennas (PIFAs) for Mobile Phone Handsets". *IEEE Antennas and Wireless Propagation Letters*, 2, pp226-229.

Antennova. (2004) High Dielectric Antennas White Paper [Online] http://www.antenova.com/media/news/Technology%20White%20Paper.pdf [13 August 2004].

Centurion wireless technologies. (2000) 'Design guide for wireless device antenna systems' [online] http://www.centurion.com/pdf/centurion-designguide-043002.pdf [23 November 2003].

Chang, F.-S., Yeh, S.-H. and Wong, K.-L. (2002) "Planar monopole in wrapped structure for low-profile GSM/DCS mobile phone antenna". *Electronics Letters,* 38 (11, May), pp499-500.

Ciais, P., Luxey, C., Diallo, A., Staraja, R. and Kossiavas, G. (2004) 'Design of internal multiband antennas for mobile phone and WLAN standards' [Online] http://www.s2.chalmers.se/costworkshop/workshop_papers/126.pdf [13 August 2004].

Guo, Y.-X., Luk, K.-M., Lee, K.-F. and Chair, R. (2002) "A Quarter-Wave U-Shape Patch Antenna With Two Unequal Arms for Wideband and Dual-Frequency Operation". *IEEE Transactions On Antennas And Propagation,* 50 (8, August), pp1082-1087.

Kerry, G. (2002) 'Stretching the limit of Embedded Antenna Design' [Online] http://www.skycross.com/WDD_032002.asp [23 November 2003].

Wong, K.-L, Lee, G.-Y. and Chiou T.-W. (2003) "A Low Profile Monopole Antenna for Multiband Operation of Mobile Handsets". *IEEE Transactions on Antennas and Propagation*, 51 (1, January) pp121-124.

Author Index

Distributor:

Network Research Group
School of Computing, Communications & Electronics
University of Plymouth
Drake Circus
Pymouth
PL4 8AA
United Kingdom

ISBN 978-1-84102-140-9 90000
9 781841 021409